

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

A Data Quality-Centric Approach for Predicting Radiology Report Delays to Support Medical Appointment Scheduling

Daniela Martins Silva



Master in Biomedical Engineering

Supervisor: Hélder Filipe P. de Oliveira

Proponent: ByMe - Information Technology, Lda

July 8, 2025

Abstract

Delays in medical exam interpretation raise costs when the subsequent medical appointments are canceled due to the absence of results. These delays can be estimated with machine learning models whose predictions support the optimal medical appointment scheduling, and thus, counter this problem. However, there is a major obstacle: the quality of hospital datasets is poor. These contain data that is tainted by outliers, missing values, duplicate data and standardization problems.

The main objective of this dissertation was to develop different preprocessing algorithms and understand the most efficient techniques to clean and improve the quality of such data. This preprocessed data, as it was being treated, was iteratively evaluated as an input for training and testing of different machine learning models, that work to predict those delays.

The dataframe used was first tailored to the specific requirements of the problem throughout a well-structured questionnaire that was designed and sent to physicians of different specialties. Then, data was acquired in collaboration with ByMe, a healthcare leading company, whose databases contain real data representative of hospital workflows in Portugal. After establishing the baseline of the dataframe, a preliminary cleaning of the dataset ensured there were no missing values by removing what was not important and imputing known internal patterns. This was followed by the construction of important time features and the development of specific algorithms to address categorical and numerical outliers, detected with both univariate and multivariate techniques. Performance and other evaluation metrics of different tree-based models served as validation metrics of the effectiveness of the algorithms created.

The results showed consistent performance improvements across the first stages: from baseline to initial preprocessing, derived feature creation, categorical outlier removal and cyclic outlier removal. However, after these stages, the results varied depending on the method used to treat the continuous numerical outliers. On one hand, the best performances were achieved with algorithms of deletion and clipping like winsorization, detected with Interquartile Range. On the other hand, for treatments such as KNN and MICE Imputation, the overall performance slightly decreased. These findings outlined that more data does not always mean better data. The relevance of these results validate the importance of exploring different data cleaning techniques in preprocessing phases to further enhance predictive modeling.

Keywords: Data Preprocessing, Data Cleaning, Data Quality, Feature Engineering, Outlier Treatment, Machine Learning, Predictive Modeling, Medical Reporting, Exam Results

Resumo

Atrasos na interpretação de exames médicos aumentam os custos quando as consultas médicas subsequentes são canceladas devido à ausência de resultados. Esses atrasos podem ser estimados com modelos de machine learning cujas previsões apoiam o agendamento ideal de consultas médicas e, assim, combatem esse problema. No entanto, há um grande obstáculo: a qualidade dos dados hospitalares é baixa. Estes contêm dados contaminados por outliers, valores ausentes, dados duplicados e problemas de padronização.

O principal objetivo desta dissertação foi desenvolver diferentes algoritmos de pré-processamento e compreender as técnicas mais eficientes para limpar e melhorar a qualidade desses dados. À medida que eram tratados, foram avaliados iterativamente como entrada para o treino e teste de diferentes modelos de machine learning, que funcionam para prever esses atrasos.

O dataframe utilizado foi primeiro adaptado aos requisitos específicos do problema através de um questionário bem estruturado que foi concebido e enviado a médicos de diferentes especialidades. Em seguida, os dados foram adquiridos em colaboração com a ByMe, uma empresa líder na área da saúde, cujas bases de dados contêm dados reais representativos dos fluxos de trabalho hospitalares em Portugal. Após estabelecer a linha de base da estrutura de dados, uma limpeza preliminar do conjunto de dados garantiu que não houvesse valores em falta, removendo o que não era importante e imputando padrões internos conhecidos. Seguiu-se a construção de variáveis temporais importantes e o desenvolvimento de algoritmos específicos para lidar com outliers categóricos e numéricos, detetados com técnicas univariadas e multivariadas. O desempenho e outras métricas de avaliação de diferentes modelos baseados em árvores serviram como métricas de validação da eficácia dos algoritmos criados.

Os resultados mostraram melhorias consistentes de desempenho nas primeiras etapas: da baseline ao pré-processamento inicial, criação de variáveis derivadas, remoção de outliers categóricos e remoção de outliers cíclicos. No entanto, após essas etapas, os resultados variaram dependendo do método usado para tratar os outliers numéricos contínuos. Por um lado, os melhores desempenhos foram alcançados com algoritmos de exclusão e recorte, como a winsorização, detetados com Interquartil Range. Por outro lado, para tratamentos como imputação KNN e MICE, o desempenho geral diminuiu ligeiramente. Essas descobertas destacaram que mais dados nem sempre significam melhores dados. A relevância desses resultados valida a importância de explorar diferentes técnicas de limpeza de dados nas fases de pré-processamento para aprimorar ainda mais a modelagem preditiva.

Keywords: Pré-processamento de Dados, Limpeza de Dados, Qualidade dos Dados, Engenharia de Variáveis, Tratamento de Outliers, Machine Learning, Modelação Preditiva, Relatórios Médicos, Resultados de Exames

UN Sustainable Development Goals

The United Nations Sustainable Development Goals, often called the SDGs, offer a road map for building a fairer, greener planet. The set covers 17 specific targets from wiping out extreme poverty and closing the inequality gap to curbing climate damage, protecting nature and strengthening peace and justice. In Table 1 are presented the contribution of the project to some SDGs and associated performance indicators and metrics.

Table 1: SDGs and contribution of the work.

SDG	Target	Contribution of the project	Performance Indicators and Metrics
3	3.8 – Achieve universal health coverage	Supports more equitable healthcare delivery by aligning model inputs with clinical priorities to reduce scheduling inefficiencies.	Inclusion of clinical feedback in model design; reduction in scheduling bottlenecks.
	3.c – Increase health financing and workforce efficiency	Facilitates better use of healthcare professionals' time by minimizing data inconsistencies and enabling operational insights.	Time saved during scheduling due to improved data; reduction in manual data cleaning needs.
9	9.5 – Enhance scientific research and technological capability	Applies machine learning and data engineering to improve data quality in a clinical context.	Model performance before vs after cleaning; number of data quality strategies tested.
	9.b – Support technology development and innovation	Drives innovation by developing and comparing preprocessing algorithms for predictive modeling.	Number of preprocessing algorithms evaluated.
12	12.6 – Encourage sustainable practices in institutions	Encourages sustainable data management practices in healthcare institutions through structured workflows and collaboration.	Adoption of data governance practices; stakeholder satisfaction with data outputs.

Acknowledgments

Firstly, I would like to express my sincere gratitude to my academic supervisor, Professor Dr. Hélder Oliveira, for his expert guidance, constructive feedback and continuous support throughout this research.

Inside ByMe, I would also like to extend my special thanks to Ana Freire for her patience, readiness and invaluable expertise, which greatly contributed to the progress of this project. Additionally, I want to thank Jorge Araújo and Ernesto Palhares, for their trust and for providing me with the opportunity to collaborate with ByMe. I also want to thank all ByMe professionals and physicians that collaborated with their insights on the topic. And because great ideas often come from collaboration, my sincere thanks goes to my colleague Pedro Fernandes, whose cooperation and insightful contributions greatly facilitated my work.

I would also like to acknowledge my fellow Master's classmates - David, Inês, Luísa, Mariana and Vânia - for their constant support and willingness to help throughout this two-year adventure. My appreciation extends to all members of JuniFEUP, both current and former, with particular thanks to my director, Kika. Thank you for your leadership, dedication and kindness, and for providing me with one of the most transformative and enriching experiences I could have asked for.

Lastly, I wish to thank my closest friends - Bia, Carol, Duda, Inês, Margarida, Mariana and Marta - for their constant friendship and encouragement, even across the miles that sometimes separate us. Their support and care have been a comforting reminder that true friendship knows no distance.

Above all, my most heartfelt gratitude goes to my family, whose unconditional love and belief in me have been the cornerstone of all my achievements. Thank you mom, dad, "mano" and Daniel. Your constant presence has been my greatest source of strength and this dissertation is as much yours as it is mine.

"It always seems impossible until it's done."

Nelson Mandela

Contents

1	Introduction	1
1.1	Current Problem	2
1.2	Motivation and Relevance	2
1.3	Goals	3
1.4	Contributions	3
1.5	Document Structure	4
2	Background	5
2.1	Foundations of Data Analysis Methodologies	5
2.1.1	Data Quality Assessment Methodologies	5
2.1.2	CRISP-DM Framework	7
2.2	Insights Through Questionnaire Structuring and Assessment	8
2.2.1	Development and Structuring of a Questionnaire	8
2.2.2	Analysis and Assessment of Questionnaire Responses	9
2.3	Data Preprocessing	10
2.3.1	Detection Techniques	10
2.3.2	Repairing Techniques	13
2.4	Data Mining Tools	15
2.5	Model Validation and Performance Metrics	18
2.6	Frameworks for Model Interpretability and Explainability	19
2.7	Summary	21
3	Literature Review	22
3.1	Search Methodology	22
3.2	Data Quality Pipelines in the Healthcare Context	23
3.3	ML Models Applications for Reducing Hospital Delays	26
3.4	Summary	27
4	Methodology and Resources	28
4.1	Overview	28
4.2	Ethical Considerations in Data Use	30

4.3	Summary	30
5	Foundation and Study Preparation	31
5.1	Requirements Gathering	31
5.2	Questionnaire Results	35
5.3	Dataset Context and Structure	39
5.4	Evaluating Data Cleaning Effectiveness	41
5.5	Summary	43
6	Non X-ray Radiological Data	44
6.1	Data Understanding	44
6.1.1	Non X-ray Baseline Performance	48
6.2	Preliminary Data Preparing and Exploratory Assessment	48
6.2.1	Variable Format Standardization	49
6.2.2	Deduplication	49
6.2.3	Case Deletion Based on Criteria	50
6.2.4	Imputation Methods Based on Known Patterns	51
6.2.5	Derived Feature Creation	51
6.2.6	Target Definition	57
6.2.7	Final Dataset Structure After Preliminary Cleaning	57
6.3	Categorical Outliers	60
6.3.1	Detection	60
6.3.2	Treatment	62
6.3.3	Performance Variability Results	62
6.4	Manual Approach: Visual-Based Numerical Outlier Removal	64
6.4.1	Detection and Treatment	64
6.4.2	Performance Variability Results	66
6.5	Automatic Approach: Statistical and ML-Based Numerical Outlier Handling	66
6.5.1	Detection	66
6.5.2	Treatment	71
6.5.3	Performance Variability Results with IQR Detection	74
6.5.4	Performance Variability Results with IF Detection	86
6.6	Summary	91
7	X-ray Radiological Data	92
7.1	Data Understanding	92
7.1.1	X-ray Baseline Performance	95
7.2	Preliminary Data Preparing and Exploratory Assessment	96
7.2.1	Variable Format Standardization	96

7.2.2	Deduplication	96
7.2.3	Case Deletion Based on Criteria	96
7.2.4	Imputation Methods Based on Known Patterns	96
7.2.5	Derived Feature Creation	96
7.2.6	Target Definition	97
7.2.7	Final Dataset Structure After Preliminary Cleaning	98
7.3	Categorical Outliers	98
7.3.1	Detection	99
7.3.2	Treatment	100
7.3.3	Performance Variability Results	100
7.4	Summary	101
8	Comparative Analysis	102
9	Conclusions and Future Work	114
9.1	Final Considerations	114
9.2	Limitations and Difficulties	114
9.3	Future work	115
	References	122
A	Questionnaire - Factors Weighting Section	123
B	Group-Weighted Questionnaire Results	131
C	Medical Acts - Complete List	133

List of Figures

2.1	The CRISP-DM process.	8
2.2	Illustration of two examples of <i>Likert Scales</i> used for data collection.	9
2.3	Overview of the data preprocessing phase, highlighting its two stages.	10
2.4	Graphical results generated by the PyOD Toolbox for outlier detection.	12
2.5	Illustrative example of a SHAP result displayed graphically.	19
2.6	Illustrative example of an ELI5 explanation result.	20
2.7	Illustrative example of a graphical LIME result.	20
2.8	Partial Dependency Plots result for three models (random forest, gradient boosted tree and DNN) in the context of loan prediction.	21
3.1	PRISMA Flow Diagram that represents how the selection of specific records was made, condensing everything explained in the previous subsections.	23
4.1	The methodology flow used: from clinical insight to iterative ML modeling, based on CRISP-DM.	28
5.1	Table made for evaluating temporal categories. Here, both reporting-prescribing and prescribing physicians evaluate the relevance of temporal factors concerning their work.	34
5.2	Professional profiles of the physicians who answered.	35
5.3	Results from hospital-related data.	36
5.4	Frequency of delays in exam report delivery perceived by two different type of physicians.	36
5.5	Representation of the most prescribed specialties.	38
5.6	Diagram illustrating the data split based on the type of exam (X-ray vs. non X-ray). Equal exam types from both sites share the same predictive target. The colors are only used to explain which subsets share the same target.	39
6.1	Age histogram of non X-ray radiological data.	47
6.2	Patient gender histogram of non X-ray radiological data.	47
6.3	Scatter plots between sine and cosine of cyclic variables from the training set of Site 2.	55
6.4	Seasonality study for Non X-ray Radiological data of Site 1.	58
6.5	Seasonality study for Non X-ray Radiological data of Site 2.	59
6.6	Frequency analysis of MedicalAct categories of Site 1.	61
6.7	Frequency analysis of MedicalAct categories of Site 2.	61

6.8	LightGBM evolution with replacing and removal of categorical outliers at Site 1.	63
6.9	XGBoost evolution with replacing and removal of categorical outliers at Site 2.	63
6.10	Site 1 scatter plots of the train set numerical features without any treatment to allow proper visualization and manual removal.	64
6.11	Site 2 scatter plots of the train set numerical features without any treatment to allow proper visualization and manual removal.	65
6.12	Distribution and outlier visualization of TotalProcessTimeout Site 1.	67
6.13	Distribution and outlier visualization of SchedulingTime Site 1.	67
6.14	Distribution and outlier visualization of SchedulingTimeDeviation Site 1.	67
6.15	Distribution and outlier visualization of ExamDelay Site 1.	67
6.16	Distribution and outlier visualization of WaitingForExam Site 1.	68
6.17	Distribution and outlier visualization of PatientAge Site 1.	68
6.18	Distribution and outlier visualization of TotalProcessTimeout Site 2.	68
6.19	Distribution and outlier visualization of SchedulingTime Site 2.	68
6.20	Distribution and outlier visualization of SchedulingTimeDeviation Site 2.	69
6.21	Distribution and outlier visualization of ExamDelay Site 2.	69
6.22	Distribution and outlier visualization of WaitingForExam Site 2.	69
6.23	Distribution and outlier visualization of PatientAge Site 2.	69
6.24	Scatter plot between sine and cosine of hour variable from the training set of Site 2 and Site 1 after removal of outliers with IQR, factor 1.5.	75
6.25	Q1 and Q3, IQR-based outlier thresholds (factor=3) and the Winsorization bounds (1st and 99th percentiles) from Site 1.	81
6.26	Q1 and Q3, IQR-based outlier thresholds (factor=1.5) and the Winsorization bounds (1st and 99th percentiles) from Site 1.	81
6.27	Correlation between different continuous numerical variables within Site 1.	84
6.28	Correlation between different continuous numerical variables within Site 2.	84
6.29	Impact of listwise deletion during preprocessing of continuous features after IQR-based outlier detection, trained and tested with XGBoost with Site 2 data.	86
6.30	Scatter plot between sine and cosine of cyclic variables from the training set of Site 1 after removal of outliers with IF.	88
6.31	Scatter plot between sine and cosine of cyclic variables from the training set of Site 2 after removal of outliers with IF.	88
6.32	Impact of listwise deletion during preprocessing of continuous features after IF-based outlier detection, trained and tested with XGBoost with Site 2 data.	91
7.1	Age histogram of X-ray radiological Data.	94
7.2	Patient gender histogram of X-ray radiological Data.	95
7.3	Frequency of medical acts from Site 1 X-ray radiological data.	99
7.4	Frequency of medical acts from Site 2 X-ray radiological data.	99
7.5	Impact of replacing or removal of categorical outliers during preprocessing, trained and tested with Site 1 data.	101

8.1	Comparative performance and MAE analysis of ML models during baseline, preliminary preprocessing and categorical outliers removal with data from Site 2.	103
8.2	Comparative RMSE analysis of ML models during baseline, preliminary preprocessing and categorical outliers removal with data from Site 2.	103
8.3	Comparative performance and MAE analysis of ML models during manual numerical outliers removal with data from Site 2.	104
8.4	Comparative RMSE analysis of ML models during manual numerical outliers removal with data from Site 2.	105
8.5	Comparative performance analysis of ML models during categorical outliers removal and different numerical outliers treatment with data from Site 2.	106
8.6	Comparative MAE analysis of ML models during categorical outliers removal and different numerical outliers treatment with data from Site 2.	107
8.7	Comparative RMSE analysis of ML models during categorical outliers removal and different numerical outliers treatment with data from Site 2.	108
8.8	SHAP summary plot of XGBoost with non X-ray data of Site 2 treated with the best combination of treatments, to understand the impact of each feature on the predictions. .	112
A.1	Introductory question for Factors Weighting section. This question is made to reporting-prescribing physicians and prescribing physicians.	123
A.2	Introductory question for Factors Weighting section. This question is made to the three types of physicians.	124
A.3	Introductory question for Factors Weighting section on Reporting Areas. This question is made to reporting-prescribing physicians and reporting physicians.	124
A.4	Introductory question for Factors Weighting section on years of work. This question is made to reporting-prescribing physicians and reporting physicians.	125
A.5	Introductory question for Factors Weighting section on work methodology. This question is made to reporting-prescribing physicians and reporting physicians.	125
A.6	Table made for evaluating reporting physician category. Here, factors that influence the physician work as a reporting physician are outstanced.	126
A.7	Table made for evaluating exam category. Here, both reporting-prescribing and prescribing physicians evaluate the relevance of exam factors concerning their work.	127
A.8	Table made for evaluating patient category. Here, the three types of physicians evaluate the relevance of patient factors concerning their work.	128
A.9	Table made for evaluating resources category. Here, both reporting-prescribing and reporting physicians evaluate the relevance of resources concerning their work.	129
A.10	Table made for evaluating resources category. Here, prescribing physicians evaluate the relevance of resources concerning their work.	130

List of Tables

1	SDGs and contribution of the work.	iii
2.1	Overview of different methodologies for data quality assessment, outlining the name, steps and explanation.	6
2.2	Comparison of the advantages and disadvantages of each data quality assessment methodology.	7
2.3	Overview of imputation methods as repairing techniques, including their names, objectives, advantages and disadvantages.	13
2.4	Overview of deletion methods as repairing techniques, including their names, objectives, advantages and disadvantages.	14
2.5	Overview of python packages for data preprocessing, including their name, type, features, limitations and license.	16
2.6	Overview of python toolboxes for data preprocessing, including their name, type, features, limitations and license.	17
5.1	Mean importance ratings and variability of factors influencing medical report delays. . .	37
5.2	Structure of the dataset of each hospital at the time of collection.	40
5.3	Actual start and end time of RealizationShift.	40
5.4	Hyperparameters adjusted to Decision Tree.	42
5.5	Hyperparameters adjusted to XGBoost.	43
5.6	Hyperparameters adjusted to LightGBM.	43
5.7	Hyperparameters adjusted to CatBoost.	43
6.1	Missing values in non X-ray radiological data.	45
6.2	Number of unique categories per variable for both Sites.	46
6.3	Percentage of exams by FamilyMedicalAct per Site.	46
6.4	Percentage of origin per Site.	47
6.5	Model validation metrics of the raw non X-ray dataset.	48
6.6	Descriptive statistics for SchedulingTimeDeviation while preprocessing Non X-ray Radiological data of Site 1 and 2.	52
6.7	Descriptive statistics for TotalProcessTimeout while preprocessing Non X-ray Radiological data of Site 1 and 2.	53
6.8	Descriptive statistics for SchedulingTime while preprocessing Non X-ray Radiological data of Site 1 and 2.	53

6.9	Descriptive statistics for ExamDelay while preprocessing Non X-ray Radiological data of Site 1 and 2.	54
6.10	Descriptive statistics for WaitingForExam while preprocessing Non X-ray Radiological data of Site 1 and 2.	54
6.11	Model metrics comparison using temporal variables represented either as continuous numerical values or as cyclic transformations (sine and cosine) for hour, day of the week, month and shift in Site 1.	56
6.12	Model metrics comparison using temporal variables represented either as continuous numerical values or as cyclic transformations (sine and cosine) for hour, day of the week, month and shift in Site 2.	56
6.13	Model validation metrics after preliminary cleaning and feature creation for non X-ray data.	60
6.14	Treating categorical outliers in non X-ray by replacing the rarest categories with "Other". The percentage of medical acts affected is next to the threshold value.	62
6.15	Treating categorical outliers in non X-ray by deleting the rarest categories of MedicalAct. The percentage of medical acts affected is next to the threshold value.	63
6.16	Model validation metrics after detection performed by plotting scatter graphs of the features TotalProcessTimeout, ExamDelay, SchedulingTime, SchedulingTimeDeviation, WaitingForExam and PatientAge, followed by manual deletion.	66
6.17	Model validation metrics after IQR-based outlier detection followed by listwise deletion during preprocessing of cyclic variables (time, weekday, shift, and month).	75
6.18	Evolution of the number of outliers after listwise deletion with IQR-based outlier detection (factor=3).	76
6.19	Evolution of the number of outliers after listwise deletion with IQR-based outlier detection (factor=1.5).	76
6.20	Model validation metrics after IQR-based outlier detection followed by listwise deletion during preprocessing of continuous features (TotalProcessTimeout, ExamDelay, SchedulingTime, SchedulingTimeDeviation, WaitingForExam).	77
6.21	Model validation metrics after IQR-based outlier detection and capping at the IQR boundaries ($Q1 - k \times IQR$ and $Q3 + k \times IQR$) during the preprocessing of continuous variables (TotalProcessTimeout, ExamDelay, SchedulingTime, SchedulingTimeMedian, and WaitingForExam).	78
6.22	Model validation metrics after IQR-based outlier detection and capping at the percentile boundaries during the preprocessing of continuous variables (TotalProcessTimeout, ExamDelay, SchedulingTime, SchedulingTimeMedian, and WaitingForExam).	79
6.23	Evolution of the number of outliers after capping at the percentile boundaries with IQR-based outlier detection (factor=3, percentile=0.01).	80
6.24	Evolution of the number of outliers after capping at the percentile boundaries with IQR-based outlier detection (factor=1.5, percentile=0.01).	80
6.25	Evolution of outliers after winsorization with IQR (factor=1.5, percentile=0.1).	81
6.26	Model validation metrics after IQR-based outlier detection and capping at the percentile boundaries during the preprocessing of continuous variables (TotalProcessTimeout, ExamDelay, SchedulingTime, SchedulingTimeMedian, and WaitingForExam) with more extreme threshold values.	82

6.27	Model validation metrics after IQR-based outlier detection and imputation with KNN algorithm during the preprocessing of continuous variables (TotalProcessTimeout, ExamDelay, SchedulingTime, SchedulingTimeMedian, and WaitingForExam).	83
6.28	Evolution of the number of outliers after KNN imputation with IQR-based outlier detection (factor=3).	83
6.29	Evolution of the number of outliers after KNN imputation with IQR-based outlier detection (factor=1.5).	83
6.30	Model validation metrics after IQR-based outlier detection and imputation with MICE algorithm during the preprocessing of continuous variables (TotalProcessTimeout, SchedulingTime, SchedulingTimeDeviation and WaitingForExam).	85
6.31	Evolution of the number of outliers after MICE imputation with IQR-based outlier detection (factor=3).	85
6.32	Evolution of the number of outliers after MICE imputation with IQR-based outlier detection (factor=1.5).	85
6.33	Model validation metrics after IF-based outlier detection followed by listwise deletion during preprocessing of cyclic variables (time, weekday, shift, and month).	87
6.34	Evolution of the number of outliers after listwise deletion with IF-based outlier detection.	87
6.35	Model validation metrics after IF-based outlier detection followed by listwise deletion during preprocessing of continuous variables (TotalProcessTimeout, ExamDelay, SchedulingTime, SchedulingTimeDeviation, WaitingForExam).	89
6.36	Evolution of the number of outliers after listwise deletion in continuous variables with IF-based outlier detection.	89
6.37	Model validation metrics after IF-based outlier detection followed by KNN imputation during preprocessing of continuous variables (TotalProcessTimeout, ExamDelay, SchedulingTime, SchedulingTimeDeviation, WaitingForExam).	90
6.38	Evolution of the number of outliers after KNN imputation in continuous variables with IF-based outlier detection.	90
7.1	Missing values in X-ray radiological data.	93
7.2	Number of unique categories per variable for both Sites.	94
7.3	Percentage of origin per Site.	94
7.4	Model validation metrics of the raw X-ray dataset.	95
7.5	Descriptive statistics for SchedulingTimeDeviation while preprocessing X-ray Radiological data of Site 1 and 2.	97
7.6	Descriptive statistics for TotalProcessTimeout while preprocessing X-ray Radiological data of Site 1 and 2.	97
7.7	Descriptive statistics for SchedulingTime while preprocessing X-ray Radiological data of Site 1 and 2.	97
7.8	Descriptive statistics for WaitingForExam while preprocessing X-ray Radiological data of Site 1 and 2.	97
7.9	Model validation metrics after preliminary cleaning and feature creation for X-ray data.	98
7.10	Treating categorical outliers in X-ray data by replacing the rarest categories with "Other". The percentage of medical acts affected is next to the threshold value.	100

7.11	Treating categorical outliers in X-ray data by deleting the rarest categories of MedicalAct. The percentage of medical acts affected is next to the threshold value.	100
8.1	Model metrics variation when using close to optimal or optimal hyperparameters to test KNN and MICE techniques with IQR detection.	109
8.2	Model metrics variation when using close to optimal or optimal hyperparameters to test Listwise Deletion and KNN techniques with IF detection.	109
B.1	Quantitative analysis of the importance physicians attributed to factors related to the reporting physician.	131
B.2	Quantitative analysis of the importance physicians attributed to factors related to the exam.	131
B.3	Quantitative analysis of the importance physicians attributed to factors related to the patient.	131
B.4	Quantitative analysis of the importance physicians attributed to factors related to resources.	132
B.5	Quantitative analysis of the importance physicians attributed to factors related to time. . .	132
C.1	Counts of medical acts of Non X-ray radiological data of Site 1, before categorical outlier removal.	133
C.2	Counts of medical acts of Non X-ray radiological data of Site 2, before categorical outlier removal.	140
C.3	Counts of medical acts of X-ray radiological data of Site 1, before categorical outlier removal.	147
C.4	Counts of medical acts of X-ray radiological data of Site 2, before categorical outlier removal.	150

List of Acronyms

AI	Artificial Intelligence
CDQ	Complete Data Quality
CIHI	Canadian Institute for Health Information
CRISP-DM	Cross-Industry Standard Process for Data Mining
DICOM	Digital Imaging and Communications in Medicine
DQA	Data Quality Assessment
DWQ	Data Warehouse Quality
ELI5	Explain Like I'm 5
eSIS	Health Information Ecosystem
HIS	Hospital Information Systems
HL7	Health Level 7
IQM	Information Quality Measurement
IF	Isolation Forest
IQR	Interquartile Range
KNN	K-Nearest Neighbor
LIME	Local Interpretable Model-Agnostic Explanations
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MICE	Multiple Imputation by Chained Equation
ML	Machine Learning
MSE	Mean Squared Error
NHS	National Health Service
PDP	Partial Dependence Plot
RMSE	Root Mean Squared Error
SHAP	SHapley Additive exPlanations
TAT	Turnaround time
TDQM	Total Data Quality Management
WA	Weighted Average

Chapter 1

Introduction

The management of diagnostic and therapeutic procedures is essential to the effectiveness and efficiency of hospital care delivery. Prescription, screening, scheduling, performing and the preparation of medical reports are only a few of the stages involved in these procedures. The increasing availability of data in these processes presents considerable potential for the application of machine learning (ML) models that can anticipate and optimize different actions. This could improve patient services and expedite treatment processes in various medical specialties.

Radiology serves a fundamental role in the day-to-day life of a hospital by making diagnosis, treatment planning and follow-up care possible in almost every type of medical specialty. Imaging studies like X-rays, CT scans, ultrasounds and MRIs are often decisive, as they can uncover potential problems in the patient's condition that no other diagnostic and therapeutic procedure can. With imaging becoming more and more an integral part of follow-up and diagnosis, radiology services also saw dramatic increases in demand with consequent increased pressure on turnaround time (TAT) and hospital workflow [1].

As radiological medical exams are so important, hospitals have to make sure the time taken to produce the corresponding report with the result is as small as possible so that care can be delivered continuously and efficiently. If this is not the case, patients may experience negative consequences such as postponements of intended treatments and appointments, prolonged stays in the hospital and, in general, reduced satisfaction [2]. From a clinical workflow perspective, the TAT of radiology reports becomes a key performance indicator. When this time is minimized, physicians can make informed decisions more rapidly, which in turn facilitates faster interventions. For efficient resource and time utilization, the scheduling of follow-up appointments should consider the completion timing of relevant radiological reports. The need to synchronize the availability of reports and the scheduling of appointments is imperative to prevent unnecessary rescheduling and ensure optimal clinical resources efficiency.

When medical appointments are scheduled without confirmation that the corresponding imaging reports will be ready, it can result in situations where patients arrive at the hospital only to find that the necessary exam results are not available, which would render the medical appointment pointless. In critical specialties like oncology, delayed reporting can result in delayed interventions and extended diagnostic uncertainty [2]. For the hospital, when follow-up appointments are scheduled without verifying whether the relevant radiology reports will be ready, it can result in inefficient resource allocation, increased administrative burden and suboptimal patient follow-up. Reporting delays can have economic implications, such as increased hospital length of stay and redundant exams due to uncoordinated care, all of which place further strain on already limited healthcare budgets [1]. To address this, hospitals should implement integrated scheduling systems that take into account the expected completion times of diagnostic reports. This synchronization would ensure that follow-up appointments or procedures are only scheduled when all necessary diagnostic data is available, thus supporting seamless care continuity and better patient flow management.

The challenges related to delayed radiology reporting and misaligned scheduling are not new issues within

healthcare systems. In fact, these problems have persisted over the years and continue to significantly impact patient care and hospital operations. In 2012, António Taveira, president of the Portuguese National Association of Laboratories, emphasized delays in the delivery of exam results as a significant issue in the Portuguese National Health Service (NHS): he stated there is a clear and significant harm to Portuguese NHS users, who previously had 80% to 90% of results available within 24 hours, at most, within 36 hours; and now, these users face extended waiting times for blood collections, analysis and receiving results [3]. Ten years later, the problem persists and in some cases it got even worse due to COVID-19 pandemic. In the 2020 Portugal Oncology Registry, it is possible to see how the pandemic "has had a negative impact on cancer care and research, causing delays in screening, diagnosis and treatment" [4]. Unfortunately, nowadays, in Portugal and even outside, this is not only a reality for cancer treatment but also for other medical areas that rely a lot on diagnostic and therapeutic procedures, turning this problem into a global challenge in health systems. Thus, as these diagnostic delays frequently occur as a result of incomplete reporting, this poses a serious issue for NHS and thereby raises the need of implementing modern solutions that help in this matter. Given the massive amount of data involved in hospital systems, working with data science and ML might provide very valuable tools to ideally manage radiology workflow much better and preemptively anticipate delays in service, thus supporting more timely care.

1.1 Current Problem

Radiological exam reporting delays are a widespread challenge faced by healthcare systems worldwide. Reporting delays happen due to a variety of factors. These include hospital demand at the time of reporting, availability of support tools to help this process (e.g. speech recognition, transcription tools), complexity of the elements being analyzed in each exam (e.g. images), among others [5].

Hospital information systems (HIS) work by coordinating the various administrative, clinical and operational processes of a hospital. Although HIS ensures the management of a great variety of general processes, it is complemented by specific systems that manage more specialized medical areas (e.g. RIS - Radiology Information System, LIS - Laboratory Information System). Together, these systems, combined with people and other technologies, give shape to the health information ecosystem (eSIS) [6] that focuses on sharing, managing and using healthcare data securely and efficiently [7]. Nowadays, the majority of hospital systems do not incorporate mechanisms to forecast report completion, making it difficult to align appointments with the diagnostic information required, as clinicians and administrative staff cannot know in advance when the reports will be completed and available. However, inside eSIS, inevitably, there is a wide quantity of data that can help us understanding better the reasons behind the high time taken to deliver a report after an exam.

Hence, predicting report turnaround times is feasible. Advances in data science and ML offer promising tools capable of analyzing data and forecast reporting delays accurately. **However, a critical barrier to deploying such predictive models lies in the quality of the underlying data. Hospital data includes problems such as missing values, outliers, duplicates and standardization problems.** These data quality issues not only limit the ability to generate reliable predictions but also inhibit broader efforts to optimize healthcare processes. This data quality-centric problem drives the core of this dissertation.

1.2 Motivation and Relevance

Delays in the readiness of radiology reports can negatively affect patients, healthcare professionals and hospital operations. For the physician and the hospital, these delays cause increased workload, waste money on empty slots and, many times, clinicians are left with limited support for decision-making regarding confirmation of diagnoses or initiation of treatments. For the patient, who waits too much for a result, it can lead to not only prolonged anxiety, but also a delay in a possible treatment initiation,

or worse, the progression of an undiagnosed condition that in some cases means longer hospital stays. These problems highlight the importance of studying different preprocessing techniques to improve the accuracy of models that predict such delays.

Predictive modeling offers a promising direction to study reporting delays and enable optimal scheduling of follow-up appointments, aligning clinical demand with actual report availability. However, to build robust ML models, it is first necessary to address the inaccurate and partial data that often characterizes real-world healthcare datasets, as these issues significantly degrade the performance of ML models. As a result, it is crucial to identify and analyze these data gaps, in order to treat potential outliers and ensure they meet the requirements for reliable and accurate results. **This obstacle gives shape to this dissertation, which has, as its main focus, treating data, to later obtain improvements in a model that handles delays in reporting times of radiological exams.**

This study is therefore timely and necessary, as it explores a dual focus on data quality improvement and predictive modeling in the context of radiology. Its practical relevance is particularly significant: the solutions proposed can be directly applied to hospital data that later can support healthcare professionals in improving care coordination and optimizing resource management.

1.3 Goals

The main goal of this dissertation is to assess how different data cleaning and quality improvement strategies impact model performance. The model aims to estimate radiology report delivery delays to later support the alignment between report availability and clinical scheduling. Inside this major goal, there are several objectives to accomplish:

- Define a data structure aligned with the goal of predicting delays in delivering medical reports;
- Ensure the adequacy of the model features by integrating feedback from physicians, promoting an approach centered on the real needs of medical practice;
- Establish collaborations with key hospital stakeholders (ByMe) to ensure the collection, viability and relevance of the data used;
- Identify data problems and prepare critical variables that represent the operational processes associated with reporting;
- Develop data preprocessing techniques that contribute to the robustness and reliability of predictive analyses;
- Systematically evaluate the impact of different data quality improvement approaches on the performance of ML models.
- Demonstrate the value of different data treatment techniques in building effective predictive solutions for hospital contexts, with an emphasis on the interpretability and relevance of the identified predictive variables.

1.4 Contributions

This research includes investigating and putting into practice strategies for treating data in ways that enhance its quality and completeness when ingested by the model. The contribution of this work concerns the enhancement of the efficiency of a ML model by assessing the impact of different combinations of outlier detection and imputation methods as well as initial data preparing. For that, incorporated cyclical transformations for time-related features (day of the week, hour, month, shift) to capture their inherent

periodic nature and enhance the model's ability to learn from temporal patterns. Regarding categorical outliers, two algorithms were designed, one that excludes based on a threshold and another that groups less significant categories. Regarding numerical outliers, algorithms of Listwise Deletion, Winsorization, K-Nearest Neighbors (KNN) and Multiple Imputation by Chained Equations (MICE) Imputation, detected with Interquartile Range (IQR) and Isolation Forest (IF), have been developed.

Overall, this dissertation addresses a concrete real-world problem that proposes a practical and transferable framework that can be adapted and applied in similar healthcare settings to improve operational efficiency and patient care.

1.5 Document Structure

Alongside the introduction, this article is organized in different chapters and at the end of the document, the list of bibliographic references that support this article is available as well as a list of complementary annexes for a good understanding of the entire report.

Chapter 2, Background, presents the most recent and impactful methods behind data quality pipelines and ML evaluation metrics are described alongside with different detection and repairing techniques.

Chapter 3, Literature Review, describes the research strategy and relies on existing work related.

Chapter 4, Methodology and Resources, details the overall approach, the adopted strategy, and ethical considerations involved in the use of data.

Chapter 5, Foundation and Study Preparation, covers the requirement gathering process, including questionnaire results, dataset characterization, and some important decisions regarding the evaluation of data cleaning effectiveness, namely the choice of ML models.

Chapter 6, Non X-ray Radiological Data, focuses on the analysis of non X-ray imaging data, describing the understanding and cleaning of data, outlier detection and treatment, and the resulting performance variations.

Chapter 7, X-ray Radiological Data, mirrors the structure of Chapter 6, but for X-ray data. It is **important to note** that these two chapters, Chapters 6 and 7, present the analysis and corresponding results separately for each type of data. This organizational strategy was considered the most appropriate to clearly distinguish between the two data types and to enhance the clarity and readability of the work. This split will also be better explained throughout the report.

Chapter 8, Comparative Analysis, discusses and compares the methodological choices, analyzes the impact of different outlier treatment approaches and studies feature importance.

Chapter 9, Final Considerations and Future Work, summarizes the main findings, highlights difficulties and limitations encountered and suggests possible directions for future research.

Chapter 2

Background

This chapter provides the theoretical foundation necessary to understand the work developed in this dissertation. Each section presents recent and relevant methodologies within key areas. The first section explores foundational concepts in data analysis and CRISP-DM framework. Then, the development, structuring, and validation of questionnaires in investigation context is analyzed in order to help in future work. Moving on, techniques for detecting and treating data quality issues are addressed as this is the main obstacle to be solved. Subsequently, reviews of python tools used for building data preprocessing pipelines and evaluation metrics used in ML are summarized. Finally, model interpretability and explainability frameworks are assessed.

2.1 Foundations of Data Analysis Methodologies

In order to start the work of studying data quality, it was important to understand which steps are normally taken to assess it. Multiple studies use different approaches to keep track of the quality of data. In this section are condensed some of the most used data quality assessment methodologies and the cross-industry standard process model for data mining, the CRISP-DM Framework.

2.1.1 Data Quality Assessment Methodologies

Many of the studies do not follow an already founded theory but rather use an order of actions that seems to make the most sense. Priestley et al. [8], for example, considers that consultation with pertinent stakeholders must be the first step in defining the ML data requirements - **Dataset Use Case and Design**. In this project's topic of research, this step ensures that physicians help deciding which features or columns of the dataset makes more sense to include. This will increase the likelihood that the output of the ML system will match the specific needs of the problem. Afterwards, regarding **Data Acquisition**, Priestley highlights that by recording the data collecting procedure (for example, with datasheets or checklists), metadata clarity can be enhanced. After documentation, **Data Cleaning and Preprocessing** activities such as deduplication, outlier removal, consistency checking, anonymization and imputation of missing values frequently come after data collection. In order to prevent knowledge leakage and overly optimistic model performance during testing, it is imperative that dependencies between features be reported. Reducing the complexity of the features and eliminating redundant cases can increase the uniqueness of data entries and characteristics. Finally, **Data Maintenance** includes keep documenting on data content (adding labels), have a maintenance plan and a mission statement, that accurately states the purpose of the dataset. In the context of this project, identifying problems and keeping them documented will support future work.

This shows a simple yet effective example of order of events regarding data quality assessment. But, when

starting cleaning and preprocessing, which is the right way to do it? There are several methodologies that help us to find inconsistencies and errors. Batini et al. [9] examined thirteen approaches for evaluating quality of data and categorized them into four groups. Of them all, those below were selected for analysis as they present scientific interest in this topic of study.

- Comprehensive methodologies: Total Data Quality Management (TDQM) and Complete Data Quality (CDQ).
- Operational methodologies: Data Warehouse Quality (DWQ), Canadian Institute for Health Information (CIHI), Data Quality Assessment (DQA) and Information Quality Measurement (IQM).
- Data-driven methodologies: CDQ.
- Process-oriented methodologies: TDQM and DWQ.

In Table 2.1 each step of each methodology is explained.

Table 2.1: Overview of different methodologies for data quality assessment, outlining the name, steps and explanation.

Methodology Acronym	Steps	Explanation	Ref
TDQM	Definition; Measurement; Analysis; Improvement.	The fundamental objective of TDQM is to extend the principles of Total Quality Management (TQM) to data quality. The definition phase identifies data as information products and establishes information quality requirements (describes, analyzes and maps the processes related to the production of information from raw data by information systems); the measurement phase assesses data quality using specific metrics; the analysis phase investigates the root causes of data quality issues; and the improvement phase develops and implements solutions to enhance data quality.	[10] [11]
DWQ	Definition; Assessment; Analysis; Improvement.	This cycle examines the connection between design decisions and quality goals when constructing data warehouses. The definition phase sets quality requirements, identifying dimensions and desired levels. The evaluation phase measures quality, storing results as metadata, including requirements, achieved metrics and causal dependencies. The analysis phase uses these results to trace and address root causes of quality issues. The improvement phase focuses on enhancing quality, though the DWQ methodology lacks specific details for this step.	[12]
CIHI	Analysis. Measurement; Documentation.	The cyclic CIHI methodology was developed to assess and improve data quality at the Canadian Institute for Health Information. The Data Quality Framework is defined using standardization of information on data quality, development of a common strategy for assessing data quality and defining a work process by identifying priorities and improvement procedures. The analysis of the most used data is carried out in three steps: analysis, evaluation and documentation of data quality.	[13]
DQA	Subjective and objective data quality measurement; Comparison; Improvement.	DQA classifies data quality metrics into two main categories: subjective and objective metrics. Subjective metrics capture stakeholders' perceptions, needs, experiences and they are often obtained through questionnaires. Objective metrics are then classified into task-independent (without context) and task-dependent (with context) and each of these two is divided in three classes: simple ratio, min or max value and weighed average. After metrics defined, it is time for measuring, comparing and finally, document .	[14]
IQM	Assessment planning; Assessment configuration; Measurement; Follow-up activities.	IQM is a valuable methodology for web information systems. The assessment planning focus on achieving an information quality audit plan. Then, in assessment configuration and measurement phases , it is important to have a list of indicators associated with information quality criteria and use them to have measurement results. Finally, follow-up activities involves keep controlling and making adjustments.	[15]
CDQ	State Reconstruction; Assessment; Improvement.	The CDQ methodology focus on three simple concepts: completeness, simplicity and flexibility. In the state reconstruction phase, organizational units, processes, services and data are mapped and modeled using matrices. Then, during assessment , establishes new quality standards that must be met in order to enhance process qualities. Finally, towards improvement , a cost-benefit analysis is used to determine the best improvement approach.	[16] [17]

Regarding advantages and disadvantages, Table 2.2 provides an overview:

Table 2.2: Comparison of the advantages and disadvantages of each data quality assessment methodology.

Methodology Acronym	Advantages	Disadvantages	Ref
TDQM	Comprehensive, with guidelines from requirements analysis to implementation; Uses a methodology and dimensions to describe information production processes. The cyclical approach works well with the dynamics of ML, which require constant adjustments.	It can be complex to implement due to its scope.	[10] [11]
DWQ	Considers the subjectivity of the concept of quality; Defines metadata corresponding to different quality objectives.	Mainly focuses on the assessment phase and offers little support for the improvement phase.	[12]
CIHI	Useful for managing large heterogeneous databases.	Validating assessment criteria can be complex and time-consuming.	[13]
DQA	Provide general principles for defining data quality metrics.	Focus on defining metrics and does not offer complete guidelines for improving data quality.	[14]
IQM	Flexible enough to span healthcare data and ML.	Specifically designed for web data; May not be suitable for structured data or information systems outside the context of the web.	[15]
CDQ	Complete methodology that addresses technical and economic aspects of data quality. Flexible, adaptable to different contexts and types of data. Emphasizes requirements gathering and context awareness;	It can be complex to implement due to its scope.	[16] [17]

2.1.2 CRISP-DM Framework

After exploring methodologies for assessing data quality, it is essential to understand the standard process of data mining. CRISP-DM (Cross-Industry Standard Process for Data Mining) is a cross-industry standard process model for data mining [18]. It is a mature approach that is widely accepted in data mining projects, including those using ML algorithms. The CRISP-DM methodology consists of the six phases presented in Figure 2.1 and explained below:

- **Business Understanding:** The project begins with defining the project objectives. In an application, this involves transforming the business objective into a specific data mining problem [18].
- **Data Understanding:** At this stage, hypotheses are formulated about hidden information relevant to the project objective based on experience and qualified assumptions [18].
- **Data Preparation:** The engineer collects the relevant data and prepares it for the data mining task, which includes preprocessing (reduction, filtering) and feature generation. A variety of techniques and tools can be used for ETL (Extraction, Transformation, Loading) processes to obtain a dataset with the most relevant attributes [18] [19].
- **Modeling:** A data mining workflow is constructed to find the desired parameter settings for the selected algorithms and perform the data mining task on the preprocessed data [18].
- **Evaluation:** The trained model is tested on real datasets and the results are evaluated according to the underlying business objectives [18] [19].
- **Deployment:** After successful evaluation, the trained model is deployed into production [18].

Although not all phases of CRISP-DM apply to the scope of this project, this methodology is a good foundation to start exploring and understanding the normal flow of data mining.

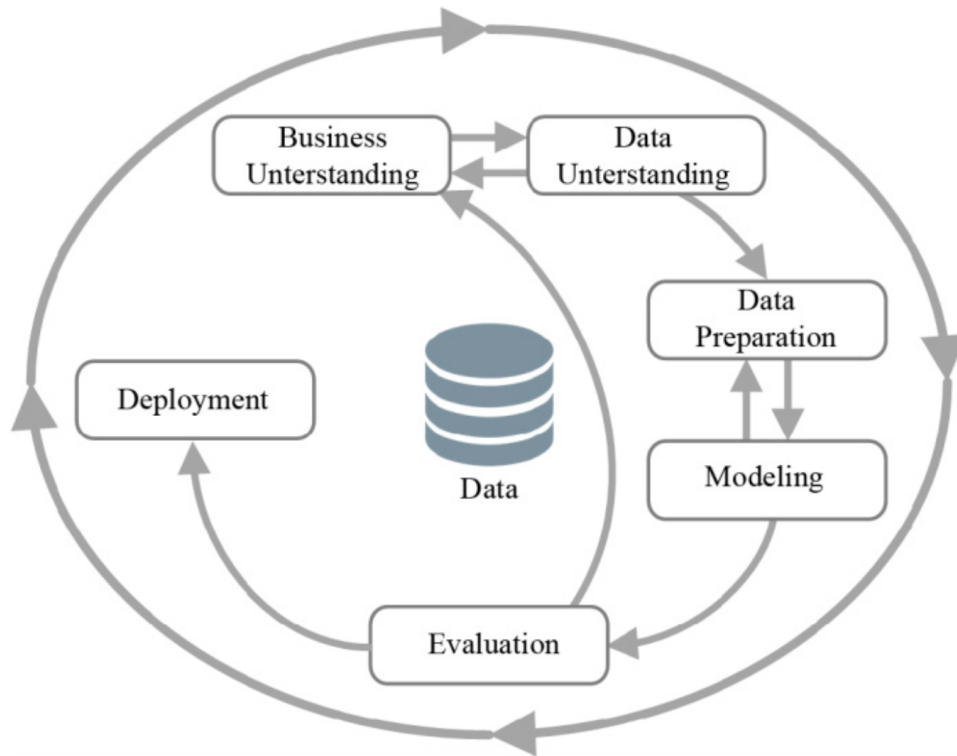


Figure 2.1: The CRISP-DM process [18].

2.2 Insights Through Questionnaire Structuring and Assessment

The baseline dataset design plays a crucial role in deriving meaningful insights and achieving the objectives of the study. This section describes good practices in the development of a questionnaire, as it was tailored to gather relevant information for the dataset design. Additionally, it describes the structured approach used to ensure the questionnaire's reliability and validity. The subsequent subsections elaborate on the creation and organization of the questionnaire, as well as the methods employed to analyze and evaluate the responses.

2.2.1 Development and Structuring of a Questionnaire

Gathering information for the construction of a well-structured dataset can be accomplished with various methods. One of the most simple ways of doing this is through a questionnaire. Hence, it was important to **understand how can a questionnaire be well-drawn and focused on the user**.

Whenever looking forward to construct a questionnaire that helps to extract information, it is important to take into account various factors such as suitability, unambiguity, impartiality and intelligibility [20] [21].

Stone [20], tried to make a description of each step whenever constructing a questionnaire. The first step highlighted was the need to **define the necessary data**: this implies articulating different goals in order to help visualizing the intended results. Then it is important to **create a list containing specific information related to the topic**, based on literature or experience of others and try to validate if each item of the list is related to, at least one, one of the project's goal. Next, it is time to **design the format**

of each question and this depends on the nature of the data (quantitative or qualitative). After the format had been decided, it is time to **compose the questions** and this must be done using brevity and precision, in order to avoid ambiguities. Finally, it is crucial to think about the **layout** (normally it is better to use lighter colors) and the sequence of questions (go from general to specific). With everything ready, a small sample testing should be the next step, followed by the evaluation of these responses, to understand what is missing and what has not been well-understood. After correcting, the inquiry is ready to be released.

Between the process of creating a questionnaire, the process of evaluation and analysis of the responses must always be in line of thought. **Codification** helps in the construction of these inquiries by transforming the answers collected into numeric data or other categories so that it can be analyzed. This can occur with **prior coding** or **autocoding**. The first one guarantees that the response categories are defined before the questionnaire is available; the second one occurs when the respondent himself codes the answers while filling out the questionnaire [20].

Many scales are used to preset the weight of different categories to be assessed in a questionnaire. *Likert Scale* (Figure 2.2), a well-known scale developed in 1932 by Rensis Likert is an example of that. This scale, with already various versions, consists of 5 or 7 ordinal scales [22].

TABLE		TYPICAL LIKERT SCALES		
1	2	3	4	5
Never	Rarely	Sometimes	Often	Always
Completely disagree	Disagree	Neutral	Agree	Completely agree

Figure 2.2: Illustration of two examples of *Likert Scales* used for data collection [22].

Moreover, it is important to acknowledge how the quality of each element directly impacts the quality of the data collected. For that, Yaddanapudi et al. [21] reinforced the need to comply with the following points:

- Clear and intelligible language adapted to the type of user who will respond;
- Unbiased questions that do not lead to distorted results;
- Well defined questions type, considering their impact on the quantity and quality of responses;
- Validated reliability ensured through different testing.

A well-established order of steps in the construction of a questionnaire, combined with a well-designed question structure, will result in maximizing response rates, reducing the distortion in estimates caused by the lack of response from some participants and consequently, increasing the accuracy of estimates [23].

2.2.2 Analysis and Assessment of Questionnaire Responses

When analyzing data obtained from a form, two types of tests can be done, depending on the type of data: parametric tests and non-parametric tests. Parametric tests evaluate quantitative data as for non-parametric tests, evaluates qualitative data. The first one uses a normal distribution and the second one a skewed distribution [24]. Due to the increased likelihood of detecting a statistically significant difference between groups (when one actually exists), parametric tests are generally more powerful than non-parametric tests [22].

As for evaluating quantitative data, some tests are available:

- T-Test focuses on small sample size (less than 30) and compares the mean of 2 samples from 1 population. In the context of the project, this test can be relevant to check if there is a significant difference between the evaluations of two specific features [24].
- F-Test focuses on small sample size (less than 30) and compares the variance of 2 samples from 1 population. In the context of the project, this test can be relevant to evaluate whether the variances of responses for different features are significantly different [24].
- ANOVA (Analysis of Variance) does not have a fix limit of sample size and compares the mean of more than 2 samples of the same population or more than 1 population. As so, ANOVA is an extension of T-Test. ANOVA allows the comparison of more than two groups at the same time to determine whether a relationship exists between them. In the context of the project, this test can be relevant to assess whether the relevance averages assigned to various features differ significantly from each other [24].
- Pearson's Correlation test varies from -1 to 1 and assesses the linear relationship between two continuous variables. In the context of the project, The Pearson's Correlation test can be an excellent tool to identify similar patterns between the responses of different physicians [24].

2.3 Data Preprocessing

Data preprocessing and data processing are closely related concepts that play distinct but complementary roles in data analysis. **Preprocessing** refers to the initial stage where raw data is prepared and cleaned to ensure it is suitable for subsequent analysis or modeling. This involves tasks such as handling missing values, correcting inconsistencies, normalizing or standardizing numerical variables, encoding categorical data and splitting the dataset into training, validation and testing sets. The goal of preprocessing is to transform messy, unstructured data into a structured and usable format. Once the data is ready, the **Data Processing** phase will focus on using the preprocessed data to train the ML model.

While looking for data preprocessing methodologies, it was important to understand how does this phase works. In Figure 2.3, it can be seen that the first step is to find detection techniques that help us finding the problems. After identifying the issue it is fundamental for the process to correct it with a repairing technique [25].

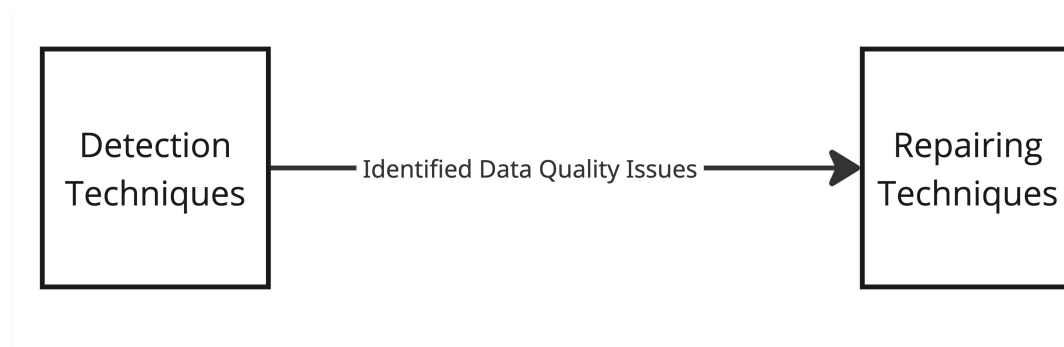


Figure 2.3: Overview of the data preprocessing phase, highlighting its two stages.

2.3.1 Detection Techniques

The detection stage in a data cleaning process refers to the process where data anomalies that could compromise the quality of the dataset are identified. This stage is crucial to find problems within the data, such as missing values, outliers, duplicates, or inconsistencies. The quality of the treatment depends heavily on the detection stage.

Missing Data

A number of methods have been put forth to identify missing values in datasets. Performing exploratory data analysis, which entails **visualizing the data and discovering missing values**, is one popular method [26]. This can be accomplished by using methods that can assist in locating patterns and outliers in the data, such as box plots, scatter plots and histograms. However, this missing data maps can't tell why the data is missing, for that statistical methods are needed.

Frequency analysis can provide statistics on the percentage of missing values for each variable. Data frequency analysis is just the first step in detecting and managing missing data. However this analysis focuses only in the quantity of missing data rather than the underlying patterns behind it.

Statistical techniques can also help to find these missing spots. Calculating percentages of missing values on relevant variables can reveal the extent of missing data [27]. This helps assess the reliability of the data collection process. Tabulating patterns of missing data or graphical tools such as UpSet charts can help identify missing values. UpSet plots are particularly useful for exploring missing data patterns in multivariate datasets [28]. Moreover, statistical methods like logistic regression or cross tabulation can help us understand the reason behind these missing values and help uncover potential patterns impossible to find with simple frequency or visualization analysis. These three concepts below explain the reasons and how absent data mechanisms work:

- **Missing completely at random (MCAR):** This absence pattern results from a process completely not related to any variables of the dataset, or from, as the name says, a completely random process [27].
- **Missing at random (MAR):** MAR is more general than MCAR. This absence pattern results from a process that is not dependent on the absent variable itself, but rather on other data that is seen in the dataset [27].
- **Missing not at random (MNAR):** This must be the hardest to handle since in reality it will be challenging to determine the possible cause of the absence [28]. Here, the probability that a variable is absent depends on the value of the missing data variable. [27].

Outliers

Outliers detections methods are divided in two main group: univariate methods, which analyze one variable at a time, and multivariate methods, which examine two or more variables together to identify unusual patterns.

Statistical methods like Interquartile Range (IQR) and Z-scores are inserted in the first group. IQR defines outliers as any data points that fall outside the range of 1.5 times the IQR below the first quartile or 1.5 times the IQR above the third quartile. Although 1.5 is the default multiplier, alternative factors can be used depending on the context. The choice of this factor is very important since it may either miss important outliers or flag too many normal points as anomalies. Z-scores, which relies on the assumption that the data is normally distributed, calculates how many standard deviations a certain data point is from the mean value. In skewed distributions z-score can lead to inaccurate identification of outliers. [25].

In the second group, multivariate methods, there are **techniques based on ML**, that have algorithms like DBSCAN clustering or Isolation Forest that help detect outliers by grouping similar data points into clusters [25]. This methods have more computational complexity than univariate methods.

For outlier detection, there are also some toolboxes known to help with this goal. PyGOD supports a broad range of popular graph-based techniques (Figure 2.4), such as sampling, mini-batch processing and deep models like deep learning and sampling [29]. For scalable outlier detection on multivariate data, PyOD is an open-source Python toolbox that gives researchers and practitioners access to a variety of

algorithms through a single, thoroughly defined API [30]. COPOD, which is implemented in PyOD, is a parameter-free, highly interpretable and computationally efficient outlier detection algorithm [31].

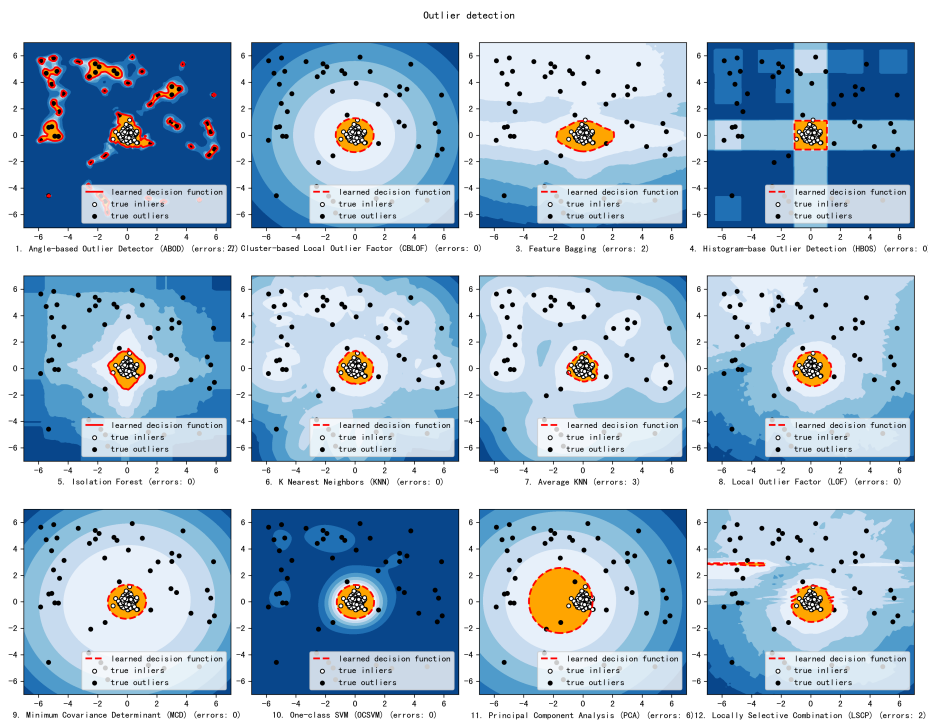


Figure 2.4: Graphical results generated by the PyOD Toolbox for outlier detection [30].

Data Duplicity

Direct comparison can be the first attempt in detecting data duplications. However, while effective for small datasets, it becomes computationally expensive as the dataset size increases. Hence, different techniques have been developed to help in those cases [32].

Blocking technique helps by reducing the number of comparisons needed by dividing the dataset into smaller blocks that are only compared with each other. **Windowing technique** sorts records based on key fields and compare only records within a fixed-size "window". Both techniques restrict the number of comparisons to increase efficiency, but this can lead to the omission of real duplicates that, because they are in different blocks or windows, are not compared [32].

Looking at other approaches, with an average detection rate of 88 out of 112 entries in a restaurant dataset, the **Levenshtein distance technique** was capable of effectively identifying duplicate records in databases [33]. This method, with a python library available, can enhance the detection of these duplicated values. However, the comparison between all possible pairs using Levenshtein can be computationally intensive.

In addition to the previous methods, duplicate entries may also show up as outliers in a dataset because they provide unusual concentrations of the same value. PyOD, as presented in the previous section, can assist in identifying these possible duplicates [30].

Standardization Problems

Standardization guarantees uniformity in the dataset's structure, which is especially important for health-care data. Errors during analysis and modeling are avoided, for example, by standardizing date formats or making sure medical diagnoses are labeled consistently.

When looking at a dataset, previously established rules can help us defining what is an inconsistency and what is not. For that, **initial rules on what is acceptable and what is not** must be defined. For example, if one of the rules states that a patient's age must be greater than 0 and less than 120, any value outside that range may be flagged as an inconsistency [34]. Another good thing to have is **cross-validation between fields**. This will guarantee that, for example, a record where the exam start timestamp is earlier than the date of birth may be flagged as inconsistent [35] [34] [36]. Regarding the validation of units and formats, it is essential to ensure that basic units such as time (seconds, minutes, hours) are used consistently, as well as, ensure that date format is consistent in every cell.

2.3.2 Repairing Techniques

Repairing the dataset comes next when problems have been identified. This step involves applying a combination of methods such as standardization adjustments, deduplication, imputation and deletion, which are not mutually exclusive but rather complementary. Using these techniques together if needed allows for a more comprehensive and balanced data cleaning process.

Missing Data

Most of the times, imputation procedures come from the need to fill missing values in a dataset or to treat an outlier. Imputation methods for missing values are divided in four: traditional and statistical, ML, deep learning and hybrid imputation methods [37]. Table 2.3 describes some of the most well-known.

Table 2.3: Overview of imputation methods as repairing techniques, including their names, objectives, advantages and disadvantages.

Imputation Procedures	Goals	Advantages	Disadvantages	Ref
Mean Imputation Technique	Fills the missing values with the mean of the existing data. Used for numeric data.	Very low computational complexity; Fast and efficient; Performs best when data is MCAR.	Not robust to outliers; Introduce bias if a pattern is found in the missing data; Inappropriate for MNAR data.	[27] [28] [38] [39]
Median Imputation Technique	Fills the missing values with the median of the available data. Used for numeric data.	Robust to outliers; Suitable for data with skewed distributions.	Worst when attributes have normally distributed trends; Distortion of variance and covariance.	[38]
Mode Imputation Technique	Fills the missing values with the mode of the available data. Used for categorical data.	Very low computational complexity; Fast and efficient;	Ignores relationships between variables; Underestimation of variability.	[28] [38] [39]
KNN Imputation Technique	Fills the missing values based on the k nearest neighbors, based on Euclidean or Manhattan distances from the nearest data points in the respective feature space.	Considers the relationships between variables; Performs best when local trends exist in the data; Preserves data structure.	More complex and computationally expensive; Challenges when using with large datasets; Inappropriate for sparse data.	[38] [39]
MICE Imputation Technique	Generates values for each variable individually, based on its relationship with other variables in the dataset.	Preserves relationships between variables; Flexible and adaptable to MCAR, MNAR, or MAR; Most suitable when there are dependencies between attributes or the dataset has high redundancy.	More complex and computationally expensive; Less effective in situations where it relies on attribute dependencies that are weak or absent.	[28] [38] [40]

Outliers Treatment

Outlier treatment can follow one of four paths. The most simple one is to **remove the outlier** by deletion [25], these methods are explored in Table 2.4. Secondly, as previously discussed and summarized in Table 2.3, once outliers are identified they can be corrected through appropriate imputation techniques [25]. **Data Transformation** is the third option, by applying logarithmic or other scaling transformations, the influence of outliers can be reduced (but not eliminated). However, after transformation, the data is no

longer in its original scale, which can make it difficult to interpret [39]. Finally **Winsorization**, consists of limiting outlier values to an "acceptable" value within a predefined range, such as for example the 1st and 99th percentile. Although this can help to reduce the impact of extreme values, it can also mask real and relevant variations [41].

Deletion Methods

Just as outliers can be replaced with other plausible values, deletion is also an option. Even though there is little agreement on optimal methodologies for treating incomplete data [27], there are some more used described in Table 2.4.

Table 2.4: Overview of deletion methods as repairing techniques, including their names, objectives, advantages and disadvantages.

Deletion Methods	Method	Vantages	Major Problems	Ref
Listwise	Prior to analysis, the dataset is cleared of each observation (row) that has at least one missing value in any variable. In other words, the entire row is rejected if one or more variables in the row contain missing values.	Simplicity: Easy to implement and understand.	Eliminates real data from incomplete observations; Biased under MAR and MNAR.	[27]
Pairwise	Instead of removing a whole row when one or more variables are absent, the analysis considers all of the data that are available for each pair of variables independently. This implies that different analyses may use different numbers of observations.	Maximizes data usage: Uses all available observations for each calculation.	Inconsistent estimates: distinct subsets mixtures are represented by distinct relationships; Biased under MAR and MNAR.	[27]

Deduplication

After identifying duplicate entries in a dataset, the next step is deduplication, the way to repair or merge those duplicates [25]. Normally, the techniques used to find these issues are specialized functions.

Most reliable value selection refers to choosing the most trustworthy value when merging partially duplicated rows. The selection is based on factors such as the source of the data, the date it was updated, the quality of the data or other context factors. Usually, **for numeric properties**, the average of the values are calculated; **for categorical properties**, the most common value is selected [25].

There are also several tools that implement the **Levenshtein distance** and the **Jaro-Winkler distance**, two methods for deduplication. ¹*Duke*, a duplicate detection tool, is available as an open-source project on GitHub and uses Jaro-Winkler distance to correct these values [42].

Standardization

Finding discrepancies between data in a dataset is common and the methods to find them were already described. Standardization is crucial for ensuring that the same terms within the dataset are being used and establishing uniform formats (e.g. date entry). After detecting abnormalities, data transformations must happen to uniformize. This includes guaranteeing that identical variables have the same type (all dates stored as datetime objects or all categorical variables as categories) and harmonize categories that might appear under multiple labels (such as gender being recorded as "M" or "Male").

¹<https://github.com/larsga/Duke>

2.4 Data Mining Tools

In the context of a data quality pipeline, data mining tools can be used for various crucial tasks. For data preprocessing, these tools assist in cleaning incomplete data and correcting inconsistencies to ensure data quality. During the initial data exploration, they help identify patterns or relevant variables that may contribute to the effectiveness of your model. Mikut et al. [43] analyzed data mining tools, tracing their historical evolution from statistical analysis programs on mainframes to modern cloud-based solutions. It categorizes these tools into nine types based on criteria such as user groups, data structures, data mining tasks and others. Of the nine types of tools outlined, three stand out as particularly relevant for this project:

- **Data mining libraries (LIBs):** integrating these libraries into Python code can help to develop a pipeline, by customizing specific functions to address the quality of hospital data. Inside of these there are **python packages** and **libraries** (a package can have multiple libraries; a library has a collection of functions);
- **Mathematical packages (MATs):** tools in this category provide support for advanced numerical routines, which can be utilized later for data validation or modeling.
- **Integration packages (INTs):** combine various open-source algorithms, providing a platform for the development and comparison of methods. Here, there are a set of tools called **python toolboxes**, often with a graphical interface, that simplifies data mining and can help identify errors.

One example of LIBs is the Sakdas Package. Sakdas is a Python module for managing data quality that is open source. Sakdas is divided into three main parts: data profiling, data quality auditing and data visualization. Furthermore, sakdas uses three dimensions for data quality metrics: completeness, integrity and consistency [44]:

- **Completeness** indicates how complete a dataset is in terms of meeting customer expectations and having the necessary features: it has in account the number of records, number of columns, number of blank columns and number and percentage of complete records.
- **Integrity** is the extent to which data conforms to data relationship standards, guaranteeing a comprehensive, consistent and legitimate depiction of ideas: it has in account number of duplicate data, number of records after removing duplicate data and number and percentage of missing data.
- **Consistency** evaluates how well data matches a comparable set of data that was typically generated throughout time using the same procedure or under comparable circumstances: data length and data pattern percentage.

This package uses other Python libraries like Pandas for core data analysis, NumPy is used for efficient numerical computation and Matplotlib for data visualization. These two and others are summarized in in Table 2.5.

Alongside with these data mining libraries, there are also different algorithms, commonly called toolboxes, that can help in preprocessing tasks. One example is PyOD, a python toolbox for the treatment of outliers, that directly depends of different libraries: NumPy, SciPy and Scikit-learn (also explored in Table 2.5). In Table 2.6 other examples of toolboxes are summarized.

Table 2.5: Overview of python packages for data preprocessing, including their name, type, features, limitations and license.

Package	Type	Features	Limitations	License	Ref
NumPy	Open Source	Provides high-performance N-dimensional array objects; Provides tools for integrating C/C++ and Fortran code; Provides tools for linear algebra, Fourier transform and random number generation.	May be slower than native code for some operations; Requires familiarity with matrix and indexing concepts.	BSD	[41]
Matplotlib	Open Source	Creates high-quality 2D charts of various types (line charts, scatter plots, histograms, bar charts and pie charts); Allows customization of all aspects of charts, including titles, labels, legends and colors; Can be used interactively and programmatically.	Creating complex 3D plots can be difficult; Requires familiarity with the matplotlib API.	Matplotlib	[41]
Pandas	Open Source	Provides high-performance and easy-to-use data structures for data analysis, such as Series and Data Frames; Provides tools for loading, cleaning, transforming, aggregating and visualizing data; Can handle data from a variety of sources, including CSV and Excel files, SQL databases and JSON data.	Can be slow to process very large datasets; Requires familiarity with the pandas API.	BSD	[41]
Scikit-learn	Open Source	Integrates a wide range of cutting-edge ML algorithms. Easy to use, with a consistent API.	May not be suitable for large scale problems; Reliance on NumPy for matrix operations can limit performance in some cases (e.g. k-means).	BSD	[41] [45]
Data Cleaner	Commercial tool with free version	Completeness analysis; Unique key verification; Time/date analysis; String analysis; Numerical analysis; Referential integrity.	Free version with limited functionality compared to the commercial version.	MIT	[46]
SciPy	Open Source	Complements NumPy library by providing additional modules for optimization, interpolation, integration and statistics.	Not easy to learn for data science beginners.	BSD	[47]
Great expectations	Open Source	Data validation through "expectations" that function as unit tests; Creating custom indicators for specific rules.	Need to read the entire table contents to run the indicators; It depends on technical and business knowledge to choose the right indicators.	Apache 2.0	[48]
Sakdas	Open Source	Data profiling; Data quality auditing; Data visualization.	The sources do not mention any specific limitations of Sakdas.	BSD	[44]

Table 2.6: Overview of python toolboxes for data preprocessing, including their name, type, features, limitations and license.

Toolbox	Type	Features	Limitations	License	Ref
nPYc-Toolbox	Open Source	Targeted preprocessing module with reporting and data consistency checks; Exploratory data analysis; Export of processed datasets; Tools for visualization.	Fewer algorithms when compared with other toolboxes.	MIT	[49]
PyOD	Open Source	Variety of algorithms for outlier detection (20+): classical and neural networks.	Neural networks require Keras (open source library that provides an interface for building and training neural networks).	CC-BY 4.0	[30]
PyGOD	Open Source	Supports a variety of outlier detection algorithms, including the latest GNN-based ones.	The sources do not mention any specific limitations of PyGOD.	BSD 2-Clause	[30]
SEAOP	Open Source	Multi-resampling strategy to reduce bias; Interactive visualization tools; Tools for data normalization; Addresses the problem of missing values.	Potential for overfitting; Study performed only on outlier detection in quantitative proteomics data.	MIT	[50]
Tensor Flow Data Val-idation	Open Source	Includes a data analyzer component, which calculates statistics on large amounts of data; Includes a data validator, which finds anomalies in the data; Includes a data visualizer, which provides visualizations of the statistics, schema and anomalies.	Features a limited schema definition language that focuses on dependable, highly precise alerts by capturing the data limitations for the majority of our consumers.	Apache 2.0	[51]
Dask	Open Source	Offers many of the same statistical and analytical features as the pandas and NumPy libraries; Divides data into smaller blocks to perform calculations in parallel.	Dask adds overhead when managing tasks, so for small datasets or simple operations, it can be slower than Pandas.	BSD	[52]
Open Refine-Py	Open Source	Toolbox with python interface that allows to interact with the OpenRefine tool programmatically; Regarding data cleaning, allows error correction, duplicate removal and format standardization; Regarding data transformation, it is possible to split and merge columns, add calculated columns and transpose rows and columns.	Depends on a running OpenRefine server.	BSD	[53]
DataRobot	Commercial tool	Support to various algorithms like Scikit-learn and Tensor Flow; Complete preprocessing and implementation documentation for all models; Explains the forecast's justifications at the forecast level, target forecast probability and decision-making criteria.	Requires product purchase.	Commercial tool	[54]

2.5 Model Validation and Performance Metrics

Since one of the project's future objectives is to iteratively evaluate the model's performance during different stages of preprocessing, it is pivotal to understand ML model validation and performance indicators. Below, y_i denotes the actual value for sample i and $\hat{y}_i$ denotes its predicted value [55]:

- Mean Absolute Error (MAE)

MAE in Equation 2.1 indicates the average significance of the estimating mistake. There is no mistake in the projected values when the MAE is 0. Since the model with the lowest MAE is regarded as the optimal model, the lower the MAE, the better the model performs [50]. MAE is easy to interpret and less affected by outliers than MSE. However, there is no difference in penalty between large errors and small errors and it does not distinguish between overprediction and underprediction. [56].

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (2.1)$$

Mean Absolute Percentage Error (MAPE) in Equation 2.2 converts MAE to a percentage which makes it easier to interpret the value [56] [50]. However, penalizes overprediction less than underprediction.

$$MAPE = \frac{100}{N} \sum_{i=1}^N \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (2.2)$$

MAE has the same units of the target variable, and it tells that if the model predicts late reporting and has an MAE of X, it misses an average of X units per forecast, regardless of whether it's too high or too low.

- Mean Squared Error (MSE)

Similar to the MAE, the MSE in Equation 2.3 is not biased but records variance in error or an average of expected and actual values. MSE is a positive real number and the smaller the value, the better [50]. It punishes big errors more than small errors, which may be a desired property in some cases. However, it is hard to interpret, since it is expressed in square units and does not distinguish between overprediction and underprediction [56].

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (2.3)$$

Because it is measured in the same units as the target variable, the Root Mean Squared Error (RMSE) is frequently used in place of the MSE to achieve an average error with the same magnitude as the original mistakes. RMSE maintains the significant error penalty while taking the square root of the MSE error to return it to the original unit.

- R-squared (R^2)

Knowing that $\bar{y}$ denotes the mean of the actual values:

$$\bar{y} = \frac{1}{N} \sum_{i=1}^N y_i \quad (2.4)$$

R-squared equals:

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (2.5)$$

Of all the previous performance measures, R^2 is the most informative in terms of how often did the model correctly predict and it is easy to interpret, as it varies between 0 and 1. R^2 evaluates how much of the variability in the data can be explained by the model [50]. However, it can be misleading in models with many outliers or with non-linear relationships [56].

2.6 Frameworks for Model Interpretability and Explainability

Since determining the contribution of each feature to the model output is important to improve future work, it became necessary to look for tools to do this. Understanding the influence of each feature helps to enhance the overall interpretability of the results. To address this need, the subsequent subsections present some well-know libraries that try to help in understanding a model result.

SHAP

In many situations, knowing why a model generates a particular forecast can be just as important as the prediction's accuracy. To effectively know how much each feature impacted the model performance there is a unique framework called SHAP (SHapley Additive exPlanations). SHAP gives an importance value to each feature in a specific prediction [57]. This approach will help to validate the choice of features early made in this project. Python has a SHAP library for understanding and debugging ML models, that can be easily installed.

An example of this application can be found in the study of Tsai et al. [55], where they attempted to use ML to produce precise TAT predictions for clinical chemical samples. During model evaluation, to examine how and to what degree the features affect TAT, SHAP values were plotted like in Figure 2.5.

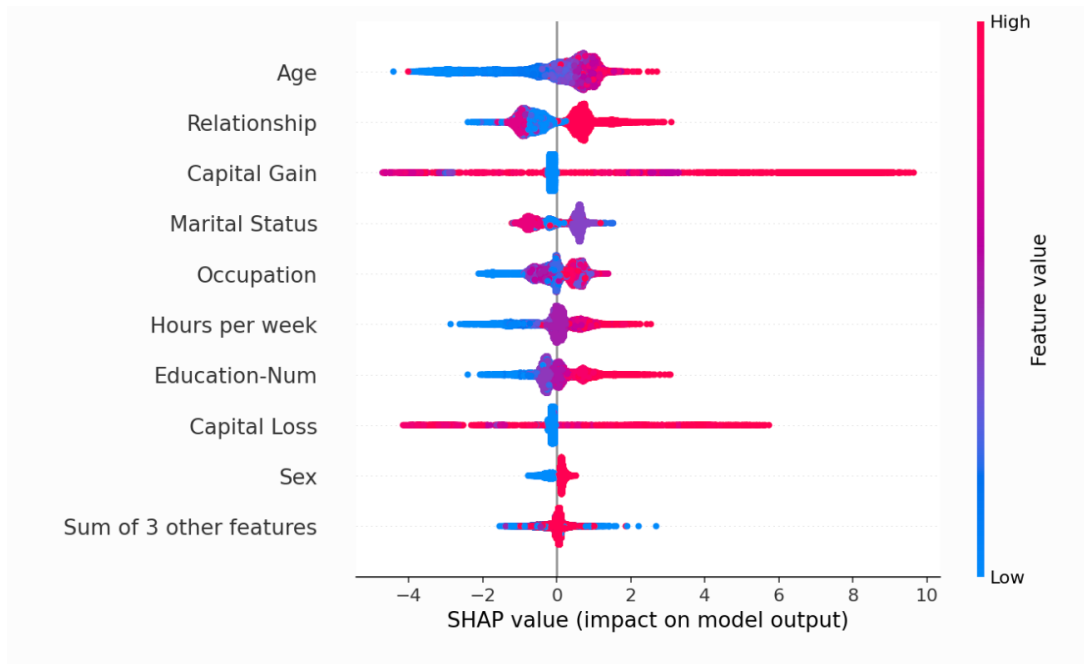


Figure 2.5: Illustrative example of a SHAP result displayed graphically [58].

ELI5

Priestley et al. [8] also present in their study an XAI (Explainable Artificial Intelligence) tool called ELI5 (Explain Like I'm 5) used in Python aimed at show feature importances and explain predictions for many different ML models, such as: XGBClassifier, XGBRegressor xgboost.Booster, LGBMClassifier, LGBMRegressor. ELI5 has both local (how and why a specific prediction is made) and global (how an overall model works) properties. In the study of Sarp et al. [59] ELI5 was used to show the importance of different characteristic (e.g. solar irradiation, temperature, humidity) in the final model prediction aimed at forecasting solar photovoltaic energy like showed in Figure 2.6.

Weight	Feature
+11.493	<BIAS>
+6.280	x
-14.140	y

Figure 2.6: Illustrative example of an ELI5 explanation result [60].

LIME

LIME (Local Interpretable Model-Agnostic Explanations) is another technique that helps to understand predictions of complex ML models. LIME is “model agnostic”, meaning it can be used with any type of ML model. Additionally, LIME provides “local” explanations, meaning that the explanations are specific to an individual predictions and may vary for different predictions.

LIME works by creating a simpler local model which can be more easily interpreted. After data sampling, samples are weighted accordingly to their proximity to the original observation. Closer samples have more weight, indicating the most important characteristics to this specific prediction [61].

Kumarakulasinghe et al. [62] evaluated LIME in the context of clinical ML classification models. They used LIME to help understand the predictions and observations of a sepsis classification model. The model was trained with data from patients in intensive care units. In the end, the study involved the participation of physicians to assess the clinical relevance of the explanations provided by LIME, shown in Figure 2.7.

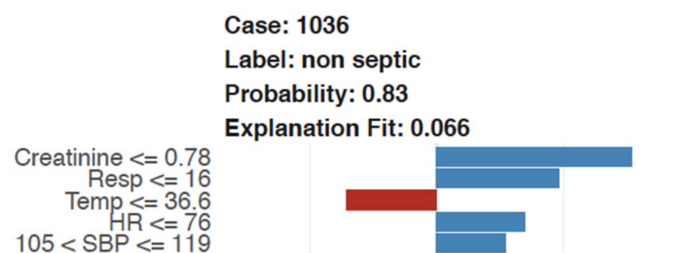


Figure 2.7: Illustrative example of a graphical LIME result [62].

PDP

Partial Dependence Plot (PDP) is a visualization tool that illustrates the effect of one or more features on the prediction of an ML model. Rationally, what PDP does is to show how the model performance is affected when a change in one feature is made while the others remain constant [63].

The degree to which the feature affects the prediction is shown by the slope of the PDP curve. A significant effect is shown by a steep slope, whereas a minor influence is indicated by a gradual slope [64] as it can be seen in Figure 2.8.

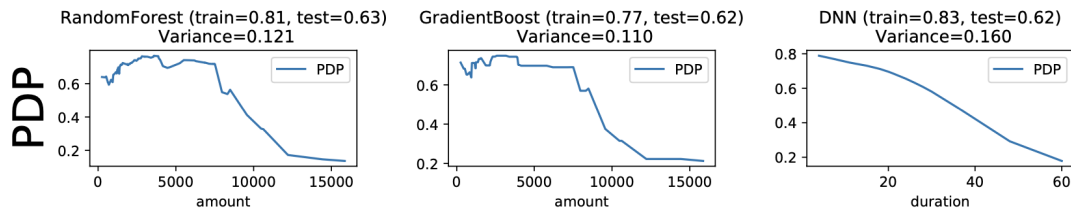


Figure 2.8: Partial Dependency Plots result for three models (random forest, gradient boosted tree and DNN) in the context of loan prediction [64].

It is crucial to see PDP limitations as a tool of insight rather than causal inference. PDP shows the relationship between the feature and the model prediction, but does not necessarily imply a cause-and-effect relationship in the real world [63].

2.7 Summary

This background investigation serves as a comprehensive foundation for understanding key topics, methods and techniques that will be later used to develop a robust methodology. The research made on Data Analysis Methodologies highlighted the importance of understanding how to identify data contamination and which set of steps are normally done to assess data quality. Additionally, with Insights Through Questionnaire Structuring and Assessment it was possible to understand critical steps on the construction of a survey and how can the data collected contribute for the study. Data Preprocessing, Data Mining Tools, Model Validation and Performance Metrics and Frameworks for Model Interpretability and Explainability stand out as essential tools that will help the development of this dissertation.

Chapter 3

Literature Review

The difficulties found by the increasing complexity and volume of healthcare data have created a need for robust data processing and advanced analytical methods. More importantly, maintaining data quality and using ML algorithms have proven useful in solving operational problems such as mitigating hospital delays.

This chapter integrates three pertinent strands of literature relevant to this work. First, the search methodology outlines the criteria and approach taken to choose the most relevant studies for analysis. Second, focus is placed on data quality in health data pipelines, considering common data preprocessing steps, imputation methods and reliable analysis frameworks. Finally, the last section reviews the application of ML models to hospital delay reduction, some specific to radiology. During this process some features that were important for these case studies, seen as critical factors leading to delays and inefficiencies in clinical workflows, are also highlighted to later help in the dataset design. This literature review consolidates findings from these domains to aid in model development aimed at optimizing healthcare delivery.

3.1 Search Methodology

The search process of related work is described below and includes eligibility criteria, information sources and explanation of the search strategy.

Regarding eligibility criteria, both inclusion and exclusion rules were taken into account.

- **Inclusion criteria:** studies on data quality components within eSIS; empirical studies that worked on the problem; published from the year 1999 to the present.
- **Exclusion criteria:** content presented in a format other than a scientific article such as conference papers, book sections and other articles; publication language not in English or Portuguese; full text not available.

This study used a systematized review approach to identify related work on methods for enhancing data quality and for evaluating ML models that aimed to minimize the time taken in medical processes. The latter is as important because, although the main objective is to address data quality during preprocessing, there is a longer-term goal of having this preprocessing support a model that predicts delays. This highlights the need to understand what usually these models have as input features, and which approaches can be followed.

The literature search has been carried out since September 2024, using the following electronic scientific databases: Science Direct, PubMed, Scopus and Google Scholar search engine as an additional source for tracing references. The following keywords are some examples used in the search strategy: data

quality, data improvement, data cleaning, preprocessing, outliers imputation, deduplication, normalization, hospital information system, machine learning, prediction, medical reports, reporting delay. These keywords were searched using various combinations in the fields of title, abstract and keyword. It was considered the search features in each database and used boolean operators (AND, OR) to combine and search selected keywords. Figure 3.1 presents the PRISMA Flow Diagram, which explains the studies selection process.

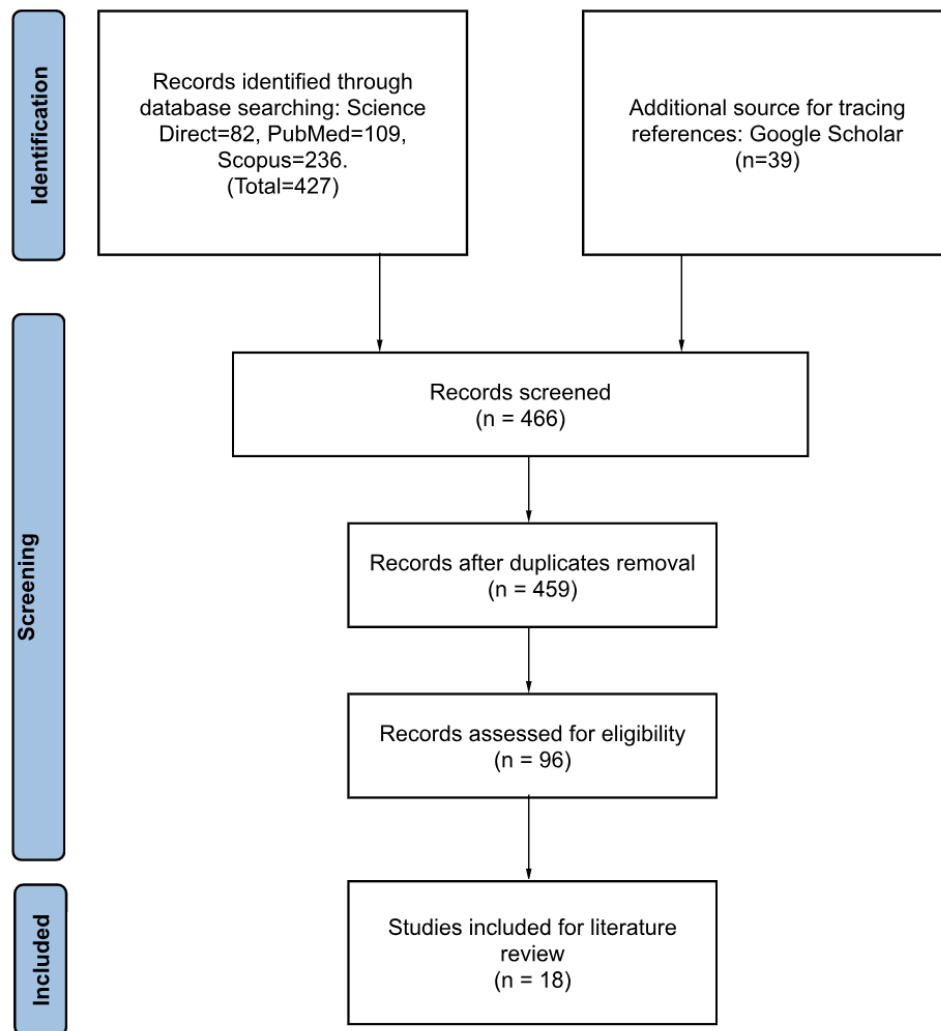


Figure 3.1: PRISMA Flow Diagram that represents how the selection of specific records was made, condensing everything explained in the previous subsections.

3.2 Data Quality Pipelines in the Healthcare Context

Guaranteeing the quality of data that is used to encompassing data collection, preprocessing, model training and validation is crucial to achieve an excellent model performance. Rangineni [65] made a research that aimed to help ML professionals improve data quality and the results showed a simple yet effective data processing structure. The first part of the structure was iterative, focusing on the data use case and design, followed by the data collection and data cleaning. Next, after building the ML model, it would be trained and tested and finally deployed and monitored. This framework gets very close to the methodology used in this project.

Data pipelines are pieces of software that automate data transformation and transfer data from various

source systems to predetermined locations. As a result, data pipelines are workflows for digital data processing that are built on a collection of simple scripts. These pipelines help to process data effectively, transport, storage, managing all data operations and encapsulating the complete data flow from point of origin to point of destination [66]. A data processing pipeline will be the starting point for a better ML model performance [8].

According to Pervaiz et al. [67] from University of Washington, data processing requires three main steps: data gathering, data cleaning and then data analytics. Nearly 80% of the overall time is spent on the second step, which takes the longest. This step includes removing duplicates, correcting inconsistencies, filling missing values, correcting unexpected formats and unit conversion. Finding outliers and verifying the quality of the data are crucial in data analytics. Contrariwise, Foidl et al. [66] from the University of Innsbruck, have a different view on what data processing includes. They divided in five main phases: data ingestion, data integration, data cleaning, data transformation and finally data loading. Dead code, duplicated code and language heterogeneity are some of the factors highlighted in the study as factors that influence the ability of data pipelines providing quality data. These factors create the need to look for methodologies to correct these errors.

The article of Breger et al. [68] addresses challenges and solutions for cleaning and preparing data from the National COVID-19 Chest Imaging Database (NCCID), a centralized UK database containing chest images and corresponding radiological clinical data. This cleaning aids in the creation of ML technologies targeted at the 2019 Coronavirus Disease (COVID-19). Although a data cleaning pipeline was published in 2021 by NHSx (NHS England), the paper highlights the need for an extension to correct additional systematic inconsistencies in the raw data such as categorical features, dates and numerical features. Regarding categorical features (columns or fields in a dataset that contain values that represent distinct categories or groups, rather than continuous numbers), one of the highlights was problems with gender. A lot of inconsistencies were found ("F/M", "1/2", "0/1") and to solve them, it was used DICOM metadata (Digital Imaging and Communications in Medicine). DICOM data is a standardized format and protocol for storing, transmitting and managing biomedical images and related information to ensure interoperability across medical systems [34]. Looking at dates, different versions were found such as, UK format, USA format and others. To standardize, all data element were converted to one only format. Concerning numerical features, a lot of medical values had wrong units, raising the need of also standardize the units used. The extended cleaning pipeline ¹*nccidxclean* is publicly available.

This article clearly states the importance of this dissertation's theme: data quality pipelines to enhance ML model performance. As the authors say, data cleaning is important to ensure the accuracy, reproducibility and interpretability of artificial intelligence models. Hence, they leave some recommendations [68]:

- Create new data standards or use ones that already exist: the authors leave the ²*Study Data Tabulation Model (SDTM)* and ³*Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT)*;
- Include real-time data validation techniques that keep checking data problems;
- Explore electronic health records (EHR) to understand what type of data exists in health.

Implementing a system that instantly verifies the entered data and highlights any potential problems is the idea. Therefore, regarding data validation techniques, the article presents some important data quality functions developed in python. For this topic of research, below are highlighted the most important function's examples given [68]:

- *inferences*: when a value is missing, the function tries to fill in based on information taken from other similar fields;

¹<https://gitlab.developers.cam.ac.uk/math/cia/covid-19-projects/nccidxclean>

²<https://www.cdisc.org/standards/foundational/sdtm>

³<https://www.snomed.org/>

- *clip_numeric*: when there is a integer field, the function removes non-integer numbers;
- *remap_sex*: the function standardizes hospital names.

Ainsworth and Buchan [69] encourage the integration of health data pipelines to enhance information quality and minimize data processing redundancy, with the goal of promoting learning and the effectiveness of the health system. In their study, it is highlighted that different HIS have different information pipelines that many times have duplicated information in them. This results in inefficiency and possibly less-than-ideal analysis because of limited access to knowledge that could be found in another health-care system pipeline. Hence, the authors suggest to improve the quality of information by combining specialized and contextual knowledge from different pipelines.

Förstel et al. [70] investigated, in 2024, the challenges of data quality in HIS, focusing on a German regional hospital and highlights the need for robust management practices and data quality assurance. The study programmatically analyzed and cleaned a dataset of over 2,000,000 patient data entries taken from the HIS using a combination of data analysis and expert interviews. Their main results were outstanding:

- An analysis revealed that approximately 25% of the dataset was unusable due to errors and inconsistencies;
- The complexity of the HIS interface resulted in frequent errors and incomplete data entries;
- Lack of adequate training on data entry and HIS functionalities contributed to inconsistencies and errors;
- Several errors were identified in key variables such as the patients unique IDs for categorizing patient records, patient types and patient category details.

This study highlights how seriously data quality in HIS enables digital transformation in the healthcare sector. The findings brought into light the need to enhance such data management procedures as system integration, real-time validation methods and more user-friendly user interfaces. It is with addressing issues on data quality and provision of accurate, reliable and useful patient information that hospitals can be able to offer better patient care and operational efficiencies. Munappy et al. [71], in turn, raise the key issues about the management of the data pipeline, including problems around: data gets lost on its path from a place of collection to transmission, hence arriving incomplete and of poor quality. The authors acknowledge that errors and discrepancies that arise from manual processing at some stages of a data pipeline should, at most, be automated.

Both Croce et al. [72] and Manias et al. [73] studied the importance of health data preparing and cleaning in the application of ML models. On the one hand, Croce et al. introduce a three component data processing pipeline that aims to create holistic health records. **Data Cleaner**, deals with cleaning the data, which has to be accurate, integral and uniform, by making use of libraries like Pandas and Scikit-learn for the detection of errors, missing values, inconsistencies and their correction. **Data Qualifier**, performs the rating on the sources in regard to reliability, classifying them as either reliable or unreliable in regards to the quality of the data that has been received and analyze the percentage of clean data versus those with errors. **Data Harmonizer**, harmonizes different data to a common format; in this case, HL7 (Health Level 7) FHIR (Fast Healthcare Interoperability Resources), a standard developed for the healthcare sector by HL7 International. On the other hand, Manias et al. describe the application of a methodology in the preparation of data for the project AMD-STITCH. In this case, the data cleaning process is also divided in three components. **Data harmonization**, focus on correcting discrepancies between different data. **Missing and incomplete data**, tries to fill in with reasonable values, even if null, everytime something is missing. **Data privacy**, ensures that data is anonymized to protect sensible information of the patient. These two very different scopes demonstrate the regular need of data preprocessing and cleaning pipelines, towards data quality, consistency and reliability. These two approaches complement each other by joining different topics of importance together, such as preprocessing and privacy issues.

As for best practices while preprocessing, Brownlee and García et al., leave some key points:

- **Training and evaluating the model on different data.** It is mandatory to split the data in train and test because when the same data is used for training and testing, the model already “knows” the answers. It may seem like the model is very good, but in reality it just memorized the answers (overfitting) and did not learn real patterns that generalize to new data [74].
- **Looking for imbalances in the distribution of target values.** Whenever there are many examples with low target values and few with high values (or vice versa), the model ends up focusing on predicting the most common values well and consistently misses the extremes. This must be corrected in the target [75].
- **Performing data preprocessing steps only after the train-test split.** One of the worst problems in ML is when the model has access to information from the future, that is, data that it should not see during training. If the split is not done before applying preprocessing methods, the model learns with “extra information” that it will not have in the real world, failing when applied to truly new data [74].

3.3 ML Models Applications for Reducing Hospital Delays

Following an examination of potential approaches to handle the dataset, it was imperative to hunt for research on ML models that sought to speed up and increase hospital efficiency. An important part of this literature review was also acknowledge which type of features can normally enhance model performance in this medical context.

A radiological study focused on learning from experience, tried to identify factors that influenced the on-time starts for Computerized Tomography (CT) procedures. To do this, the team performed a Root Cause Analysis (RCA). One of the major problems they found as they tested different scenarios was that regarding measurements, many times, results would take too long, making it impossible to make the CT on the scheduled date [36]. Likewise, a study that aimed to find causes of appointment nonattendance and cancellation at a university hospital, found that out of 147 people that had the appointment canceled, 25 (17%) of them was due to not having completed diagnostic studies in time. This number shows the huge need to improve hospital TAT, since these canceled appointments result in costs not only for the hospital but also for the patient [76].

Tsai et al. [55] tried to achieve accurate TAT predictions for clinical chemistry samples using ML. As they explained, TAT is an excellent performance forecaster that help healthcare professionals to save time and make an efficient planning of resources. In this study, TAT equals to the time of generation of results minus the time of arrival of the sample. The list of features obtained from the dataset given by the lab included the sample log files, priority level, required tests, date and time of the various activities and some numbers on the samples. Some of these features were binary and other features values had a wider range of integer values. The ratio for splitting data to be tested that showed better MAE, MAPE and R^2 results was 95% training and 5% testing. Extra Trees (ET) Regressor was the model tested that showed greater performance. This turned out to be true due to the high dimensional feature space and high degree of ET randomization, that happens to reduce variance drastically and also the probability of overfitting. After model evaluation, the most outlined features were the number of modules a sample is allocated to, the workload in pre-analysis and whether the sample needs a CO2-L test or not. These three variables showed the following results: higher the workload, higher the waiting times; higher the quantity of modules, higher the waiting time; and samples requiring a CO2-L test, higher waiting time.

Similarly, Shah et al. [77] made a retrospective study to investigate TAT at two hospital campuses during non-business hours. Their outcomes were measured with two very important variables: the total time, that represents the first time any type of information about the exam was given (even if verbally) minus

the scanning order time; and the interpretation time, that equals to the first communication time minus the scan completion time. Regarding factors that could be indispensable for the model training, six features were outlined: the inpatient status (inpatient vs emergency), the time of day, the body part, the study description and the post-graduate year (PGY) level in training. For free-text, a Random Forest model was combined with a Bag of Words model and for the metadata features a XGBoost Classification model was trained. Regarding data preprocessing, categorical variable were one-hot-encoded, sentences were tokenized and specific terminology was used to map different terms used in medicine (CLEVER and RadLex). After preprocessing, the model training started. It was created and evaluated a ML model that, with the help of NLP components, could forecast when on-call radiology cases would be interpreted slowly. Combining the two models mentioned above, a final prediction was achieved. Looking at results, the study description, time of day and PGY level in training presented values of importance between 0.43 and 0.49, 0.18 and 0.16 and 0.18 and 0.17, respectively. While Tsai's research focuses on laboratory process optimization, Shah's work emphasizes the significance of anticipating delays and modifying operations during poor availability, using ML techniques to enhance resource planning.

Patients' waiting time in the Canadian healthcare system is a major obstacle for every hospital's efficiency. Contrariwise to Shah, Mizan et al. [78] aimed to reduce waiting times through a trained model that would forecast not only the miss-TAT rate but also the patient arrival and the workload count at six different imaging centers with twelve different modalities each. As these three goals are directly correlated, a multi-target time series forecasting technique was developed. From a dataset of sixty four attributes, three target variables represented the three goals mentioned above. The rest attributes were input features for the model training and evaluation. This dataset represents two years of instances that were 80% used for training and 20% used for testing. The project was divided in three main phases: the initial one where offline training was made; the second one, the online prediction phase, focused on the selection of the higher accuracy model and prediction of incoming data; and the last one, that used the predictions of phase two to allocate the workload with optimized waiting time. To have a more efficient resources allocation as well as workload distribution, two fundamental techniques were introduced: the weighted-sum technique and ϵ -constraint technique. The first one, combines different weights proportional to their relative importance with the model variables. The second one, focus on a unique objective to improve, having other problems as restrictions to this main objective. The two models were tested and showed insignificant efficiency changes in the solution. The overall forecasting showed that the rate of miss-TAT depends on the quarter of the day. The optimized framework improved scheduling by producing 33.3% of reduced workload and it was possible to minimize the waiting times 396 minutes.

These experiments demonstrate how ML models can maximize healthcare efficiency and produce genuinely positive outcomes.

3.4 Summary

In summary, the search methodology for this literature review, alongside with a robust background search, ensured a structured and comprehensive identification of relevant studies. The review in data quality pipelines highlighted the critical role of preprocessing techniques in ensuring reliable and interpretable ML outcomes, particularly in clinical environments. Finally, the analysis of ML model applications for reducing hospital delays revealed the growing potential of these models to uncover actionable insights and support more efficient healthcare delivery. Building upon this foundation, a structured methodology is proposed in the subsequent chapter.

Chapter 4

Methodology and Resources

This chapter gives an overview of the methodology as well as an explanation of adopted resources and strategies for the well development of this dissertation. Here it is also explained the approach selected to present the results and how were data privacy policies handled.

4.1 Overview

The main focus of this dissertation was to develop algorithms to treat the quality of data that later was used to train and test a model that studies delays in reporting times. In Figure 4.1 the process is outlined.

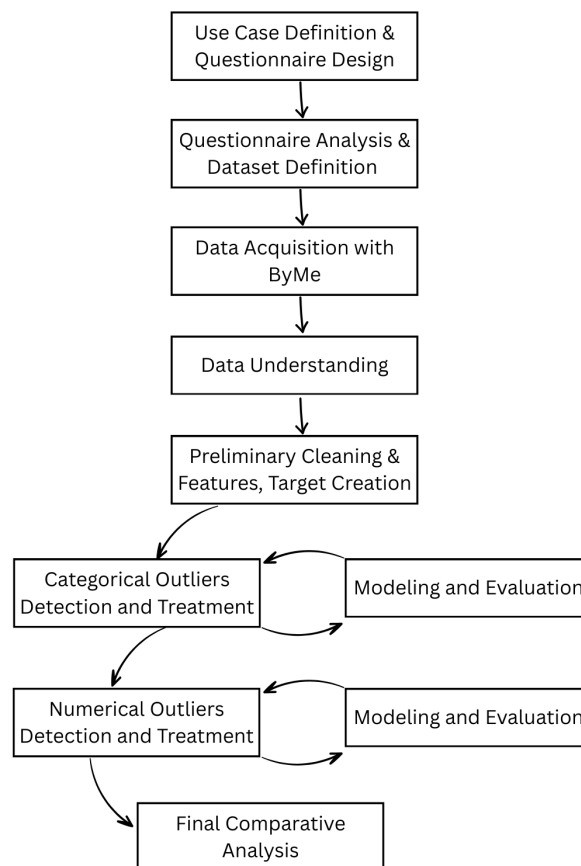


Figure 4.1: The methodology flow used: from clinical insight to iterative ML modeling, based on CRISP-DM.

1. Use Case Definition & Questionnaire Design

During this first phase, the real problem and the short and long-term goals are defined. Here, the development of a questionnaire aimed at physicians begins, in order to collect feedback on what are the root causes of report delays. The goal is to identify relevant factors in the process of delivering reports that can help a model to predict delays.

2. Questionnaire Analysis & Dataset Definition

The questionnaire responses support the construction of a customized initial dataset, by helping define the input variables.

3. Data Acquisition with ByMe

Based on the previous point, there is data extraction from real-world hospital data (with the help of ByMe) and organization according to the insights collected.

4. Data Understanding

A detailed exploration of the dataset is performed to understand its structure, data types, value distributions and initial quality issues. A baseline model is trained to evaluate the initial performance of the dataset without any preprocessing.

5. Preliminary Cleaning & Features, Target Creation

Initial cleaning involves removing irrelevant columns, duplicates and standardizing data formats. New features, such as calculating time intervals between key clinical events and the target are also defined.

6. Categorical Outliers Detection and Treatment

Categorical features are analyzed to detect rare or inconsistent categories. The same ML algorithm and validation strategy are used before and after this step to allow direct performance comparison. This enables an iterative decision-making process: if performance does not improve or decreases, the cleaning strategy is revised.

7. Numerical Outliers Detection and Treatment

Numerical features are assessed for outliers using univariate and multivariate detection techniques. After detection, different treatment algorithms are tested (deletion, winsorization, KNN imputation and MICE imputation). The same iterative modeling and evaluation process applied to categorical outliers is performed in numerical outliers.

8. Final Comparative Analysis

In the final stage, model performances across the different preprocessing stages are compared to assess which combination of treatments yielded the best predictive capability.

As further detailed in Section 5.3, ByMe provided data from two hospitals. The dataset of each hospital had to undergo a split into two distinct subsets (Non X-ray Radiological Data and X-ray Radiological Data), reflecting the differences in their underlying workflows and associated prediction targets. Consequently, phases 4, 5, 6 and 7 of the methodology presented above, are applied separately to each subset and are therefore presented in the two dedicated chapters (Chapter 6 and 7, respectively). This organization allows for a focused analysis and preparation tailored to the specific characteristics of each subset, as well as a separation of results according to the type of data and specific target.

Besides CRISP-DM, the iterative approach of TDQM is also used to assess data quality. TDQM is a recognized and researched framework that gives credibility and robustness to the methodological approach. It offers tools that can be adapted to meet particular challenges in healthcare data, including handling missing or incomplete fields, standardizing heterogeneous data and ensuring reproducibility

and reliability [10]. One of the most important strengths of TDQM is the iterative improvement cycle: Define, Measure, Analyze and Improve, fitting seamlessly with a pipeline development process [10]. This process allows continuous improvement of data quality, meaning algorithms are developed paired with the needs of the project. Besides solving current needs for data cleaning, TDQM provides a framework for improvement and scaling. That way, it will keep the effectiveness of the pipeline when either the size or the variety of datasets increases, being this way the perfect choice for those projects with long-term objectives. In defining the methodology, particular attention was paid to several critical aspects of data quality, including:

- Understand what the quality of data means for who uses systems within eSIS. This first issue is protected by the requirement gathering with physicians.
- Evaluate intrinsic data quality such as: believability, reputation, accuracy and objectivity. The first two are secured by a data acquisition with ByMe, a healthcare leading company with reliable solutions; and the last two are taken into account during pipeline development.
- Evaluate contextual data quality such as: value-added, relevancy, timeliness, completeness, appropriate amount of data. The first two are secured by the requirement gathering with physicians; timeliness guaranteed by a recent data acquisition; completeness is assessed during data understanding and later treated; and appropriate amount of data is secured by ByMe, with around a million of rows, each hospital.
- Evaluate accessibility data quality: accessibility and access security. Data acquisition made with ByMe, with data use access covered in Section 4.2.

4.2 Ethical Considerations in Data Use

As explained before, the dataset is provided by ByMe. This dataset has already been properly anonymized and in compliance with applicable data protection regulations, such as GDPR (General Data Protection Regulation), a European certificate (EU 2016/679), directly applied throughout countries in the European Union, that determines the rules regarding the protection, processing and free movement of people's personal data [79]. Anonymization was performed in a way that prevents re-identification of individuals.

The work in this project does not directly generate any predictions or influence clinical decisions. It rather prepares the data for subsequent use in ML models, with the ultimate goal of support the optimal scheduling of medical appointments. The models used for evaluation are only to test whether the performance metrics have been enhanced by the preprocessing pipeline.

A Non-Disclosure Agreement (NDA) was signed between the author and ByMe to ensure confidentiality and restrict the use and sharing of the data. This agreement guarantees that all data access and preprocessing remain strictly within the scope of this research and under the terms defined by ByMe. The data is processed in accordance with the applicable data protection law.

4.3 Summary

This chapter is fundamental to understand the methodology used to assess data quality and explore different treatment techniques. Additionally, as is only appropriate given the sensitive nature of the data involved, ethical considerations in data use were a priority to support the smooth and responsible execution of the project.

Chapter 5

Foundation and Study Preparation

This chapter focuses on the preliminary phase, where the foundational steps for dataset structuring and model selection were undertaken. Initially, the focus is on requirements gathering through a questionnaire and then the responses analysis. This analysis assisted in determining important dataset variables which are helpful and provide useful understanding into the study. With those insights, the structure of the datasets used was created with other considerations such as data quality as well as its applicability to the goals of the project.

In addition, this chapter considers the models selected for the later stages of analysis, stressing their relevance to the given problem. This chapter aims to provide evidence on the chosen models for validation and how was that evaluation proceeded.

5.1 Requirements Gathering

To ensure the effective identification and prioritization of factors influencing medical reporting time, it was essential to consult with practicing physicians. Their insights provide valuable context to create a dataset that accurately reflects the real-world clinical environment and reporting workflows.

Consulting with active medical physicians was crucial to guaranteeing the efficient identification and prioritization of elements impacting medical report generation timelines. Hence, in order to collect relevant information, a structured questionnaire was designed in Google Forms in Portuguese and sent to various physicians with experience in medical reporting and medical prescription in Portugal. This choice could guarantee accessibility and convenience. For the purposes of this report, the questionnaire has been translated into English using DeepL Translate web tool, solely for demonstration and alignment with the language of the document.

This questionnaire was created in cooperation with ByMe professionals, who regularly interact with these physicians and through state-of-the-art guidance on questionnaire design. All questions were designed to be concise and straightforward, covering a mix of multiple-choice, ranking and open-ended responses to gather both quantitative and qualitative data. The collected data will guide the creation of a robust dataset for preprocessing and analysis.

The main goal was to use important background information they had to offer for developing a dataset that faithfully captures the actual situation regarding medical prescription and medical exams reporting. The questionnaire can be accessed through the ¹*link*.

¹<https://forms.gle/NDYdfebqiq4chFMX9>

1st Section - Current Professional Status: In this part of the questionnaire, the main goal was to locate the physician in his or her actual professional situation. The three available options include identifying their specific involvement in the prescription and reporting stages of medical exams:

- Reporting and prescribing physician;
- Prescribing physician;
- Reporting physician;

This information was essential to contextualize the answers and adapt the study lessons to their professional profile. Moreover, in this part, region and frequency of use of specific systems within the eSIS are also taken.

2nd Section - Factors Weighting: The second questionnaire section, had as main goal evaluating the contribution of different factors to predict the delivery time of exam reports. This causes the respondent to encode the answers while completing the questionnaire (autocoding), making the coding process easier and faster. There are 3 versions of this section. Depending on the type of physician selected in the first part, different categories were evaluated:

Factors Weighting **Adapted to the Reporting and Prescribing Physician**

Before evaluating the level of relevance of each factor, the physician answers to 5 questions:

- What departments do people request most medical procedures from? (checkboxes question, shown in Figure A.1)
- What is your experience with delays in delivery of exam reports? (multiple choice question, shown in Figure A.2)
- What are your reporting areas? (checkboxes question, shown in Figure A.3)
- How many years have you worked as a reporting physician? (multiple choice question, shown in Figure A.4)
- What would you say is your working methodology when it comes to writing reports? (multiple choice question, shown in Figure A.5)

Afterwards, the physician evaluates the relevance of 5 categories:

- *Reporting Physician:* Factors such as the specialization, workload, availability, age and the time of receipt of the report request (shown in Figure A.6).
- *Exam:* Attributes like priority, specialty, type (e.g. X-ray), number of elements to be analyzed (e.g. images), image complexity and the body part being examined (shown in Figure A.7).
- *Patient:* Patient-specific variables including age, gender, medical history, next appointment date and clinical context (shown in Figure A.8).
- *Resources:* Considerations of infrastructural and systemic factors, including hospital unit where the exam was performed, origin of the patient visit (e.g. emergency, outpatient clinic), room availability, technical issues, availability of specific tools such as speech recognition, transcription and AI tools (e.g. for image processing) (shown in Figure A.9).

- *Temporal Variables*: Time-related aspects like the day of the week and month, shift during which the report was made and level of hospital demand during the reporting period (shown in Figure 5.1).

Factors Weighting **Adapted to the Prescribing Physician**

Before evaluating the level of relevance of each factor, the physician answers to 2 questions:

- What departments do people request most medical procedures from? (checkboxes question, shown in Figure A.1)
- What is your experience with delays in delivery of exam reports? (multiple choice question, shown in Figure A.2)

Afterwards, the physician evaluates the relevance of 3 categories:

- *Exam*: Attributes like priority, specialty, type (e.g. X-ray), number of elements to be analyzed (e.g. images), image complexity and the body part being examined (shown in Figure A.7).
- *Patient*: Patient-specific variables including age, gender, medical history, next appointment date and clinical context (shown in Figure A.8).
- *Resources*: Hospital unit where the exam was performed, origin of the patient visit (e.g. emergency, outpatient clinic), room availability and technical issues (shown in Figure A.10).

Factors Weighting **Adapted to the Reporting Physician**

Before evaluating the level of relevance of each factor, the physician answers to 4 questions:

- What is your experience with delays in delivery of exam reports? (multiple choice question, shown in Figure A.2)
- What are your reporting areas? (checkboxes question, shown in Figure A.3)
- How many years have you worked as a reporting physician? (multiple choice question, shown in Figure A.4)
- What would you say is your working methodology when it comes to writing reports? (multiple choice question, shown in Figure A.5)

Afterwards, the physician evaluates the relevance of 4 categories:

- *Reporting Physician*: Factors such as the specialization, workload, availability, age and the time of receipt of the report request (shown in Figure A.6).
- *Patient*: Patient-specific variables including age, gender, medical history, next appointment date and clinical context (shown in Figure A.8).
- *Resources*: Considerations of infrastructural and systemic factors, including hospital unit where the exam was performed, origin of the patient visit (e.g. emergency, outpatient clinic), room availability, technical issues, availability of specific tools such as speech recognition, transcription and AI tools (e.g. for image processing) (shown in Figure A.9).

- *Temporal Variables*: Time-related aspects like the day of the week and month, shift during which the report was made and level of hospital demand during the reporting period (shown in Figure 5.1).

Each factor evaluated in each category/table is presented in depth in the respective figures, available in Appendix A. Each of these categories is represented by a table. The table has evaluation values dispersed in the columns and different factors related to the category dispersed across the table's lines. In Figure 5.1 is an example.

With regard to time variables, how do you rate the relevance of the following factors in predicting the turnaround time for exam reports? (Rate each item from 1 to 4, where 1 = Not at All Relevant and 4 = Extremely Relevant, or N/A) *

	1	2	3	4	N/A
Level of hospital demand during the reporting period (normal, peak)	☐	☐	☐	☐	☐
Shift during which the report is made (morning, afternoon, evening)	☐	☐	☐	☐	☐
Day of the week the report is made	☐	☐	☐	☐	☐
Month in which the report is made	☐	☐	☐	☐	☐

Figure 5.1: Table made for evaluating temporal categories. Here, both reporting-prescribing and pre-prescribing physicians evaluate the relevance of temporal factors concerning their work.

It is important to note that, each physician has an open-ended question at the end of each category (table) about additional factors that are omitted from the table but the physician believes to be pertinent to that topic.

The decision regarding the evaluation parameters was made based on the use of *Likert Scale* [80]. However, an adaptation of this scale was implemented, evaluating relevance using a '1 to 4' scale, with the central element of the traditional *Likert Scale* (commonly representing neutrality) being replaced with a "N/A" option and moved to the last column of the table. The decision to separate the central

value was made after gathering feedback from professionals in the area. This change was performed in order to prevent users from choosing this option arbitrarily, simply because it is the middle option, without carefully reflecting on the relevance of the factors being evaluated. This approach aims to encourage a more accurate and considered assessment, forcing the respondent to take a more conscious position regarding the weighting of each factor. Additionally, the "N/A" option ensures that users who are unfamiliar with a particular topic are able to refrain from providing misleading answers, thereby preserving the integrity of the results [81].

3rd Section - Target Variable Definition: In the third part of the questionnaire, since the long-term objective of the project is to predict delays, it was important to understand within the questionnaire how would physicians perceive this variable taking into account the information system they used. Hence, in this third part, open questions on what would be the best way to define the target variable for this problem were made.

4th Section - Open Suggestions: The last part of the questionnaire focused on understanding, with an open text question, how do physicians believe that the efficiency of the medical exam reporting process can be improved using ML. Any relevant open suggestion was taken into account to the construction of the dataset.

5.2 Questionnaire Results

As the construction of this questionnaire required careful considerations to ensure it aligned with the practical needs of each physician, the dissemination of the questionnaire started in end of November. The questionnaire was closed in February and had 32 responses. Below, in Figure 5.2 and 5.3 are presented results concerning type of physician, frequency of use of specific systems within the eSIS and region of work.

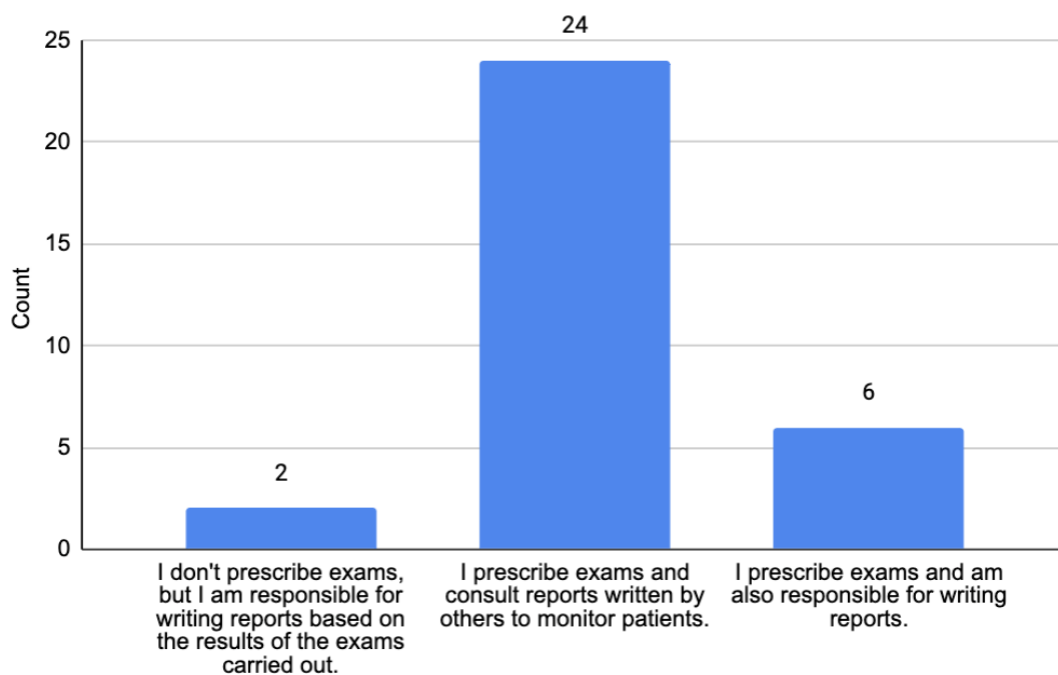


Figure 5.2: Professional profiles of the physicians who answered.

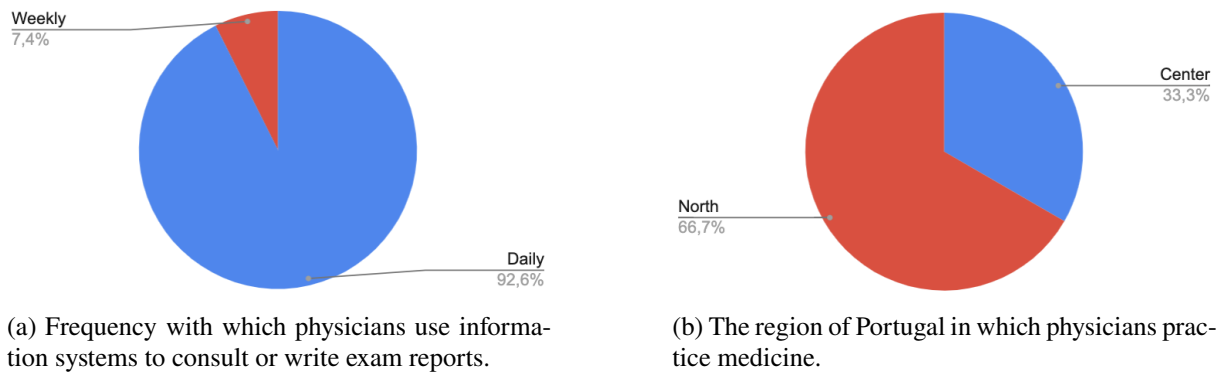


Figure 5.3: Results from hospital-related data.

As can be perceived, a large percentage of doctors use information systems on a daily basis to report and prescribe exams. This shows how the data obtained within the eSIS can be very useful to solve problems such as delays in delivering results. In Figure 5.4, it is possible to understand that delays in report delivery are perceived to reporting and prescribing physicians as frequent (the maximum value of frequency given within the options).

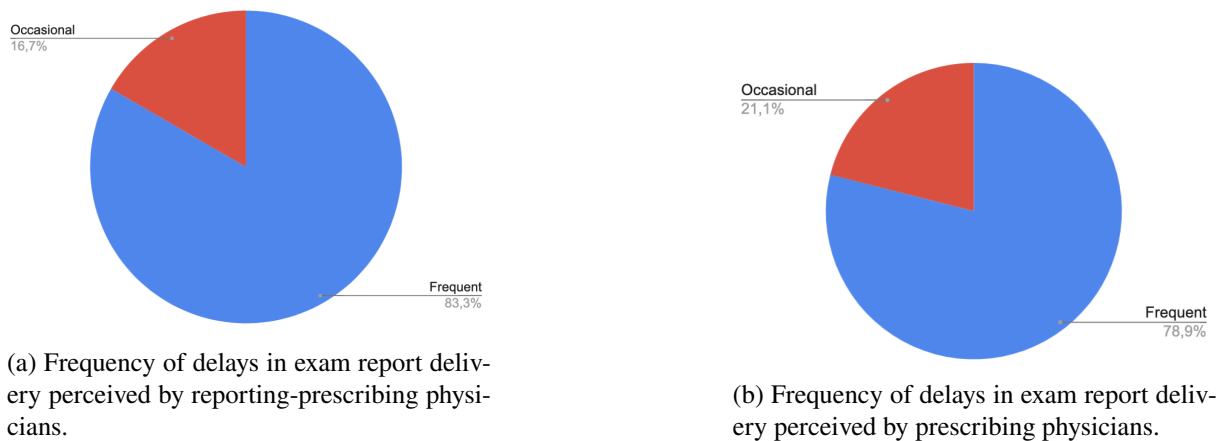


Figure 5.4: Frequency of delays in exam report delivery perceived by two different type of physicians.

In order to understand which factors are perceived by physicians as most influential in the time taken to issue medical reports, the responses were analyzed from two complementary perspectives:

1. Descriptive analysis of general perception

For each factor assessed, the mean and standard deviation of the responses were calculated and are presented in Table 5.1, ordered in descending order of the mean. This approach allowed to identify factors with the greatest “perceived weight”, meaning those consistently considered most relevant among physicians. Factors with high means and low variability were interpreted as having a strong consensus among physicians and were therefore potentially more critical in the reporting process.

Table 5.1: Mean importance ratings and variability of factors influencing medical report delays.

Variable	Mean	Standard Deviation	Number of Responses
RepPhysicianAvailability	4.00	0.00	8
ExamPriority	3.90	0.31	30
RepPhysicianSpecialization	3.88	0.35	8
TemporalHospitalDemand	3.88	0.35	8
PatientClinicalStatus	3.87	0.35	30
ResourceRequestOrigin	3.78	0.42	32
ExamType	3.59	0.63	29
ResourceTechnicalIssues	3.53	0.78	30
PatientNextAppointment	3.52	0.63	31
RepPhysicianPendingReports	3.50	0.53	8
ResourceVoiceRecognitionTools	3.50	0.76	8
ExamComplexity	3.45	0.69	29
ExamElementsToAnalyze	3.40	0.67	30
RepPhysicianAge	3.25	0.89	8
TemporalReportShift	3.25	1.04	8
TemporalMonth	3.25	1.16	8
ExamSpecialty	3.20	0.85	30
ResourceExamRoomAvailability	3.20	0.92	30
ResourceUnit	3.13	1.07	30
PatientExamHistory	3.13	0.73	30
TemporalDayOfWeek	3.12	1.25	8
RepPhysicianRequestTime	3.00	0.93	8
ExamBodyPart	2.97	0.89	30
ResourceAI_Tools	2.86	1.21	7
PatientAge	2.81	1.14	31
PatientGender	2.13	1.02	31

In the analysis of the responses, each question was processed individually to compute its mean rating, standard deviation, and number of valid responses. "N/A" was treated as a missing value, so it would not influence the result.

The most impactful factors have high mean and low standard deviation, showing strong agreement on operational and clinical priority factors as main contributors to delays. These include, physician availability, exam priority, reporting physician specialization, hospital demand and patient's clinical condition. Factors with mean between 3.2 and 3.7, can influence efficiency, but perceptions vary. Means below 3.0 may reflect either genuine lower impact or divergence in awareness and use among physicians, especially for AI tools.

2. Analysis of variance between groups (ANOVA)

In parallel, the one-way ANOVA statistical test was applied to detect possible significant differences in responses (dependent variable) between the three different groups of physicians (independent variable). The aim of this analysis was to understand whether there were significant variations in opinion among doctors from different contexts.

Of the factors analyzed, only two showed statistically significant differences between groups ($p < 0.05$):

- Patient age – suggesting that the perception of the impact of age may vary depending on the doctor's profile.
- Patient gender – with a marginal p-value (≈ 0.0506), indicating a possible, albeit weak, divergence of opinions.

In contrast, for most variables, no statistically significant differences were observed between groups, which reinforces the idea that the perception of critical factors is relatively consensual among professionals.

Additionally, after this statistical analysis, it was essential to evaluate the importance assigned to each doctor-factor pair. To achieve this, a weighted average (WA) was calculated to aggregate the overall opinion. This step was crucial in ensuring the creation of a robust dataset. These results, presented in Appendix B, were used in collaboration with ByMe, to understand which variables among those available in the existing internal data sources could be most useful for the development of the dataset. This process ensured that the initial dataset structure aligned both with clinical relevance and data availability constraints.

In addition to identify factors that have the greatest impact on reporting delays, it was also provided a foundation for selecting radiology as the specialty to focus on.

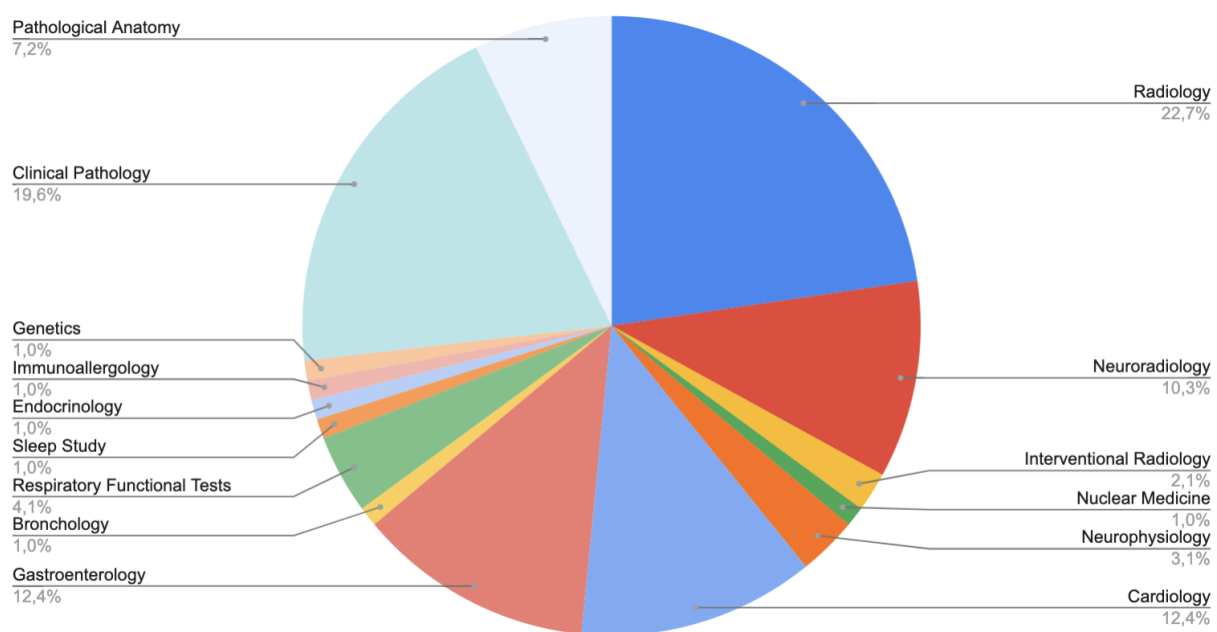


Figure 5.5: Representation of the most prescribed specialties.

By observing Figure 5.5, it was possible to narrow down the dataset, focusing primarily on some of the most frequently requested specialties: radiology and neuroradiology.

5.3 Dataset Context and Structure

The process of extraction of data was made by ByMe from their databases. Data was collected from two different hospitals (named Site 1 and Site 2), each following distinct work methodologies. Therefore, later, the analysis is conducted separately for each one to account for these operational differences. Both Site 1 and Site 2 include medical exams for which the requests were made between 01/01/2023 at 00:00 and 31/12/2024 at 23:59. The goal involves developing preprocessing algorithms scalable that can be easily adapted to process data from other years of these Sites.

As the data exploration phase began and with insights provided by ByMe professionals, it **became clear that each Site would need to be split into two sets: one containing only X-ray exams and another with all other types of medical acts (Non X-ray)**. This decision was based on the fact that the process for X-rays is completed immediately after the exam, as it only requires the images to be considered finalized. To preserve the temporal integrity of the target variable, this separation was essential. As explained in Section 4.1, this split brings the need to separate analysis and results into two chapters: Non X-ray Radiological Data (Chapter 6) and X-ray Radiological Data (Chapter 7), which differ in both the type of data and the predictive target, as graphically explained in Figure 5.6. Nevertheless, and as already mentioned, the analysis per Site in each chapter is made independently, as they have distinct work methodologies.

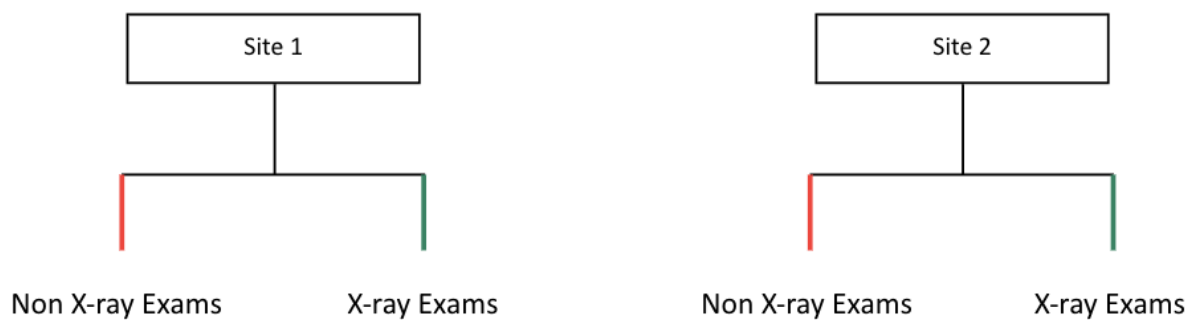


Figure 5.6: Diagram illustrating the data split based on the type of exam (X-ray vs. non X-ray). Equal exam types from both sites share the same predictive target. The colors are only used to explain which subsets share the same target.

Each feature of the dataset was carefully collected after analysis of the questionnaire insights; however, some requested data was not available and as a result, certain columns were substituted. In particular, variables such as the age and specialty of the physician responsible for the report were not available, as the integrated human resources' systems that provide this information store it as optional fields and often lack input. Additionally, clinical data and physicians' shift schedules, which could have enriched the analysis, were excluded due to their sensitive nature and the fact that such information is often not properly parameterized in the hospital systems. Based on the insights gathered from the questionnaire, the dataset is structured to include critical factors identified by the physicians. Each entry/row in the dataset represents a single medical act and its associated metadata, structured as it is presented in Table 5.2. This means that the dataset is structured at the level of medical exam order, with each row capturing details such as the patient characteristics, the scheduling and execution information, the type of exam and the associated resources. The same structure was used in data extraction when collecting from both Sites.

Table 5.2: Structure of the dataset of each hospital at the time of collection.

Variable	Type of Variable	Variable Interpretation
<i>RequestDate</i>	d/M/yy h:mm:ss	Refers to the exact date and time of when the request of a medical act was made.
<i>SchedulingDate</i>	d/M/yy h:mm:ss	Refers to the exact date and time of when the scheduling of a request was made.
<i>ExamDate</i>	d/M/yy h:mm:ss	Refers to the exact date and time the exam was scheduled for.
<i>AdmissionDate</i>	d/M/yy h:mm:ss	Refers to the exact date and time that the patient is admitted into the hospital.
<i>CallDate</i>	d/M/yy h:mm:ss	Refers to the exact date and time when the patient is called to perform the respective medical procedure.
<i>RealizationDay</i>	String	Refers to the day of the week the exam was actually performed.
<i>RealizationShift</i>	String	Refers to the shift where the exam was actually performed.
<i>RealizationDate</i>	d/M/yy h:mm:ss	Refers to the exact date and time the exam was carried out.
<i>ReportDate</i>	d/M/yy h:mm:ss	Refers to the exact date and time of when the report was open for the first time to be written.
<i>ValidationDate</i>	d/M/yy h:mm:ss	Refers to the exact date and time of when the report was closed and validated.
<i>ValidationShift</i>	String	Refers to the shift where the report was validated.
<i>ValidationDay</i>	Integer	Refers to the day of the week the report was actually validated.
<i>DateNextAppointment</i>	d/M/yy h:mm:ss	Refers to the exact date and time of the next medical appointment.
<i>AppointmentSpecialty</i>	String	Refers to the specialty of the medical appointment.
<i>AutomaticScheduling</i>	Boolean	Tells if the scheduling of that exam was made automatically.
<i>DirectAdmission</i>	Boolean	Tells if the admission of the patient was direct.
<i>Rescheduled</i>	Boolean	Tells if the exam has been rescheduled.
<i>PreviousSchedulingDate</i>	d/M/yy h:mm:ss	Refers to the exact date and time of when the scheduling of the previous request was made.
<i>PreviousExamDate</i>	d/M/yy h:mm:ss	Refers to the exact date and time of the previous medical appointment.
<i>ResultsDate</i>	d/M/yy h:mm:ss	Refers to the exact deadline for having a validated report; this date is defined by the referring physician at the time of requisition.
<i>Origin</i>	String	Refers to the origin of that request.
<i>RealizationUnit</i>	String	Unit where the respective medical exam is carried out.
<i>Resource</i>	String	Room where the respective medical exam is carried out.
<i>ExternalResource</i>	Boolean	Tells if the exam was made in an external resource (outside that hospital).
<i>PatientAge</i>	Integer	Age of the patient doing the medical exam.
<i>PatientGender</i>	String	Gender of the patient doing the medical exam.
<i>Priority</i>	String	Priority of doing the medical exam.
<i>Specialty</i>	String	Specialty of the exam.
<i>MedicalAct</i>	String	Medical exam.
<i>TypeMedicalAct</i>	String	Type of medical exam, a more general category than the medical act itself.
<i>FamilyMedicalAct</i>	String	Family of medical exam, a more general category than type.
<i>HasImages</i>	Boolean	Tells if the exam included images in the report.
<i>PatientCode</i>	Integer	Code of the patient.
<i>ReportCode</i>	Integer	Code of the report of the exam.
<i>UsesVoiceRecognition</i>	Boolean	Tells if the report of the exam was written by a voice recognition tool.
<i>ValidatedBy</i>	String	Code of the physician that validates the report.
<i>AccessionNumber</i>	String	Code specific to that request.

Additionally, to map categories of RealizationShift to the actual timeline, there was also an auxiliary dataset that gave important information about the duration of shifts and its presented in Table 5.3.

Table 5.3: Actual start and end time of RealizationShift.

Shift	Start Time	End Time
Morning 1	08:00:00.00	09:59:59.99
Morning 2	10:00:00.00	11:59:59.99
Morning 3	12:00:00.00	13:59:59.99
Afternoon 1	14:00:00.00	16:59:59.99
Afternoon 2	17:00:00.00	19:59:59.99
Other	20:00:00.00	07:59:59.99

5.4 Evaluating Data Cleaning Effectiveness

The process of evaluating the data cleaning algorithms is made iteratively, allowing to decide between different options of treatment along the process. For doing this, different ML models were used, in order to obtain insights on variability of R^2 , MAE and RMSE.

The main task of the model is to forecast the difference between the due date and actual report delivery date. Therefore, a supervised regression approach was deemed appropriate, as it enables us to predict a continuous outcome, the expected delay in hours, that is critical to inform decision-making regarding the optimal time to book future appointments so that the necessary exam results are available in time. This was the rationale behind the regression approach.

Since the problem under study consists of predicting a continuous value, only supervised regression algorithms were explored. Between linear models, neural networks, distance models or tree-based models the decision of what to use was made upon the nature of the data, the objectives of the study and the fact that this is the first time such analysis is being performed using data provided by ByMe.

Tree-base models were chosen to test the quality of the pipeline for several reasons:

1. High performance on hospital tabular data, which stands in contrast to distance models [82] [83];
2. Flexibility to handle categorical variables with many classes, which stands in contrast to linear models [83] [84];
3. Ability to serve as “sensors” for data quality, by observing the importance attributed to features, it is possible to see whether the information is well represented [85];
4. Offer a good balance between performance and training time, a particularly important advantage in iterative development processes, where fast model retraining is essential. This stands in contrast to neural networks, which often require significantly longer training times and extensive tuning [83].

During literature review, several tree-based models were highlighted and frequently discussed. Listed below are the ones chosen to be used.

- XGBoost
- LightGBM
- CatBoost
- Decision Tree

Some of these models are naturally robust to the presence of outliers, meaning their performance does not suffer large variations due to extreme values [86]. Therefore, if there was an improvement in performance in these models after outlier treatment, this is a strong indication that the outliers were really negatively affecting the quality of the data. Then, using tree models as a “validation filter” to test different hypotheses for treating outliers is a solid approach. If performance improves even in robust models, it is a sign that the treatment had a beneficial effect and that the data became more consistent and representative of reality.

During training and testing, categorical columns have been properly converted to the category type, allowing XGBoost to efficiently handle this type of variables by enabling categorical option. The tree method used was histogram-based, useful when using big datasets and fast in training. In the case of LightGBM, categorical columns were converted to the category type in pandas and were explicitly indicated in the fit method through a categorical feature parameter. CatBoost was also configured with categorical columns. The loss function used for gradient boosting models was the standard mean squared error. For the Decision Tree model, a different approach from the previous ones was used, given that this

algorithm does not deal with categorical variables. A pipeline was implemented with OneHotEncoder from scikit-learn, in order to transform categorical variables into binary variables, ensuring that the model could interpret them correctly. Binary variables, in turn, have been transformed into integers (0 or 1) for all models, ensuring their correct interpretation as numeric entries. The remaining numerical variables were maintained as int or floats.

The datasets were split into training and testing subsets using 80/20 hold-out validation. A fixed random state (42) was used to ensure reproducibility, making it possible to obtain the same split across different runs and allowing for consistent comparison across models and preprocessing.

The development of algorithms for cleaning was performed exclusively in Python. This was used as the primary programming language due to its extensive capabilities in data manipulation and analysis. Some libraries mentioned in Section 2.4 were used, such as:

- Pandas, NumPy (data manipulation)
- Scikit-learn (modeling)
- Seaborn, Matplotlib, SHAP (visualization and explainability)

First, to ensure a fair comparison between different data preprocessing strategies (namely missing value imputation and outlier correction), a broad hyperparameter search was conducted using a reference version of the dataset. This hyperparameter optimization process was performed using the dataset with initial preprocessing and all the features needed already created (specified in Section 6.2). This version included essential treatments, such as simple imputation and basic removal, in order to avoid biases caused by specific transformations. The same set of optimized hyperparameters was subsequently applied across all experiments involving different preprocessing techniques. This methodological choice was made in light of the wide range of techniques employed, which required conducting multiple training and validation tests, and were computationally and time intensive. Moreover, given that the proportion of data flagged as outliers across different detection thresholds was always relatively small, the underlying data structure remained largely stable. Therefore, it was considered methodologically acceptable to apply the same hyperparameters across preprocessing variants without significantly compromising the validity of the comparisons. It is acknowledged that this approach may limit the absolute performance potential of each configuration, but it enables a controlled and scalable comparison across preprocessing strategies, with close to optimal hyperparameters. Nevertheless, in the comparative analysis, the best techniques obtained are re-tuned to ensure a more equitable assessment for each model.

To select the best hyperparameters for each model, a search method with 3-fold cross-validation on the training set was used. This procedure allows a robust evaluation of the model's performance on different subsets of the data, helping to mitigate the risk of overfitting during the tuning process. In addition, the test set remained separate, ensuring that the model's generalization is properly verified on data not seen during training or optimization. In this way, it is considered that the hyperparameter adjustment performed is sufficient to minimize the risk of overfitting. The set of hyperparameters tried are in Tables 5.4, 5.5, 5.6 and 5.7. The best hyperparameters found are in bold.

Table 5.4: Hyperparameters adjusted to Decision Tree.

Parameter	Non X-ray		X-ray	
	Site 1	Site 2	Site 1	Site 2
Max Depth	[5, 10, 15, 20, 25 , None]	[5, 10, 15, 20, 25 , None]	[5, 10, 15, 20, 25 , None]	[5, 10, 15, 20 , 25, None]
Min Samples Split	[2 , 5, 10, 20]	[2 , 5, 10, 20]	[2, 5 , 10, 20]	[2 , 5, 10, 20]
Min Samples Leaf	[1, 2, 5 , 10, 20]	[1, 2 , 5, 10, 20]	[1, 2, 5 , 10, 20]	[1 , 2, 5, 10, 20]
Max Features	[None , 'sqrt', 'log2']	[None , 'sqrt', 'log2']	[None , 'sqrt', 'log2']	[None , 'sqrt', 'log2']

Table 5.5: Hyperparameters adjusted to XGBoost.

Parameter	Non X-ray		X-ray	
	Site 1	Site 2	Site 1	Site 2
Number of estimators	[300, 500, 700, 1000, 1200 , 1500, 1700]	[100, 300, 500, 700, 1000, 1200 , 1500]	[100, 300, 500, 700, 1000, 1200 , 1500]	[100, 300, 500, 700, 1000, 1200 , 1500]
Learning Rate	[0.01 , 0.03, 0.05, 0.1, 0.2, 0.3]	[0.01, 0.05 , 0.1, 0.2, 0.3]	[0.01, 0.03 , 0.05, 0.1, 0.2, 0.3]	[0.01, 0.05 , 0.1, 0.2, 0.3]
Max Depth	[3, 5, 8, 10, 12 , 15]	[3, 5, 10, 12, 15]	[3, 5, 8 , 10, 12, 15]	[3, 5, 8, 10, 12 , 15]
Subsample	[0.7, 0.8, 0.9 , 1]	[0.7, 0.8, 0.9 , 1]	[0.7, 0.8, 0.9 , 1]	[0.7, 0.8, 0.9 , 1]
Reg Alpha	[0.5, 1 , 5]	[0.5, 1 , 5]	[0.5, 1 , 5]	[0.5, 1 , 5]
Reg lambda	[0, 0.5, 1, 2 , 5]	[0, 0.5, 1, 2 , 5]	[0, 0.5, 1, 2 , 5]	[0, 0.5, 1, 2 , 5]
Gamma	[0, 0.1 , 0.3, 0.5, 1]	[0, 0.1 , 0.3, 0.5, 1]	[0, 0.1 , 0.3, 0.5, 1]	[0, 0.1 , 0.3, 0.5, 1]
Colsample By Tree	[0.7, 0.8 , 0.9, 1]	[0.7, 0.8, 0.9 , 1]	[0.7, 0.8, 0.9 , 1]	[0.7, 0.8, 0.9 , 1]

Table 5.6: Hyperparameters adjusted to LightGBM.

Parameter	Non X-ray		X-ray	
	Site 1	Site 2	Site 1	Site 2
Number of Estimators	[100, 300, 500, 700, 1000, 1200, 1500]	[100, 300, 500, 700, 1000, 1200 , 1500]	[100, 300, 500, 700, 1000, 1200 , 1500]	[100, 300, 500, 700, 1000, 1200 , 1500]
Learning Rate	[0.01 , 0.05, 0.1, 0.2, 0.3]	[0.01, 0.05 , 0.1, 0.2, 0.3]	[0.01, 0.05 , 0.1, 0.2, 0.3]	[0.01, 0.05 , 0.1, 0.2, 0.3]
Max Depth	[3, 5, 10, 12 , 15]	[3, 5, 10, 12 , 15]	[3, 5, 10, 12 , 15]	[3, 5, 10, 12 , 15]
Number of Leaves	[31, 50, 100 , 150]	[31, 50, 100 , 150]	[31, 50, 100 , 150]	[31, 50, 100 , 150]
Subsample	[0.7 , 0.8, 0.9]	[0.7 , 0.8, 0.9]	[0.7 , 0.8, 0.9]	[0.7 , 0.8, 0.9]
Colsample By Tree	[0.7, 0.8 , 0.9, 1]	[0.7, 0.8, 0.9 , 1]	[0.7, 0.8, 0.8 , 1]	[0.7, 0.8, 0.9 , 1]

Table 5.7: Hyperparameters adjusted to CatBoost.

Parameter	Non X-ray		X-ray	
	Site 1	Site 2	Site 1	Site 2
Number of Iterations	[300, 500, 700, 1000 , 1200, 1500]	[300, 500, 700, 1000 , 1200, 1500]	[300, 500, 700, 1000 , 1200, 1500]	[300, 500, 700, 1000 , 1200, 1500]
Learning Rate	[0.01, 0.05, 0.1, 0.2 , 0.3]	[0.01, 0.05, 0.1 , 0.2, 0.3]	[0.01, 0.05, 0.1, 0.2 , 0.3]	[0.01, 0.05, 0.1 , 0.2, 0.3]
Depth	[3, 5, 7 , 10]	[3, 5, 7, 10]	[3, 5, 7 , 10]	[3, 5, 7, 10]
L2 Leaf Reg	[1, 3, 5, 7, 9]	[1, 3, 5 , 7, 9]	[1, 3, 5, 7, 9]	[1, 3, 5 , 7, 9]
Subsample	[0.7 , 0.8, 0.9, 1.0]	[0.7, 0.8 , 0.9, 1.0]	[0.7 , 0.8, 0.9, 1.0]	[0.7, 0.8 , 0.9, 1.0]
Colsample By Level	[0.7, 0.8, 0.9, 1.0]	[0.7, 0.8, 0.9 , 1.0]	[0.7, 0.8, 0.9 , 1.0]	[0.7, 0.8 , 0.9, 1.0]

5.5 Summary

In this chapter, physicians' insights were analyzed, providing important feedback for data acquisition with stakeholders. Here the first dataset is defined and the strategic split into Non X-ray data and X-ray data marks an important point towards subsequent analysis. Additionally, decisions were taken on what models should be used to serve as validation metrics. Now that there is data from two different hospitals and evaluation metrics are well-defined, everything is ready to start exploring data preparation and treatments.

Chapter 6

Non X-ray Radiological Data

Chapter 6 focuses on handling non X-ray radiological data, beginning with an all-around understanding of the dataset and evaluation of the baseline performance of the model in this subset. The following section presents the preliminary data preparation steps, including standardization of variable formats, deduplication, selective elimination of rows based on specified criteria, imputation methods, feature engineering, and selection of the prediction target. This is followed by the presentation of the final dataset structure after initial cleaning. The chapter then addresses the detection and management of categorical outliers and the consequent impact on model performance. Numerical outliers are explored using two complementary methods: a direct visual inspection method and an automated approach that uses statistical and ML based methods for detection. For each approach, the variability in model performance is analyzed, providing insights into the effects of different preprocessing strategies on predictive outcomes. Hence, this chapter combines the explanation of the preprocessing methods with the presentation of the results obtained at each stage.

Given that this study was performed in data from different hospitals, with distinct workflows and specific associated categories, it was decided to analyze the data separately. This approach allows a more precise understanding of the particularities of each group and avoids biases that could result from the aggregation of heterogeneous operational realities.

6.1 Data Understanding

Before any data preprocessing happened, it was essential to gain a comprehensive understanding of both Sites. This phase involved exploring the structure and content of the data available, identifying key variables, and uncovering initial issues. Several exploratory analyses were carried out to interpret the context and nature of the variables, assess their distributions, and evaluate potential correlations relevant to the prediction of report delays. Not all exploratory visualizations are presented within this section. Instead, they are placed throughout the chapter, integrated into the most relevant contexts, to promote a more intuitive reading flow. In this section, particular attention was given to identifying inconsistencies, missing values and frequency analysis as well as an overall perspective of the whole dataset in terms of size and missing values. This understanding served as the foundation for the data cleaning steps that followed.

- Site 1 represents an unit of a hospital. Initially, this dataset consisted of 491023 rows and 37 columns. The total percentage of missing values in the dataset is 19.61%.
- Site 2 represents four units of a hospital. Initially, this dataset consisted of 554054 rows and 37 columns. The total percentage of missing values in the dataset is 14.24%.

One of the major problems found in both Sites as data understanding started, was the large percentages of missing values for several variables, which are presented in Table 6.1.

Table 6.1: Missing values in non X-ray radiological data.

	Site 1 Non X-ray		Site 2 Non X-ray	
	Missing Values	Percentage of Missing Values (%)	Missing Values	Percentage of Missing Values (%)
RequestDate	0	0.00	0	0.00
SchedulingDate	90304	18.39	45657	8.24
ExamDate	67897	13.83	25350	4.58
AdmissionDate	109930	22.39	70710	12.76
CallDate	118952	24.23	65322	11.79
RealizationDate	101799	20.73	52650	9.50
RealizationShift	0	0.00	0	0.00
RealizationDay	101799	20.73	52650	9.50
ReportDate	103248	21.03	92883	16.76
ValidationDate	105921	21.57	103226	18.63
ValidationShift	0	0.00	0	0.00
ValidationDay	290243	59.11	103226	18.63
DateNextAppointment	454386	92.54	314157	56.70
AppointmentSpecialty	458790	93.44	362462	65.42
AutomaticScheduling	0	0.00	0	0.00
DirectAdmission	0	0.00	0	0.00
Rescheduled	0	0.00	0	0.00
PreviousSchedulingDate	416691	84.86	495296	89.39
PreviousExamDate	416690	84.86	495296	89.39
ResultsDate	2080	0.42	165952	29.95
Origin	0	0.00	39	0.01
RealizationUnit	67897	13.83	25350	4.58
Resource	67979	13.84	25365	4.58
ExternalResource	0	0.00	0	0.00
PatientAge	101799	20.73	52650	9.50
PatientGender	0	0.00	0	0.00
Priority	0	0.00	0	0.00
Specialty	0	0.00	0	0.00
MedicalAct	65	0.01	0	0.00
FamilyMedicalAct	0	0.00	0	0.00
TypeMedicalAct	0	0.00	0	0.00
HasImages	67897	13.83	25350	4.58
PatientCode	0	0.00	0	0.00
ReportCode	103248	21.03	92883	16.76
UsesVoiceRecognition	140766	28.67	123866	22.36
ValidatedBy	106306	21.65	103392	18.66
AccessionNumber	67897	13.83	25350	4.58

Variables like DateNextAppointment, AppointmentSpecialty, PreviousSchedulingDate and PreviousExamDate, with percentage above 80% of missing values are immediately target to be further excluded from Site 1 since there was no viable way to impute these dates. The same happened for Site 2, in the same variables, but in this case, the missing values percentages are lower, however insufficient yet.

Table 6.2: Number of unique categories per variable for both Sites.

	Unique Values Site 1	Unique Values Site 2
RealizationShift	6	6
RealizationDay	7	7
ValidationShift	6	6
ValidationDay	7	7
AutomaticScheduling	2	2
DirectAdmission	2	2
Rescheduled	2	2
Origin	7	6
RealizationUnit	1	4
Resource	36	39
ExternalResource	2	2
PatientGender	3	3
Priority	3	3
Specialty	2	1
MedicalAct	534	508
FamilyMedicalAct	7	11
TypeMedicalAct	24	16
HasImages	2	2
UsesVoiceRecognition	2	2

Since non X-ray data included different types of exams, in Table 6.2 it is possible to see the distribution of the Variable FamilyMedicalAct. Each row of this wider variable has different TypeMedicalAct, and TypeMedicalAct had different MedicalAct. Here it is possible to note that Site 2 has more wider categories. Additionally, CT Scans seem to be the family of exams most used in Table 6.3. Site 1, as it is perceived below, has less families covered in the dataset. This can be important to later assess generalization.

Table 6.3: Percentage of exams by FamilyMedicalAct per Site.

Medical Imaging Family	Percentage Site 1 (%)	Percentage Site 2 (%)
CT Scan	42.71	48.92
Ultrasound	31.90	19.48
Magnetic Resonance Imaging	21.80	17.23
Other Imaging	-	4.74
Angiography	0.84	4.38
Mammography	2.62	1.78
Nuclear Medicine	-	1.48
Orthopantomogram	-	0.82
Bone Densitometry	-	0.81
Interventional Radiology	0.13	0.31
Fluoroscopy	-	0.05
Cardiac MRI Test	0.00	-

Since different backgrounds bring sometimes different workflows, Table 6.4 presents the origin of the request. More than half of the medical acts comes from outpatient appointments. For both Sites it is also possible to see that emergency covers a great part of percentage.

Table 6.4: Percentage of origin per Site.

Origin	Percentage Site 1 (%)	Percentage Site 2 (%)
Outpatient Appointments	58.86	58.24
Emergency	25.06	25.97
Inpatient	15.92	14.42
Radiology	0.07	1.21
Operating Room	0.04	0.13
Day Hospital	0.05	0.04
Laboratory	0.00	0.00

Looking at the distribution of age in Figure 6.1, it is possible to recognize the impact of missing values for Site 1. For both Sites, the highest frequencies are around 40 and 80 years old, which is expected, since in these ages people tend to go more to the hospital than when they are younger.

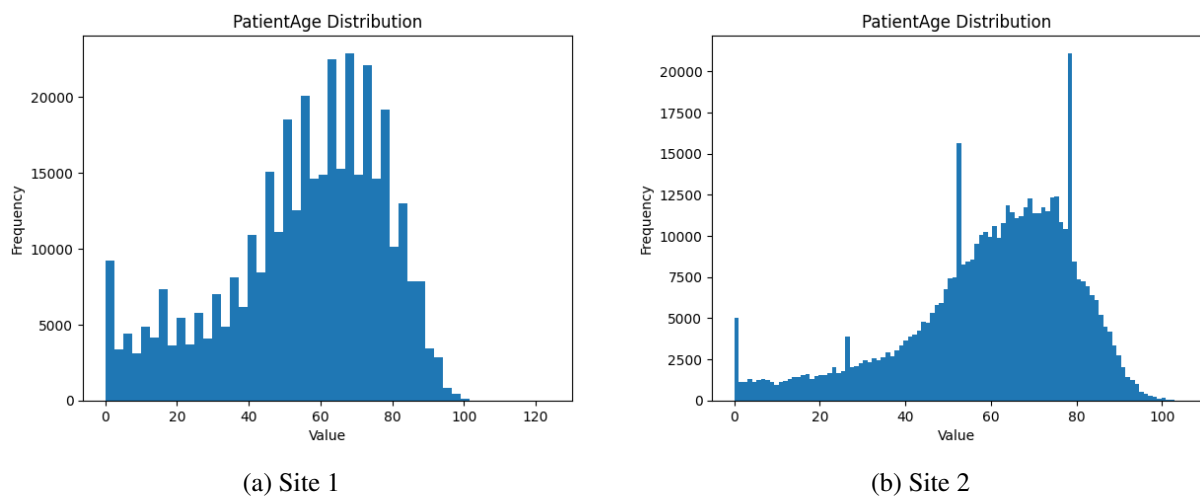


Figure 6.1: Age histogram of non X-ray radiological data.

Still focusing on demographic factors, gender distribution was also analyzed. Looking at Figure 6.2, the datasets show a relatively balanced representation between male and female patients.

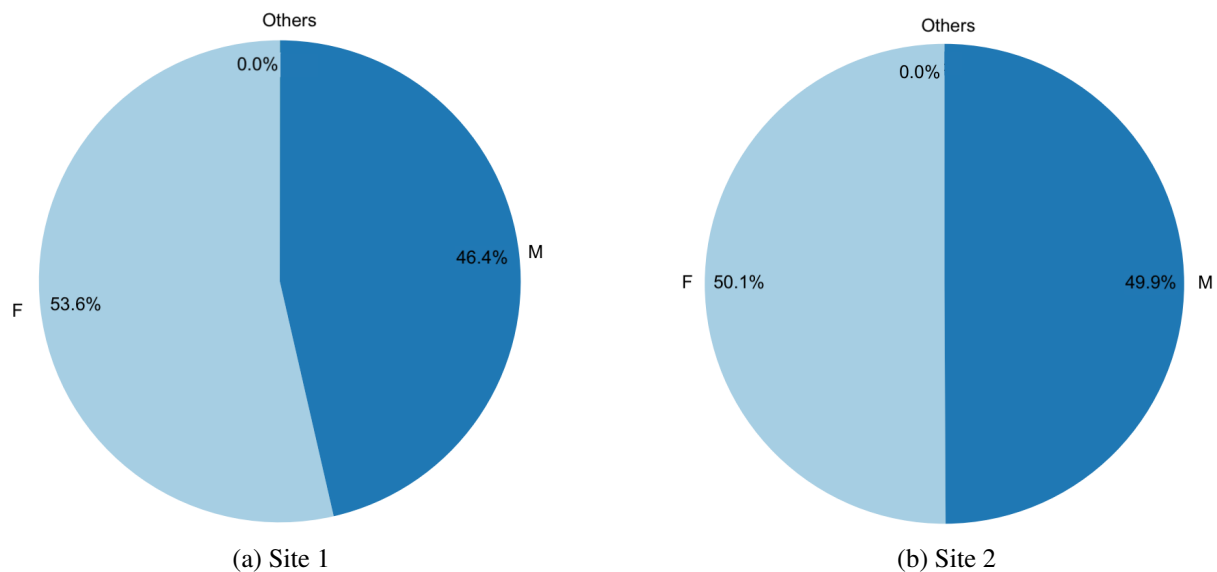


Figure 6.2: Patient gender histogram of non X-ray radiological data.

Other analysis were conducted regarding data understanding, namely percentages of rescheduled appointments, external resources and priority levels. This first part of the project was essential to identify possible error patterns and discuss internally to understand the reason.

6.1.1 Non X-ray Baseline Performance

One of the first steps done as data was collected was to acknowledge how was the performance of the dataframe without any preprocessing treatment. Hence, an initial version of the dataset was constructed in which only the most basic inconsistencies were addressed, namely the removal of null values essential for the functioning of the models. In addition, the target, was defined without any creation of new explanatory features. The main objective was to obtain a first performance metric, which would serve as an objective reference and baseline to evaluate future improvements resulting from data transformations, variable selection, creation of new features or adjustments to the algorithms used.

The baseline model was trained and tested using the raw dataset (with only minimal preprocessing), in order to serve as a reference for further comparisons. The corresponding results are presented in Table 6.5.

Table 6.5: Model validation metrics of the raw non X-ray dataset.

Model	Site 1 Non X-ray	Site 2 Non X-ray
XGBoost	$R^2 = 0.099$ MAE = 229.471 MSE = 329615.127 RMSE = 574.121	$R^2 = 0.240$ MAE = 291.720 MSE = 416714.956 RMSE = 645.534
LightGBM	$R^2 = 0.168$ MAE = 214.251 MSE = 304194.044 RMSE = 551.538	$R^2 = 0.221$ MAE = 288.523 MSE = 427036.560 RMSE = 653.480
CatBoost	$R^2 = 0.171$ MAE = 212.472 MSE = 302953.266 RMSE = 550.412	$R^2 = 0.238$ MAE = 283.522 MSE = 417672.133 RMSE = 646.276
Decision Tree	$R^2 = 0.120$ MAE = 220.952 MSE = 321919.110 RMSE = 567.379	$R^2 = 0.201$ MAE = 288.278 MSE = 438090.865 RMSE = 661.884

As expected, the values of performance are too low, as the model cannot explain much of the variability in the actual values. The error values are also relatively high, suggesting that the model's predictions deviate significantly from the actual values.

After understanding, preparing started.

6.2 Preliminary Data Preparing and Exploratory Assessment

In order to establish a solid foundation for treating outliers, a preliminary data preparation was essential. This preliminary preprocessing was made before the train and test split, since the methods applied did not compromise any data leakage. The preparing included variable format standardization, deduplication, row deletion based on criteria, simple imputation methods, derived feature creation and target definition. It is important to note that during this preparation phase, different types of variable were tested, and new derived features were created. To better understand the behavior of these new features, their distributions are also presented in this section. Additionally, auxiliary results are presented in this section, as they contributed to building and preparing the dataset in the most effective way. Finally, this section concludes by presenting the structured dataset that will be used in the subsequent outlier detection and cleaning procedures.

6.2.1 Variable Format Standardization

Variable format standardization refers to ensuring that data is represented consistently, regardless of its source or input, for different types of data. It is an essential step in data cleaning and preparation.

Date Data Types

The standardization of variables began with those of date type, as the extracted data required the calculation of various time intervals. To ensure these calculations were accurate, it was essential to convert all date-related variables to a consistent datetime format. Hence, to each date column (RequestDate, SchedulingDate, ExamDate, AdmissionDate, CallDate, RealizationDate, ReportDate, ValidationDate, DateNextAppointment, PreviousSchedulingDate, PreviousExamDate, ResultsDate), the following approach was followed:

1. Convert column values to dates, accepting multiple formats;
2. If there is an error in the conversion, replace with a null value;
3. Ensure column is in datetime format with seconds precision 'datetime64[s]'

Converting to 'datetime64[s]' removed unnecessary precision by storing dates at the second level (rather than microseconds or nanoseconds), which saved memory. This allowed consistent and accurate analysis of relevant times.

Numerical Data Types

To ensure all numerical features had the right type, all numerical relevant values were converted to numeric type. If conversion was not possible, the value would be replaced with a null value.

Boolean Data Types

As for boolean data types, the majority of boolean features had "Yes" or "No" values, so it was standardized to 0 and 1. For this, 'No' was replaced with 0 and 'Yes' with 1 in each row that it was verified. Then, the type was converted to boolean.

6.2.2 Deduplication

The deduplication process began by eliminating lines that were 100% duplicates, keeping only one of these versions. After some time analyzing the dataset, a very specific case of deduplication was also found.

During data cleaning, rows were identified as fully duplicated across all columns **except** for UsesVoiceRecognition feature, specifically in Site 1. These duplicate records were detected based on a defined subset of relevant columns, excluding UsesVoiceRecognition, which showed inconsistent values (1 or missing). This inconsistency was promptly identified and discussed internally. It was concluded that the issue was caused by the existence of two user accounts associated with the same staff member. According to internal rules, the correct record is the one where UsesVoiceRecognition is set to 1. Therefore, the other duplicates were removed from the dataset.

To address this:

1. All duplicate rows were extracted and removed from the original dataset. A new dataframe was created by selecting rows where there are duplicate values in specified columns (all except UsesVoiceRecognition). Remove all duplicate rows from the dataframe based on the specified columns.
2. Within the group of duplicates, rows with missing values in UsesVoiceRecognition were excluded.
3. The remaining rows (with valid values for UsesVoiceRecognition) were reinserted into the dataset by concatenation.

This approach ensured that only the correct entries, those with a defined value for UsesVoiceRecognition, were retained.

6.2.3 Case Deletion Based on Criteria

As the work progressed, it became clear that some values lacked significance, were not sufficiently representative, or could not be reliably imputed. For this reason, some of these rows were eliminated from the dataset (examples bellow).

FamilyMedicalAct

It was found a type of FamilyMedicalAct called "Cardiac MRI Test", as showed in Table 6.3. This was a clinical trial made by ByMe that did not progress and therefore did not generate tangible results. Knowing this, it was eliminated from the dataset.

MedicalAct

Additionally, as the preprocessing progressed, additional medical acts were found and realized it lacked significance for this context. This type of MedicalAct are additional tests, with no values relevant to the study. Therefore, it was decided that all additional categories would be automatically excluded, using the following process:

The goal of the procedure was to remove from the dataset all rows in which the value of the MedicalAct column begins with the expression "RM, ADDITIONAL".

1. For each row in the dataset, the value present in the MedicalAct column is checked.
2. If this value begins with the string "RM, ADDITIONAL", this row is marked for removal.
3. After going through all the rows, the dataset is updated to contain only the remaining rows, that is, those in which the value of the MedicalAct column does not begin with "RM, ADDITIONAL".

Rows without a MedicalAct were deleted since there was also no way to know what to impute.

Priority

As for underrepresented categories, the Priority feature had only one case "very urgent" within 491023 rows from Site 1, so it was excluded.

Origin

To better align with the clinical workflow under analysis at ByMe, it was decided to exclude rows containing medical acts originating from the emergency department. These cases were not compatible with the long-term objective of the study, which is to support the optimal timing for scheduling medical appointments based on when the corresponding exam results are predicted to be ready.

Age

For both sites, due to the very few cases with ages above 100, those rows were removed. Although most values are concentrated between 50 and 80 years, all age groups were considered relevant to the study. Therefore, all ages from 0 to 100 were retained in the dataset.

Others

Requests with a null resource, were excluded, as there was no viable imputation method.

Columns like ValidatedBy, AccessionNumber, PatientCode and ReportCode were excluded from the analysis, as they did not provide any added value to the study. They were only useful for identifying and filtering out specific cases with data inconsistencies.

AppointmentSpecialty variable was also excluded due to the high percentages of missing values for both Sites, and no viable way to impute it.

6.2.4 Imputation Methods Based on Known Patterns

Based on internal knowledge from ByMe it was possible to impute some boolean variables, UsesVoiceRecognition and HasImages. With knowledge of the processes and flows of the ByMe software, when these two features are null, it means that their value is false, so their value was replaced by 0 and applied again the variable standardization to type boolean.

During the preliminary preprocessing process, a situation has been identified where the ResultsDate column, which represents the due date set by the referring physician for the report to be delivered, is not populated in some cases. However, when the ResultsDate is missing but the DateNextAppointment is available, the latter is considered as the deadline for the report to be delivered, assuming that the report should be ready by the patient's next appointment.

Hence, an algorithm was performed according to the following steps:

1. Creates a logical mask that identifies all rows where ResultsDate is absent and DateNextAppointment is not null.
2. Counts how many rows meet this condition to know how many cases will be updated.
3. Replaces the missing values in ResultsDate with the respective dates of DateNextAppointment in those rows only.

This procedure ensures that there is a consistent reference date for the delivery of the medical report, even when the original date (ResultsDate) was not explicitly recorded.

6.2.5 Derived Feature Creation

After knowing the data, having a preliminary cleaning and understanding what the model needs to help predict exam reporting delays, it was necessary to start creating other features with the data available.

These variables were all created prior to the deletion of date variables. The statistics shown for each variable (Tables 6.6, 6.7, 6.8, 6.9, 6.10) reflect their distribution after preliminary cleaning.

SchedulingTimeDeviation

SchedulingTimeMedian was design to give the median within the data available, of the time it takes to schedule exams (SchedulingDate-RequestDate), for a specific (Origin, FamilyMedicalAct) tuple.

To address this:

1. First, it started by creating a unique identifier for each unique pair (Origin, FamilyMedicalAct). For each row in the dataset, this pair was created. Then all unique combinations of these pairs were identified and a numeric identifier (starting from 1) was assigned to each distinct combination. Then this identifier was stored in a new 'idAssignment' column.
2. Calculated the time it takes to schedule. For each row in the dataset, RequestDate was subtracted from SchedulingDate and the time difference was converted to seconds. Then, the result was divided by 3600 to get the value in hours. The result was stored in a new column 'delete'.
3. Calculated the median for each tuple. For each group Origin and FamilyMedicalAct the median of the values in the 'delete' column were calculated. Then the results were saved as a new structure where the index is the combination (Origin, FamilyMedicalAct), and the proper value is the corresponding median.
4. Assigned the median value to a new column, SchedulingTimeMedian, for each corresponding row in the dataset. For each row in the dataset the combination of Origin and FamilyMedicalAct was identified and the median of the 'delete' column for all rows with that same combination was calculated. Then, that median was assigned to the new SchedulingTimeMedian column.
5. Removed all columns containing the word "delete" (case insensitive) and dropped the idAssignment column from the dataframe.

After creating SchedulingTimeMedian, it was created a new column SchedulingTimeDeviation that stores the difference between SchedulingTime (SchedulingDate-RequestDate) and SchedulingTimeMedian. Then, SchedulingTimeMedian column was removed from the dataframe.

Table 6.6: Descriptive statistics for SchedulingTimeDeviation while preprocessing Non X-ray Radiological data of Site 1 and 2.

	Site 1 (in hours)	Site 2 (in hours)
Mean	636.466	563.748
STD	2124.883	1882.034
Min	-1917.360	-3207.182
25%	-716.968	-647.344
50%	0.160	3.838
75%	1153.129	1089.134
Max	16781.859	18847.356

A positive value of SchedulingTimeDeviation shows that the exam was scheduled later than usual (compared to the median for that combination of Origin and FamilyMedicalAct). This may indicate an abnormal delay or inefficiency in the scheduling process. As for a negative value of SchedulingTimeDeviation shows that the exam was scheduled more quickly than normal for that type of exam and origin, which may indicate a special priority, efficiency or deviations that deserve analysis (like automatic rescheduling).

TotalProcessTimeout

To quantify the total limit duration of the diagnostic process, the variable TotalProcessTimeout was created by calculating the time difference, in hours, between the request date (RequestDate) and the due date for the results to be available (ResultsDate), established by the referring physician at the time of request. This metric serves as a temporal baseline for understanding the expected waiting time for the delivery of exam results.

Some negative values were found within this variable, accounting for approximately 4% of the data. Since almost all of these values had a decimal scale below 0.001, suggesting that they likely resulted from workflow inconsistencies, these values were normalized to zero instead of row removal. This was made in order to preserve logical coherence and avoid misleading the model, as this would require more time to study the origin of the error.

Table 6.7: Descriptive statistics for TotalProcessTimeout while preprocessing Non X-ray Radiological data of Site 1 and 2.

	Site 1 (in hours)	Site 2 (in hours)
Mean	2082.109	1855.555
STD	2409.617	2121.994
Min	0.000	0.000
25%	158.793	225.029
50%	1326.986	1115.439
75%	3012.443	2720.220
Max	26303.979	18468.670

SchedulingTime

To quantify the total duration of the scheduling process, the variable SchedulingTime was created by calculating the time difference, in hours, between the date of scheduling (SchedulingDate) and the date the request was made (RequestDate). This metric is essential since while doing data understanding it was easy to see that there was a considerable amount of time between these two variables.

Table 6.8: Descriptive statistics for SchedulingTime while preprocessing Non X-ray Radiological data of Site 1 and 2.

	Site 1 (in hours)	Site 2 (in hours)
Mean	1660.752	1588.759
STD	2310.278	2029.390
Min	0.001	0.000
25%	25.542	125.016
50%	597.987	720.618
75%	2455.259	2326.591
Max	17974.452	19132.393

ExamDelay

To quantify the delay in realization, the variable ExamDelay was created by calculating the time difference, in hours, between the date of realization (RealizationDate) and the date for which the exam was scheduled (ExamDate). This metric is important to understand if a delay in reporting might be caused by a delay in the realization.

Since this variable was only meant to give the idea of the delay, any negative case (representing exams that began before the supposed) was normalized to 0, indicating only that the delay was 0.

Table 6.9: Descriptive statistics for ExamDelay while preprocessing Non X-ray Radiological data of Site 1 and 2.

	Site 1 (in hours)	Site 2 (in hours)
Mean	18.822	11.826
STD	114.639	85.985
Min	0.000	0.000
25%	0.108	0.045
50%	0.466	0.308
75%	0.965	0.813
Max	11348.515	8791.266

WaitingForExam

To quantify the time between the day of the request and the day of the actual exam, the variable WaitingForExam was created by calculating the time difference, in hours, between the date the exam was scheduled for (ExamDate) and the date of request (RequestDate). This metric is important because it allows for the analysis of the efficiency and responsiveness of the scheduling system.

Table 6.10: Descriptive statistics for WaitingForExam while preprocessing Non X-ray Radiological data of Site 1 and 2.

	Site 1 (in hours)	Site 2 (in hours)
Mean	1973.552	1744.022
StD	2402.242	2083.212
Min	0.000	0.000
25%	169.878	238.719
50%	1029.067	902.535
75%	2889.066	2543.479
Max	28344.199	19132.393

Hour, Day of the Week, Month and Shift of Realization

The Month_RealizationDate column was created to capture the month in which a medical exam was performed, for the purpose of performing temporal analysis. This can be useful for identifying seasonal patterns or variations in volume within certain months, which may be related to factors such as resource availability or hospital workflows during different times of the year.

This variable was initially a continuous numerical variable. To properly represent the cyclical nature of time-related variables, the original features were transformed using sine and cosine functions. This encoding captures the circular continuity of time (e.g., 11:00 pm and 12:00 am are close in time but far apart numerically) and allows ML models to better understand temporal patterns. After this transformation, only the resulting features were retained (Month_cos and Month_sin), as together they preserve the complete cyclic information of the original variable. This can help temporal feature engineering and avoids misleading the model with artificial discontinuities in the data. The original features were then eliminated.

The same approach was used in the remaining temporal variables:

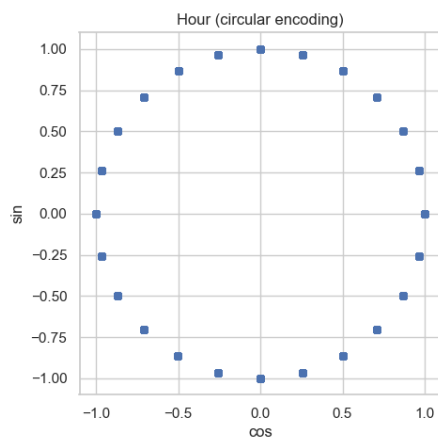
- DayWeek_RealizationDate
- Hour_RealizationDate
- RealizationShift

The DayWeek_RealizationDate column is designed, in this case, to capture the day of the week in which a medical exam was performed, for the same purpose as above. This will help to identify different patterns along the week. The creation of this feature, which directly extracts the day of the week from the date, made the variable RealizationDay, which had many missing values (Table 6.1), unnecessary and thus it was removed. After the transformation, two features were created (DayWeek_cos and DayWeek_sin) and the original removed.

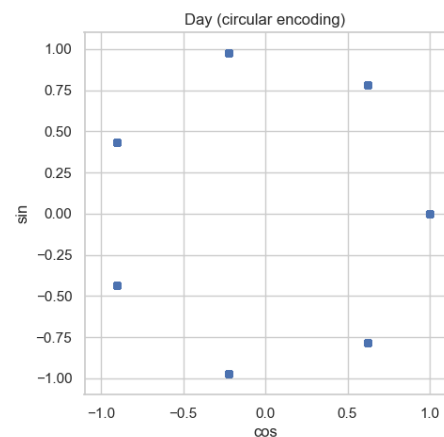
The Hour_RealizationDate column is designed to capture the hour of the day in which a medical exam was performed, for the same purpose as above. This will help to identify different patterns along the day. After the transformation, two features were created (Hour_cos and Hour_sin) and the original removed.

The RealizationShift column was initially a categorical variable that just indicated the number of the shift (explained in Table 5.3). Since it was possible to obtain an auxiliary dataframe that indicated the hours of each shift, each shift category was associated with an approximate average time that exams occur within that period. For this, it was used a map function that replaces the strings in the RealizationShift column with the corresponding mean time, converting the categorical variable into a continuous numeric feature that can be directly used. After this, the same cyclic transformation was used to test, creating RealizationShift_cos and RealizationShift_sin and the original was removed.

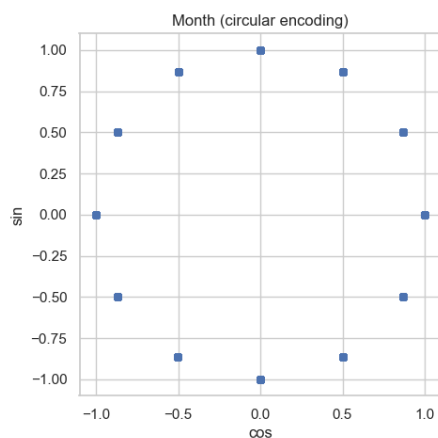
The final encoding of these four variables is plotted in Figure 6.3. This figure corresponds to Site 2 but Site 1 was equal.



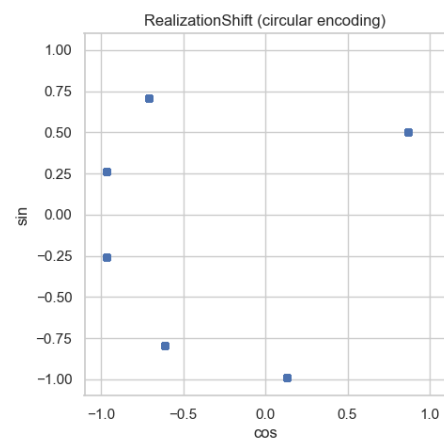
(a) Circular encoding of the hour of realization.



(b) Circular encoding of the day of the week of realization.



(c) Circular encoding of the month of realization.



(d) Circular encoding of the shift of realization.

Figure 6.3: Scatter plots between sine and cosine of cyclic variables from the training set of Site 2.

To test the hypothesis of cyclic variables being better than continuous, both were tested and the results are represented below for Site 1 (Table 6.11) and Site 2 (Table 6.12).

Table 6.11: Model metrics comparison using temporal variables represented either as continuous numerical values or as cyclic transformations (sine and cosine) for hour, day of the week, month and shift in Site 1.

Model	Continuous Variables	Cyclic Variables
XGBoost	$R^2 = 0.550$ MAE = 254.091 MSE = 235619.079 RMSE = 485.406	$R^2 = 0.557$ MAE = 246.550 MSE = 231764.291 RMSE = 481.419
LightGBM	$R^2 = 0.722$ MAE = 162.012 MSE = 145464.575 RMSE = 381.398	$R^2 = 0.720$ MAE = 162.452 MSE = 146389.288 RMSE = 382.609
CatBoost	$R^2 = 0.731$ MAE = 164.454 MSE = 140625.934 RMSE = 375.001	$R^2 = 0.732$ MAE = 165.075 MSE = 140091.573 RMSE = 374.288
Decision Tree	$R^2 = 0.609$ MAE = 193.330 MSE = 204219.726 RMSE = 451.907	$R^2 = 0.606$ MAE = 192.994 MSE = 205978.685 RMSE = 453.849

For Site 1, the performance difference between the approaches is small across all models. This shows that no single transformation is dramatically superior, which can be explained by the fact that data has other strong variables.

Table 6.12: Model metrics comparison using temporal variables represented either as continuous numerical values or as cyclic transformations (sine and cosine) for hour, day of the week, month and shift in Site 2.

Model	Continuous Variables	Cyclic Variables
XGBoost	$R^2 = 0.502$ MAE = 290.505 MSE = 336802.336 RMSE = 580.347	$R^2 = 0.511$ MAE = 287.617 MSE = 330938.297 RMSE = 575.272
LightGBM	$R^2 = 0.816$ MAE = 138.581 MSE = 124458.269 RMSE = 352.786	$R^2 = 0.817$ MAE = 137.321 MSE = 123699.311 RMSE = 351.709
CatBoost	$R^2 = 0.801$ MAE = 146.123 MSE = 134565.379 RMSE = 366.832	$R^2 = 0.809$ MAE = 143.963 MSE = 129300.618 RMSE = 359.584
Decision Tree	$R^2 = 0.667$ MAE = 112.728 MSE = 224909.529 RMSE = 474.246	$R^2 = 0.712$ MAE = 106.267 MSE = 194350.600 RMSE = 440.852

The tests performed at Site 2 reinforce the decision to use temporal variables with cyclic representation instead of continuous values since all models performed better with cyclical variables, even if in some cases the difference was small.

From this stage of the work, it was decided to represent these temporal variables as cyclical components instead of continuous numerical values. This was decided because this type of variable is more natural and gives an accurate representation of time, it gives greater ability to the model to capture real temporal patterns and the results will be more robust, especially because this is a context where time and seasonality influences system behavior.

6.2.6 Target Definition

Given that both Sites had identical feature sets, the definition of the target variable was aligned and applied equally to both hospital datasets. As explained in Section 5.3, the process for X-ray exams differs significantly from that of other imaging exams. Hence, it was necessary to address that issue while creating the non X-ray target.

The delay in obtaining the report results for non X-ray data, considering the initially proposed limit, was calculated as:

$$\text{Delay in Conclusion (in hours)} = \text{ResultsDate} - \text{ValidationDate}$$

This variable was used as the target in a regression model, with the goal of predicting the deviation (either a delay or an earliness) between the due date to obtain the results and the date where the report results become available. By modeling this delay, it became possible to identify patterns and contributing factors that can support more efficient scheduling of medical appointments, taking into account the expected completion time of the report.

If the model predicts a positive value (e.g. +200h), can be expected the report to be ready 8 days in advance, allows to schedule the appointment earlier, if needed. If it predicts a negative value (e.g. -120h), indicates that the report will arrive 5 days after the deadline, allowing the hospital to schedule the appointment with that margin.

To preserve the **temporal causality** and ensure the model is realistic and applicable in a real-world scenario, all variables that were only known after the target date (after the report results are available) were removed from the dataset after target creation (ValidationDay, ValidationShift, ValidationDate). This is essential to avoid data leakage and to ensure that the model only uses information that would have been available at the time of prediction.

Importantly, ResultsDate is defined by the physician at the time of the exam request as the intended deadline for availability of the exam report and is thus known *a priori*. As such, variables such as TotalProcessTimeout do not violate the principle of causality when used to compute the delay. Features that do not incorporate future information related to the actual validation process, do not introduce data leakage into the model.

After all features created, all date variables (RequestDate, SchedulingDate, ExamDate, AdmissionDate, CallDate, RealizationDate, ReportDate, ValidationDate, DateNextAppointment, PreviousSchedulingDate, PreviousExamDate, ResultsDate) were removed from the dataframe after temporal features creation.

6.2.7 Final Dataset Structure After Preliminary Cleaning

After preliminary preparing, the dataset from Site 1 contains 251249 rows and 30 columns, while the dataset from Site 2 consists of 295169 rows and the same 30 columns. Hence, the structure of both datasets contains the following columns:

- AutomaticScheduling
- DirectAdmission
- Rescheduled
- Origin
- RealizationUnit
- Resource
- ExternalResource
- PatientAge
- PatientGender
- Priority
- Specialty
- MedicalAct
- FamilyMedicalAct
- TypeMedicalAct
- HasImages
- UsesVoiceRecognition
- TotalProcessTimeout
- SchedulingTime

- *WaitingForExam*
- *SchedulingTimeDeviation*
- *ExamDelay*
- *RealizationShift_sin*
- *RealizationShift_cos*
- *Hour_sin*
- *Hour_cos*
- *Month_sin*
- *Month_cos*
- *Day_sin*
- *Day_cos*
- *target_delay_conclusion*

In order to study seasonality, graphs of average times per weekday and average times per month were plotted, in order to correctly understand patterns. Since the report is supposed to start being written after the exam being taken, the month of realization is the key event that marks the beginning of the report issuance process. The performing of the exam defines the starting point of the timeframe for the construction of the medical report. Hence, by grouping data by month of realization, it is possible to identify seasonal patterns that directly impact the performance of reporting, such as volume increases, holiday periods or operational limitations. The same happens for weekday.

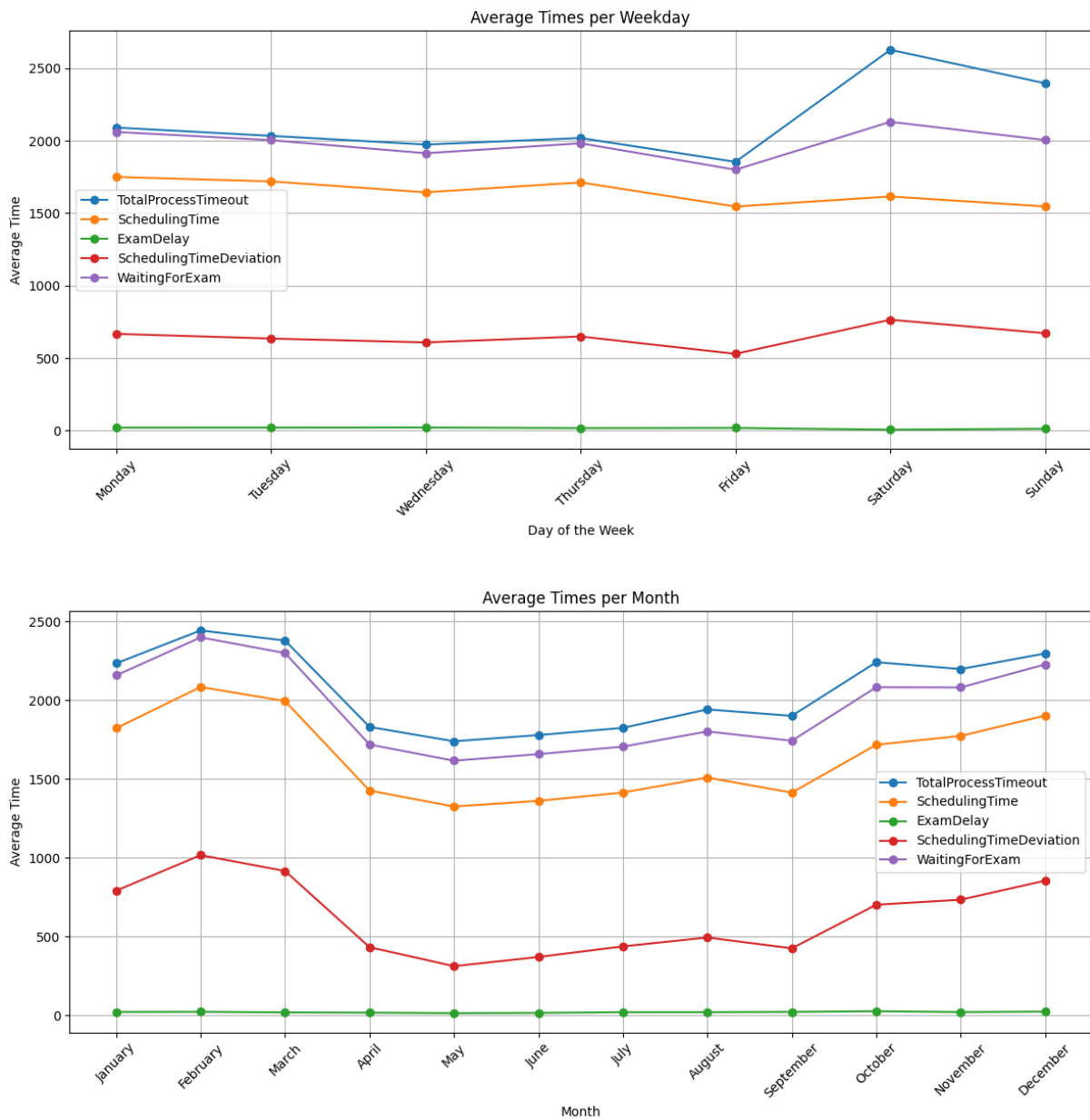


Figure 6.4: Seasonality study for Non X-ray Radiological data of Site 1.

Observing Figures 6.4 and 6.5, the graphs show the average variation of different temporal metrics over the months, considering the month in which the exam was performed as a temporal reference. The month was reconstructed using cyclic coding (Month_sin and Month_cos), allowing the original value of the month to be recovered. The same was made for weekday, but using Day_sin and Day_cos, in this case.

In the spring and summer months, it is possible to observe that, in both sites, the processing times plotted are generally lower between April and September. This is also in line with expectations, as these months typically correspond to lower seasonal demand, for example, fewer cases of flu and respiratory infections, which reduces overall pressure on the system.

On weekdays, the processing times tend to be lower, which is expected given that during weekends there are fewer medical exams, reduced staff availability, and consequently, increased delays.

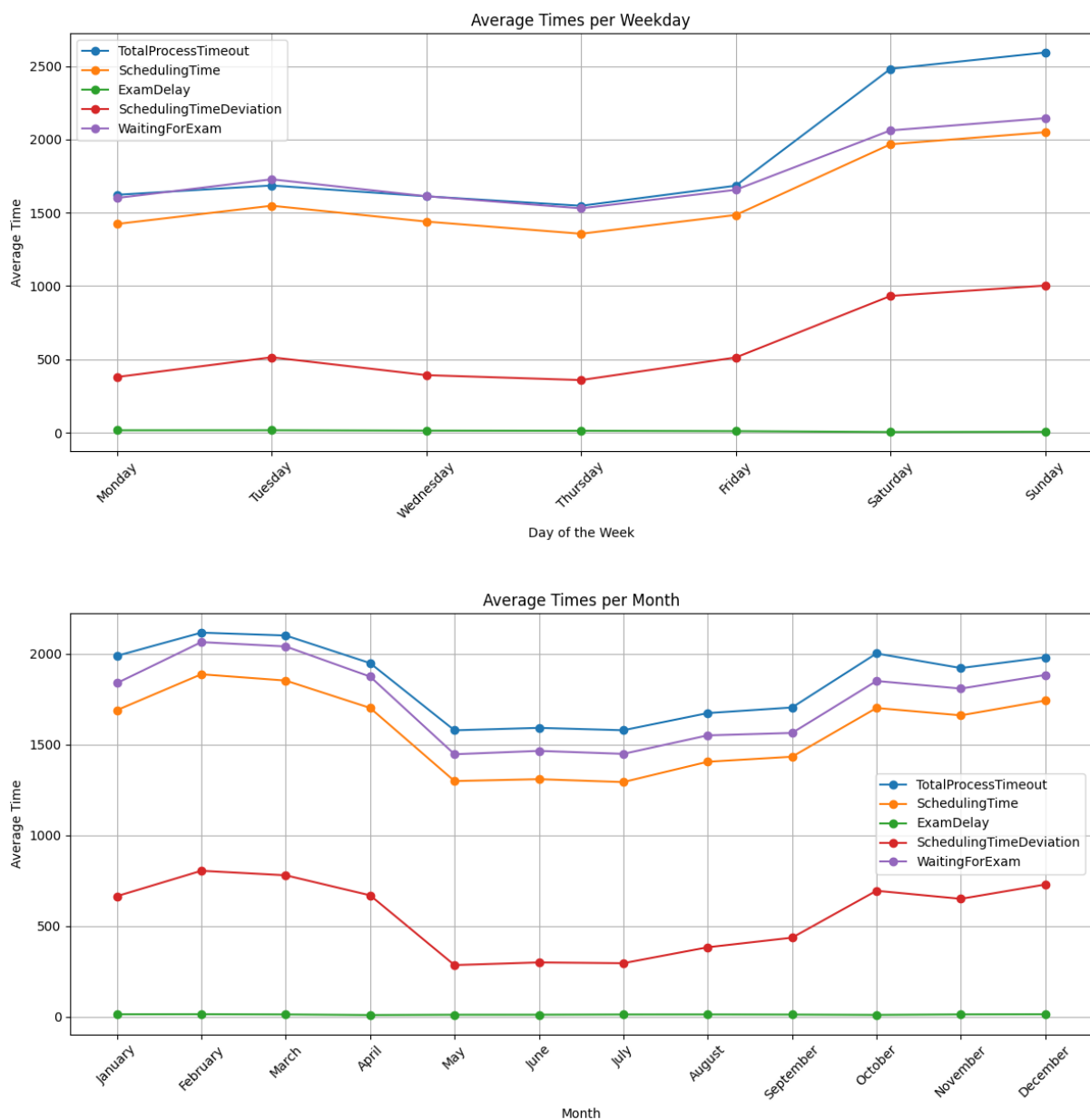


Figure 6.5: Seasonality study for Non X-ray Radiological data of Site 2.

It was important to understand how has the preliminary cleaning and the creation of carefully chosen features affect the model performance and error. As it can be seen in Table 6.13, the results improve drastically.

Table 6.13: Model validation metrics after preliminary cleaning and feature creation for non X-ray data.

Model	Site 1 Non X-ray	Site 2 Non X-ray
XGBoost	$R^2 = 0.557$ MAE = 246.550 MSE = 231764.291 RMSE = 481.419	$R^2 = 0.511$ MAE = 287.617 MSE = 330938.297 RMSE = 575.272
LightGBM	$R^2 = 0.720$ MAE = 162.452 MSE = 146389.288 RMSE = 382.609	$R^2 = 0.817$ MAE = 137.321 MSE = 123699.311 RMSE = 351.709
CatBoost	$R^2 = 0.732$ MAE = 165.075 MSE = 140091.573 RMSE = 374.288	$R^2 = 0.809$ MAE = 143.963 MSE = 129300.618 RMSE = 359.584
Decision Tree	$R^2 = 0.606$ MAE = 192.994 MSE = 205978.685 RMSE = 453.849	$R^2 = 0.712$ MAE = 106.267 MSE = 194350.600 RMSE = 440.852

The results show XGBoost with the worst result. Even with the categorical enable option set at true, the model seems to be limited when using rare categorical variables.

After establishing a solid structure, free of missing values, containing the features relevant to the model's objective, and with a well-defined target, the next step is to handle outliers. The following three sections (6.3, 6.4 and 6.5) focus on treating different kinds of outliers. Each section includes:

1. A subsection explaining the detection approaches used;
2. A subsection detailing the treatment strategies used;
3. A subsection presenting the results for different pair of detection and treatment technique.

6.3 Categorical Outliers

Categorical outliers are instances where the values in categorical columns have too little significance. These outliers may represent rare events or errors in data entry. The detection and treatment of categorical outliers play a crucial role in improving the overall quality of the data and ensuring the robustness of the predictive models.

6.3.1 Detection

To address categorical outliers, the categorical variable with the highest number of unique categories was selected. This choice was made because a variable with many distinct values increases the likelihood of containing rare or anomalous categories that could be considered outliers. MedicalAct, as it can be seen in Table 6.1, has for both sites >500 unique values, so it was chosen to treat outliers.

Site 1 - Medical Act

In order to detect underrepresented features a frequency analysis was performed and it is presented in Figure 6.6. For this visualization, categories with counts below 0.7% of the total were grouped into a

single class labeled "Others" in order to improve readability and interpretability. The complete list of all original categories, including those grouped under "Others", can be found in Appendix C, Table C.1.

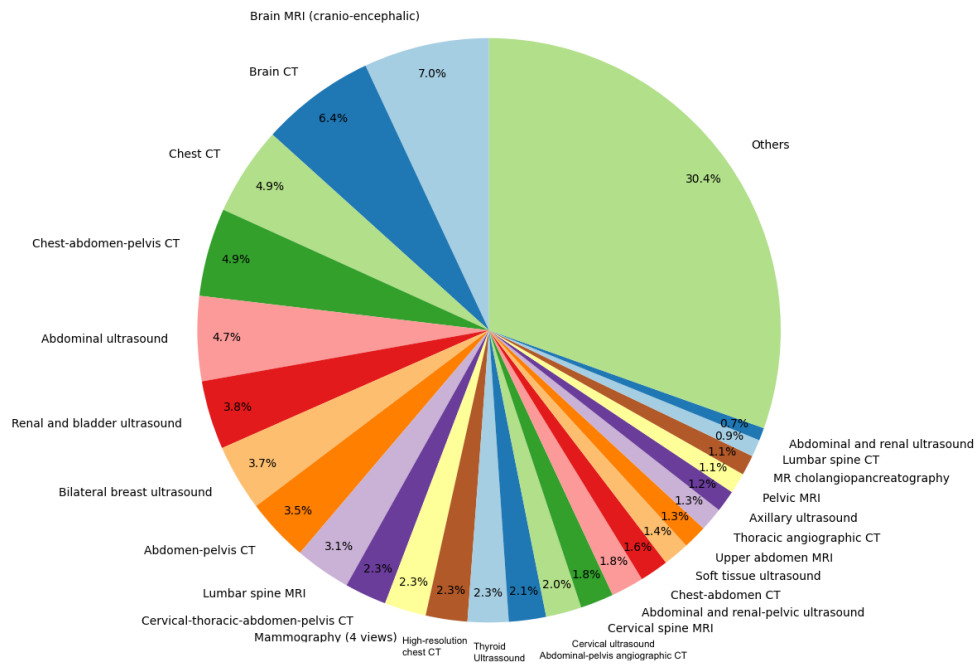


Figure 6.6: Frequency analysis of MedicalAct categories of Site 1.

Site 2 - Medical Act

As it can be seen in Figure 6.7, even after a preliminary cleaning, there are categories with very low significance. For this visualization, categories with counts below 0.7% of the total were grouped into a single class labeled "Others" in order to improve readability and interpretability. The complete list of all original categories, including those grouped under "Others", can be found in Appendix C, Table C.2.

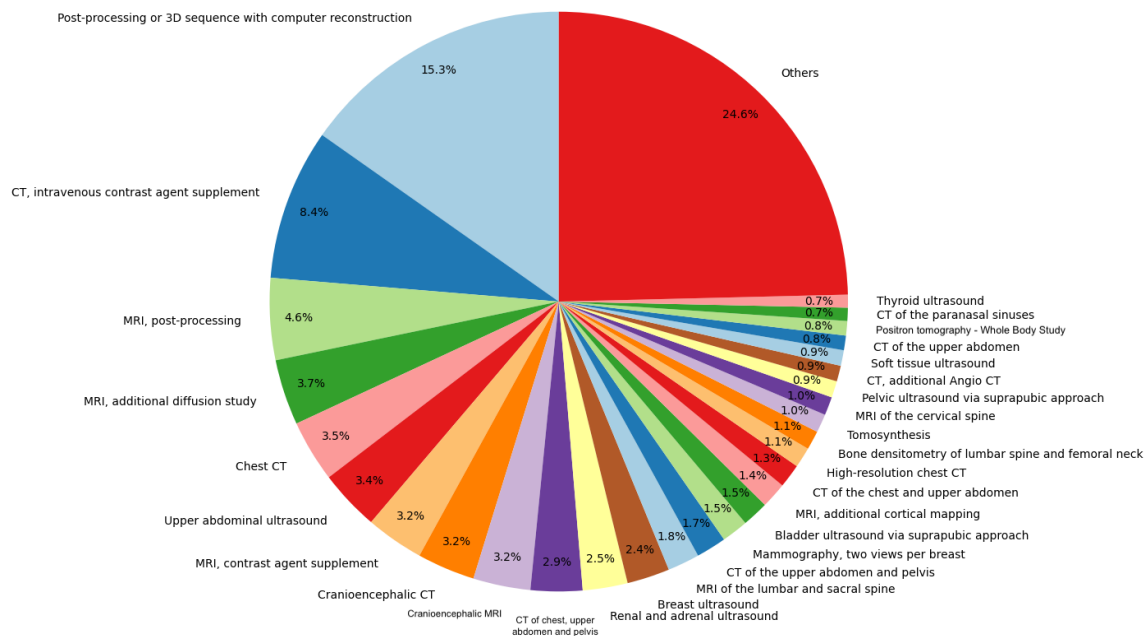


Figure 6.7: Frequency analysis of MedicalAct categories of Site 2.

For the purposes of this report, the medical acts presented above have been translated into English using DeepL Translate web tool, solely for demonstration and alignment with the language of the document. The same happens in the Appendix. The medical acts presented in this dataset are institution-specific and defined according to the internal parametrization of the hospital. As such, any apparent redundancy or similarity between acts cannot be arbitrarily grouped or simplified, since each corresponds to a distinct clinical, administrative, or billing context defined by the institution.

6.3.2 Treatment

In order to prepare the pipeline to handle underrepresented features, two approaches were tried.

The first function processes a given categorical column based on a defined frequency threshold. It identifies categories whose number of occurrences is less than or equal to the threshold and considers them as rare. The function receives as input the column, the threshold, an option of treatment, and a tag.

- If option = 0, these rare categories are replaced with NaN, treating them as missing values.
- If option = 1, the rare categories are replaced with a predefined label (by default, "Other").
- If an invalid option is provided, an error message is printed, and the column is returned unchanged.

The second function replaces the original column in the dataframe with the updated version processed by the previous function. If the column contains any missing values (as a result of option 0), the corresponding rows are removed from the dataframe. The cleaned dataframe is then returned.

Both approaches were compared and evaluated.

6.3.3 Performance Variability Results

After having a measurement of performance of the preliminary cleaned dataset, it was time to address categorical outliers. In order to assume a single approach for the remaining preprocessing, both options (replacing - Table 6.14, and removing - Table 6.15) were tested, each with two different thresholds.

Table 6.14: Treating categorical outliers in non X-ray by replacing the rarest categories with "Other". The percentage of medical acts affected is next to the threshold value.

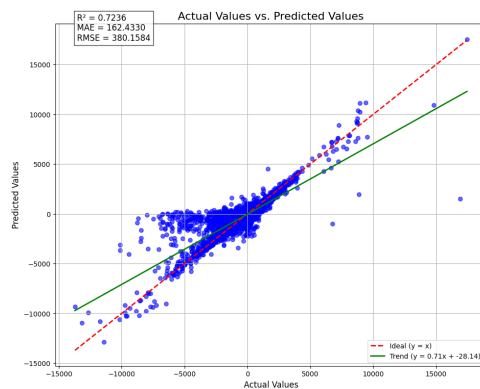
Model	Site 1 Non X-ray		Site 2 Non X-ray	
	threshold=100 ($\approx 2.68\%$)	threshold=1000 ($\approx 17.98\%$)	threshold=100 ($\approx 2.07\%$)	threshold=1000 ($\approx 14.82\%$)
XGBoost	$R^2 = 0.656$ MAE = 199.741 MSE = 179877.259 RMSE = 424.119	$R^2 = 0.613$ MAE = 213.345 MSE = 202305.010 RMSE = 449.783	$R^2 = 0.777$ MAE = 145.506 MSE = 150637.679 RMSE = 388.121	$R^2 = 0.763$ MAE = 151.258 MSE = 160363.510 RMSE = 400.454
LightGBM	$R^2 = 0.723$ MAE = 162.467 MSE = 145067.710 RMSE = 380.878	$R^2 = 0.724$ MAE = 162.433 MSE = 144520.447 RMSE = 380.158	$R^2 = 0.821$ MAE = 136.491 MSE = 121319.126 RMSE = 348.309	$R^2 = 0.831$ MAE = 131.906 MSE = 114563.024 RMSE = 338.472
CatBoost	$R^2 = 0.730$ MAE = 165.341 MSE = 141096.138 RMSE = 375.628	$R^2 = 0.730$ MAE = 165.615 MSE = 141285.217 RMSE = 375.879	$R^2 = 0.813$ MAE = 143.720 MSE = 126653.005 RMSE = 355.883	$R^2 = 0.813$ MAE = 143.163 MSE = 126589.019 RMSE = 355.794
Decision Tree	$R^2 = 0.613$ MAE = 192.905 MSE = 202437.723 RMSE = 449.931	$R^2 = 0.614$ MAE = 193.040 MSE = 201676.149 RMSE = 449.084	$R^2 = 0.712$ MAE = 107.320 MSE = 195145.192 RMSE = 441.752	$R^2 = 0.723$ MAE = 104.968 MSE = 187361.275 RMSE = 432.852

By looking at XGBoost previous results (Table 6.13) and the results after applying a replace approach, with a wider category, it is possible to understand that now with less insignificant categories, XGBoost improves both in performance as in error.

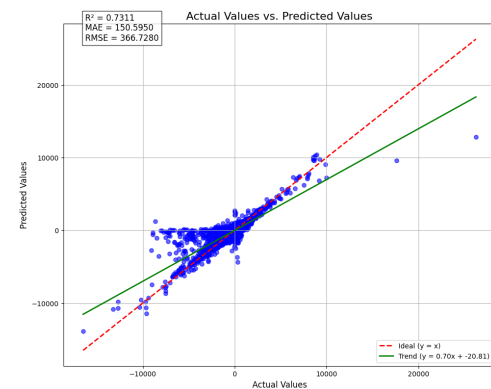
Table 6.15: Treating categorical outliers in non X-ray by deleting the rarest categories of MedicalAct. The percentage of medical acts affected is next to the threshold value.

Model	Site 1 Non X-ray		Site 2 Non X-ray	
	threshold=100 ($\approx 2.68\%$)	threshold=1000 ($\approx 17.98\%$)	threshold=100 ($\approx 2.07\%$)	threshold=1000 ($\approx 14.82\%$)
XGBoost	$R^2 = 0.716$ MAE = 162.510 MSE = 151279.156 RMSE = 388.946	$R^2 = 0.689$ MAE = 164.744 MSE = 155690.528 RMSE = 394.576	$R^2 = 0.817$ MAE = 132.055 MSE = 123245.287 RMSE = 351.063	$R^2 = 0.858$ MAE = 104.106 MSE = 77871.089 RMSE = 279.054
LightGBM	$R^2 = 0.724$ MAE = 161.378 MSE = 146697.996 RMSE = 383.012	$R^2 = 0.731$ MAE = 150.595 MSE = 134489.445 RMSE = 366.728	$R^2 = 0.830$ MAE = 135.315 MSE = 114494.014 RMSE = 338.370	$R^2 = 0.831$ MAE = 126.223 MSE = 92506.720 RMSE = 304.149
CatBoost	$R^2 = 0.726$ MAE = 164.074 MSE = 145716.910 RMSE = 381.729	$R^2 = 0.730$ MAE = 153.960 MSE = 134704.543 RMSE = 367.021	$R^2 = 0.825$ MAE = 141.318 MSE = 118084.140 RMSE = 343.634	$R^2 = 0.812$ MAE = 134.842 MSE = 103195.929 RMSE = 321.241
Decision Tree	$R^2 = 0.608$ MAE = 190.534 MSE = 208566.524 RMSE = 456.690	$R^2 = 0.579$ MAE = 189.129 MSE = 210730.947 RMSE = 459.054	$R^2 = 0.696$ MAE = 118.625 MSE = 204937.227 RMSE = 452.700	$R^2 = 0.692$ MAE = 116.400 MSE = 168732.563 RMSE = 410.771

To better understand why deleting has given better results, a plot was generated for each approach to provide a clearer visual representation. In Figures 6.8 and 6.9 it is possible to observe both Sites.

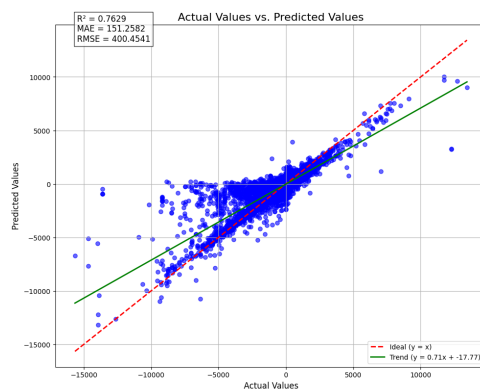


(a) Replacing of MedicalActs that have <1000 counts with a wider class "Other".

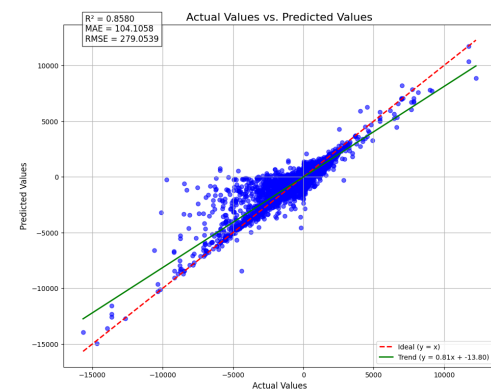


(b) Removing of MedicalActs that have <1000 counts.

Figure 6.8: LightGBM evolution with replacing and removal of categorical outliers at Site 1.



(a) Replacing of MedicalActs that have <1000 counts with a wider class "Other".



(b) Removing of MedicalActs that have <1000 counts.

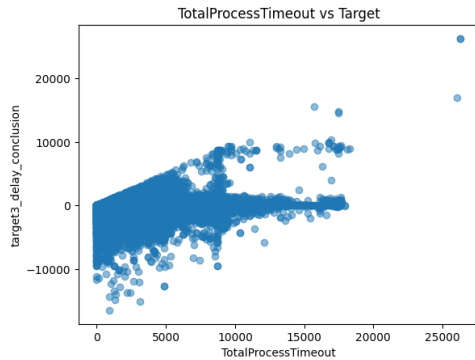
Figure 6.9: XGBoost evolution with replacing and removal of categorical outliers at Site 2.

6.4 Manual Approach: Visual-Based Numerical Outlier Removal

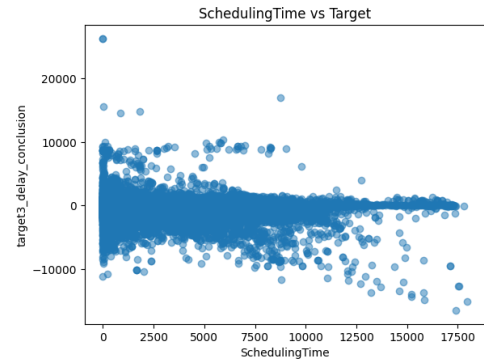
After treating categorical outliers, a manual approach was performed with the goal of identifying potential patterns and inconsistencies, supported by visual analysis.

6.4.1 Detection and Treatment

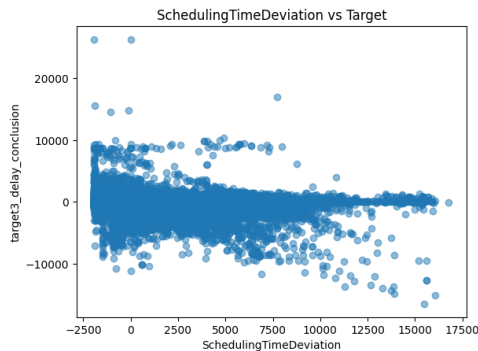
The aim was to uncover relationships between the new features created in preliminary preprocessing and understand how they are connected to the target. Figure 6.10 presents distributions of Site 1.



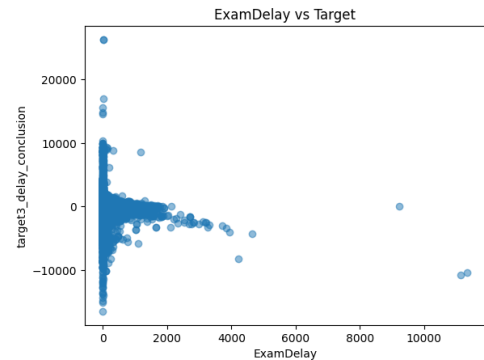
(a) In TotalProcessTimeout, only values below 9000 were maintained.



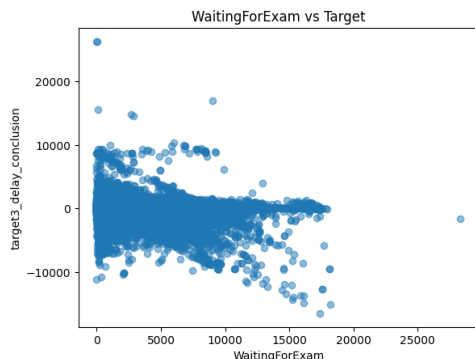
(b) In SchedulingTime, only values below 12500 were maintained.



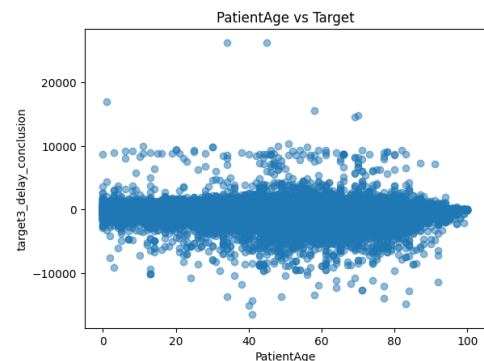
(c) In SchedulingTimeDeviation, only values below 12000 were maintained.



(d) In ExamDelay, only values below 2000 were maintained.



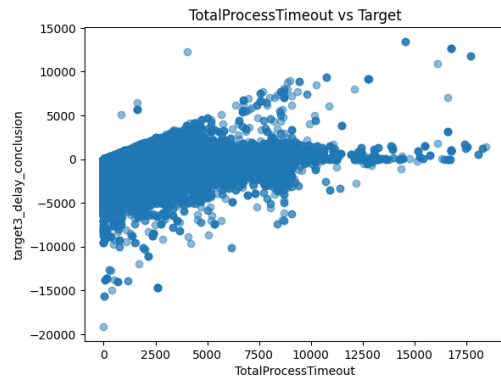
(e) In WaitingForExam, only values below 12500 were maintained.



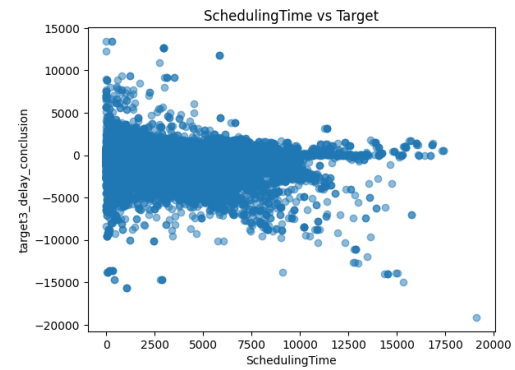
(f) In PatientAge, although there were less significant values, all were relevant, so every row was maintained.

Figure 6.10: Site 1 scatter plots of the train set numerical features without any treatment to allow proper visualization and manual removal.

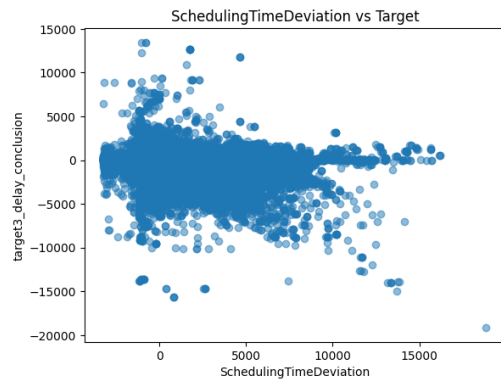
Several scatter plots were designed to all numerical continuous features. At this stage, cyclical transformations of temporal variables, such as shift, time, month or day of the week, were not changed. **Based on the plots, extreme outlier points were visually identified and, the respective rows, removed to improve data quality.** Below each plot, the number of removed points is indicated. These thresholds were defined after multiple iterations aimed at improving model performance and reducing error.



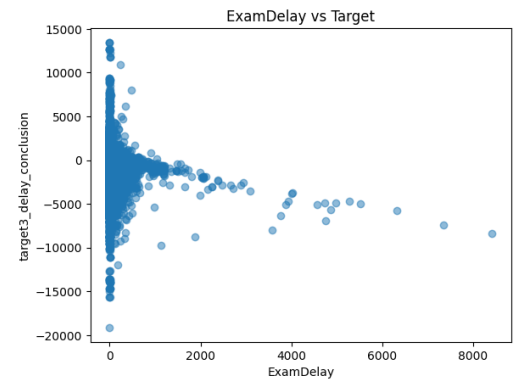
(a) In TotalProcessTimeout, only values below 10000 were maintained.



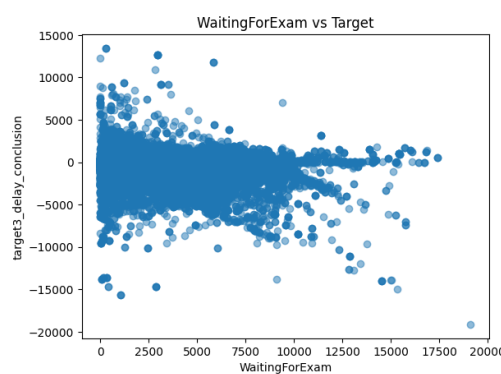
(b) In SchedulingTime, only values below 11000 were maintained.



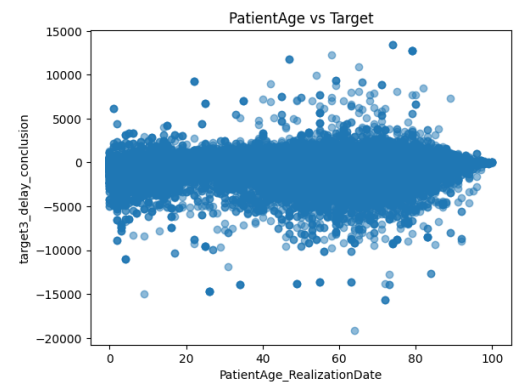
(c) In SchedulingTimeDeviation, only values below 9000 were maintained.



(d) In ExamDelay, only values below 1000 were maintained.



(e) In WaitingForExam, only values below 9500 were maintained.



(f) PatientAge, although there were less significant values, all were relevant, so every row was maintained.

Figure 6.11: Site 2 scatter plots of the train set numerical features without any treatment to allow proper visualization and manual removal.

This approach was mainly applied to the numerical features established in preliminary cleaning, such as the ones in Figure 6.11. Values considered inconsistent or distant from the general trend observed in the

graphs were manually removed, based on thresholds defined by visual analysis.

6.4.2 Performance Variability Results

In this iteration, to analyze the change in performance, the preliminary basis of the dataset was used along with the best option of treatment within categorical outliers (explored in the previous section), the removal of categorical outliers with threshold 1000. The results are presented in Table 6.16.

Table 6.16: Model validation metrics after detection performed by plotting scatter graphs of the features TotalProcessTimeout, ExamDelay, SchedulingTime, SchedulingTimeDeviation, WaitingForExam and PatientAge, followed by manual deletion.

Model	Site 1 Non X-ray	Site 2 Non X-ray
XGBoost	$R^2 = 0.672$ MAE = 160.963 MSE = 137145.363 RMSE = 370.331	$R^2 = 0.849$ MAE = 101.883 MSE = 74039.745 RMSE = 272.102
LightGBM	$R^2 = 0.709$ MAE = 147.008 MSE = 121631.643 RMSE = 348.757	$R^2 = 0.816$ MAE = 123.573 MSE = 89904.678 RMSE = 299.841
CatBoost	$R^2 = 0.710$ MAE = 150.158 MSE = 121457.564 RMSE = 348.508	$R^2 = 0.800$ MAE = 132.033 MSE = 98230.320 RMSE = 313.417
Decision Tree	$R^2 = 0.557$ MAE = 182.223 MSE = 185371.757 RMSE = 430.548	$R^2 = 0.648$ MAE = 116.272 MSE = 172297.895 RMSE = 415.088

For both sites, although performance slightly worsened when compared to only removing categorical outliers (Table 6.15), all type of errors decreased significantly, contributing to a more balanced model, with lower overall error and more stable.

6.5 Automatic Approach: Statistical and ML-Based Numerical Outlier Handling

In a different approach, data was analyzed with more complex methods in this section. In order to prevent data leakage, the dataframe was divided in train and test prior to the numerical outlier repairing. All steps involving statistical computation, such as outlier detection thresholds, were **performed exclusively on the training set**. The same thresholds were then applied to the test set, ensuring no information from the test set influenced model training or preprocessing decisions.

6.5.1 Detection

For outliers treatment, different methods were applied to **detect potential outliers**. The process started by a visual analysis of frequency and boxplots, in the training set. The boxplots shown in this section use IQR method with a factor of 1.5 to identify potential outliers. Alternative thresholds were also considered and are explored and evaluated in the Performance Variability Results sections. Presented in Figure 6.12, 6.13, 6.14, 6.15, 6.16 and 6.17 are the distributions of Site 1.

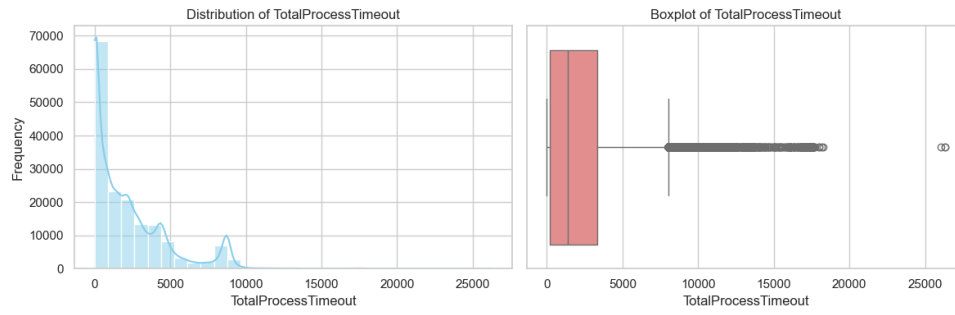


Figure 6.12: Distribution and outlier visualization of TotalProcessTimeout Site 1.

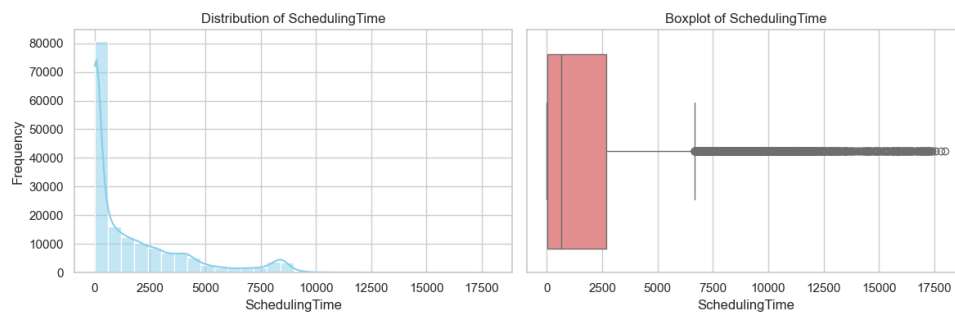


Figure 6.13: Distribution and outlier visualization of SchedulingTime Site 1.

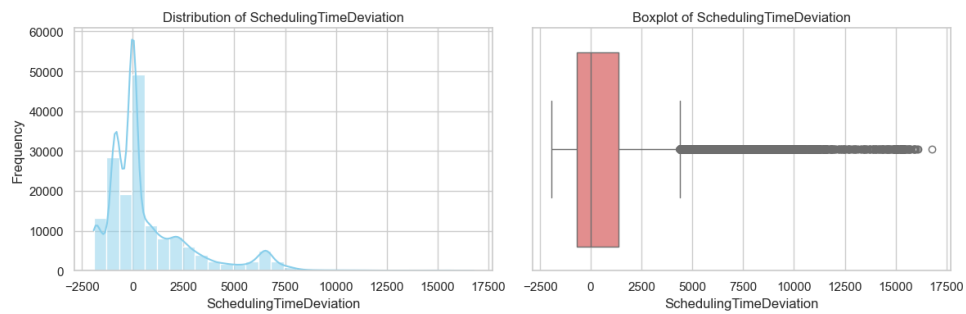


Figure 6.14: Distribution and outlier visualization of SchedulingTimeDeviation Site 1.

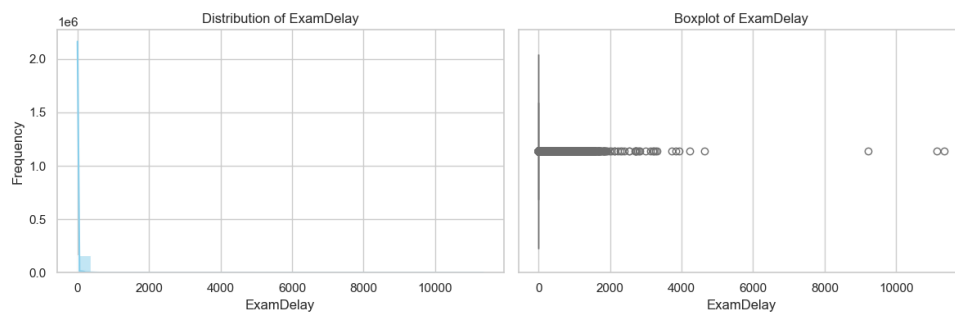


Figure 6.15: Distribution and outlier visualization of ExamDelay Site 1.

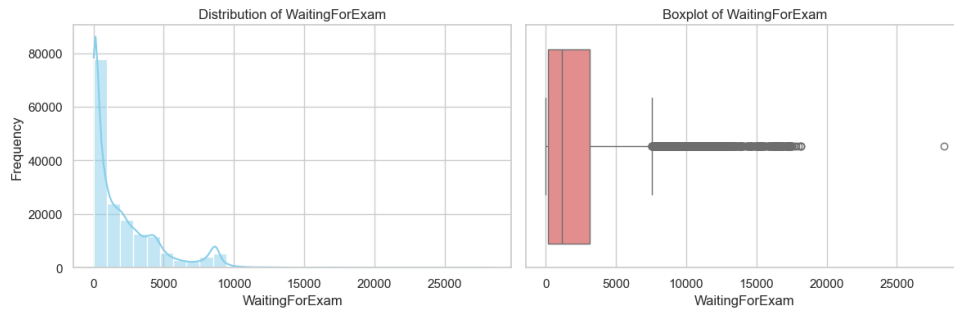


Figure 6.16: Distribution and outlier visualization of WaitingForExam Site 1.

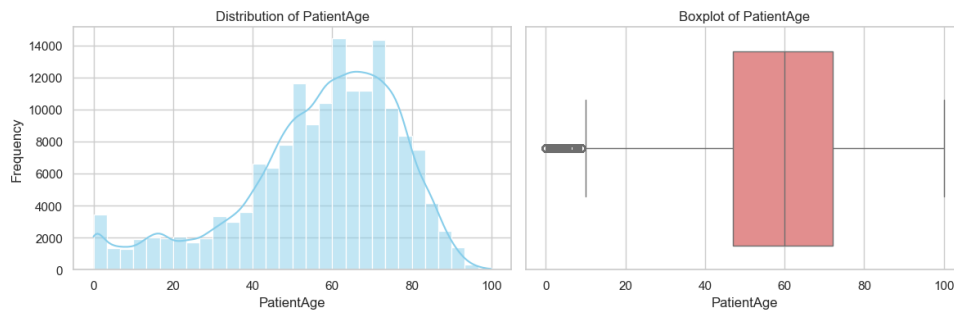


Figure 6.17: Distribution and outlier visualization of PatientAge Site 1.

Presented in Figure 6.18, 6.19, 6.20, 6.21, 6.22 and 6.23 are the distributions of Site 2.

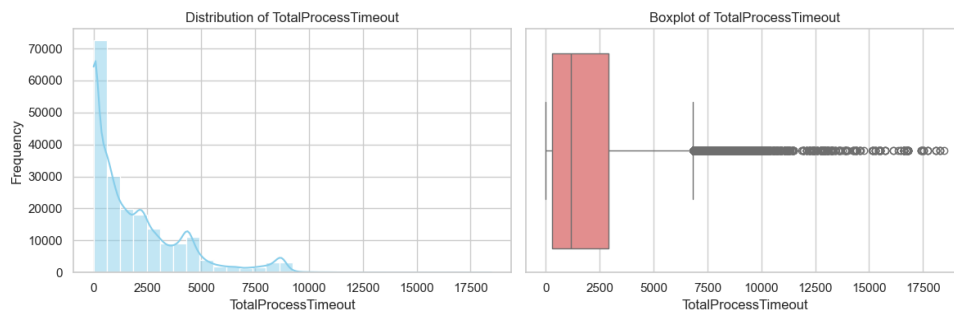


Figure 6.18: Distribution and outlier visualization of TotalProcessTimeout Site 2.

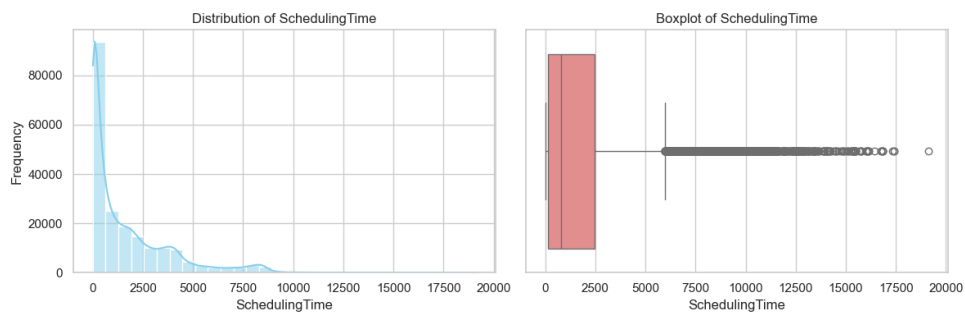


Figure 6.19: Distribution and outlier visualization of SchedulingTime Site 2.

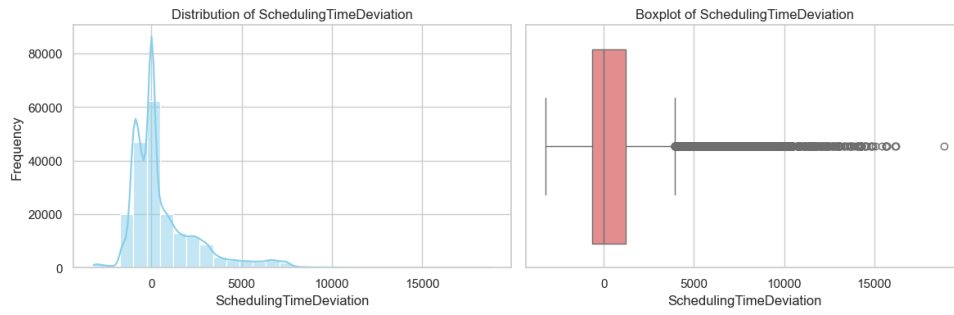


Figure 6.20: Distribution and outlier visualization of SchedulingTimeDeviation Site 2.

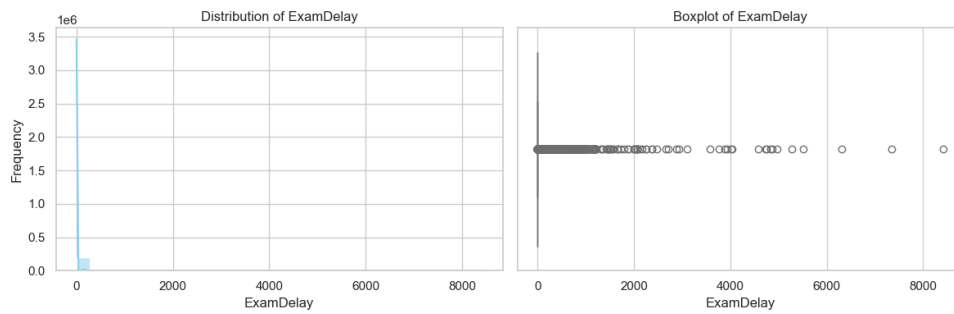


Figure 6.21: Distribution and outlier visualization of ExamDelay Site 2.

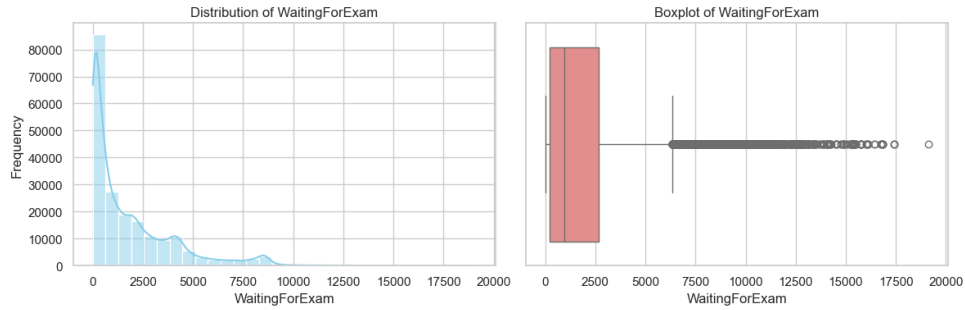


Figure 6.22: Distribution and outlier visualization of WaitingForExam Site 2.

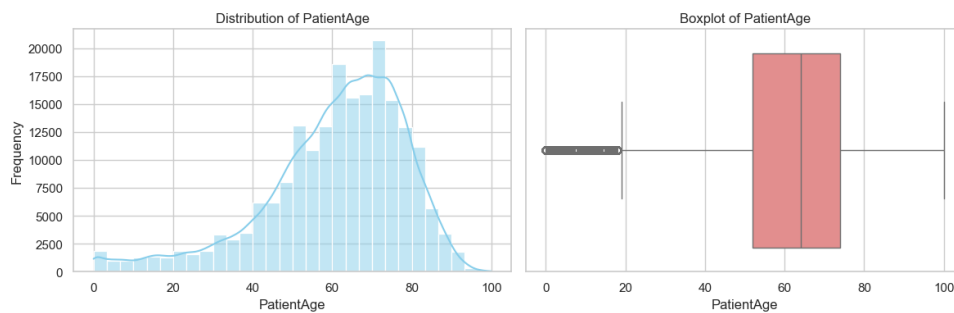


Figure 6.23: Distribution and outlier visualization of PatientAge Site 2.

After having a more quantitative visualization of the data, different detection methods were analyzed.

Interquartile Range (IQR)

IQR method identifies outliers based on the range between the first (Q1) and third (Q3) quartile, making it relatively robust to the presence of skewed data or outliers. It defines boundaries outside of k times the IQR from Q1 to Q3 as possible outliers. If the data are skewed, have heavy tails, or do not follow a normal distribution, the IQR method may be better suited to identify outliers [87].

A value x is considered an outlier if it satisfies [87]:

$$x < Q_1 - k \cdot IQR \quad \text{or} \quad x > Q_3 + k \cdot IQR$$

where:

- Q_1 is the first quartile (25th percentile)
- Q_3 is the third quartile (75th percentile)
- $IQR = Q_3 - Q_1$ is the interquartile range
- k is a constant factor, typically 1.5 or 3

IQR method is a univariate technique that identifies outliers independently for each variable, based on how far values deviate from the central 50% of the data (between the first and third quartiles). While simpler and more interpretable, IQR does not capture relationships between variables [87].

Z-Score

Z-score method detects outliers based on how far a value is from the mean in terms of standard deviations. It assumes that the data follows a normal distribution. If the data is approximately normal, the Z-score method is likely to be more accurate and reliable [88].

To test if the sample follows a normal distribution (a bell-shaped curve) a histogram was made to have a visual assessment of data distribution and Shapiro-Wilk test was used, which makes a formal statistical assessment of normality. Since no feature followed a normal distribution, this method was not used to detect.

Isolation Forest (IF)

IF is a tree-based ML algorithm that is particularly effective for detecting outliers in large, high-dimensional datasets. Its core principle is to identify anomalies by recursively partitioning the data using randomly selected features and split values. Because it operates in a multivariate way, considering multiple features simultaneously, it is able to detect outliers that may not appear anomalous in any single variable but are inconsistent when viewed across multiple dimensions. This makes it especially suitable for datasets with complex, non-linear interactions [89].

In the end, the two methods decided to use as outliers detectors were IQR and IF. This decision happened because in the context of hospital data, many variables had different distributions, some very asymmetric. This choice helped to maintain interpretability of the data preprocessing made.

6.5.2 Treatment

The treatment of numerical features included imputing and removing outliers. Different methods were tested and every approach had in mind the principle of data leakage and the best hyperparameters found to perform the model as a validation metric of the preprocessing.

Listwise Deletion

Listwise Deletion removes from the analysis entire cases (rows) with any missing values. This process was applied once with **IQR** detection, for each continuous feature, calculate the lower and upper limits of the IQR using a multiplicative factor (1.5 or 3) applied to the IQR (as explained in the previous Subsection). Then cases outside this range were removed. With **IF** detection, it detects anomalous patterns and removes. Listwise Deletion was used to eliminate outliers that are not critically important cases but anomalous or unrepresentative.

IQR + Listwise Deletion = univariate detection, case removal.

IF + Listwise Deletion = multivariate detection, case removal.

Winsorization

In order to mitigate the impact of outliers in continuous variables, the winsorize technique was applied. Winsorize has as main goal to limit the influence of extreme values without removing them.

With **IQR** detection the process started by, for each continuous feature, calculating the limits. Values that fall outside this wide range are considered outliers. The first approach clipped values that were outliers to the IQR limits. Then, other approach was used, and instead outliers detected with IQR are replaced by values corresponding to the 1% and 99% percentiles (or other %) of the variable's distribution. This allows us to reduce the influence of the extremes without eliminating them, while retaining the global sample organization. The IQR employed boundaries were calculated only on the training set and applied then to the test set (applying a clip function), without any information leak between stages. For each variable, the number of outliers before and after winsorization was also recorded, in order to assess the impact of the method. This strategy offers a compromise between statistical robustness and preservation of the original data and is particularly useful when the aim is to mitigate the effect of extreme values without resorting to eliminating or introducing new values [90].

IQR + Winsorization = univariate detection, univariate imputation.

As winsorization is, by definition, a univariate technique that consists of limiting the extreme values of a variable to its lower and upper percentiles, this technique acts only on the individual distribution of each variable, ignoring relationships with others. Therefore, it makes sense that outlier detection must also follow this univariate logic, to ensure consistency between the identification and treatment of extreme values. **IF** is a multivariate outlier detection method. It employs the joint distribution of all features simultaneously for the detection of outliers, which may not be spotted by examining variables in isolation. The multivariate aspect of IF allows it to spot subtle interactions and outliers that do not exist in one of the variables but are outlier-like when the collection of variables is considered [89]. If winsorization, a purely univariate intervention, were applied after outlier detection using IF, the key advantage of employing the multivariate detection would be forfeit. The intervention would disregard the multivariate context that defined these outliers and instead deal with each feature in isolation, potentially distorting the interfeature relations and making the reason for the employment of IF moot. These are the main reasons why multivariate IF detection with univariate winsorization treatment was not performed.

Instead of that, outliers detected by IF were addressed with other methods that respect the multivariate nature of the anomaly and thus preserve the integrity and advantages of the multivariate approach.

K-Nearest Neighbors (KNN)

To deal with extreme values in continuous variables, other method used combines outlier detection using the IQR or IF criterion with value imputation based on the KNN Imputer algorithm from the scikit-learn library, which replaces missing values with the mean of the nearest neighbors.

An outlier treatment strategy based on KNN imputation was implemented, applied after univariate detection of extreme values using **IQR** method. For all numeric variables, the lower and upper limits were calculated as mentioned earlier and values outside the limits were marked as outliers. The values were temporarily assigned as NaN to allow the imputation to be performed later by the imputation algorithm. Imputation was performed by the KNNImputer algorithm using data from nearest observations (neighbors) to predict missing values. This method is multivariate, that is, it uses several variables simultaneously to infer the imputed values, which allows better preservation of the joint structure of the data. Again, IQR limits were calculated exclusively based on the training set and imputation was performed using an imputer adjusted only on that set, in order to avoid data leakage. In the test set, outliers were identified using the same previously defined limits and imputation was performed with the same trained model. As in the other approaches attempted, the number of outliers before and after treatment was recorded so that the impact of the approach could be monitored. This method is a compromise between complete removal of extreme data and their explicit transformation (e.g., winsorization), being more of a flexible way to deal with outliers in terms of pattern found in the data itself.

IQR + KNN = univariate detection, multivariate imputation.

On the other hand, an outlier treatment strategy was tested based on the combination of multivariate detection with **IF** and subsequent imputation with KNN. IF algorithm was adjusted based on the training set, considering the selected numerical variables. This model, as explained above, allows identifying anomalous observations by considering the joint pattern of the variables and is especially useful for data with complex relationships between attributes. After identifying the outliers, these observations were treated as NaN. Subsequently, the KNNImputer method was used, which fills in the missing values based on the neighbor values, simultaneously considering all the columns, which reinforces the multivariate nature of the approach. In the test set, the previously trained IF model was used to detect new outliers, ensuring the consistency of the strategy and avoiding data leakage. Imputation was then performed with the same imputer adjusted with the training set. This technique allows taking advantage of the benefits of multivariate outlier detection, which is more sensitive to anomalous patterns that occur only when several variables are considered together. Thus, it differs from more simplistic approaches such as removal or univariate imputations, offering greater sophistication and the ability to adapt to the real behavior of the data.

IF + KNN = multivariate detection, multivariate imputation.

The number of neighbors for the KNN Imputer was selected based on different attempts, choosing the value that minimized the model's validation error.

Multiple Imputation by Chained Equations (MICE)

In this case, outliers were handled by MICE algorithm imputation that was offered by IterativeImputer from the scikit-learn library. In the training phase, univariate outlier detection was applied using the **IQR** criterion. Values outside the limits were marked as NaN, allowing them to be subsequently imputed. After

this marking, the IterativeImputer was applied, which iteratively estimates the missing values of each variable based on regression models trained with other variables in the dataset, thus allowing multivariate imputation, more robust than univariate approaches. IterativeImputer used as initial strategy the median, meaning that initial missing values were filled with the median of the variable before starting the iterative process. Then, for each variable with missing values, the imputer trained a regression model (by default, a Bayesian linear regression) with other variables as predictors, repeating this process iteratively (number of max iterations as input of the function) until convergence was achieved. Bayesian linear regression in MICE serves to impute continuous values in a more statistically realistic way, incorporating model uncertainty rather than assuming deterministic predictions [91]. The imputer was adjusted exclusively with the training data (to prevent data leakage) and, subsequently, applied to the test set. Then, in the test set, thresholds that were created in the training were reused to identify outliers. These were also marked as NaN and imputed by the same trained imputer object, so that there was consistency in transformation between the training and test sets. Apart from excluding the impact of outliers, this method also preserves the multivariate nature of the data, so it is particularly suitable in environments where variables are correlated with each other. Although missing values created by IQR-based outlier imputation may not satisfy MCAR or MAR MICE assumptions [92], their imputation via a multivariate method like MICE is still a reasonable and valid approach. Since such outliers are likely to be noise or wrong values, their replacement via context-sensitive predictions based on the joint variable distribution acts to reduce their pernicious influence on the model while preserving the multivariate data structure.

IQR + MICE = univariate detection, multivariate imputation.

Although the MICE imputation method is a widely used technique for treating missing values in multivariate contexts, its application after multivariate outlier detection with IF proves to be methodologically incoherent. IF identifies outliers based on complex relationships between variables, values that individually could appear normal, but which stand out negatively when considered in a multivariate context. This is precisely the great advantage of this technique over univariate methods such as IQR. However, MICE, although being a multivariate approach, assumes that missing values (in this case, the outliers detected and marked as NaN) are distributed similarly to the rest of the data and can be imputed based on statistical models fitted to the “clean” dataset. MICE assumes that missing values follow a pattern such as MAR and MCAR [91] [92]. However, this assumption conflicts with the logic of IF since the values marked as outliers were precisely identified as being inconsistent with existing patterns. Imputing them assuming these same patterns partially nullifies the value of the multivariate detection that was performed. This limitation is similar to that found in the application of Winsorization after multivariate detection. Winsorization, by definition, is a univariate technique that acts only on the extremes of the distribution of a single variable, ignoring relationships between variables. As in the case of MICE, the application of a treatment technique with univariate assumptions or based on “expected” patterns becomes inadequate after a detection that breaks precisely these assumptions. MICE imputes each variable individually, using others as predictors. When IF is used to detect anomalous combined points and then ask MICE to fill in based on “general trends”, the multivariate character of the anomaly is completely lost. For this reason, it was decided to restrict the use of MICE to scenarios in which outlier detection also followed a univariate logic, ensuring methodological coherence between the way in which outliers are identified and the way in which they are subsequently treated.

Normalization

One of the steps frequently considered in data preprocessing is normalization. Continuous variables, like PatientAge, were normalized after outliers treatment and prior to the training and testing. Despite being a common practice, it was found that normalizing did not have impact on the final results of tree-based models, since results did not improve or got worse with normalized data, when compared to non-normalized. This conclusion is in line with Cabello-Solorzano et al. [93] study, which compared the

impact of data normalization on the accuracy of ML algorithms, concluding that models based on trees are not sensitive to the scale of the variables, since it works with data divisions based on thresholds and not on distances or magnitudes.

However, during the outlier treatment, some imputation methods were sensitive to the scale of the data, namely KNN and MICE, when the latter uses models based on distance or regression. Hence, before imputation with these two methods, data was scaled, ensuring that all variables had the same influence during the imputation process [94]. After imputation, values were transformed back to the original scale, and the same outlier thresholds and normalization parameters adjusted on the training data were applied to the test set, ensuring process consistency. In the subsequent Subsections the methods for normalization for each method are explained. It is important to note that this normalization was applied only in the context of imputation, and the data was subsequently restored to their original scale to ensure compatibility with tree-based predictive models, which showed to neither require nor benefit from normalization.

Across all methods, certain best practices were consistently followed:

1. Consistent treatment of outliers in training and test sets: calculation of limits or training of the algorithm exclusively on the training set; after defining in the training set, they were applied directly to the test set.
2. Joint processing of all considered variables: rather than processing each variable separately and independently, all chosen numeric variables were processed jointly.
3. In removal methods, conformity with the corresponding target: after data filtering in input data (X, for train and test), the output vector (y, for train and test) was also filtered with the same indexing, so the relation between instances and their corresponding labels is maintained.

6.5.3 Performance Variability Results with IQR Detection

Assuming the best option of treatment within categorical outliers, the cleaning of numerical outliers continues upon it. The results presented in this Subsection show the variability of results for both continuous and cyclical variables which had outliers detected by the IQR method. Each approach is also explained in detail.

It was decided to treat the cyclical variables first and only then the continuous ones, a decision based in the minimization of undesirable effects between stages. The cyclical variables were initially treated by removing outliers based on the IQR method. This approach is simple, direct and does not involve transformations or imputations that could alter the structure of the data. Since these variables have a well-defined periodic behavior, their prior cleaning ensures that they do not negatively affect the subsequent treatment of other variables. On the other hand, the cleaning of continuous variables was not only removal but also more advanced methods (like the ones analyzed previously).

1 - Treat outliers in cyclic variables using IQR bounds as detection criteria.

As explained the processes started by detecting and treating outliers in cyclic temporal variables. For this, outliers were identified in each cyclic numeric column using the IQR method and the corresponding lower and upper bounds calculated within the training set were saved to later reuse the exact same bounds to clean the test set. After cleaning, it was verified how many outliers still remain after applying the cleaning process (in this case removing). IQR bounds on cyclic variables were calculated once and the results are presented in Table 6.17.

Table 6.17: Model validation metrics after IQR-based outlier detection followed by listwise deletion during preprocessing of cyclic variables (time, weekday, shift, and month).

Model	Site 1 Non X-ray		Site 2 Non X-ray	
	factor=3	factor=1.5	factor=3	factor=1.5
XGBoost	$R^2 = 0.689$ MAE = 164.744 MSE = 155690.528 RMSE = 394.576	$R^2 = 0.694$ MAE = 165.416 MSE = 155157.776 RMSE = 393.901	$R^2 = 0.858$ MAE = 104.106 MSE = 77871.089 RMSE = 279.054	$R^2 = 0.866$ MAE = 99.455 MSE = 72572.652 RMSE = 269.393
LightGBM	$R^2 = 0.731$ MAE = 150.595 MSE = 134489.445 RMSE = 366.728	$R^2 = 0.740$ MAE = 147.299 MSE = 132087.476 RMSE = 363.438	$R^2 = 0.831$ MAE = 126.222 MSE = 92506.707 RMSE = 304.149	$R^2 = 0.838$ MAE = 122.112 MSE = 87424.675 RMSE = 295.677
CatBoost	$R^2 = 0.730$ MAE = 153.960 MSE = 134704.543 RMSE = 367.021	$R^2 = 0.740$ MAE = 151.699 MSE = 132108.488 RMSE = 363.467	$R^2 = 0.812$ MAE = 134.842 MSE = 103195.929 RMSE = 321.241	$R^2 = 0.805$ MAE = 133.981 MSE = 105072.284 RMSE = 324.148
Decision Tree	$R^2 = 0.578$ MAE = 189.162 MSE = 210863.451 RMSE = 459.199	$R^2 = 0.599$ MAE = 183.143 MSE = 203460.674 RMSE = 451.066	$R^2 = 0.691$ MAE = 116.474 MSE = 169005.839 RMSE = 411.103	$R^2 = 0.764$ MAE = 84.258 MSE = 127453.067 RMSE = 357.006

For Site 1 and 2, the behavior was the same. With a factor of 3, no outliers were found, with a factor of 1.5, outliers were found in the Hour variable (6438 in the training set and 1640 in the test set). This process removes all rows (in both training and testing sets) that contain at least one outlier in any of the analyzed columns. The approach selected to continue was the one with factor=1.5. As it can be seen in Figure 6.24, this approach removes hours between 8p.m. and 4a.m., highlighting a known operational pattern: no medical procedures are performed during these nighttime hours.

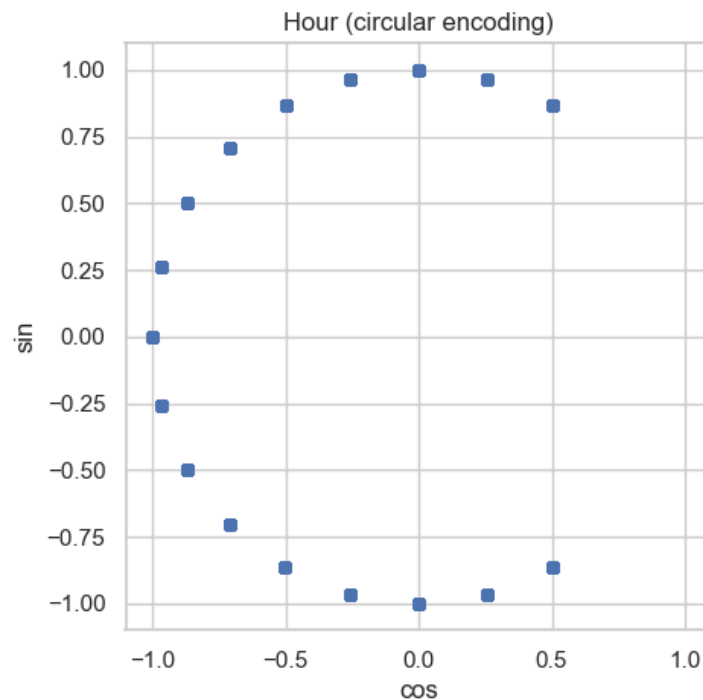


Figure 6.24: Scatter plot between sine and cosine of hour variable from the training set of Site 2 and Site 1 after removal of outliers with IQR, factor 1.5.

2 - Treat outliers in continuous variables using IQR bounds as detection criteria.

Below are presented the most effective results regarding different combinations of treatment of continuous variables by detecting outliers with IQR method. From now on, every iteration below used the correction of time features that showed better results (the one with factor=1.5).

2.1 - The first approach involved removing outliers from continuous variables using IQR bounds as detection criteria.

When removing outliers rows based on continuous variables, cyclic variable bounds were not recalculated or more cyclic outliers were removed. In Table 6.18 and 6.19 the number of outliers before and after listwise deletion are presented for factor 3 and 1.5, respectively.

Table 6.18: Evolution of the number of outliers after listwise deletion with IQR-based outlier detection (factor=3).

Variable	Site 1				Site 2			
	X_Train		X_Test		X_Train		X_Test	
	Before	After	Before	After	Before	After	Before	After
TotalProcess Timeout	268	0	64	0	304	0	73	0
ExamDelay	11382	0	2852	0	10959	0	2855	0
SchedulingTime	434	0	127	0	544	0	143	0
SchedulingTime Deviation	1406	0	356	0	4048	0	992	0
WaitingForExam	311	0	83	0	494	0	126	0

Table 6.19: Evolution of the number of outliers after listwise deletion with IQR-based outlier detection (factor=1.5).

Variable	Site 1				Site 2			
	X_Train		X_Test		X_Train		X_Test	
	Before	After	Before	After	Before	After	Before	After
TotalProcess Timeout	9927	0	2502	0	9903	0	2442	0
ExamDelay	14384	0	3594	0	16420	0	4134	0
SchedulingTime	11254	0	2838	0	11133	0	2716	0
SchedulingTime Deviation	13426	0	3346	0	13509	0	3306	0
WaitingForExam	10339	0	2618	0	10648	0	2616	0

In Table 6.20 it is possible to see the results.

Table 6.20: Model validation metrics after IQR-based outlier detection followed by listwise deletion during preprocessing of continuous features (TotalProcessTimeout, ExamDelay, SchedulingTime, SchedulingTimeDeviation, WaitingForExam).

Model	Site 1 Non X-ray		Site 2 Non X-ray	
	factor=3	factor=1.5	factor=3	factor=1.5
XGBoost	$R^2 = 0.705$ MAE = 153.554 MSE = 115206.540 RMSE = 339.420	$R^2 = 0.651$ MAE = 148.323 MSE = 109933.863 RMSE = 331.563	$R^2 = 0.880$ MAE = 85.900 MSE = 50872.781 RMSE = 225.550	$R^2 = 0.847$ MAE = 88.642 MSE = 53837.233 RMSE = 232.029
LightGBM	$R^2 = 0.739$ MAE = 137.047 MSE = 101825.321 RMSE = 319.101	$R^2 = 0.682$ MAE = 136.711 MSE = 100103.839 RMSE = 316.392	$R^2 = 0.830$ MAE = 113.739 MSE = 72094.615 RMSE = 268.504	$R^2 = 0.800$ MAE = 111.364 MSE = 70313.950 RMSE = 265.168
CatBoost	$R^2 = 0.740$ MAE = 139.785 MSE = 101299.482 RMSE = 318.276	$R^2 = 0.680$ MAE = 140.732 MSE = 100700.119 RMSE = 317.333	$R^2 = 0.811$ MAE = 121.791 MSE = 80174.442 RMSE = 283.151	$R^2 = 0.771$ MAE = 119.615 MSE = 80232.838 RMSE = 283.254
Decision Tree	$R^2 = 0.592$ MAE = 170.605 MSE = 159227.288 RMSE = 399.033	$R^2 = 0.538$ MAE = 167.656 MSE = 145452.422 RMSE = 381.382	$R^2 = 0.745$ MAE = 75.188 MSE = 108203.426 RMSE = 328.943	$R^2 = 0.699$ MAE = 73.193 MSE = 105550.741 RMSE = 324.886

The results obtained showed that, for both Sites, the use of factor 3 led to a better overall performance of the model, although with slightly higher errors. This behavior can be explained by the fact that a more restrictive factor, such as 1.5, tends to eliminate a greater number of points that are outside the range defined as "normal". However, some of them are not anomalous values, but valid and useful information. Their removal caused the loss of relevant information, weakening the ability of the model to capture real patterns and, therefore, reducing its performance. Conversely, factor 3 provides more tolerance to extreme values, which led to a more complete training set. Even though there was an increase in the absolute errors (because there were values that were harder to predict), the overall performance was better, which implies that keeping more extreme cases helped the model.

2.2 - The second approach involved applying winsorization while using IQR bounds as detection criteria.

The methods were designed to apply winsorization only on the upper side for variables that contain only values greater than or equal to zero, while it performs two-sided winsorization for variables that include negative values.

2.2.1 - The first attempt while studying the effect of winsorization include understanding how outliers are handled using the IQR method. Specifically:

- Any value below the lower bound, defined as $Q1 - k \times IQR$, is replaced with that lower bound.
- Any value above the upper bound, defined as $Q3 + k \times IQR$, is replaced with that upper bound.

This first attempt ensures that **all extreme values detected by IQR are brought within a statistically reasonable range**, without removing data points [90]. It is a simple and robust technique for reducing the influence of outliers while preserving the overall structure of the dataset and the results in performance are presented in Table 6.21.

Table 6.21: Model validation metrics after IQR-based outlier detection and capping at the IQR boundaries ($Q1 - k \times IQR$ and $Q3 + k \times IQR$) during the preprocessing of continuous variables (TotalProcessTimeout, ExamDelay, SchedulingTime, SchedulingTimeMedian, and WaitingForExam).

Model	Site 1 Non X-ray		Site 2 Non X-ray	
	factor=3	factor=1.5	factor=3	factor=1.5
XGBoost	$R^2 = 0.674$ MAE = 169.476 MSE = 165314.335 RMSE = 406.589	$R^2 = 0.624$ MAE = 178.927 MSE = 190648.569 RMSE = 436.633	$R^2 = 0.864$ MAE = 100.252 MSE = 73396.426 RMSE = 270.918	$R^2 = 0.848$ MAE = 106.874 MSE = 82278.726 RMSE = 286.843
LightGBM	$R^2 = 0.732$ MAE = 151.636 MSE = 135785.553 RMSE = 368.491	$R^2 = 0.665$ MAE = 164.030 MSE = 169729.897 RMSE = 411.983	$R^2 = 0.833$ MAE = 123.996 MSE = 89916.303 RMSE = 299.860	$R^2 = 0.810$ MAE = 132.760 MSE = 102433.153 RMSE = 320.052
CatBoost	$R^2 = 0.733$ MAE = 155.961 MSE = 135290.884 RMSE = 367.819	$R^2 = 0.664$ MAE = 169.605 MSE = 170501.739 RMSE = 412.919	$R^2 = 0.806$ MAE = 134.643 MSE = 104637.976 RMSE = 323.478	$R^2 = 0.786$ MAE = 142.872 MSE = 115548.711 RMSE = 339.925
Decision Tree	$R^2 = 0.588$ MAE = 188.011 MSE = 209130.165 RMSE = 457.308	$R^2 = 0.522$ MAE = 201.062 MSE = 242578.074 RMSE = 492.522	$R^2 = 0.769$ MAE = 85.480 MSE = 124825.854 RMSE = 353.307	$R^2 = 0.709$ MAE = 96.875 MSE = 157007.132 RMSE = 396.241

Again, looking at the results, it is possible to conclude that factor = 3 preserves more important data for the model than the results with factor = 1.5.

2.2.2 - In the second attempt while studying the effect of winsorization, two types of limits are defined for each variable. The IQR limits are calculated using $Q1$ and $Q3$ and a chosen factor of the IQR. Additionally, winsorization limits are defined based on different percentiles of the variable's distribution. **In practice, the code identifies outliers as values outside the IQR limits and replaces these outliers by clipping them to the winsorization limits based on the percentiles.**

Hence, two distributions can occur:

$$liq \leq lw \leq Q1 \leq Q3 \leq uw \leq uiq$$

where:

- lw – Lower Winsorization Limit
- liq – Lower IQR Limit ($Q1 - k \cdot IQR$)
- $Q1$ – First Quartile (25th percentile)
- $Q3$ – Third Quartile (75th percentile)
- uiq – Upper IQR Limit ($Q3 + k \cdot IQR$)
- uw – Upper Winsorization Limit

For skewed distributions or those with heavy tails, the IQR-based limits can fall outside the range defined by the percentiles. This means that when IQR limits extend beyond the percentile limits, the actual values are restricted to remain within the [lower percentile, upper percentile] range after winsorization. Hence, this process ensures that the outliers are clipped within a robust interval that encompasses the central spread of the data, essentially limiting the influence of outliers. The approach highlights the sensitivity of IQR outlier detection with the practical limitation of percentile clipping to provide controlled winsorization.

$$lw \leq liq \leq Q1 \leq Q3 \leq uiq \leq uw$$

where:

- lw – Lower Winsorization Limit
- liq – Lower IQR Limit ($Q_1 - k \cdot IQR$)
- $Q1$ – First Quartile (25th percentile)
- $Q3$ – Third Quartile (75th percentile)
- uiq – Upper IQR Limit ($Q_3 + k \cdot IQR$)
- uw – Upper Winsorization Limit

When the IQR limit is less extreme than the corresponding winsorize percentile, winsorization also applies the cutoff directly to the winsorization threshold. This means that values between the winsorization threshold and the IQR threshold (that are considered outliers by the IQR definition) are not adjusted. In this case, the cleaning of outliers is less extreme.

These examples illustrate specific cases only. In practice, the treatment of the lower and upper tail may be different, depending on the characteristics of the distribution. The method adapts the winsorization thresholds for each tail, applying the most conservative threshold between the IQR and the percentiles, ensuring a flexible approach to skewed or heavy-tailed data. The results of this version of winsorization are presented in Table 6.22.

Table 6.22: Model validation metrics after IQR-based outlier detection and capping at the percentile boundaries during the preprocessing of continuous variables (TotalProcessTimeout, ExamDelay, SchedulingTime, SchedulingTimeMedian, and WaitingForExam).

Model	Site 1 Non X-ray		Site 2 Non X-ray	
	factor=3; percentile=0.01	factor=1.5; percentile=0.01	factor=3; percentile=0.01	factor=1.5; percentile=0.01
XGBoost	$R^2 = 0.675$ MAE = 167.293 MSE = 164685.845 RMSE = 405.815	$R^2 = 0.665$ MAE = 168.017 MSE = 169950.057 RMSE = 412.250	$R^2 = 0.864$ MAE = 99.867 MSE = 73441.212 RMSE = 271.000	$R^2 = 0.860$ MAE = 100.876 MSE = 75818.374 RMSE = 275.351
LightGBM	$R^2 = 0.725$ MAE = 149.984 MSE = 139278.562 RMSE = 373.200	$R^2 = 0.720$ MAE = 150.937 MSE = 141876.878 RMSE = 376.665	$R^2 = 0.832$ MAE = 123.926 MSE = 90317.732 RMSE = 300.529	$R^2 = 0.832$ MAE = 123.752 MSE = 90819.529 RMSE = 301.363
CatBoost	$R^2 = 0.722$ MAE = 154.205 MSE = 140917.767 RMSE = 375.390	$R^2 = 0.709$ MAE = 155.468 MSE = 147593.433 RMSE = 384.179	$R^2 = 0.805$ MAE = 133.910 MSE = 105052.303 RMSE = 324.118	$R^2 = 0.802$ MAE = 135.146 MSE = 106808.640 RMSE = 326.816
Decision Tree	$R^2 = 0.588$ MAE = 183.043 MSE = 209007.793 RMSE = 457.174	$R^2 = 0.586$ MAE = 186.496 MSE = 210087.761 RMSE = 458.353	$R^2 = 0.760$ MAE = 85.646 MSE = 129326.906 RMSE = 359.621	$R^2 = 0.751$ MAE = 87.957 MSE = 134346.104 RMSE = 366.533

For a better understanding of what this approach actually does to the data, a printout of the number of outliers was made, before and after treatment (Table 6.23 and 6.24). **Before treatment**, any value below the lower IQR limit or above the upper IQR limit was considered an outlier. **After treatment**, the definition of outliers was adjusted based on the comparison between the IQR limits and the winsorization limits:

- If the IQR upper bound is greater than the winsorization upper bound, outliers after treatment are defined as values greater than the IQR upper bound.
- If the upper winsorization limit is greater than the upper IQR limit, outliers after treatment are defined as values greater than the upper winsorization limit.
- If the IQR lower bound is less than the winsorization lower bound, outliers after treatment are defined as values less than the IQR lower bound.
- If the winsorization lower bound is less than the IQR lower bound, outliers after treatment are defined as values less than the winsorization lower bound.

Table 6.23: Evolution of the number of outliers after capping at the percentile boundaries with IQR-based outlier detection (factor=3, percentile=0.01).

Variable	Site 1				Site 2			
	X_Train		X_Test		X_Train		X_Test	
	Before	After	Before	After	Before	After	Before	After
TotalProcess Timeout	268	0	64	0	304	0	73	0
ExamDelay	11382	0	2852	0	10959	0	2855	0
SchedulingTime	434	0	127	0	544	0	143	0
SchedulingTime Deviation	1406	0	356	0	4048	0	992	0
WaitingForExam	311	0	83	0	494	0	126	0

Table 6.24: Evolution of the number of outliers after capping at the percentile boundaries with IQR-based outlier detection (factor=1.5, percentile=0.01).

Variable	Site 1				Site 2			
	X_Train		X_Test		X_Train		X_Test	
	Before	After	Before	After	Before	After	Before	After
TotalProcess Timeout	9927	0	2502	0	9903	0	2442	0
ExamDelay	14384	0	3594	0	16420	0	4134	0
SchedulingTime	11254	0	2838	0	11133	0	2716	0
SchedulingTime Deviation	13426	0	3346	0	13509	0	3306	0
WaitingForExam	10339	0	2618	0	10648	0	2616	0

In this attempt, two multiplicative factors of the IQR were also considered: 1.5 (traditional value) and 3 (more conservative). However, it was observed that the use of factor 3, despite detecting fewer outliers, resulted in greater effectiveness of winsorization. This occurred because many of the values detected as outliers with factor 1.5 were still within the limits of uw and uiq (or lw and liq), and therefore were not modified. On the other hand, the outliers identified with factor 3, being more extreme, were often outside these limits and were in fact all adjusted. Furthermore, as previously observed, using a factor of 1.5 identifies values as outliers that are actually important for the model's performance. As a result, it leads to a decrease in performance as when compared to using a factor of 3. This can be easily seen in Figure 6.25 and 6.26, that illustrates what occurred to Site 1. For Site 2, the behavior was the same, although, of course, with different boundaries.

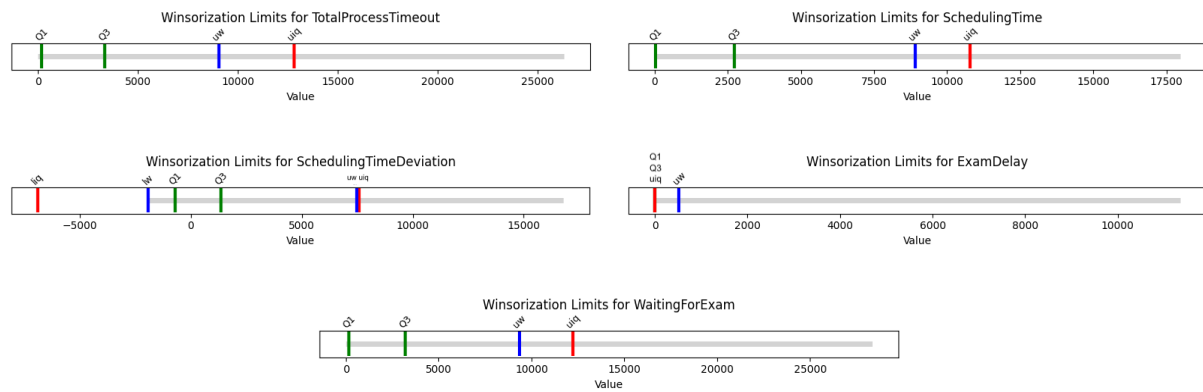


Figure 6.25: Q1 and Q3, IQR-based outlier thresholds (factor=3) and the Winsorization bounds (1st and 99th percentiles) from Site 1.

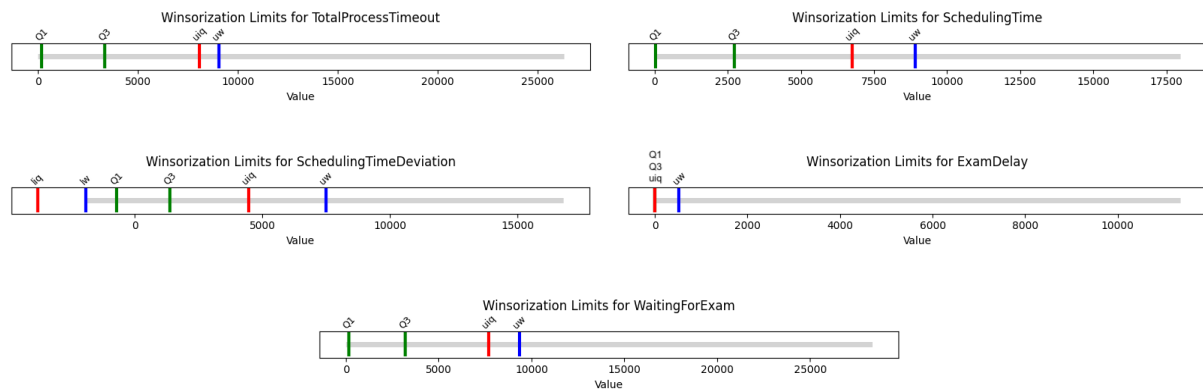


Figure 6.26: Q1 and Q3, IQR-based outlier thresholds (factor=1.5) and the Winsorization bounds (1st and 99th percentiles) from Site 1.

Thus, factor 3 proved to be more effective in practice, even though it is theoretically less sensitive, showing that the interaction between the detection criterion (IQR) and the substitution criterion (percentiles) directly influences the results.

Knowing this, a more drastic winsorization approach was tested with different values in order to see what happens when a great amount of data is capped by winsorization percentiles. For this, a winsorization percentile of 0.1 was used and the results are both in Table 6.25 and 6.26.

Table 6.25: Evolution of outliers after winsorization with IQR (factor=1.5, percentile=0.1).

Variable	Site 1				Site 2			
	X_Train		X_Test		X_Train		X_Test	
	Before	After	Before	After	Before	After	Before	After
TotalProcessTimeout	9927	0	2502	0	9903	0	2442	0
ExamDelay	14384	0	3594	0	16420	0	4134	0
SchedulingTime	11254	0	2838	0	11133	0	2716	0
SchedulingTimeDeviation	13426	0	3346	0	13509	0	3306	0
WaitingForExam	10339	0	2618	0	10648	0	2616	0

Table 6.26: Model validation metrics after IQR-based outlier detection and capping at the percentile boundaries during the preprocessing of continuous variables (TotalProcessTimeout, ExamDelay, SchedulingTime, SchedulingTimeMedian, and WaitingForExam) with more extreme threshold values.

Model	Site 1 Non X-ray	Site 2 Non X-ray
	factor=1.5; percentile=0.1	factor=1.5; percentile=0.1
XGBoost	$R^2 = 0.614$ MAE = 181.357 MSE = 195773.074 RMSE = 442.463	$R^2 = 0.841$ MAE = 108.845 MSE = 85763.731 RMSE = 292.854
LightGBM	$R^2 = 0.662$ MAE = 165.051 MSE = 171192.000 RMSE = 413.754	$R^2 = 0.804$ MAE = 134.311 MSE = 105579.435 RMSE = 324.930
CatBoost	$R^2 = 0.660$ MAE = 170.128 MSE = 172365.398 RMSE = 415.169	$R^2 = 0.781$ MAE = 143.588 MSE = 118304.880 RMSE = 343.955
Decision Tree	$R^2 = 0.498$ MAE = 204.111 MSE = 254497.738 RMSE = 504.478	$R^2 = 0.697$ MAE = 97.308 MSE = 163510.454 RMSE = 404.364

The results show that the overall performance did not improve. This suggests that values detected as outliers with IQR factor=1.5, may have real importance for the model, and clipping them brings no better results.

By comparing Table 6.21 with Table 6.22 it is possible to see that the approach that showed to be more cautious was the one with two limits: IQR bounds and winsorize percentiles. This makes sense from a statistical and practical point of view: clipping with IQR limits cut all outliers to IQR limit values; conditional winsorize with IQR bounds and extreme percentiles instead of cutting them directly to the IQR limit, reassigns a more realistic (and less aggressive) value, based on the percentiles.

2.3 - The third approach involved applying KNN imputation while using IQR bounds as detection criteria.

Prior to the use of the KNN Imputer approach, numerical continuous variables had to be initially normalized. This transformation was necessary to place the variables all on a uniform scale since normalization is particularly important in this instance because the KNN algorithm depends on the calculation of distances (most usually Euclidean) amongst points. Without normalization, variables of different scales may dominate the distance calculation and bias the imputation process. With all the variables being equally important in the distance, there is a promotion of more balanced and equitable imputation [94].

While training different approaches of normalization, the use of MinMaxScaler led to better model performance compared to RobustScaler (which is robust to outliers). Scaling all features to the same [0, 1] range using MinMaxScaler ensured that the influence of each variable is balanced when computing distances between samples [95]. The results of the approach are presented in Table 6.27.

Table 6.27: Model validation metrics after IQR-based outlier detection and imputation with KNN algorithm during the preprocessing of continuous variables (TotalProcessTimeout, ExamDelay, SchedulingTime, SchedulingTimeMedian, and WaitingForExam).

Model	Site 1 Non X-ray			Site 2 Non X-ray		
	factor=3, neighbors=3	factor=3, neighbors=7	factor=1.5, neighbors=3	factor=3, neighbors=3	factor=3, neighbors=7	factor=1.5, neighbors=3
XGBoost	$R^2 = 0.607$ MAE = 175.841 MSE=199436.410 RMSE = 446.583	$R^2 = 0.608$ MAE = 174.217 MSE=198717.042 RMSE = 445.777	$R^2 = 0.541$ MAE = 195.024 MSE=232654.725 RMSE = 482.343	$R^2 = 0.830$ MAE = 105.797 MSE=91677.561 RMSE = 302.783	$R^2 = 0.829$ MAE = 106.580 MSE=92336.237 RMSE = 303.869	$R^2 = 0.754$ MAE = 129.622 MSE=133051.410 RMSE = 364.762
LightGBM	$R^2 = 0.640$ MAE = 159.595 MSE=182553.171 RMSE = 427.262	$R^2 = 0.645$ MAE = 159.860 MSE=179813.872 RMSE = 424.045	$R^2 = 0.599$ MAE = 178.370 MSE=203354.714 RMSE = 450.949	$R^2 = 0.795$ MAE = 131.389 MSE=110869.451 RMSE = 332.971	$R^2 = 0.793$ MAE = 132.495 MSE=111625.943 RMSE = 334.105	$R^2 = 0.721$ MAE = 153.698 MSE=150753.418 RMSE = 388.270
CatBoost	$R^2 = 0.644$ MAE = 162.996 MSE=180709.970 RMSE = 425.099	$R^2 = 0.643$ MAE = 164.289 MSE=181226.418 RMSE = 425.707	$R^2 = 0.598$ MAE = 184.662 MSE=204137.521 RMSE = 451.816	$R^2 = 0.763$ MAE = 142.392 MSE=127901.820 RMSE = 357.634	$R^2 = 0.760$ MAE = 142.769 MSE=129619.535 RMSE = 360.027	$R^2 = 0.686$ MAE = 163.488 MSE=169683.992 RMSE = 411.927
Decision Tree	$R^2 = 0.506$ MAE = 194.053 MSE=250305.886 RMSE = 500.306	$R^2 = 0.511$ MAE = 195.087 MSE=248106.896 RMSE = 498.103	$R^2 = 0.431$ MAE = 208.750 MSE=288485.241 RMSE = 537.108	$R^2 = 0.712$ MAE = 92.213 MSE=155379.546 RMSE = 394.182	$R^2 = 0.686$ MAE = 94.354 MSE=169490.339 RMSE = 411.692	$R^2 = 0.536$ MAE = 123.790 MSE=250508.271 RMSE = 500.508

For a better understanding of the approach, a printout of the number of outliers was made, before and after treatment, with the different factors (Table 6.28 and 6.29).

Table 6.28: Evolution of the number of outliers after KNN imputation with IQR-based outlier detection (factor=3).

Variable	Site 1				Site 2 (neighbors=3,7)			
	X_Train		X_Test		X_Train		X_Test	
	Before	After	Before	After	Before	After	Before	After
TotalProcessTimeout	268	0	64	0	304	0	73	0
ExamDelay	11382	0	2852	0	10959	0	2855	0
SchedulingTime	434	0	127	0	544	0	143	0
SchedulingTime Deviation	1406	0	356	0	4048	0	992	0
WaitingForExam	311	0	83	0	494	0	126	0

Table 6.29: Evolution of the number of outliers after KNN imputation with IQR-based outlier detection (factor=1.5).

Variable	Site 1				Site 2 (neighbors=3)			
	X_Train		X_Test		X_Train		X_Test	
	Before	After	Before	After	Before	After	Before	After
TotalProcessTimeout	9927	0	2502	0	9903	0	2442	0
ExamDelay	14384	0	3594	0	16420	0	4134	0
SchedulingTime	11254	0	2838	0	11133	0	2716	0
SchedulingTime Deviation	13426	0	3346	0	13509	0	3306	0
WaitingForExam	10339	0	2618	0	10648	0	2616	0

As expected due to previous approach's results, factor = 1.5 really does not contribute to model performance. The choice between number of neighbors seems to be dependent of the dataset, although for Site 1 it seems to be better 7 neighbors, for Site 2, 3 neighbors are enough.

2.4 - The fourth approach involved applying MICE imputation while using IQR bounds as detection criteria.

MICE uses the given columns as a basis to predict the NaN imputed values in each of the others [92]. Hence, as this experiment started, it was important to study correlations between numerical continuous variables, as done in Figure 6.27 and 6.28.

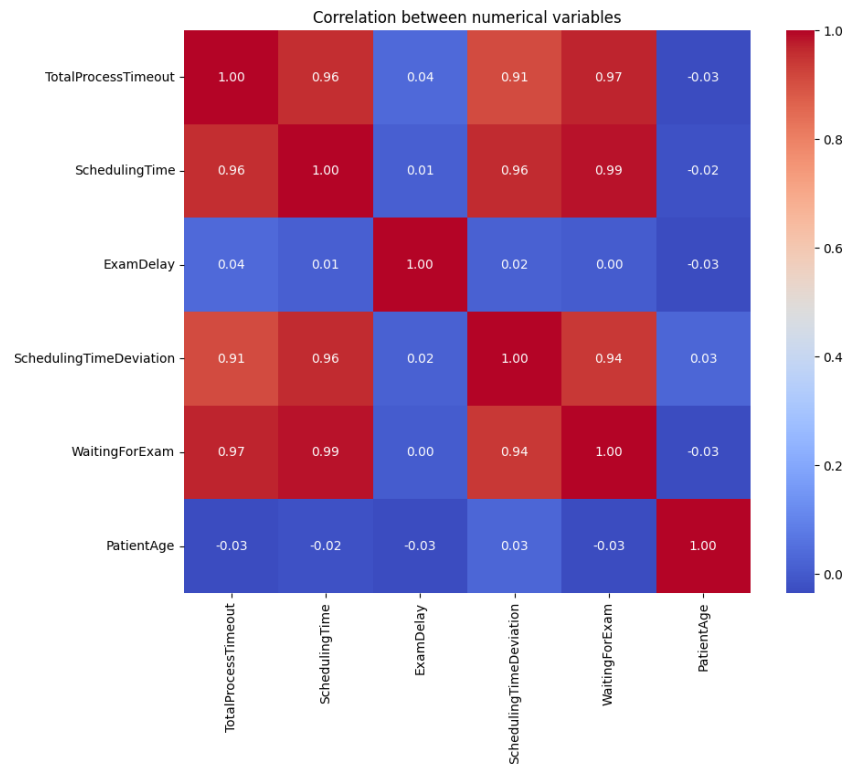


Figure 6.27: Correlation between different continuous numerical variables within Site 1.

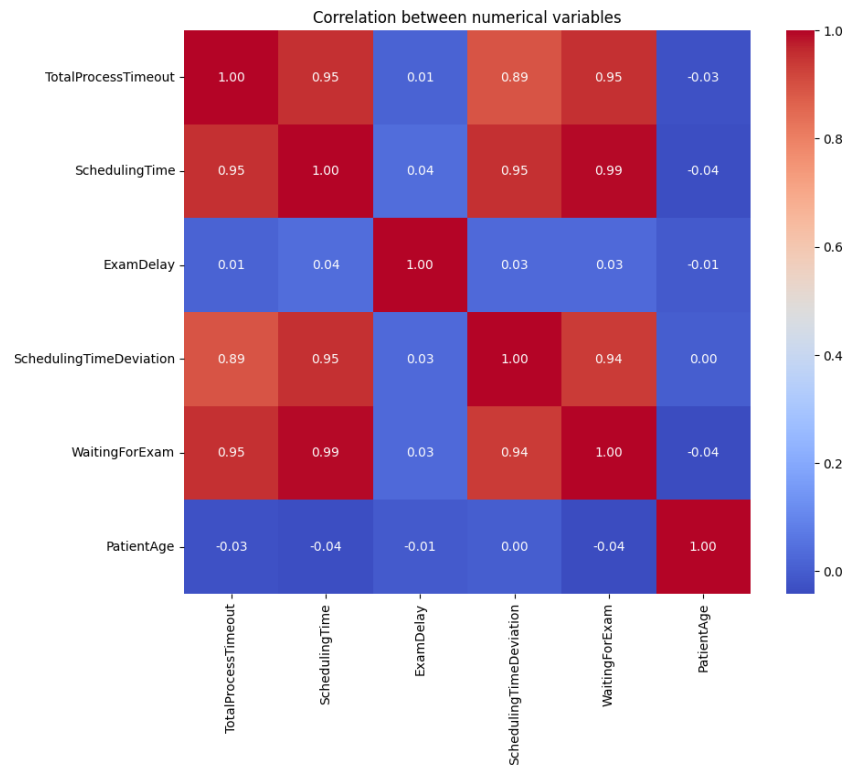


Figure 6.28: Correlation between different continuous numerical variables within Site 2.

For this treatment, ExamDelay was not included because the correlation with the other features was ≤ 0.04 and PatientAge with negative correlations. All other features had correlation between each other ≥ 0.89 , making them ideal to learn from each other and help MICE algorithm to impute. No other technique was applied to the remaining features, so the effect of MICE would not be masked. MinMaxScaler was used internally in the imputation process with IterativeImputer to normalize continuous variables to the interval $[0, 1]$. This transformation improves the numerical stability and efficiency of imputation, preventing variables with different orders of magnitude from dominating the iterative regression process. The results of the approach are presented in Table 6.30.

Table 6.30: Model validation metrics after IQR-based outlier detection and imputation with MICE algorithm during the preprocessing of continuous variables (TotalProcessTimeout, SchedulingTime, SchedulingTimeDeviation and WaitingForExam).

Model	Site 1 Non X-ray		Site 2 Non X-ray	
	factor=3	factor=1.5	factor=3	factor=1.5
XGBoost	$R^2 = 0.594$ MAE = 173.147 MSE = 205972.517 RMSE = 453.842	$R^2 = 0.452$ MAE = 205.819 MSE = 278066.906 RMSE = 527.320	$R^2 = 0.842$ MAE = 103.584 MSE = 85445.959 RMSE = 292.311	$R^2 = 0.774$ MAE = 121.979 MSE = 122250.364 RMSE = 349.643
LightGBM	$R^2 = 0.628$ MAE = 156.996 MSE = 188744.590 RMSE = 434.447	$R^2 = 0.511$ MAE = 179.788 MSE = 247951.606 RMSE = 497.947	$R^2 = 0.801$ MAE = 130.026 MSE = 107208.483 RMSE = 327.427	$R^2 = 0.728$ MAE = 150.363 MSE = 146741.294 RMSE = 383.068
CatBoost	$R^2 = 0.614$ MAE = 160.696 MSE = 195928.173 RMSE = 442.638	$R^2 = 0.513$ MAE = 183.270 MSE = 246849.441 RMSE = 496.839	$R^2 = 0.769$ MAE = 139.822 MSE = 124587.886 RMSE = 352.970	$R^2 = 0.680$ MAE = 160.746 MSE = 172686.874 RMSE = 415.556
Decision Tree	$R^2 = 0.487$ MAE = 192.049 MSE = 260284.008 RMSE = 510.180	$R^2 = 0.330$ MAE = 216.290 MSE = 340024.847 RMSE = 583.116	$R^2 = 0.690$ MAE = 92.489 MSE = 167240.909 RMSE = 408.951	$R^2 = 0.576$ MAE = 109.748 MSE = 228880.880 RMSE = 478.415

A printout of the number of outliers was made with the different factors (Table 6.31 and 6.32).

Table 6.31: Evolution of the number of outliers after MICE imputation with IQR-based outlier detection (factor=3).

Variable	Site 1				Site 2			
	X_Train		X_Test		X_Train		X_Test	
	Before	After	Before	After	Before	After	Before	After
TotalProcessTimeout	268	0	64	0	304	0	73	0
SchedulingTime	434	98	127	36	544	24	143	5
SchedulingTimeDeviation	1406	453	356	137	4048	920	992	239
WaitingForExam	311	6	83	0	494	0	126	0

Table 6.32: Evolution of the number of outliers after MICE imputation with IQR-based outlier detection (factor=1.5).

Variable	Site 1				Site 2			
	X_Train		X_Test		X_Train		X_Test	
	Before	After	Before	After	Before	After	Before	After
TotalProcessTimeout	9927	0	2502	0	9903	0	2442	0
SchedulingTime	11254	191	2838	62	11133	0	2716	0
SchedulingTimeDeviation	13426	697	3346	194	13509	1154	3306	288
WaitingForExam	10339	0	2618	0	10648	3	2616	0

It is possible to see that for both Sites there were values imputed that still fall outside the defined IQR boundaries and are therefore considered outliers.

Conclusion of the best treatment approach after IQR-based outlier detection.

Of all the experiments, the method to correct continuous numerical variables that gave the best results (highest performance, lowest errors) was the one where the outliers detected by IQR were only removed.

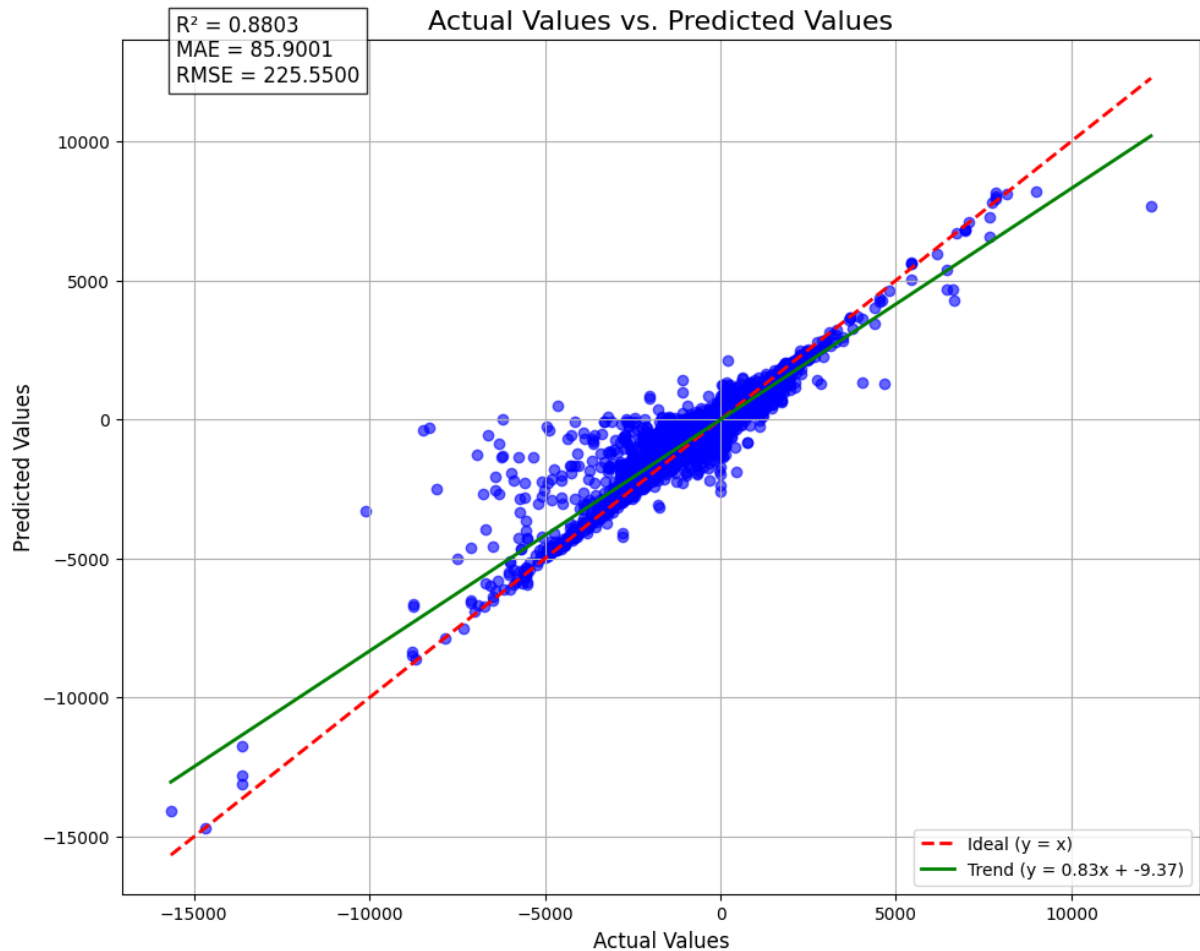


Figure 6.29: Impact of listwise deletion during preprocessing of continuous features after IQR-based outlier detection, trained and tested with XGBoost with Site 2 data.

In Figure 6.29, each point represents a model prediction compared to the actual value. Since the target is the deviation in hours, a positive value tells the report was delivered before the deadline, a negative value shows a delay (report delivered after the deadline). The line $y = x$ shows where the points would fall if the model was perfect.

6.5.4 Performance Variability Results with IF Detection

In outlier detection methodology taken with IF algorithm, it should be noted that detection of anomalous values happens in a multivariate sense, meaning that the method is looking at a group of variables at the same time. In contrast to univariate methods like IQR, which look at the values of each variable separately, IF examines how the variables behave together to decide whether an observation differs significantly from the overall pattern of data [89].

1 - Treat outliers in cyclic variables using IF algorithm as detection criteria.

To ensure the quality of the input data and avoid the negative influence of atypical values on the predictive models, outliers were detected and removed in the cyclical variables of the dataset. Here, IF simultaneously considers variables such as the sine and cosine transformations of hours, days, months and shifts. The expected percentage of outliers (known as contamination for IF) was set at different levels. After detection, all rows that were identified as outliers in at least one of the cyclical variables were removed, thus ensuring listwise deletion and maintaining data integrity. The results of the approach are presented in Table 6.33.

Table 6.33: Model validation metrics after IF-based outlier detection followed by listwise deletion during preprocessing of cyclic variables (time, weekday, shift, and month).

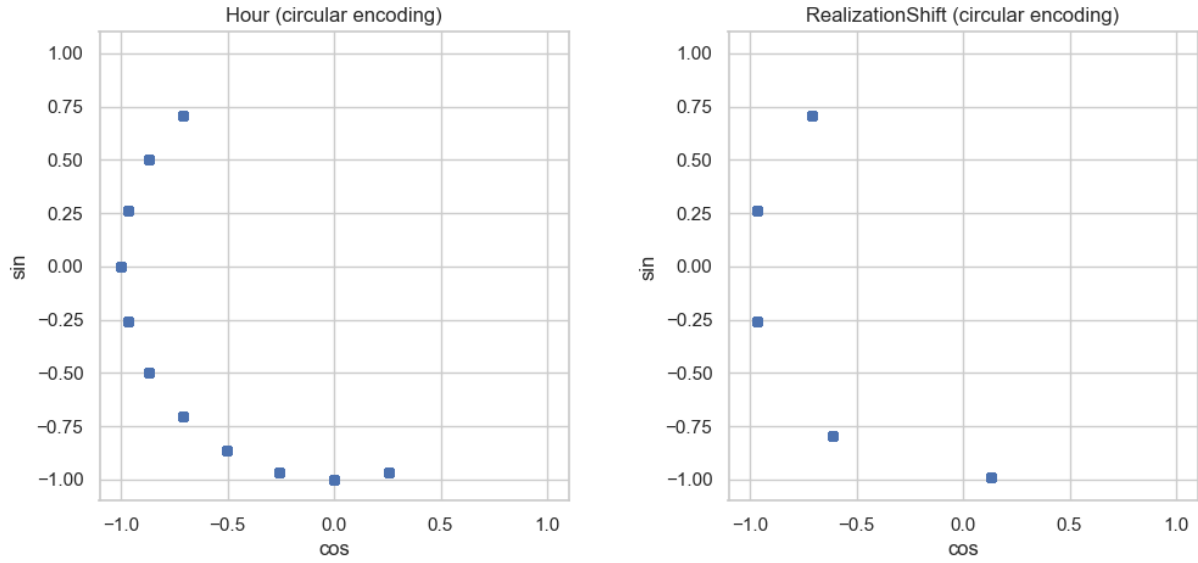
Model	Site 1 Non X-ray			Site 2 Non X-ray		
	contamination=0.05	contamination=0.1	contamination=0.15	contamination=0.05	contamination=0.1	contamination=0.15
XGBoost	$R^2 = 0.698$ MAE = 164.550 MSE=152585.385 RMSE = 390.622	$R^2 = 0.699$ MAE = 163.622 MSE=153871.442 RMSE = 392.265	$R^2 = 0.698$ MAE = 162.904 MSE=156785.675 RMSE = 395.961	$R^2 = 0.871$ MAE = 97.658 MSE=69314.022 RMSE = 263.276	$R^2 = 0.869$ MAE = 96.671 MSE=69829.197 RMSE = 264.252	$R^2 = 0.869$ MAE = 95.618 MSE=70133.485 RMSE = 264.827
LightGBM	$R^2 = 0.740$ MAE = 148.048 MSE=131370.689 RMSE = 362.451	$R^2 = 0.746$ MAE = 144.559 MSE=129985.715 RMSE = 360.535	$R^2 = 0.746$ MAE = 142.594 MSE=132036.725 RMSE = 363.369	$R^2 = 0.838$ MAE = 121.112 MSE=87143.802 RMSE = 295.201	$R^2 = 0.839$ MAE = 119.131 MSE=85812.668 RMSE = 292.938	$R^2 = 0.841$ MAE = 116.924 MSE=84834.264 RMSE = 291.263
CatBoost	$R^2 = 0.740$ MAE = 150.516 MSE=131097.739 RMSE = 362.074	$R^2 = 0.745$ MAE = 146.156 MSE=130547.313 RMSE = 361.313	$R^2 = 0.755$ MAE = 145.562 MSE=127366.674 RMSE = 356.885	$R^2 = 0.818$ MAE = 130.937 MSE=98309.049 RMSE = 313.543	$R^2 = 0.817$ MAE = 128.649 MSE=97922.407 RMSE = 312.926	$R^2 = 0.823$ MAE = 125.356 MSE=94574.898 RMSE = 307.530
Decision Tree	$R^2 = 0.589$ MAE = 185.157 MSE=207288.585 RMSE = 455.290	$R^2 = 0.632$ MAE = 180.0602 MSE=188348.495 RMSE = 433.991	$R^2 = 0.634$ MAE = 178.380 MSE=189821.866 RMSE = 435.686	$R^2 = 0.780$ MAE = 80.562 MSE=118830.697 RMSE = 344.718	$R^2 = 0.778$ MAE = 80.857 MSE=118388.800 RMSE = 344.077	$R^2 = 0.767$ MAE = 81.208 MSE=124501.823 RMSE = 352.848

For a better understanding of the approach, a printout of the number of outliers was made, before and after treatment, with different contamination levels (Table 6.34).

Table 6.34: Evolution of the number of outliers after listwise deletion with IF-based outlier detection.

Variable	Site 1				Site 2			
	X_Train		X_Test		X_Train		X_Test	
	Before	After	Before	After	Before	After	Before	After
contamination= 0.05	8240	0	2065	0	10040	0	2530	0
contamination= 0.1	16414	0	4089	0	20094	0	4977	0
contamination= 0.15	24710	0	6190	0	30058	0	7554	0

For Site 1, the approach selected to continue was the one with a contamination of 15% due to its great results. As it can be seen in Figure 6.30, this approach removes hours between 7p.m. and 9a.m., highlighting again a known operational pattern: no medical procedures are performed during these nighttime hours. Accordingly, the shift of night, is also removed, due to very few significant compared to the rest of the shifts in the dataset.

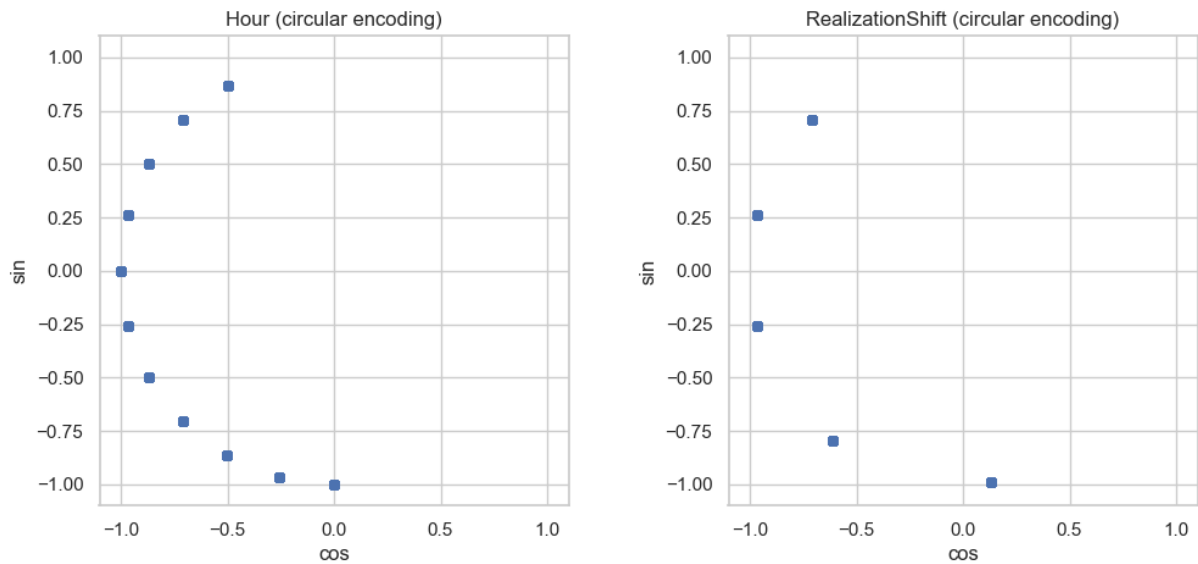


(a) Scatter plot between sine and cosine of Hour variable after outlier removal.

(b) Scatter plot between sine and cosine of Realization-Shift variable after outlier removal.

Figure 6.30: Scatter plot between sine and cosine of cyclic variables from the training set of Site 1 after removal of outliers with IF.

For Site 2, a contamination of 15%, also helped to have the best combination of high performance and low errors. As it can be seen in Figure 6.31, this approach removes hours between 6p.m. and 8a.m., highlighting again the same known operational pattern: no medical procedures are performed during these nighttime hours. Accordingly, the shift of night, is also removed, due to very few significant compared to the rest of the shifts in the dataset.



(a) Scatter plot between sine and cosine of Hour variable after outlier removal.

(b) Scatter plot between sine and cosine of Realization-Shift variable after outlier removal.

Figure 6.31: Scatter plot between sine and cosine of cyclic variables from the training set of Site 2 after removal of outliers with IF.

2 - Treat outliers in continuous variables using IF algorithm as detection criteria.

Below are presented the most effective results regarding different combinations of treatment by detecting numerical outliers with IF method. From now on, every iteration below used the correction of time features that showed better results (contamination 15%).

2.1 - The first approach involved removing outliers from continuous variables using IF algorithm as detection criteria.

Removing outliers on continuous variables with IF detection considers all variables specified as input, and it is precisely this interaction between variables that makes the detection robust, but at the same time makes it impossible to directly attribute the cause of the outlier to a single variable. As a consequence, it is not possible, with this approach, to determine in which variable (or variables) a given point was considered anomalous, only that its combination of values is atypical in relation to the rest of the dataset. This prevents a detailed analysis of outliers by variable, as it was possible in the univariate method IQR [89]. The results of the approach are presented in Table 6.35.

Table 6.35: Model validation metrics after IF-based outlier detection followed by listwise deletion during preprocessing of continuous variables (TotalProcessTimeout, ExamDelay, SchedulingTime, Scheduling-TimeDeviation, WaitingForExam).

Model	Site 1 Non X-ray			Site 2 Non X-ray		
	contamination=0.01	contamination=0.05	contamination=0.1	contamination=0.01	contamination=0.05	contamination=0.1
XGBoost	$R^2 = 0.665$ MAE = 159.439 MSE=153659.356 RMSE = 391.994	$R^2 = 0.608$ MAE = 153.715 MSE=142568.031 RMSE = 377.582	$R^2 = 0.591$ MAE = 147.834 MSE=128405.722 RMSE = 358.337	$R^2 = 0.861$ MAE = 91.083 MSE=64845.509 RMSE = 254.648	$R^2 = 0.849$ MAE = 87.932 MSE=63078.966 RMSE = 251.155	$R^2 = 0.828$ MAE = 86.369 MSE=58324.588 RMSE = 241.505
LightGBM	$R^2 = 0.707$ MAE = 140.862 MSE=134103.934 RMSE = 366.202	$R^2 = 0.661$ MAE = 136.825 MSE=123246.675 RMSE = 351.065	$R^2 = 0.642$ MAE = 133.581 MSE=112243.455 RMSE = 335.028	$R^2 = 0.822$ MAE = 115.058 MSE=83024.072 RMSE = 288.139	$R^2 = 0.811$ MAE = 112.160 MSE=78936.790 RMSE = 280.957	$R^2 = 0.781$ MAE = 109.004 MSE=74166.083 RMSE = 272.335
CatBoost	$R^2 = 0.715$ MAE = 143.438 MSE=130718.419 RMSE = 361.550	$R^2 = 0.657$ MAE = 140.149 MSE=124706.588 RMSE = 353.138	$R^2 = 0.637$ MAE = 137.990 MSE=113769.201 RMSE = 337.297	$R^2 = 0.797$ MAE = 125.205 MSE = 94830.876 RMSE = 307.946	$R^2 = 0.780$ MAE = 122.606 MSE=91668.397 RMSE = 302.768	$R^2 = 0.753$ MAE = 118.349 MSE=83681.147 RMSE = 289.277
Decision Tree	$R^2 = 0.606$ MAE = 175.925 MSE=180427.465 RMSE = 424.768	$R^2 = 0.499$ MAE = 173.402 MSE=182156.907 RMSE = 426.798	$R^2 = 0.483$ MAE = 159.667 MSE=162307.888 RMSE = 402.875	$R^2 = 0.726$ MAE = 78.980 MSE=128072.937 RMSE = 357.873	$R^2 = 0.706$ MAE = 80.415 MSE=122571.887 RMSE = 350.103	$R^2 = 0.672$ MAE = 74.917 MSE=111257.443 RMSE = 333.553

For a better understanding of the approach, a printout of the number of outliers was made, before and after treatment, with the different contamination levels (Table 6.36).

Table 6.36: Evolution of the number of outliers after listwise deletion in continuous variables with IF-based outlier detection.

Variable	Site 1				Site 2			
	X_Train		X_Test		X_Train		X_Test	
	Before	After	Before	After	Before	After	Before	After
contamination= 0.01	1402	0	369	0	1711	0	447	0
contamination= 0.05	7008	0	1672	0	8554	0	2166	0
contamination= 0.1	14015	0	3414	0	17108	0	4314	0

Looking at both Sites, better results in performance were always associated with low levels of contamination, proving that there are extreme values that really contribute to the model performance. Here the best approach seems to be contamination=0.05, since it gives a good balance between error and R^2 .

2.2 - The second approach involved applying KNN imputation while using IF algorithm as detection criteria.

To evaluate the impact of the number of neighbors in the KNN imputation after outlier detection using IF, different contamination levels were tested, each combined with two configurations: 3 and 7 neighbors. The results of the approach are presented in Table 6.37.

Contrary to KNN treatment with IQR detection, numerical continuous variables were normalized using RobustScaler instead of MinMaxScaler, because it showed greater results. Scaling is critical for the KNN Imputer to ensure that all variables have equal weight in determining nearest neighbors. This RobustScaler relies on robust statistics, namely the median and IQR, which are less sensitive to outliers. In this way, it maintains the scale of the data while preserving important differences without allowing outliers to distort the process [96].

Table 6.37: Model validation metrics after IF-based outlier detection followed by KNN imputation during preprocessing of continuous variables (TotalProcessTimeout, ExamDelay, SchedulingTime, Scheduling-TimeDeviation, WaitingForExam).

Model	Site 1 Non X-ray			Site 2 Non X-ray		
	contamination=0.01 neighbors=3	contamination=0.03 neighbors=3	contamination=0.01 neighbors=7	contamination=0.01 neighbors=3	contamination=0.03 neighbors=3	contamination=0.01 neighbors=7
XGBoost	$R^2 = 0.621$ MAE = 169.129 MSE=196802.536 RMSE = 443.624	$R^2 = 0.533$ MAE = 181.862 MSE=242236.587 RMSE = 492.175	$R^2 = 0.621$ MAE = 169.129 MSE=196802.536 RMSE = 443.624	$R^2 = 0.822$ MAE = 102.823 MSE=94966.519 RMSE = 308.166	$R^2 = 0.768$ MAE = 112.446 MSE=124173.863 RMSE = 352.383	$R^2 = 0.822$ MAE = 102.823 MSE=94966.519 RMSE = 308.166
LightGBM	$R^2 = 0.667$ MAE = 150.915 MSE=172682.137 RMSE = 415.550	$R^2 = 0.588$ MAE = 161.144 MSE=213735.980 RMSE = 462.316	$R^2 = 0.667$ MAE = 150.915 MSE=172682.137 RMSE = 415.550	$R^2 = 0.790$ MAE = 127.017 MSE=112209.099 RMSE = 334.976	$R^2 = 0.734$ MAE = 138.971 MSE=142377.408 RMSE = 377.329	$R^2 = 0.790$ MAE = 127.017 MSE=112209.099 RMSE = 334.976
CatBoost	$R^2 = 0.670$ MAE = 154.733 MSE=171285.224 RMSE = 413.866	$R^2 = 0.574$ MAE = 167.924 MSE=221066.365 RMSE = 470.177	$R^2 = 0.670$ MAE = 154.733 MSE=171285.224 RMSE = 413.866	$R^2 = 0.772$ MAE = 136.852 MSE=121743.639 RMSE = 348.918	$R^2 = 0.712$ MAE = 149.561 MSE=153826.410 RMSE = 392.207	$R^2 = 0.772$ MAE = 136.852 MSE=121743.639 RMSE = 348.918
Decision Tree	$R^2 = 0.537$ MAE = 189.259 MSE=240398.650 RMSE = 490.305	$R^2 = 0.430$ MAE = 201.585 MSE=295802.164 RMSE = 543.877	$R^2 = 0.537$ MAE = 189.259 MSE=240398.650 RMSE = 490.305	$R^2 = 0.693$ MAE = 87.516 MSE=164017.447 RMSE = 404.990	$R^2 = 0.566$ MAE = 105.273 MSE=231918.058 RMSE = 481.579	$R^2 = 0.693$ MAE = 87.516 MSE=164017.447 RMSE = 404.990

For a better understanding of the approach, a printout of the number of outliers was made, before and after treatment, with the different contamination levels (Table 6.38).

Table 6.38: Evolution of the number of outliers after KNN imputation in continuous variables with IF-based outlier detection.

Variable	Site 1				Site 2			
	X_Train		X_Test		X_Train		X_Test	
	Before	After	Before	After	Before	After	Before	After
contamination= 0.01	1402	0	369	0	1711	0	447	0
contamination= 0.03	4205	0	996	0	5133	0	1318	0

The increase in contamination showed worse results. This may indicate that, as more values were identified as outliers, more important values have been excluded, reducing performance and increasing error.

Conclusion of the best treatment approach after IF-based outlier detection.

Removal and KNN imputation were found to effectively preserve the original detection's multivariate coherence, since KNN facilitates imputation along multiple variables, without eliminating the relationship structures present in the data, and simple pruning avoids causing distortions by simply removing offending

cases. With regard to results, removal directly allowed us to preserve the integrity of the data, avoiding artificial values from being inserted that could distort the real patterns present.

The effectiveness of the removal is that the majority of the detected outliers are actually rare cases which do not follow the overall patterns of the dataset. Therefore, their removal helps to improve the quality and consistency of the data for modeling, as seen in Figure 6.32, where each point represents a model prediction compared to the actual value.

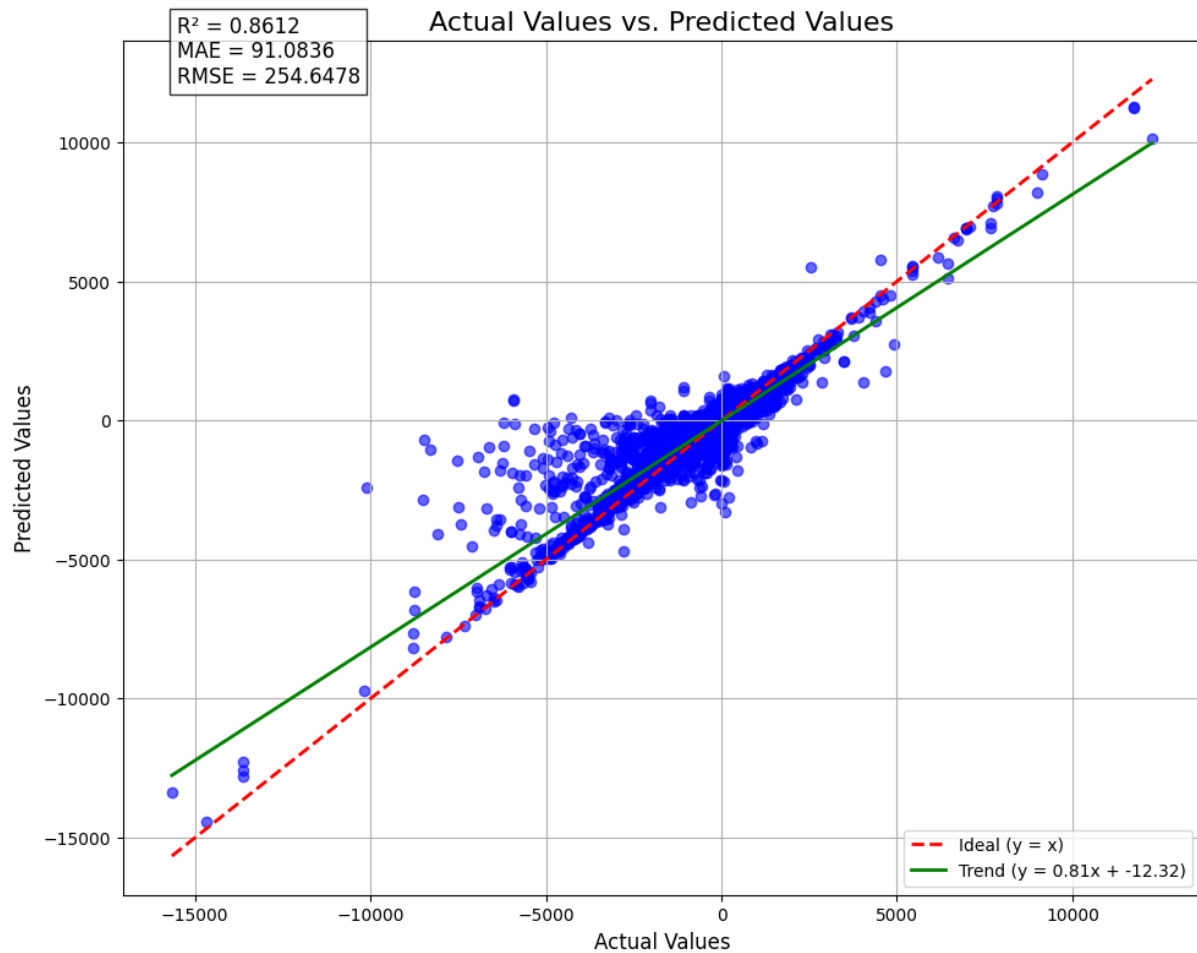


Figure 6.32: Impact of listwise deletion during preprocessing of continuous features after IF-based outlier detection, trained and tested with XGBoost with Site 2 data.

6.6 Summary

This chapter addresses Non X-ray radiological exams of the two hospitals. In this chapter was possible to understand the skewed distributions of different variables referent to these exams, prepare data with initial preprocessing and derived feature creation, and then address categorical e numerical outliers, with different approaches. For the latter, the manual approach showed insignificant results but in the automatic approach, two detection methods were tested showing IQR with better values of overall performance. Different levels of contamination and factors were tried in order to understand the impact in the model. In Chapter 8 a comparative analysis of these results will be made.

Chapter 7

X-ray Radiological Data

Chapter 7 is dedicated to the analysis of X-ray radiological data, following a structure similar to the one used for the non X-ray data to enable comparative insights. It begins with an overview of the dataset and an evaluation of baseline model performance for this specific subset. The next sections describe the preliminary data preparation steps and explain the target for this type of exams. The chapter then defines the prediction target and presents the final structure of the dataset after these initial preprocessing steps. It proceeds with the detection and treatment of categorical outliers, analyzing their impact on model performance.

As made for Non X-ray data, due to distinct hospital flows and specific associated categories, it was decided to analyze the X-ray data from two different hospitals separately. This entire chapter shows how X-ray data was handled.

7.1 Data Understanding

As previously made for non X-ray, it was essential to gain a comprehensive understanding of the data related to X-ray exams from both Sites, prior to any data preprocessing steps. This included exploring the structure and content of the available information, identifying key variables, and uncovering initial data quality issues.

- Site 1 represents two units of a hospital. Initially, this dataframe consisted of 467498 rows and 37 columns. The total percentage of missing values in the dataset is 28.65%.
- Site 2 represents four units of a hospital. Initially, this dataframe consisted of 421235 rows and 37 columns. The total percentage of missing values in the dataset is 27.87%.

Similar to other types of exams, one of the major problems found in both Sites as data understanding started, was the large percentages of missing values for several variables, which are presented in Table 7.1.

Table 7.1: Missing values in X-ray radiological data.

	Site 1 X-ray		Site 2 X-ray	
	Missing Values	Percentage of Missing Values (%)	Missing Values	Percentage of Missing Values (%)
RequestDate	0	0.00	0	0.00
SchedulingDate	47789	10.22	31798	7.55
ExamDate	26732	5.72	15914	3.78
AdmissionDate	48031	10.27	32224	7.65
CallDate	54752	11.71	32680	7.76
RealizationDate	57064	12.21	37101	8.81
RealizationShift	0	0.00	0	0.00
RealizationDay	57064	12.21	37101	8.81
ReportDate	457484	97.86	402623	95.58
ValidationDate	460437	98.49	406519	96.51
ValidationShift	0	0.00	0	0.00
ValidationDay	463537	99.15	406519	96.51
DateNextAppointment	459692	98.33	335202	79.58
AppointmentSpecialty	460448	98.49	354271	84.10
AutomaticScheduling	0	0.00	0	0.00
DirectAdmission	0	0.00	0	0.00
Rescheduled	0	0.00	0	0.00
PreviousSchedulingDate	410322	87.77	361697	85.87
PreviousExamDate	410322	87.77	361697	85.87
ResultsDate	486	0.10	211540	50.22
Origin	0	0.00	13	0.00
RealizationUnit	26732	5.72	15914	3.78
Resource	26732	5.72	15915	3.78
ExternalResource	0	0.00	0	0.00
PatientAge	57064	12.21	37101	8.81
PatientGender	0	0.00	0	0.00
Priority	0	0.00	0	0.00
Specialty	0	0.00	0	0.00
MedicalAct	0	0.00	0	0.00
FamilyMedicalAct	0	0.00	0	0.00
TypeMedicalAct	0	0.00	0	0.00
HasImages	26732	5.72	15914	3.78
PatientCode	0	0.00	0	0.00
ReportCode	457484	97.86	402623	95.58
UsesVoiceRecognition	460402	98.48	406525	96.51
ValidatedBy	460437	98.49	406519	96.51
AccessionNumber	26732	5.72	15914	3.78

Variables like ReportDate, ValidationDate and ValidationDay have values of missingness close to 100%. This happens because of the traditional workflow of X-rays in hospitals explained in Section 5.3. X-ray exams do not have a text report, as it only take the images obtained during the exam to complete the process and deliver results.

DateNextAppointment, AppointmentSpecialty, PreviousSchedulingDate and PreviousExamDate, with percentages above 75% of missing values were also immediately target to be further excluded from Site 1 since there was no viable way to impute these dates. The same happened for Site 2.

In Table 7.2 it is possible to see that the variable FamilyMedicalAct was restricted to just one category, X-rays.

Table 7.2: Number of unique categories per variable for both Sites.

	Unique Values Site 1	Unique Values Site 2
RealizationShift	6	6
RealizationDay	7	7
ValidationShift	6	6
ValidationDay	7	7
AutomaticScheduling	2	2
DirectAdmission	2	2
Rescheduled	2	2
Origin	7	6
RealizationUnit	2	4
Resource	20	26
ExternalResource	1	1
PatientGender	3	2
Priority	3	3
Specialty	1	1
MedicalAct	217	171
FamilyMedicalAct	1	1
TypeMedicalAct	14	1
HasImages	2	2
UsesVoiceRecognition	2	2

In Table 7.3, as made for non X-ray, it is possible to understand that the majority of work comes from the emergency department. This contrasts with the results for non X-ray, where outpatient appointments stood out.

Table 7.3: Percentage of origin per Site.

Origin	Percentage Site 1 (%)	Percentage Site 2 (%)
Outpatient Appointments	25.55	32.02
Emergency	55.84	49.90
Inpatient	18.40	16.40
Radiology	0.05	1.42
Operating Room	0.07	0.20
Day Hospital	0.09	0.06
Laboratory	0.00	0.00

In Figure 7.1, the distribution of age is similar to the one found in non X-ray data, although values of frequency are also high for younger people.

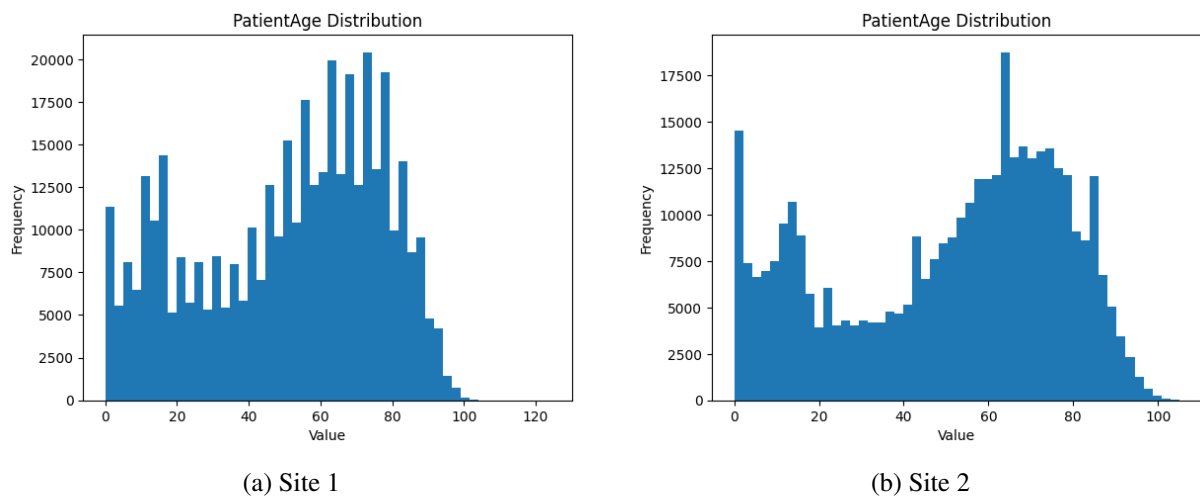


Figure 7.1: Age histogram of X-ray radiological Data.

Still in demographics factors, Figure 7.2, shows a balanced distribution of gender for both sites, as already seen also for Non X-ray radiological data.

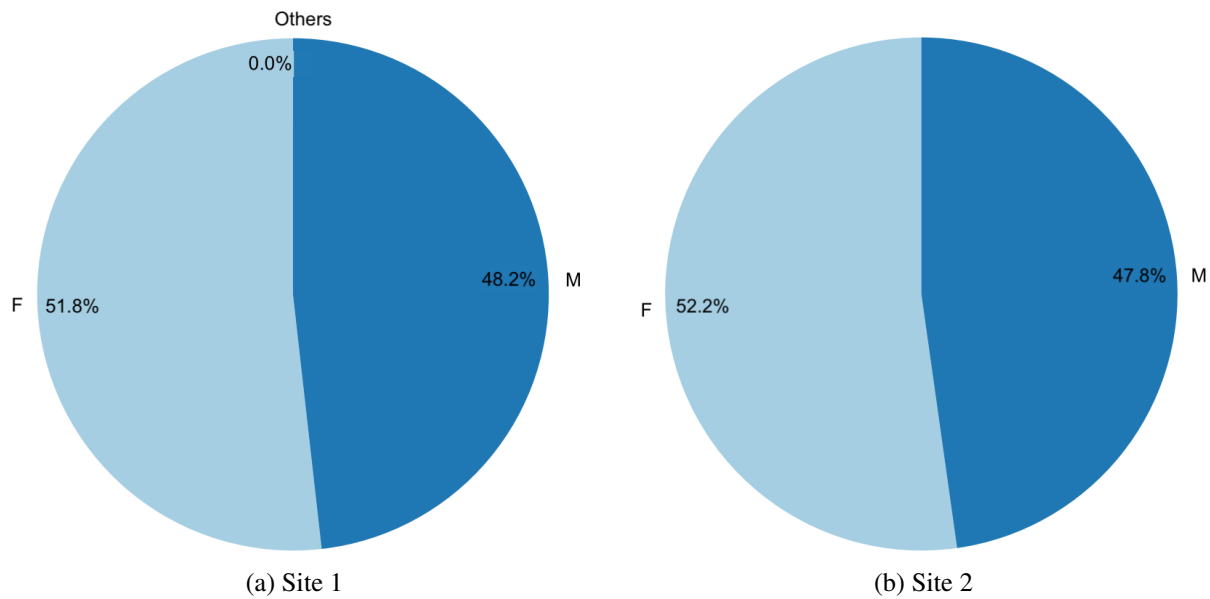


Figure 7.2: Patient gender histogram of X-ray radiological Data.

As made for non X-ray, other analysis were also conducted regarding data understanding. This first part of the project was also essential to identify possible error patterns within X-ray exams, and discuss internally to understand the reason.

7.1.1 X-ray Baseline Performance

As made for non X-ray data in Subsection 6.1.1, a baseline model was trained and tested using the raw dataset (with only minimal preprocessing), in order to serve as a reference for further comparisons. The results are in Table 7.4.

Table 7.4: Model validation metrics of the raw X-ray dataset.

Model	Site 1 X-ray	Site 2 X-ray
XGBoost	$R^2 = -0.111$ MAE = 440.390 MSE = 901008.552 RMSE = 949.215	$R^2 = 0.055$ MAE = 265.161 MSE = 350248.656 RMSE = 591.818
LightGBM	$R^2 = -0.020$ MAE = 441.789 MSE = 827288.530 RMSE = 909.554	$R^2 = 0.343$ MAE = 239.467 MSE = 243343.073 RMSE = 493.298
CatBoost	$R^2 = 0.045$ MAE = 384.978 MSE = 773900.114 RMSE = 879.716	$R^2 = 0.345$ MAE = 223.733 MSE = 242687.085 RMSE = 492.633
Decision Tree	$R^2 = -0.058$ MAE = 421.954 MSE = 857592.440 RMSE = 926.063	$R^2 = 0.220$ MAE = 238.831 MSE = 289253.812 RMSE = 537.823

Also, as expected, all metrics showed a poor dataset. After understanding, preparing started.

7.2 Preliminary Data Preparing and Exploratory Assessment

In order to establish a solid foundation for treating outliers, a preliminary data preparation was essential. This preliminary preprocessing was made before the train and test split, since the methods applied did not compromise any data leakage. The preparing included variable format standardization, deduplication, row deletion based on criteria, simple imputation methods, derived feature creation and target definition. To better understand the behavior of these new features, their distributions are also presented in this section. Additionally, auxiliary results are presented in this section, as they contributed to building and preparing the dataset in the most effective way. Finally, this section concludes by presenting the structured dataset that will be used in the subsequent outlier detection and cleaning procedures.

7.2.1 Variable Format Standardization

This process closely resembled the one applied to non X-ray data, as described in Subsection 6.2.1.

7.2.2 Deduplication

The deduplication process for X-ray datasets began by eliminating lines that were 100% duplicates, keeping only one of these versions. After some time analyzing the dataset, the very specific case of deduplication found in non X-ray data was also found but in fewer cases (4 cases for site 1 and 0 cases for site 2). The same method described in SubSection 6.2.2 was used.

7.2.3 Case Deletion Based on Criteria

X-ray set showed less rows that needed deletion, hence some of the cases of non X-ray data were also verified here, such as the issue addressed to the very high priority (that lacked significance), age, emergency department origin and others subsections, explained in Subsection 6.2.3.

Since for X-ray data the RealizationDate becomes very important for target definition and since there is no viable imputation method for these cases, the respective rows with this value missing were eliminated.

7.2.4 Imputation Methods Based on Known Patterns

This process was equal to Non X-ray Radiological data, which is explained in SubSection 6.2.4.

7.2.5 Derived Feature Creation

This process differed from non X-ray data due to preservation of temporal causality. As it will be explained in the following section, the target defined for X-ray data involves the realization date. Hence, every feature that was only known after the target date (after the exam realization) were removed from the dataset after target creation. The features created were: SchedulingTimeDeviation, TotalProcessTimeout, SchedulingTime and WaitingForExam and are all explained in SubSection 6.2.5, with respective statistics describe in Tables 7.5, 7.6, 7.7 and 7.8.

Table 7.5: Descriptive statistics for SchedulingTimeDeviation while preprocessing X-ray Radiological data of Site 1 and 2.

	Site 1 (in hours)	Site 2 (in hours)
Mean	244.314	469.253
STD	1387.106	1343.252
Min	-1222.020	-290.648
25%	-312.180	-197.135
50%	0.023	0.357
75%	318.602	383.821
Max	17044.586	18214.331

Table 7.6: Descriptive statistics for TotalProcessTimeout while preprocessing X-ray Radiological data of Site 1 and 2.

	Site 1 (in hours)	Site 2 (in hours)
Mean	1224.217	1088.809
STD	1721.337	1646.120
Min	0.000	0.000
25%	1.253	0.000
50%	324.246	225.571
75%	2052.449	1739.232
Max	26303.971	17456.211

Table 7.7: Descriptive statistics for SchedulingTime while preprocessing X-ray Radiological data of Site 1 and 2.

	Site 1 (in hours)	Site 2 (in hours)
Mean	956.160	662.600
STD	1588.330	1382.007
Min	0.000	0.000
25%	0.903	1.625
50%	21.144	24.997
75%	1535.617	672.290
Max	18266.609	18504.979

Table 7.8: Descriptive statistics for WaitingForExam while preprocessing X-ray Radiological data of Site 1 and 2.

	Site 1 (in hours)	Site 2 (in hours)
Mean	1201.266	1131.848
STD	1724.102	1668.084
Min	0.001	0.000
25%	9.454	2.440
50%	246.138	360.088
75%	2015.853	1795.490
Max	19346.310	18893.327

7.2.6 Target Definition

The delay in obtaining the X-rays report results, considering the initially proposed limit defined at the time of requisition, is calculated as:

$$\text{Delay in conclusion (in hours)} = \text{ResultsDate} - \text{RealizationDate}$$

This distinction is necessary because X-ray results typically contain only the exam images, meaning the exam is considered complete on the same day and hour it is performed, as soon as the imaging is concluded. In order to preserve **temporal causality**, as made to Non X-ray Radiological data, all variables that are only known after the target date were removed from the dataset.

7.2.7 Final Dataset Structure After Preliminary Cleaning

After preliminary preparing the dataset from Site 1 contains 168646 rows and 21 columns, while the dataset from Site 2 consists of 178222 rows and the same 21 columns. The composition of both datasets is as follows:

- *AutomaticScheduling*
- *DirectAdmission*
- *Rescheduled*
- *Origin*
- *RealizationUnit*
- *Resource*
- *ExternalResource*
- *PatientAge*
- *PatientGender*
- *Priority*
- *Specialty*
- *MedicalAct*
- *FamilyMedicalAct*
- *TypeMedicalAct*
- *HasImages*
- *UsesVoiceRecognition*
- *TotalProcessTimeout*
- *SchedulingTime*
- *WaitingForExam*
- *SchedulingTimeDeviation*
- *target_delay_conclusion*

Table 7.9 shows model validation metrics after preliminary cleaning and features creation.

Table 7.9: Model validation metrics after preliminary cleaning and feature creation for X-ray data.

Model	Site 1 X-ray	Site 2 X-ray
XGBoost	$R^2 = 0.891$ MAE = 42.439 MSE = 18067.841 RMSE = 134.417	$R^2 = 0.924$ MAE = 25.298 MSE = 26839.259 RMSE = 163.827
LightGBM	$R^2 = 0.963$ MAE = 15.662 MSE = 6136.064 RMSE = 78.333	$R^2 = 0.977$ MAE = 17.058 MSE = 8077.050 RMSE = 89.872
CatBoost	$R^2 = 0.967$ MAE = 15.144 MSE = 5500.346 RMSE = 74.164	$R^2 = 0.978$ MAE = 16.036 MSE = 7657.693 RMSE = 87.508
Decision Tree	$R^2 = 0.971$ MAE = 9.546 MSE = 4821.069 RMSE = 69.434	$R^2 = 0.978$ MAE = 6.889 MSE = 7858.783 RMSE = 88.650

After establishing a solid structure as made for non X-ray data, the next step is to handle outliers.

7.3 Categorical Outliers

Once more, the detection and treatment of categorical outliers played a crucial role in improving the overall quality of the data and ensuring the robustness of the predictive models. This section presents 3 main subsections: detection methods, treatment methods and the corresponding results.

7.3.1 Detection

Site 1 - Medical Act

In order to detect underrepresented feature a frequency analysis was performed and it can be seen in Figure 7.3. Categories with counts below 1.2% were grouped in a category named "Others" for better visualization. The complete list of medical acts of Site 1 is in Appendix C, Table C.3.

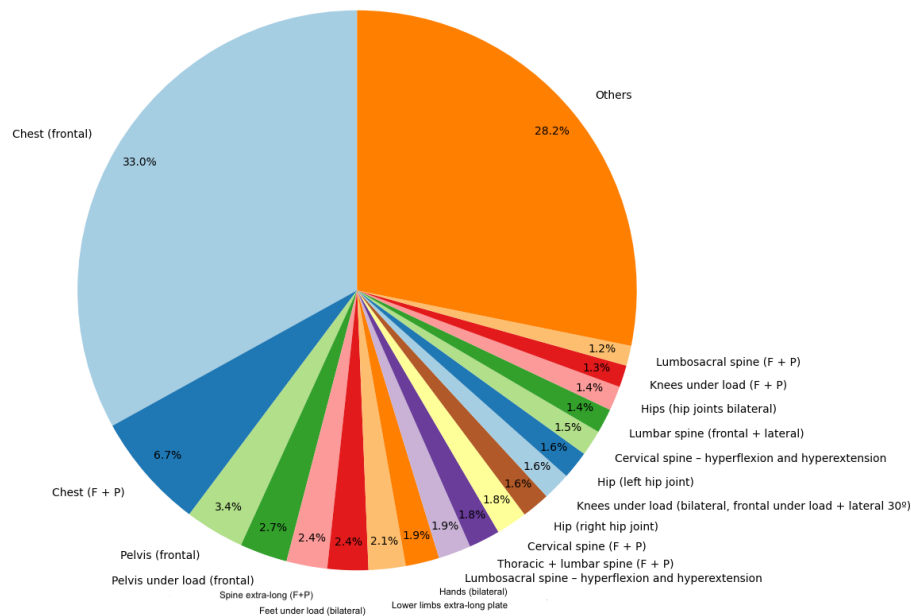


Figure 7.3: Frequency of medical acts from Site 1 X-ray radiological data.

Site 2 - Medical Act

In Figure 7.4, categories with counts below 1% were grouped in a category named "Others" for better visualization. The complete list of medical acts of Site 2 is in Appendix C, Table C.4.

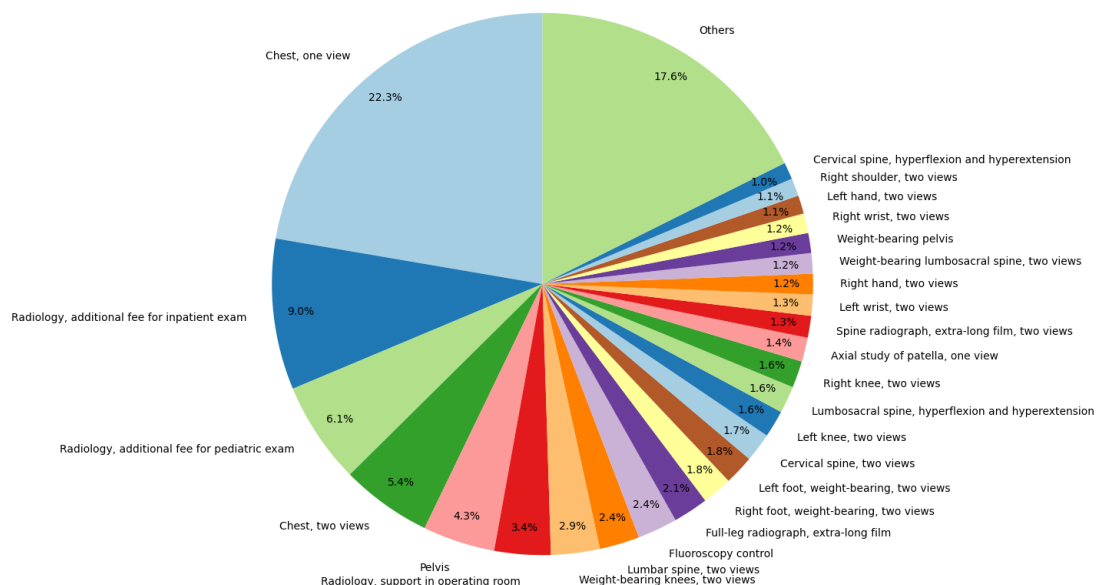


Figure 7.4: Frequency of medical acts from Site 2 X-ray radiological data.

7.3.2 Treatment

Used the same functions as in non X-ray data, each explained in Subsection 6.3.2.

7.3.3 Performance Variability Results

Both options (replacing in Table 7.10, deleting in Table 7.11) were tested, each with two different thresholds.

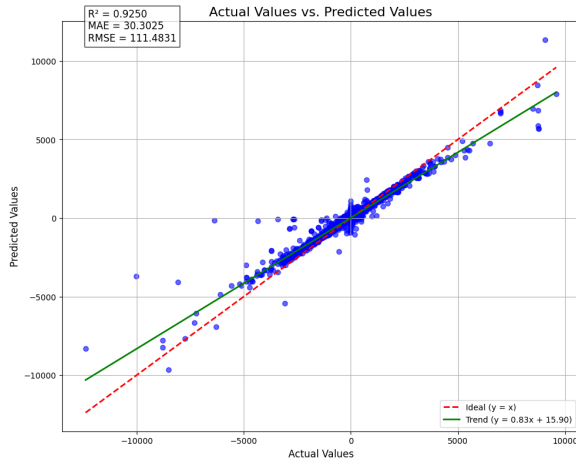
Table 7.10: Treating categorical outliers in X-ray data by replacing the rarest categories with "Other". The percentage of medical acts affected is next to the threshold value.

Model	Site 1 X-ray		Site 2 X-ray	
	threshold=100 ($\approx 1.44\%$)	threshold=1000 ($\approx 15.36\%$)	threshold=100 ($\approx 1.14\%$)	threshold=1000 ($\approx 9.23\%$)
XGBoost	$R^2 = 0.923$ MAE = 28.651 MSE = 12765.497 RMSE = 112.985	$R^2 = 0.925$ MAE = 30.302 MSE = 12428.471 RMSE = 111.483	$R^2 = 0.955$ MAE = 15.884 MSE = 15723.213 RMSE = 125.392	$R^2 = 0.960$ MAE = 14.770 MSE = 14235.171 RMSE = 119.311
LightGBM	$R^2 = 0.963$ MAE = 15.328 MSE = 6093.790 RMSE = 78.063	$R^2 = 0.964$ MAE = 15.203 MSE = 6041.378 RMSE = 77.726	$R^2 = 0.977$ MAE = 16.799 MSE = 8234.820 RMSE = 90.746	$R^2 = 0.978$ MAE = 16.596 MSE = 7785.009 RMSE = 88.233
CatBoost	$R^2 = 0.965$ MAE = 14.880 MSE = 5806.792 RMSE = 76.202	$R^2 = 0.966$ MAE = 14.995 MSE = 5556.429 RMSE = 74.541	$R^2 = 0.977$ MAE = 16.158 MSE = 8146.130 RMSE = 90.256	$R^2 = 0.979$ MAE = 15.928 MSE = 7377.454 RMSE = 85.892
Decision Tree	$R^2 = 0.971$ MAE = 9.552 MSE = 4822.923 RMSE = 69.447	$R^2 = 0.971$ MAE = 9.524 MSE = 4806.308 RMSE = 69.328	$R^2 = 0.979$ MAE = 6.774 MSE = 7277.642 RMSE = 85.309	$R^2 = 0.978$ MAE = 6.905 MSE = 7747.322 RMSE = 88.019

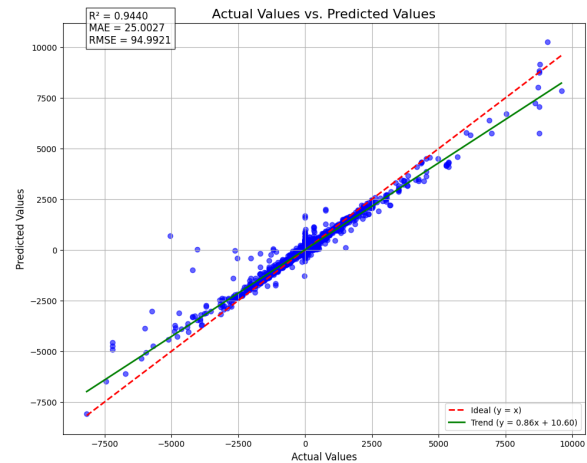
Table 7.11: Treating categorical outliers in X-ray data by deleting the rarest categories of MedicalAct. The percentage of medical acts affected is next to the threshold value.

Model	Site 1 X-ray		Site 2 X-ray	
	threshold=100 ($\approx 1.44\%$)	threshold=1000 ($\approx 15.36\%$)	threshold=100 ($\approx 1.14\%$)	threshold=1000 ($\approx 9.23\%$)
XGBoost	$R^2 = 0.938$ MAE = 28.167 MSE = 11104.925 RMSE = 105.380	$R^2 = 0.944$ MAE = 25.003 MSE = 9023.507 RMSE = 94.992	$R^2 = 0.961$ MAE = 17.175 MSE = 14301.582 RMSE = 119.589	$R^2 = 0.950$ MAE = 16.357 MSE = 18979.610 RMSE = 137.767
LightGBM	$R^2 = 0.971$ MAE = 15.598 MSE = 5238.415 RMSE = 72.377	$R^2 = 0.971$ MAE = 14.548 MSE = 4612.634 RMSE = 67.916	$R^2 = 0.973$ MAE = 17.461 MSE = 9912.796 RMSE = 99.563	$R^2 = 0.968$ MAE = 17.589 MSE = 12366.419 RMSE = 111.204
CatBoost	$R^2 = 0.976$ MAE = 14.553 MSE = 4297.954 RMSE = 65.559	$R^2 = 0.975$ MAE = 13.637 MSE = 3967.595 RMSE = 62.989	$R^2 = 0.977$ MAE = 16.812 MSE = 8538.578 RMSE = 92.404	$R^2 = 0.973$ MAE = 17.251 MSE = 10355.365 RMSE = 101.761
Decision Tree	$R^2 = 0.965$ MAE = 11.067 MSE = 6310.943 RMSE = 79.441	$R^2 = 0.968$ MAE = 13.970 MSE = 5184.685 RMSE = 72.005	$R^2 = 0.973$ MAE = 7.611 MSE = 9680.123 RMSE = 98.388	$R^2 = 0.964$ MAE = 8.461 MSE = 13854.613 RMSE = 117.706

By looking at values before treatment and after the two options of treatment given above, it is clear the effect categorical outliers can have in model performance. Below, in Figure 7.5, it is possible to observe the difference.



(a) XGBoost evolution with the replacing of MedicalActs that have <1000 counts with a wider class "Other".



(b) XGBoost evolution with the removing of MedicalActs that have <1000 counts.

Figure 7.5: Impact of replacing or removal of categorical outliers during preprocessing, trained and tested with Site 1 data.

X-rays have an important particularity compared to other imaging examinations: they are completed immediately after the images of the exam are captured, since, in these specific cases, they do not require the manual preparation of a detailed textual report to further analysis. In these cases, the system considers the results "completed" as soon as the images are made available, which means that the time between the exam being carried out and the completion of the report is practically zero or non-existent. This makes the target for X-ray data has very little variation and presents low values, close to zero. As a result, the model is able to predict this time with high accuracy, resulting in significantly reduced error metrics and higher performance rates. Given that the results obtained for X-ray are already quite satisfactory, the additional application of methods for detecting and treating numerical outliers in this specific case is not justified, since no relevant performance gains are expected.

7.4 Summary

This chapter addresses the X-ray radiological data exams of the two hospitals. In this chapter it was possible to understand the distributions of different variables referent to these exams, prepare data with initial preprocessing and derived feature creation, and then address categorical outliers which ensured great results. The absence of real delays or significant anomalies in the X-ray subset makes the prediction task simpler, which naturally improves the performance of the models.

Chapter 8

Comparative Analysis

As proposed in the methodology, the main goal of this chapter is to compare the effects of different data cleaning methods through different ML models, used as metrics of evaluation. The discussion of results is based on performance, interpretability, and generalizability metrics. Since, in comparison to non X-ray data, the X-ray data did not undergo so many preprocessing techniques and still showed high accuracy (likely due to the absence of real delays), this comparative analysis will focus primarily on the non X-ray data, where delays were more representative and required more sophisticated handling.

Starting to compare Site 1 with Site 2, looking at the results of the previous chapters, the improved performance of the models in Site 2 dataset is rather obvious. There are several notable aspects that were tested in Section 6.1 which could be accountable for this. First, Site 2 had fewer missing values (14.24%) compared to Site 1 (19.61%), and that also would have meant fewer data changed and left out in the initial cleaning. Site 2 collects data from four hospital units, which comes with higher variability and representativeness to the data. The diversity in this context might have assisted the model in learning relevant patterns and generalizing more. On the other hand, Site 1 captures a single unit of data, that could lead to lower variability and less generalizable outcomes. Because of the extensive range of the results obtained from both datasets, and in the perspective of gaining a more focused and in-depth analysis, it was decided to place **Site 2 non X-ray data at the center of the analysis**. Not only were more stable and promising predictive results achieved with this dataset, but it was also of better data quality in terms of completeness and representativeness. By concentrating on Site 2, the results can be explored in greater detail and more meaningful conclusions established that are potentially transferable to the broader hospital environment, since Site 2 is a composite of four various hospital units.

As mentioned in previous chapters, the cleaning process began with a well-defined order, and its analysis will follow accordingly:

1. Establish the baseline model performance using a minimally preprocessed version of the Site 2 non X-ray dataset;
2. Apply initial preprocessing steps and perform derived feature engineering to enrich the dataset;
3. Correct categorical outliers to ensure data consistency across nominal features;
4. Address numerical outliers using two different approaches:
 - (a) Manual removal of continuous numerical outliers based on domain knowledge and distribution analysis;
 - (b) Application of established automated methods based on IQR and IF detection to treat both cyclic and continuous numerical outliers.

The initial steps (1,2,3 in the order above) had a significant impact on performance and error variability and this can be observed in the subsequent figures.

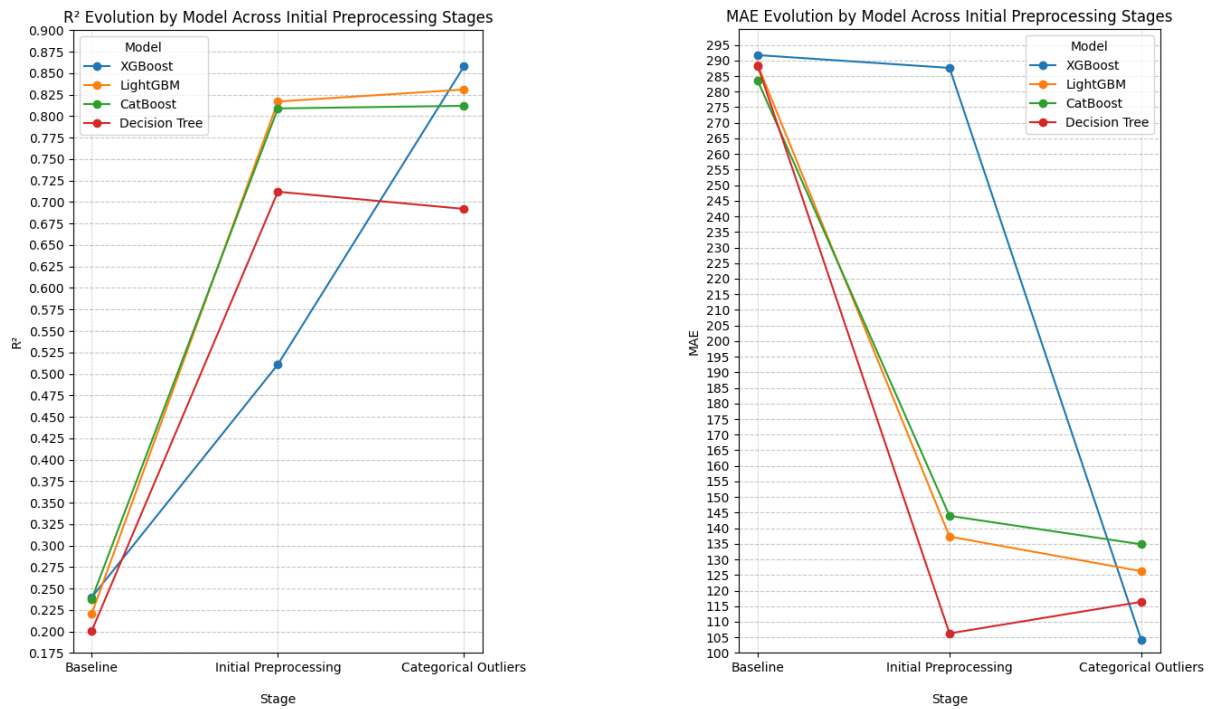


Figure 8.1: Comparative performance and MAE analysis of ML models during baseline, preliminary preprocessing and categorical outliers removal with data from Site 2.

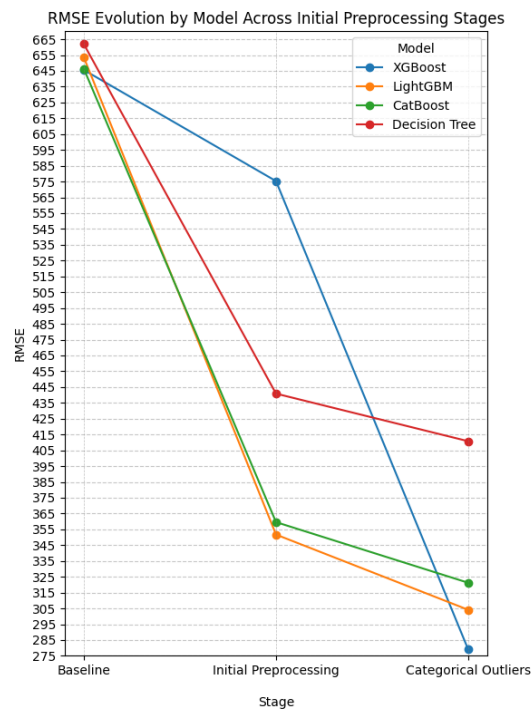


Figure 8.2: Comparative RMSE analysis of ML models during baseline, preliminary preprocessing and categorical outliers removal with data from Site 2.

Looking at Figures 8.1 and 8.2, it is possible to observe that all models started with a limited predictive power at baseline, highlighting the poor quality of the raw dataset. XGBoost's behavior is particularly notable: while other models reach performance levels over 0.8 already in the preprocessing stage, XGBoost remains at 0.52, suggesting that this model requires full treatment of categorical outliers to reach its potential, as it reaches 0.86 in the end. Looking at MAE, XGBoost demonstrates again initial resistance to preprocessing improvements, maintaining MAE virtually unchanged (approximately 290) between baseline and preprocessing, before a substantial drop to approximately 150 in the final stage, this reinforces again how categorical outliers can impact XGBoost. The other three models respond immediately to preprocessing, with Decision Tree showing the greatest improvement, jumping from 285 to 107, and later stabilizing at approximately 118. At baseline, all models start with high RMSE (650-660), but there are significant differences in the rate of improvement. XGBoost shows a more gradual improvement, remaining relatively high in the preprocessing stage before dropping sharply to in the categorical outliers phase. In contrast, LightGBM, CatBoost, and Decision Tree show more dramatic improvements in the first preprocessing stage, converging to similar final values.

Thereby, it is possible to take some preliminary conclusions:

- XGBoost demonstrates a greater dependence of the treatment of categorical outliers, while LightGBM and CatBoost respond more quickly to the first steps.
- LightGBM and CatBoost show greater robustness to initial data conditions, achieving good performance with less intensive preprocessing.
- For scenarios where quick improvements are needed with basic preprocessing, Decision Tree offers excellent return on initial investment, but it is not the best choice for maximum performance.

After the first stages analysis, the treatment of numerical outliers started. Starting with the manual approach (4.(a) in the order specified above), in Figure 8.3 and 8.4 is presented the variability of different metrics to evaluate the method.

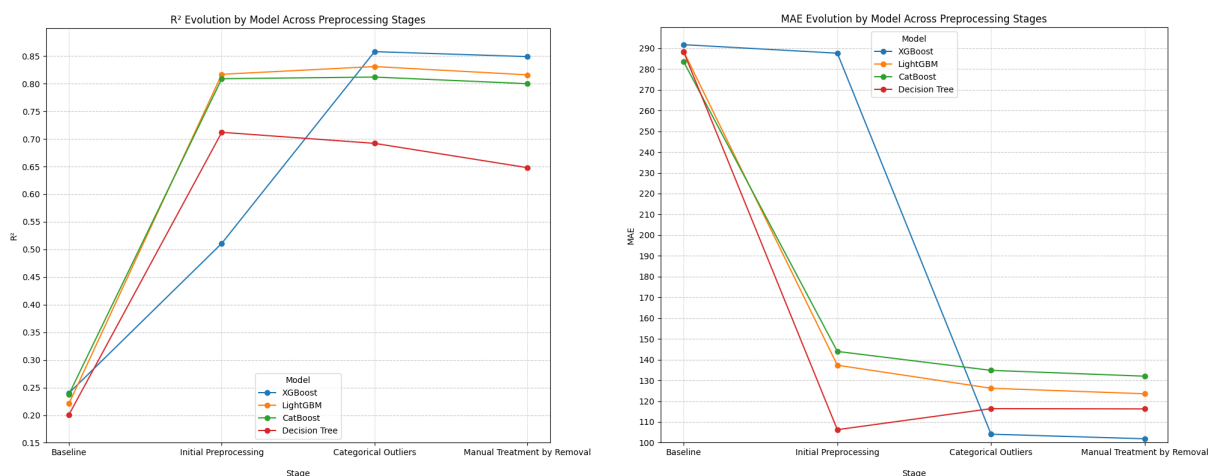


Figure 8.3: Comparative performance and MAE analysis of ML models during manual numerical outliers removal with data from Site 2.

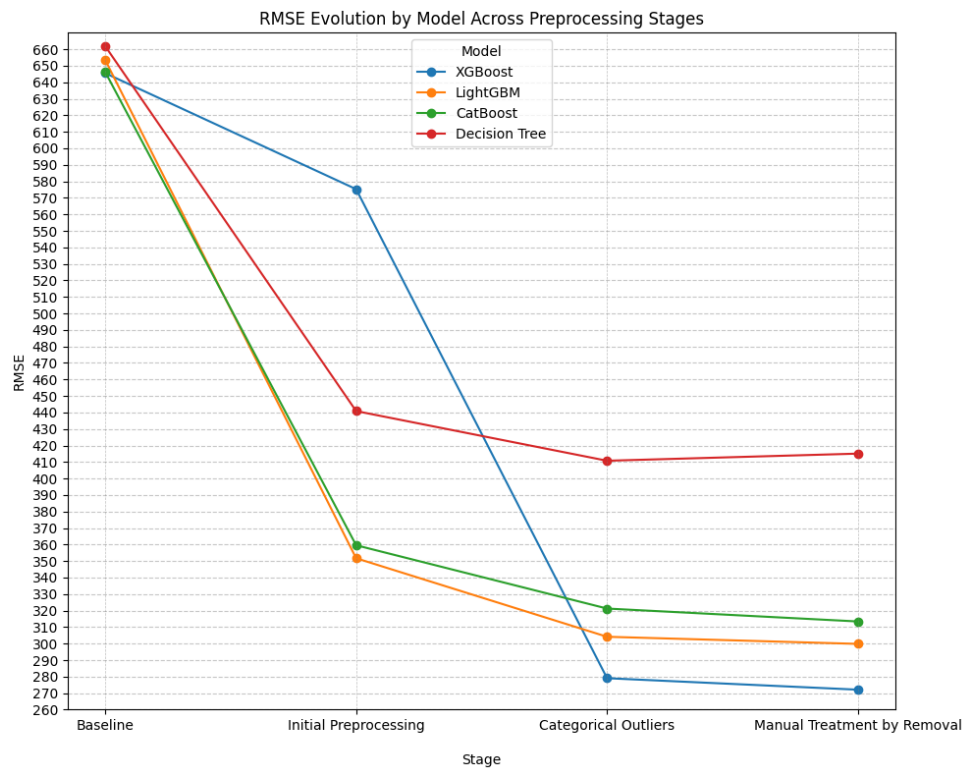


Figure 8.4: Comparative RMSE analysis of ML models during manual numerical outliers removal with data from Site 2.

Looking at the results, it is clear to see that the manual removal had very low or even zero effect across all metrics, for every model. This analysis demonstrates that data science benefits more from informed automated methods (as will be seen below) than from manual intervention, even when that intervention is done with careful visual analysis.

After manual removal analysis, in the second approach (4.(b) in the order specified above), different known techniques for handling numerical outliers were explored, using also two different methods of detection. The analysis of the three following graphs presented in Figures 8.5, 8.6 and 8.7 revealed a complex pattern of how different ML algorithms respond to sequential outlier treatments. To enhance readability, the following graphs present the remainder of the experimental pipeline: the cleaning of categorical outliers (repeated from the previous graphs for reference), followed by the treatment of cyclic and numerical outliers using two distinct detection techniques (IQR and IF). This allows to understand not only which model performed best, but also how each algorithm behaved when faced with progressive data cleaning. To carry out the comparative analysis, the threshold values or factors that yielded the best results in the previous chapter were chosen for each technique (based on the best combination of high performance and low error). This comparison between different thresholds for the same technique was conducted in Subsections 6.5.3 and 6.5.4.

A broader hyperparameter tuning was conducted (as explained in Section 5.4). This was also because performing individual hyperparameter optimization for each preprocessing technique would be computationally and time intensive, as there are 4 models, 4 techniques with IQR and 2 with IF (each with different levels of contamination), all for 2 hospital datasets. Nevertheless, to validate this and ensure a equitable comparison, a analysis on hyperparameters was made with the datasets that had gone through the best techniques found.

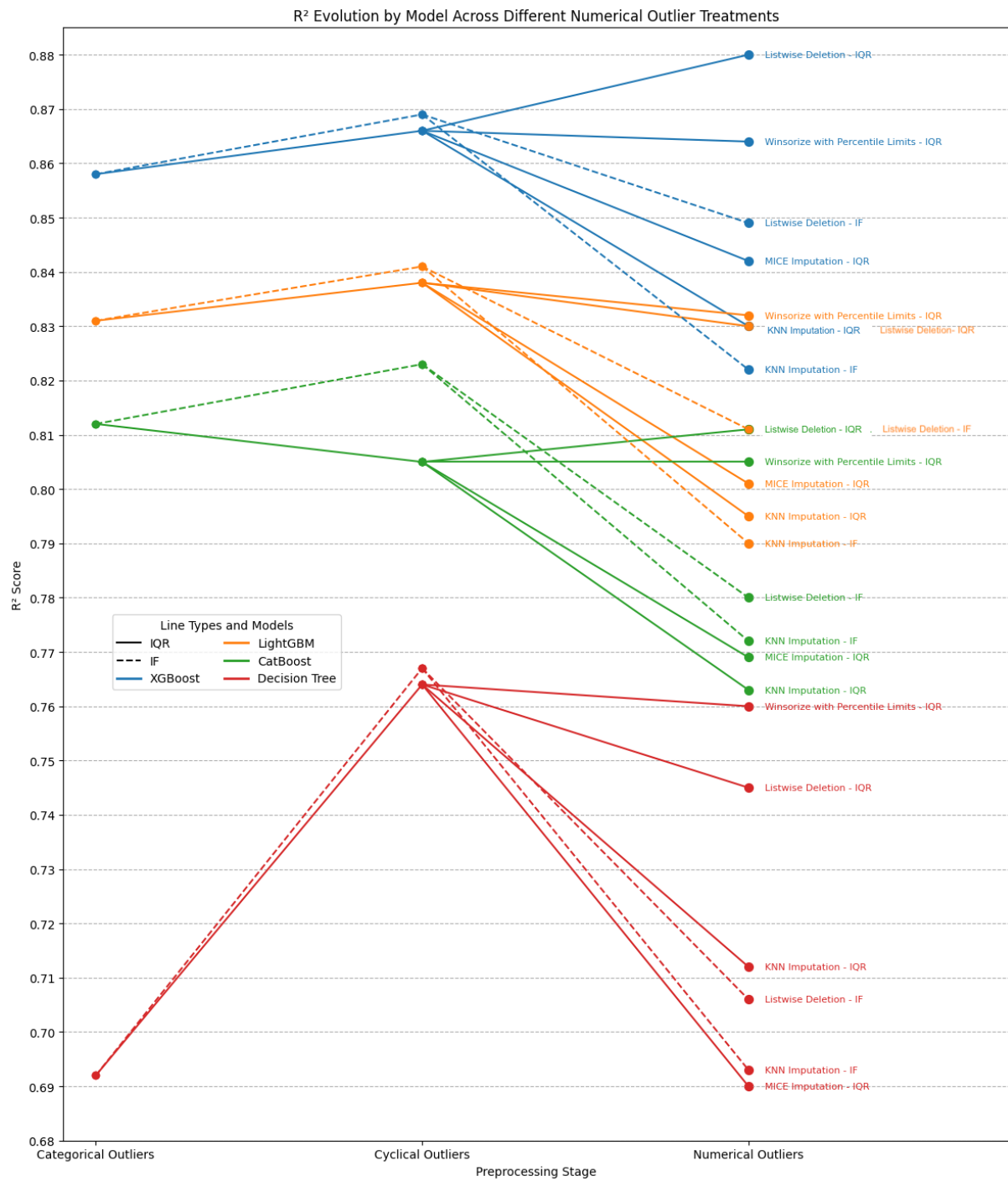


Figure 8.5: Comparative performance analysis of ML models during categorical outliers removal and different numerical outliers treatment with data from Site 2.



Figure 8.6: Comparative MAE analysis of ML models during categorical outliers removal and different numerical outliers treatment with data from Site 2.

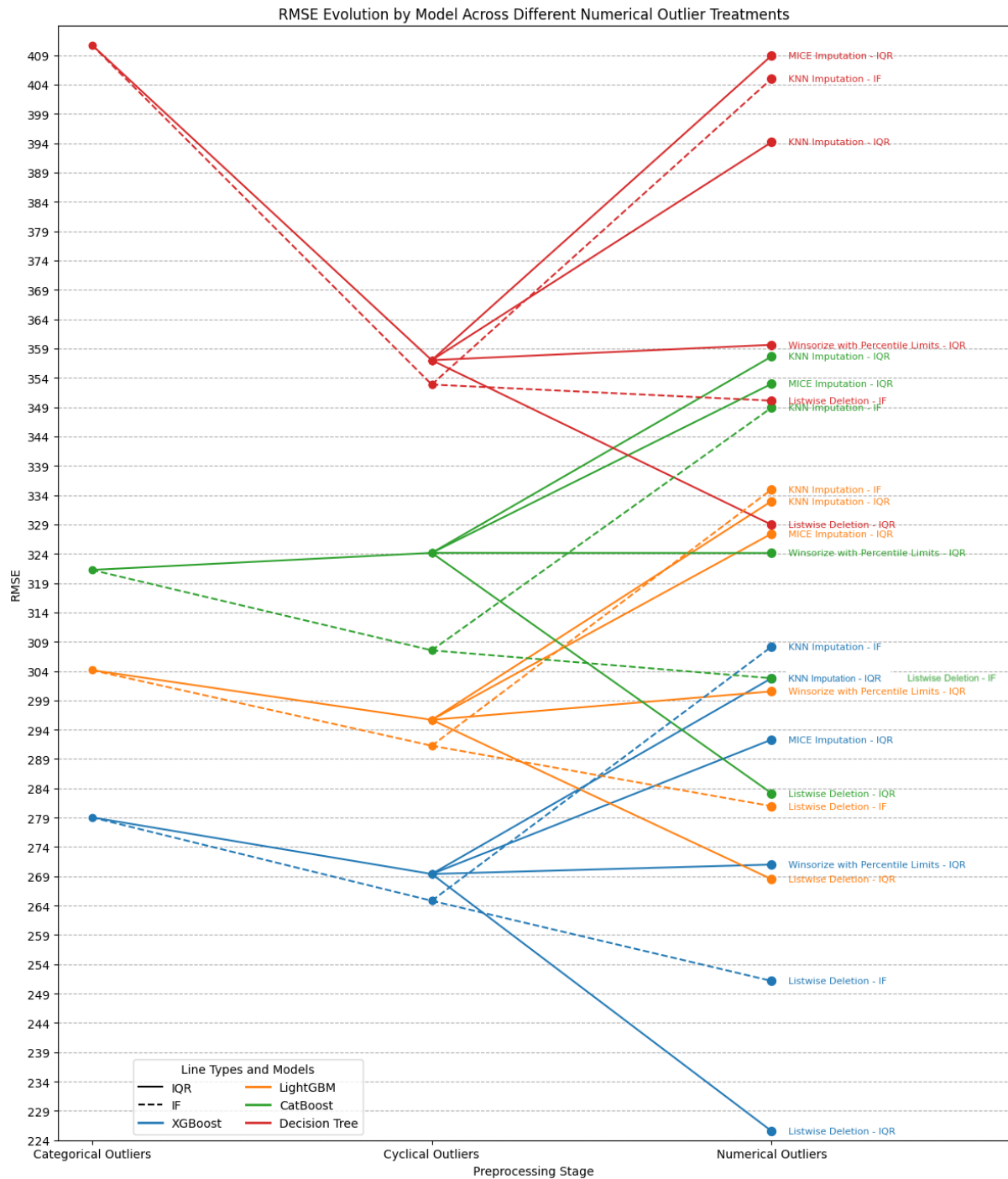


Figure 8.7: Comparative RMSE analysis of ML models during categorical outliers removal and different numerical outliers treatment with data from Site 2.

Listwise Deletion and Winsorization with IQR detection datasets were re-tuned to ensure optimal hyperparameters did not deviate from the reference used, and the set of hyperparameters did not change, validating the results above as optimal. The same approach with 3-fold CV was applied to these specific preprocessing configurations.

To ensure that the techniques that performed worse were only due to the technique and not the hyperparameters, a broader hyperparameter tuning was also re-performed on the datasets after going through those methods. The comparison is made in Tables 8.1 and 8.2.

Table 8.1: Model metrics variation when using close to optimal or optimal hyperparameters to test KNN and MICE techniques with IQR detection.

Model	KNN		MICE	
	After Reference Tuning	After Dataset-Specific Tuning	After Reference Tuning	After Dataset-Specific Tuning
XGBoost	$R^2 = 0.830$ MAE = 105.797 MSE=91677.561 RMSE = 302.783	$R^2 = 0.833$ MAE = 105.275 MSE = 89643.965 RMSE = 299.406	$R^2 = 0.842$ MAE = 103.584 MSE = 85445.959 RMSE = 292.311	$R^2 = 0.843$ MAE = 104.610 MSE = 84327.669 RMSE = 290.392
LightGBM	$R^2 = 0.795$ MAE = 131.389 MSE=110869.451 RMSE = 332.971	$R^2 = 0.795$ MAE = 131.389 MSE=110869.451 RMSE = 332.971	$R^2 = 0.801$ MAE = 130.026 MSE = 107208.483 RMSE = 327.427	$R^2 = 0.801$ MAE = 130.026 MSE = 107208.483 RMSE = 327.427
Catboost	$R^2 = 0.763$ MAE = 142.392 MSE=127901.820 RMSE = 357.634	$R^2 = 0.763$ MAE = 142.392 MSE=127901.820 RMSE = 357.634	$R^2 = 0.769$ MAE = 139.822 MSE = 124587.886 RMSE = 352.970	$R^2 = 0.769$ MAE = 139.822 MSE = 124587.886 RMSE = 352.970
Decision Tree	$R^2 = 0.712$ MAE = 92.213 MSE=155379.546 RMSE = 394.182	$R^2 = 0.712$ MAE = 92.213 MSE=155379.546 RMSE = 394.182	$R^2 = 0.690$ MAE = 92.489 MSE = 167240.909 RMSE = 408.951	$R^2 = 0.701$ MAE = 104.223 MSE = 161272.231 RMSE = 401.587

The difference in results for XGBoost was minimal for the KNN and MICE imputation with IQR Detection (less than 0.005 in R^2), as the difference between optimal hyperparameters and the already used was small (only the number of estimators decreased to 1000 for KNN and to 500 for MICE). LightGBM and Catboost showed equal hyperparameters as when tested with a refined range, for the two techniques. Decision Tree showed a minimal difference with MICE, increasing performance in 0.01 and decreasing RMSE (by changing Min Samples Leaf to 1). This shows that, with both imputation techniques, the results above could not have improved any more.

Table 8.2: Model metrics variation when using close to optimal or optimal hyperparameters to test Listwise Deletion and KNN techniques with IF detection.

Model	Listwise Deletion		KNN	
	After Reference Tuning	After Dataset-Specific Tuning	After Reference Tuning	After Dataset-Specific Tuning
XGBoost	$R^2 = 0.849$ MAE = 87.932 MSE=63078.966 RMSE = 251.155	$R^2 = 0.849$ MAE = 89.108 MSE = 62719.673 RMSE = 250.439	$R^2 = 0.822$ MAE = 102.823 MSE=94966.519 RMSE = 308.166	$R^2 = 0.823$ MAE = 103.848 MSE = 94679.542 RMSE = 307.700
LightGBM	$R^2 = 0.811$ MAE = 112.160 MSE=78936.790 RMSE = 280.957	$R^2 = 0.811$ MAE = 112.160 MSE=78936.790 RMSE = 280.957	$R^2 = 0.790$ MAE = 127.017 MSE=112209.099 RMSE = 334.976	$R^2 = 0.790$ MAE = 127.017 MSE=112209.099 RMSE = 334.976
Catboost	$R^2 = 0.780$ MAE = 122.606 MSE=91668.397 RMSE = 302.768	$R^2 = 0.780$ MAE = 122.606 MSE=91668.397 RMSE = 302.768	$R^2 = 0.772$ MAE = 136.852 MSE=121743.639 RMSE = 348.918	$R^2 = 0.772$ MAE = 136.852 MSE=121743.639 RMSE = 348.918
Decision Tree	$R^2 = 0.706$ MAE = 80.415 MSE=122571.887 RMSE = 350.103	$R^2 = 0.706$ MAE = 80.415 MSE=122571.887 RMSE = 350.103	$R^2 = 0.693$ MAE = 87.516 MSE=164017.447 RMSE = 404.990	$R^2 = 0.693$ MAE = 87.516 MSE=164017.447 RMSE = 404.990

The difference in results for XGBoost was also minimal for the Listwise Deletion and KNN Imputation techniques with IF Detection (less than 0.001 in R^2 and minimal variation in error), as the difference between optimal hyperparameters and the ones already used was small (only the number of estimators decreased, to 500 for both). LightGBM, Catboost and Decision Tree showed equal hyperparameters as when tested with reference, for the two techniques.

Hence, the decision of using reference hyperparameters based on the initial preprocessed dataset enabled the model to be tuned on a realistic baseline, by saving time and resources and maintained robustness.

The reference tuning achieved very close to optimal results and ensured that the results presented above are accurate and can be compared. This validation provides additional confidence to proceed with a comparative analysis of the techniques, knowing that they are optimized.

Models Comparison

Looking at Figures 8.5, 8.6 and 8.7, XGBoost demonstrated exceptional robustness and adaptability. Starting with a R^2 of approximately 0.858 after treatment of categorical outliers, the model showed a consistent improvement until reaching a peak of 0.88 after treating continuous numerical outliers using listwise deletion. What is particularly notable is its ability to maintain this high performance even when moving on to other options of numerical outlier treatments, where it only experiences a slight degradation. This stability is especially significant in the measures of error. XGBoost's MAE remained in between 86-106, while its RMSE within 225-308.

LightGBM performed similarly to XGBoost, but consistently inferior across all metrics. With R^2 ranging from 0.79-0.84, the model followed a similar pattern of improvement up to cyclical outliers, followed by stabilization or a slight decline across numerical outliers.

CatBoost exhibits paradoxical behavior. Traditionally known for its robustness to categorical outliers [97], it showed a more erratic performance than its gradient boosting competitors. R^2 varies significantly between 0.76-0.82, and more crucially, the model showed marked degradation when moving towards numerical outliers. This instability is particularly evident in MAE and RMSE, where CatBoost exhibits some of the largest variations between treatments.

Decision Tree served as an excellent control case in this analysis, demonstrating exactly the problems that ensemble methods seek to solve. Starting with a R^2 of approximately 0.69, it jumps to 0.76 for the cyclic outliers, only to return to 0.69-0.76 during numerical outliers treatment. Volatility is so extreme that it ends up showing the natural instability of basic decision trees with respect to change in data. The model may be memorizing patterns within each dataset, and failing to learn generalizable patterns.

With a few exceptions, the degradation trend when treating numerical outliers is consistent across models, but varies in magnitude. This would suggest that the treatment of numerical outliers may be overly aggressive. However, different factors and boundaries were tested to ensure important information would not be cut off. Nevertheless, some numerical outliers may be genuine observations at the extremes of the distribution, causing changes to cause lower overall performance. Other theory is the robustness of ensemble models, as explored in Section 5.4. XGBoost, LightGBM and CatBoost can be robust enough to handle naturally this type of extreme numerical outliers, and its treatment ruins it.

Detection IQR vs Detection IF

Looking at treatment pairs (same method, different detection) the IQR method was more stable and more predictable. The solid lines on the graphs showed more smoother gaps, less treatment variability, and overall improved performance. This reflects the built-in statistical character of IQR. It is an established method that captures outliers based on the natural data distribution. IF, capturing multidimensional outliers, had greater variability in results. The broken lines were prone to branching off from the continuous lines, especially towards the latter part of the treatment. This is an indication that IF was picking up on complex patterns rather than necessarily outliers in the traditional sense.

Imputation Method KNN vs Imputation Method MICE

Imputation methods showed mixed results, and made it difficult to draw conclusions. MICE generally outperforms KNN and there was a noticeable degradation for CatBoost with imputation. XGBoost maintains robustness even with imputed data, although errors slightly increase. However, this imputation methods bring a questionable benefit, since there is always potential introduction of artificial artifacts, and from that can be explained their overall worse performance.

Deleting vs Clipping

More preprocessing did not always lead to better performance. The optimal point lies in the treatment that reaches the final stage of the preprocessing path with the listwise deletion and IQR detection, using XGBoost. This treatment (made with a factor of 3) completely removes outlier observations, resulting in a clean dataset that allows XGBoost to achieve exceptional performance, and the lowest of all MAE and RMSE. Although this approach may lead to underrepresentation of some rare medical acts, it can be a strategic way to first focus on the most critical cases and later expand the analysis by incorporating more representative ones.

Nevertheless, for scenarios where preserving all observations is critical, Winsorization offered a performance close to listwise deletion method, where it maintains the dataset size and the robustness across different models, and not letting MAE and RMSE increase.

Generalization

With the optimal preprocessing pipeline established for these specific models, the capacity of the model to generalize was examined. Due to limited data, variations in types of medical exams from one Site to another and variations in the structure of the corresponding hospital workflows, the amount of data available to train each individual group became progressively limited. These facts prevented a compelling evaluation of the degree to which the model would generalize to other clinical settings or types of exams. Therefore, generalization was not prioritized at this stage of the work.

Nevertheless, although this was not on the scope of this project, it was important to acknowledge that the high performance achieved in the best approach and others may also reflect a degree of overfitting to the specific dataset, even after performing cross validations. Hence, an additional experiment was conducted to test cross-hospitals generalization by training the model on one and testing on the other and vice versa. Results showed that training on Site 2 and testing on Site 1 yielded better performance than the reverse. This finding is in line with previous observations, where Site 2 had better data quality and representativeness, which was likely helpful to the model's ability to generalize better when used as a source for training.

Thereby, although full generalization remains an open challenge, the cleaning pipeline developed in this work represents a significant step towards that goal, by improving data quality and consistency across different hospital units and exam types.

SHAP Value Analysis

Furthermore, after assessing the most promising treatments and its potential problems, it was made a study on feature importance of the model that showed better results (Figure 8.8). This analysis was important to understand how the requirements gathering with physicians has helped to give the model what he needs. Positive SHAP values indicate factors that contribute to closer (or even earlier) deliveries, while negative values indicate factors that contribute to delays.

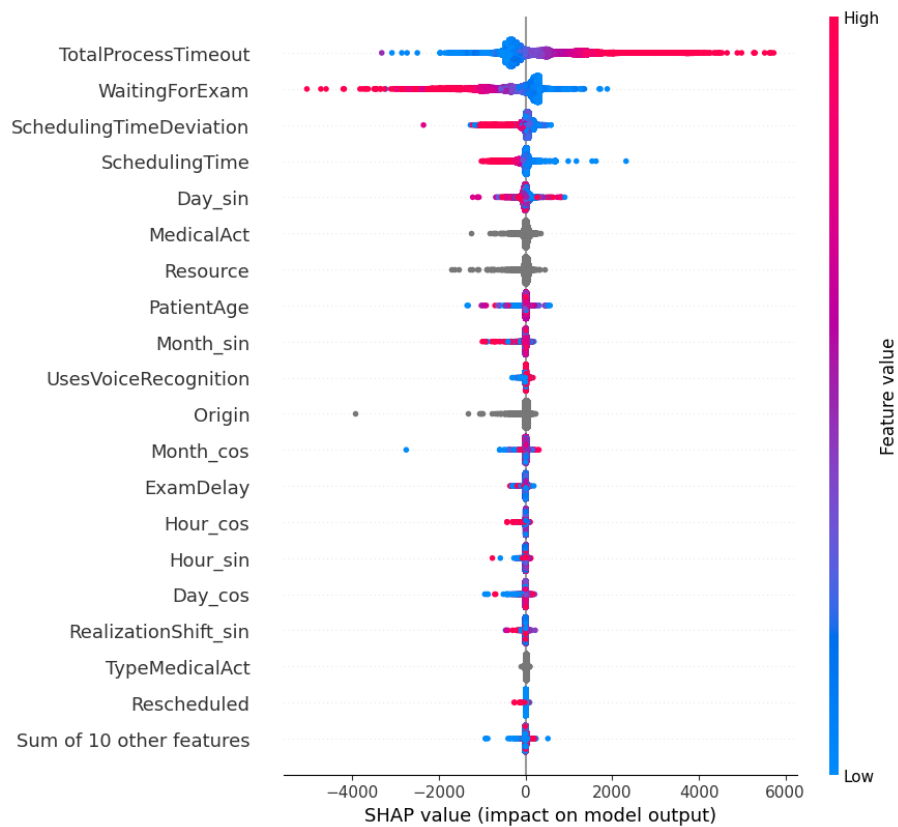


Figure 8.8: SHAP summary plot of XGBoost with non X-ray data of Site 2 treated with the best combination of treatments, to understand the impact of each feature on the predictions.

As it is seen at the top of the y axis, TotalProcessTimeout is the most critical predictor. Higher values of this feature (pink) contribute to earlier deliveries, suggesting that reports with due dates set far in the future tend to be delivered before the deadline, as expected. When the due date is close (blue), the report may be delayed or have a smaller delivery margin. WaitingForExam shows a worrying pattern: longer wait times (high pink values) consistently contribute to delayed submissions. This could indicate that a long wait for the exam also delays the issuance of the report. SchedulingTimeDeviation shows that the further the scheduling time deviates from the ideal of that tuple (origin, family), the greater the delay. SchedulingTime suggests that greater scheduling times significantly tend to deliver delayed reports. Looking at MedicalActs, certain exams or medical procedures are more prone to delays, perhaps due to their complexity. The same happens to resources and origin, perhaps due to availability. PatientAge reveals that some extremely low values are associated with delays, suggesting that younger patients might be given lower priority, which results in longer waiting times. Logically, and as expected, true values of UsesVoiceRecognition can help deliver the report sooner. As it was anticipated, delays in the realization of the exam, propagate to the report, and, rescheduled exams tend to delay more the delivery of the report, since the limit day to deliver is not changed accordingly.

When Hour_cos has high values (close to 1), SHAP values are negative. This indicates that certain specific times of the day are associated with a higher probability of late report delivery. Hour_cos is maximum around midnight. This is expected since midnight may have reduced staff availability. Hour_sin with low values represent late afternoon (around 6p.m.), and apparently there is a high probability of delays. This suggests what was already seen during preprocessing: during night shift, reporting exams are more likely to be delayed, probably because there is no longer enough time on the same day for validation.

Day_sin is maximum around Tuesdays and Wednesdays. High values of this variable appear with both positive and negative SHAP. This means that beginning of the week has a strong bidirectional impact on reporting deadlines, sometimes contributing to on-time deliveries and sometimes contributing to delays.

These days can have high operational variability due to weekend demand buildup. Day_cos is minimum around Thursdays and Fridays, and these low values are associated with negative SHAP values, indicating a tendency toward delays. This pattern may reflect a late-week workload peak, where accumulated tasks start to overload the reporting capacity, leading to more frequent delays in these days.

Month_sin is maximum between the month of April. If these high values generate negative SHAP, it suggests that medical acts carried out during these months tend to result in delayed reporting. This period comes after the peak of colds and other illnesses, which typically occur during the winter. The delays observed now may be due to increased demand in the previous months. Month_cos with low values also contribute to delays. This variable is minimum around July, suggesting that delays can also occur during this time. This period coincides with the holiday season, during which staffing levels are reduced and workload tends to accumulate.

SHAP values analysis is important to validate patterns. By comparing variables considered important for physicians (through the questionnaire) with those available in the datasets at the time of acquisition, there were some relevant variables that were not available but that can be included in a future work. These include: a proper variable that reflect the availability of the reporting physician or his number of pending reports at the time of realization; the specialty of the referring and reporting doctor; the patient clinical status and exam history; and the resource availability. The addition of these can help to have other perspective on this study.

Summary

This study presents promising insights into how different data preprocessing strategies can significantly impact the model performance, positively and negatively. The accuracy of the prediction of these delays depend on which techniques are used to prepare this data. Hereafter, these findings can also indirectly support more informed decision-making in medical scheduling, by allowing hospitals to better anticipate report delays and optimize appointment planning accordingly.

Chapter 9

Conclusions and Future Work

In this final chapter, conclusions on the work made are drawn, the limitations and difficulties found are outlined together with next steps for future work.

9.1 Final Considerations

The results presented throughout this work allowed to draw several key conclusions. Reflecting on the overall pipeline of treatment, there was an undeniable improvement in performance and error from baseline to last stage. This highlights the importance of careful preprocessing steps before predictive modeling and validates the importance of this experimental work.

Upon closer examination it was also notable that more data does not always indicate better model performance. In fact, what was found, for these specific datasets, was that cleaned data was always better than more data, and most times, complete outlier removal or clipping outperformed sophisticated preservation methods of imputation. The same happened when studying different detection methods: simple outperformed complex methods. IQR, a univariate method, consistently outperformed Isolation Forest, a multivariate method. When analyzing ML models, XGBoost's superiority was consistent across all scenarios. These results reflect what worked best for the specific datasets used in this project. Other hospitals may have different data characteristics, and what proved effective here should not be generalized without careful validation in other contexts.

Additionally, the results showed that the selection of carefully chosen features in collaboration with healthcare physicians added important value to construct the best dataset possible. This selection was later validated, by evaluating how the feature importance of the model corresponds to the actual physician's concerns collected by the questionnaire. Studying the importance of each variable to the outcome of the model has helped to understand and think about what other features have made sense to include in a future approach.

This work takes another step forward in understanding how can data quality be improved with diverse techniques, to help predict delays in reporting more accurately, and thereby reach our long-term goal of supporting optimal scheduling of medical appointments.

9.2 Limitations and Difficulties

This study faced several limitations that should be acknowledged. First, the timeframe of the project. Although the planning has always been completed on time, it was not possible to explore additional potential preprocessing methods, modeling approaches, optimizations, or generalization tests in as much depth as desired. With more time, broader experimentation and more comprehensive validation could

have been performed.

Furthermore, the range of available variables was limited to what was recorded in the datasets at the time of data extraction. Although efforts were made to align the features with physician feedback through a structured questionnaire, several relevant variables were not available, such as clinical condition of the patient, resource usage at the time of exam, or physician workload. These could have further enriched the model and improved predictive accuracy.

Regarding the models, no more extensive cross-validation or other robust techniques were implemented to ensure that the models were not overfitting the training data. This limitation may affect the generalizability and predictive performance of the models when applied to previously unseen real-world data. Although that was not the main scope of this project, in the future, more rigorous validation strategies, such as higher k-fold cross-validation, are recommended to mitigate this risk.

Lastly, data fragmentation and low representation of some medical acts made it difficult to ensure fair training and testing across all types of exams. This, combined with the constraints mentioned above, created challenges for both model robustness and generalization. Despite these difficulties, the study lays important groundwork and highlights key areas for improvement and future research.

9.3 Future work

Future work can build upon this study in several key directions. First, increasing the volume and diversity of data, by including more years and integrating datasets from additional hospital units or other specialties that also stood out from the questionnaire like cardiology, gastroenterology and clinical pathology. This can help to improve the representativeness of medical exams that are currently underrepresented and were considered as outliers, thus enabling more reliable model training and evaluation. As this was a preliminary study based on ByMe data, the amount of data used was intentionally limited to allow for simpler conclusions to be drawn, with the aim of progressively scaling up the analysis in future work. Furthermore, this broader dataset base would allow a more robust exploration of generalization, including the use of synthetic data or additional cross-hospital testing. This data increase would also be relevant to implement a strongest normalization strategy for medical acts, in order to ensure consistency across institutions.

From a modeling perspective, experimenting with non-tree-based ML models and alternative preprocessing strategies could uncover additional insights on different treatment needs and performance gains. Additionally, in the future, trying this approach with specific tuning hyperparameters for each technique, can help to better improve a fair choice between different limits within the same technique, even though the variations for this case were minimal.

Looking at what was lacking, besides all the important variables highlighted in the previous sections that were not included this time, time features that calculate the time between this and the previous exam, may be tested in future modeling pipelines due to their potential relevance in capturing clinical workflow patterns. This feature was initially created but due to missing dates values it had to be excluded in this work.

Due to time limitations, it is also important to address specific cases observed during this study, such as negative values in the TotalProcessTimeout variable, through a more refined error-handling approach, it may be possible to understand the root-causes behind these values and understand how can these affect the quality of data and consequently, the final result.

These are some suggestions for future work on this research topic. Nevertheless, the results and insights obtained already contribute valuable knowledge regarding robust preprocessing techniques in ML models.

References

- [1] Sandeep Dalal, Vadiraj Hombal, Wei Hung Weng, Gabe Mankovich, Thusitha Mabotuwana, Christopher S. Hall, Joseph Fuller, Bruce E. Lehnert, and Martin L. Gunn. Determining follow-up imaging study using radiology reports. *Journal of Digital Imaging*, 33:121–130, 2 2020.
- [2] Caroline E. Sloan, Seetharam C. Chadalavada, Tessa S. Cook, Curtis P. Langlotz, Mitchell D. Schnall, and Hanna M. Zafar. Assessment of follow-up completeness and notification preferences for imaging findings of possible cancer: What happens after radiologists submit their reports? *Academic Radiology*, 21:1579–1586, 12 2014.
- [3] Miguel Sousa Neves. Atrasos nos exames médicos estão a ter “consequências gravíssimas” na saúde dos utentes. *Jornal de Notícias*, 6 2012. Accessed: 2024-10-30.
- [4] Pedro Leite Silva and Rita Calisto. Registo oncológico nacional de todos os tumores na população residente em portugal, em 2020. *Registo Oncológico Nacional*, 11 2023. Accessed: 2024-10-30.
- [5] Ghazala Wahid, Ammara Haroon, Mehreen Samad, and Naila Tamkeen. Causes of delay in radiological reporting and ways to reduce them. *Journal of Saidu Medical College Swat*, 12:133–137, 9 2022.
- [6] SNS Digital. Sistemas de informação em saúde - conceito ecossistema de informação de saúde. *SNS*, 2017. Accessed: 2024-12-26.
- [7] SPMS. Preliminary version for public consultation - national strategy for the health information ecosystem. *SNS*, 2019. Accessed: 2024-12-26.
- [8] Maria Priestley, Fionntán O’Donnell, and Elena Simperl. A survey of data quality requirements that matter in ml development pipelines. *Journal of Data and Information Quality*, 15, 6 2023.
- [9] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. Methodologies for data quality assessment and improvement. *ACM Computing Surveys (CSUR)*, 41, 7 2009.
- [10] Richard Y. Wang. A product perspective on total data quality management. *Communications of the ACM*, 41:58–65, 2 1998.
- [11] Richard Y Wang and Diane M Strong. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12:5–33, 1996.
- [12] Manfred A. Jeusfeld, Christoph Quix, and Matthias Jarke. Design and analysis of quality information for data warehouses. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1507:349–362, 1998.
- [13] Craig E. Seko and Jennifer A. Long. A cyclic-hierarchical method for database data-quality evaluation and improvement. In *Information Quality*, pages 52–66. Routledge, 2005.
- [14] Leo L. Pipino, Yang W. Lee, and Richard Y. Wang. Data quality assessment. *Communications of the ACM*, 45:211–218, 4 2002.

- [15] Martin J. Eppler and Peter Muenzenmayer. Measuring information quality in the web context: A survey of state-of-the-art instruments and an application methodology. *MIT International Conference on Information Quality*, 2002.
- [16] Carlo Batini, Federico Cabitza, Cinzia Cappiello, and Chiara Francalanci. A comprehensive data quality methodology for web and structured data. *International Journal of Innovative Computing and Applications*, 1:205–218, 7 2008.
- [17] Carlo Batini and Monica Scannapieco. Methodologies for data quality measurement and improvement. In *Data Quality: Concepts, Methodologies and Techniques*, pages 161–200. Springer Berlin Heidelberg, 2006.
- [18] Steffen Huber, Hajo Wiemer, Dorothea Schneider, and Steffen Ihlenfeldt. Dmme: Data mining methodology for engineering applications – a holistic extension to the crisp-dm model. *Procedia CIRP*, 79:403–408, 1 2019.
- [19] Jairo Acosta Solano, Diana Janeth Lancheros Cuesta, Samir F. Umaña Ibáñez, and Jairo R. Coronado-Hernández. Predictive models assessment based on crisp-dm methodology for students performance in colombia - saber 11 test. *Procedia Computer Science*, 198:512–517, 1 2022.
- [20] D H Stone. Design a questionnaire. *British Medical Journal*, 307:1264–1266, 11 1993.
- [21] Sandhya Yaddanapudi and LN Yaddanapudi. How to design a questionnaire. *Indian Journal of Anaesthesia*, 63:335, 2019.
- [22] Gail M. Sullivan and Anthony R. Artino. Analyzing and interpreting data from likert-type scales. *Journal of Graduate Medical Education*, 5:541–542, 12 2013.
- [23] Roger Tourangeau, J. Michael Brick, Sharon Lohr, and Jane Li. Adaptive and responsive survey designs: A review and assessment. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 180:203–223, 1 2017.
- [24] Hafiz Muhammad Salman and Dr.Noreen Aleem. An easy way to understand the research tools, scales and parametric non-parametric tests. *SSRN Electronic Journal*, 10 2024.
- [25] O Alotaibi, S Tomy, and E Pardede. A framework for cleaning streaming data in healthcare: A context and user-supported approach. *Computers*, 13, 2024.
- [26] Sanju, Vinay Kumar, and Deepender. Evaluation of imputation techniques for genotypic data of soybean crop under missing completely at random mechanism. *Indian Journal of Agricultural Research*, 57:701–705, 2023.
- [27] Daniel A. Newman. Missing data: Five practical guidelines. *Organizational Research Methods*, 17:372–411, 10 2014.
- [28] Tra My Pham, Nikolaos Pandis, and Ian R. White. Missing data: Issues, concepts, methods. *Seminars in Orthodontics*, 30:37–44, 2 2024.
- [29] Kay Liu, Yingdong Dou, Yue Zhao, Xueying Ding, Xiyang Hu, Ruitong Zhang, Kaize Ding, Canyu Chen, Hao Peng, Kai Shu, George H. Chen, Zhihao Jia, and Philip S. Yu. Pygod: A python library for graph outlier detection. *Journal of Machine Learning Research* 25 1-9, 4 2024.
- [30] Yue Zhao, Zain Nasrullah, and Zheng Li. Pyod: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research*, 20, 1 2019.
- [31] Zheng Li, Yue Zhao, Nicola Botta, Cezar Ionescu, and Xiyang Hu. Copod: Copula-based outlier detection. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2020-November:1118–1123, 9 2020.

- [32] Uwe Draisbach and Felix Naumann. A generalization of blocking and windowing algorithms for duplicate detection. *Proceedings - 2011 International Conference on Data and Knowledge Engineering, ICDKE 2011*, pages 18–24, 2011.
- [33] Eugene S. Valeriano. Deduplication methods using levenshtein distance algorithm. *Journal of Electrical Systems*, 20:997–1006, 5 2024.
- [34] W. Dean Bidgood, Steven C. Horii, Fred W. Prior, and Donald E. Van Syckle. Understanding and using dicom, the data interchange standard for biomedical imaging. *Journal of the American Medical Informatics Association*, 4:199–212, 5 1997.
- [35] Sara Khalid, Cynthia Yang, Clair Blacketer, Talita Duarte-Salles, Sergio Fernández-Bertolín, Chungsoo Kim, Rae Woong Park, Jimyung Park, Martijn J. Schuemie, Anthony G. Sena, Marc A. Suchard, Seng Chan You, Peter R. Rijnbeek, and Jenna M. Reips. A standardized analytics pipeline for reliable and rapid development and validation of prediction models using observational health data. *Computer Methods and Programs in Biomedicine*, 211:106394, 11 2021.
- [36] Sadhna B. Nandwana, Gail Walls, and Susan Reich. Learning from experience: "minimizing patient delays in radiology: Optimizing on-time starts for ct procedures". *Current problems in diagnostic radiology*, 50:11–15, 1 2021.
- [37] M Afkanpour, E Hosseinzadeh, and H Tabesh. Identify the most appropriate imputation method for handling missing values in clinical structured datasets: a systematic review. *BMC Medical Research Methodology*, 24, 2024.
- [38] V Jain, S Shukla, and N Khare. Analysis of various data imputation techniques for diabetes classification on pima dataset. In *2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science, SCEECS 2024*, 2024.
- [39] M S Santos, J P Soares, P H Abreu, H Araújo, and J Santos. Influence of data distribution in missing data imputation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10259 LNAI, pages 285–294, 2017.
- [40] K M Varma, S Nayak, and P M Khilar. Imice: An improved missing data imputation using machine learning. In *Lecture Notes in Networks and Systems*, volume 873, pages 159–167, 2024.
- [41] Fabio Nelli. Python data analytics: With pandas, numpy, and matplotlib: Second edition. *Python Data Analytics: With Pandas, NumPy, and Matplotlib: Second Edition*, pages 1–569, 9 2018.
- [42] Saleh Rehiel Alenazi, Kamsuriah Ahmad, and Akeem Olowolayemo. A review of similarity measurement for record duplication detection. *Proceedings of the 2017 6th International Conference on Electrical Engineering and Informatics: Sustainable Society Through Digital Innovation, ICEEI 2017*, 2017-November:1–6, 7 2017.
- [43] Ralf Mikut and Markus Reischl. Data mining tools. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1:431–443, 9 2011.
- [44] Sakda Loetpipatwanich and Preecha Vichitthamaros. Sakdas: A python package for data profiling and data quality auditing. *2020 1st International Conference on Big Data Analytics and Practices, IBDAP 2020*, 9 2020.
- [45] Fabian Pedregosa, Vincent Michel, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Jake Vanderplas, David Cournapeau, Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Bertrand Thirion, Olivier Grisel, Vincent Dubourg, Alexandre Passos, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 11 2011.

- [46] Venkata Sai Venkatesh Pulla, Cihan Varo, and Murat Al. Open source data quality tools: Revisited. *Advances in Intelligent Systems and Computing*, 448:893–902, 2016.
- [47] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, 17:261–272, 3 2020.
- [48] Michele A. Brandão, Arthur P.G. Reis, Bárbara M.A. Mendes, Clara A.Bacha De Almeida, Gabriel P. Oliveira, Henrique Hott, Larissa D. Gomide, Lucas L. Costa, Mariana O. Silva, Anisio Lacerda, and Gisele L. Pappa. Plus: A semi-automated pipeline for fraud detection in public bids. *Digital Government: Research and Practice*, 5, 3 2024.
- [49] Caroline J. Sands, Arnaud M. Wolfer, Gonçalo D.S. Correia, Nouredin Sadawi, Arfan Ahmed, Beatriz Jiménez, Matthew R. Lewis, Robert C. Glen, Jeremy K. Nicholson, and Jake T.M. Pearce. The npyc-toolbox, a python module for the pre-processing, quality-control and analysis of metabolic profiling datasets. *Bioinformatics (Oxford, England)*, 35:5359–5360, 12 2019.
- [50] Sicong Huang, Roozbeh Jafari, and Bobak J. Mortazavi. Pulse2ai: An adaptive framework to standardize and process pulsatile wearable sensor data for clinical applications. *IEEE Open Journal of Engineering in Medicine and Biology*, 5:330 – 338, 1 2024.
- [51] Emily Caveness, G. C. Paul Suganthan, Zhuo Peng, Neoklis Polyzotis, Sudip Roy, and Martin Zinkevich. Tensorflow data validation: Data analysis and validation in continuous ml pipelines. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 2793–2796, 6 2020.
- [52] Marta Moreno, Ricardo Vilaça, and Pedro G. Ferreira. Scalable transcriptomics analysis with dask: applications in data science and machine learning. *BMC Bioinformatics*, 23:1–20, 12 2022.
- [53] Ruben Verborgh and Max De Wilde. *Using OpenRefine: The essential OpenRefine guide that takes you from data analysis and error fixing to linking your dataset to the Web*. Packt Publishing, 9 2013.
- [54] Hourieh Khalajzadeh, Mohamed Abdelrazek, John Grundy, John Hosking, and Qiang He. Survey and analysis of current end-user data analytics tool support. *IEEE Transactions on Big Data*, 8:152–165, 2 2022.
- [55] Eline R. Tsai, Derya Demirtas, Nick Hoogendijk, Andrei N. Tintu, and Richard J. Boucherie. Turnaround time prediction for clinical chemistry samples using machine learning. *Clinical chemistry and laboratory medicine*, 60:1902–1910, 11 2022.
- [56] Marcia L. Baptista, Sahil Panse, and Bruno F. Santos. Revision and implementation of metrics to evaluate the performance of prognostics models. *Measurement*, 236:115038, 8 2024.
- [57] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Neural Information Processing Systems*, 2017.
- [58] Welcome to the shap documentation. *Shap library’s documentation*, 2018. Accessed: 2024-12-13.
- [59] Salih Sarp, Murat Knzlu, Umit Cali, Onur Elma, and Ozgur Guler. An interpretable solar photovoltaic power generation forecasting approach using an explainable artificial intelligence tool. *2021 IEEE Power and Energy Society Innovative Smart Grid Technologies Conference, ISGT 2021*, 2 2021.
- [60] Welcome to the eli5 documentation. *ELI5 library’s documentation*, 2017. Accessed: 2024-12-14.
- [61] Pooja Thakkar. Drug classification using black-box models and interpretability. *International Journal for Research in Applied Science and Engineering Technology*, 9:1518–1529, 9 2021.

- [62] Nesaretnam Barr Kumarakulasinghe, Tobias Blomberg, Jintai Liu, Alexandra Saraiva Leao, and Panagiotis Papapetrou. Evaluating local interpretable model-agnostic explanations on clinical machine learning classification models. *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, 2020-July:7–12, 7 2020.
- [63] Christoph Molnar, Timo Freiesleben, Gunnar König, Giuseppe Casalicchio, Marvin N. Wright, and Bernd Bischl. Relating the partial dependence plot and permutation feature importance to the data generating process. *Communications in Computer and Information Science*, 1901 CCIS:456–479, 9 2021.
- [64] David I Inouye, Liu Leqi, Joon Sik Kim, Bryon Aragam, and Pradeep Ravikumar. Automated dependence plots. *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, pages 1238–1247, 8 2020.
- [65] Sandeep Rangineni. An analysis of data quality requirements for machine learning development pipelines frameworks. *International Journal of Computer Trends and Technology*, 71:16–27, 2023.
- [66] Harald Foidl, Valentina Golendukhina, Rudolf Ramler, and Michael Felderer. Data pipeline quality: Influencing factors, root causes of data-related issues, and processing problem areas for developers. *Journal of Systems and Software*, 207:111855, 1 2024.
- [67] Fahad Pervaiz, Aditya Vashistha, and Richard Anderson. Examining the challenges in development data pipeline. *COMPASS 2019 - Proceedings of the 2019 Conference on Computing and Sustainable Societies*, pages 13–21, 7 2019.
- [68] Anna Breger, Ian Selby, Michael Roberts, Judith Babar, Effrossyni Gkrania-Klotsas, Jacobus Preller, Lorena Escudero Sánchez, Sören Dittmer, Matthew Thorpe, Julian Gilbey, Anna Korhonen, Emily Jefferson, Georg Langs, Guang Yang, Xiaodan Xing, Yang Nan, Ming Li, Helmut Prosch, Jan Stanczuk, Jing Tang, Philip Teare, Mishal Patel, Marcel Wassink, Markus Holzer, Eduardo González Solares, Nicholas Walton, Pietro Liò, Tolou Shadbahr, James H.F. Rudd, John A.D. Aston, Jonathan R. Weir-McCall, Evis Sala, and Carola Bibiane Schönlieb. A pipeline to further enhance quality, integrity and reusability of the nccid clinical data. *Scientific data*, 10, 12 2023.
- [69] J. Ainsworth and I. Buchan. Combining health data uses to ignite health system learning. *Methods of information in medicine*, 54:479–487, 2015.
- [70] Stefan Förstel, Markus Förstel, Markus Gallistl, Dario Zanca, Bjoern M. Eskofier, and Eva M. Rothgang. Data quality in hospital information systems: Lessons learned from analyzing 30 years of patient data in a regional german hospital. *International Journal of Medical Informatics*, 192:105636, 12 2024.
- [71] Aiswarya Raj Munappy, Jan Bosch, and Helena Homström Olsson. Data pipeline management in practice: Challenges and opportunities. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12562 LNCS:168–184, 2020.
- [72] Federico Croce, Riccardo Valentini, Marianna Maranghi, Giorgio Grani, Maurizio Lenzerini, and Riccardo Rosati. Ontology-based data preparation in healthcare: The case of the amd-stitch project. *SN Computer Science*, 5, 2024.
- [73] George Manias, Eleftheria Kouremenou, Ainhoa Azqueta Alzúaz, Pavlos Kranas, Fabio Melillo, and Dimosthenis Kyriazis. An optimized pipeline for the processing of healthcare data towards the creation of holistic health records. In *Proceedings - 2023 International Conference on Applied Mathematics and Computer Science, ICAMCS 2023*, pages 50–56, 2023.

- [74] Jason Brownlee. *Data Preparation for Machine Learning - Data Cleaning, Feature Selection, and Data Transforms in Python*. Machine Learning Mastery, 2020.
- [75] Salvador García, Sergio Ramírez-Gallego, Julián Luengo, José Manuel Benítez, and Francisco Herrera. Big data preprocessing: methods and prospects. *Big Data Analytics*, 1, 12 2016.
- [76] Institute of Medicine (US) Committee on Quality of Health Care in America. Crossing the quality chasm: A new health system for the 21st century. 7 2001.
- [77] Vaibhavi Shah, Yeshwant R. Chillakuru, Alex Rybkin, Youngho Seo, Thienkhai Vu, and Jae Ho Sohn. Algorithmic prediction of delayed radiology turn-around-time during non-business hours. *Academic Radiology*, 29:e82–e90, 5 2022.
- [78] Tasquia Mizan and Sharareh Taghipour. Medical resource allocation planning by integrating machine learning and optimization models. *Artificial Intelligence in Medicine*, 134:102430, 12 2022.
- [79] John Mark Michael Rumbold and Barbara Pierscionek. The effect of the general data protection regulation on medical research. *Journal of Medical Internet Research*, 19:e7108, 2 2017.
- [80] John Robinson. Likert scale. *Encyclopedia of Quality of Life and Well-Being Research*, pages 3917–3918, 2023.
- [81] John T. Kulas, Alicia A. Stachowski, and Brad A. Haynes. Middle response functioning in likert-responses to personality items. *Journal of Business and Psychology*, 22:251–259, 3 2008.
- [82] Borim Ryu, Sooyoung Yoo, Seok Kim, and Jinwook Choi. Thirty-day hospital readmission prediction model based on common data model with weather and air quality data. *Scientific Reports*, 11:1–9, 12 2021.
- [83] Leo Grinsztajn, Edouard Oyallon, and Gael Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, 35:507–520, 12 2022.
- [84] Lingjian Yang, Songsong Liu, Sophia Tsoka, and Lazaros G. Papageorgiou. A regression tree approach using mathematical programming. *Expert Systems with Applications*, 78:347–357, 7 2017.
- [85] Guang yao Zhao, Kenji Ohsu, Henry Kasmanhadi Saputra, Teruhisa Okada, Jumpei Suzuki, Yuji Kuwahara, and Masafumi Fujita. Enhancing interpretability of tree-based models for downstream salinity prediction: Decomposing feature importance using the shapley additive explanation approach. *Results in Engineering*, 23:102373, 9 2024.
- [86] Mateusz Kalisch, Marcin Michalak, Marek Sikora, Łukasz Wróbel, and Piotr Przyszałka. Influence of outliers introduction on predictive models quality. *Communications in Computer and Information Science*, 613:79–93, 2016.
- [87] Cynthia P. Haanappel and Anne F. Voor in ‘t holt. Using the interquartile range in infection prevention and control research. *Infection Prevention in Practice*, 6:100337, 3 2024.
- [88] Stephen Adjei, Deon Yeboah Takyi-Amponsem, Jonathan Atuquaye Quaye, Yen Adams Sokama-Neuyam, John Ojuu Oleka, George Kobla Asigbi, and Turkson Kwesi Duodu. Automated hydraulic flow unit determination using fzi-z-score probability, k-means clustering with autoencoders, and machine learning classification: A case study of the subei basin, china. *Journal of Applied Geophysics*, 240:105778, 9 2025.
- [89] Véronne Yepmo, Grégory Smits, Marie Jeanne Lesot, and Olivier Pivert. Leveraging an isolation forest to anomaly detection and data clustering. *Data Knowledge Engineering*, 151:102302, 5 2024.

- [90] Ch Sanjeev Kumar Dash, Ajit Kumar Behera, Satchidananda Dehuri, and Ashish Ghosh. An outliers detection and elimination framework in classification task of data mining. *Decision Analytics Journal*, 6:100164, 3 2023.
- [91] Abdullah Z. Alruhaymi, Charles J. Kim, Abdullah Z. Alruhaymi, and Charles J. Kim. Why can multiple imputations and how (mice) algorithm work? *Open Journal of Statistics*, 11:759–777, 9 2021.
- [92] Kimseth Seu, Mi Sun Kang, and Hwamin Lee. An intelligent missing data imputation techniques: A review. *JOIV : International Journal on Informatics Visualization*, 6:278–283, 5 2022.
- [93] Kelsy Cabello-Solorzano, Isabela Ortigosa de Araujo, Marco Peña, Luís Correia, and Antonio J. Tallón-Ballesteros. The impact of data normalization on the accuracy of machine learning algorithms: A comparative analysis. *Lecture Notes in Networks and Systems*, 750 LNNS:344–353, 2023.
- [94] Shichao Zhang. Nearest neighbor selection for iteratively knn imputation. *Journal of Systems and Software*, 85:2541–2552, 11 2012.
- [95] I Nyoman Switrayana, Rifqi Hammad, Pahrul Irfan, Tomi Tri Sujaka, and Muhammad Haris Nasri. Comparative analysis of stock price prediction using deep learning with data scaling method. *JTIM : Jurnal Teknologi Informasi dan Multimedia*, 7:78–90, 1 2025.
- [96] Rakhi Wajgi, Himanshu Agarkar, Rohan Patil, Harshit Rao, and Nipun Petkar. Enhancing credit card transaction fraud detection with random forest and robust scaling. *AIP Conference Proceedings*, 3188, 12 2024.
- [97] Ahmad Cahyono Adi, Aditya Ahmad Zein, and Ibnu Amri Thaher. Investigating anomalies in diabetes time-series data using boosting algorithms family: A comparative study of ensemble approaches. *7th International Seminar on Research of Information Technology and Intelligent Systems: Advanced Intelligent Systems in Contemporary Society, ISRITI 2024 - Proceedings*, pages 1012–1017, 2024.

Appendix A

Questionnaire - Factors Weighting Section

The following figures illustrate relevant questions made in the questionnaire used in this study. The subsequent figures are organized to reflect the main section of this questionnaire: factors weighting section.

What departments do people request most medical procedures from? *

- ☐ Radiology
- ☐ Neuroradiology
- ☐ Interventional Radiology
- ☐ Nuclear Medicine
- ☐ Neurophysiology
- ☐ Neurosonology
- ☐ Cardiology
- ☐ Gastroenterology
- ☐ Bronchology
- ☐ Respiratory Functional Tests
- ☐ Sleep Study
- ☐ Dermatology
- ☐ Endocrinology
- ☐ Immunoallergology
- ☐ Rheumatology
- ☐ Urology
- ☐ Genetics
- ☐ Clinical Pathology
- ☐ Pathological Anatomy
- ☐ Other: _____

Figure A.1: Introductory question for Factors Weighting section. This question is made to reporting-prescribing physicians and prescribing physicians.

What is your experience with delays in delivery of exam reports? *

- ☐ Frequent
- ☐ Occasional
- ☐ Rare
- ☐ I have never experienced delays

Figure A.2: Introductory question for Factors Weighting section. This question is made to the three types of physicians.

What are your reporting areas? *

- ☐ Radiology
- ☐ Neuroradiology
- ☐ Interventional Radiology
- ☐ Nuclear Medicine
- ☐ Neurophysiology
- ☐ Neurosonology
- ☐ Cardiology
- ☐ Gastroenterology
- ☐ Bronchology
- ☐ Respiratory Functional Tests
- ☐ Sleep Study
- ☐ Dermatology
- ☐ Endocrinology
- ☐ Immunoallergology
- ☐ Rheumatology
- ☐ Urology
- ☐ Genetics
- ☐ Clinical Pathology
- ☐ Pathological Anatomy
- ☐ Other: _____

Figure A.3: Introductory question for Factors Weighting section on Reporting Areas. This question is made to reporting-prescribing physicians and reporting physicians.

How many years have you worked as a reporting physician? *

- ☐ I'm still an internist
- ☐ Less than 5 years
- ☐ 5-10 years
- ☐ 11-20 years
- ☐ More than 20 years

Figure A.4: Introductory question for Factors Weighting section on years of work. This question is made to reporting-prescribing physicians and reporting physicians.

What would you say is your working methodology when it comes to writing reports?

- ☐ Oriented towards the date of the next appointment
- ☐ Oriented to the workload of the day
- ☐ Report priority oriented
- ☐ Guided by the date requested by the prescribing physician
- ☐ Other: _____

Figure A.5: Introductory question for Factors Weighting section on work methodology. This question is made to reporting-prescribing physicians and reporting physicians.

The subsequent figures are organized to reflect the tables that were made to evaluate each category.

As a reporting physician, with regard to your work, how do you rate the relevance of the following factors in predicting the turnaround time for exam reports? (Rate each item from 1 to 4, where 1 = Not at All Relevant and 4 = Extremely Relevant, or N/A)

	1	2	3	4	N/A
Specialization of the reporting physician vs Type of examination: Reporting an examination from a specialty other than our own can bring more complexity to the reporting process.	☐	☐	☐	☐	☐
Availability of the reporting physician: The workload or availability of the physician (shifts and days off) may affect the time it takes to issue the report.	☐	☐	☐	☐	☐
Age of the reporting physician: Younger or older physicians may have different levels of experience when it comes to interpreting.	☐	☐	☐	☐	☐
Number of pending reports per physician: A physician with many pending reports may take longer to issue new reports.	☐	☐	☐	☐	☐
Time of request for examination report: A request for a report made at the end of the day may influence the performance of the report.	☐	☐	☐	☐	☐

Figure A.6: Table made for evaluating reporting physician category. Here, factors that influence the physician work as a reporting physician are outstanced.

With regard to the exam, how do you rate the relevance of the following factors in predicting the turnaround time for exam reports? (Rate each item from 1 to 4, where 1 = Not at All Relevant and 4 = Extremely Relevant, or N/A)

	1	2	3	4	N/A
Exam Priority: Some exams may have a higher priority (urgent vs routine), which may affect report turnaround time.	☐	☐	☐	☐	☐
Exam specialty (e.g., imaging): Some specialties may require more time to analyze and report an exam.	☐	☐	☐	☐	☐
Type of examination (e.g. X-ray): Different types of examinations may have varying interpretation times.	☐	☐	☐	☐	☐
Body part examined: Different body parts may have varying interpretation times.	☐	☐	☐	☐	☐
Number of elements to be analyzed (Ex: Images): Exams with more images may take longer to be interpreted.	☐	☐	☐	☐	☐
Image complexity: Classifying complexity (if possible) can help predict the interpretation time required.	☐	☐	☐	☐	☐

Figure A.7: Table made for evaluating exam category. Here, both reporting-prescribing and prescribing physicians evaluate the relevance of exam factors concerning their work.

With regard to the type of patient, how do you rate the relevance of the following *
factors in predicting the turnaround time for exam reports? (Rate each item from 1
to 4, where 1 = Not at All Relevant and 4 = Extremely Relevant, or N/A)

	1	2	3	4	N/A
Patient age: Younger or older patients may have special needs, which can impact the complexity of interpretation.	☐	☐	☐	☐	☐
Patient gender: Some conditions or types of examination may vary between genders.	☐	☐	☐	☐	☐
Number of exams previously performed by the patient (history): Patients who have already performed several exams may need more time for comparative analysis.	☐	☐	☐	☐	☐
Next appointment date: The proximity of the appointment date may impact the priority of issuing the report.	☐	☐	☐	☐	☐
Clinical status (reason for request): The patient's condition may indicate more or less urgency in the examination report.	☐	☐	☐	☐	☐

Figure A.8: Table made for evaluating patient category. Here, the three types of physicians evaluate the relevance of patient factors concerning their work.

As a reporting physician, with regard to resources, how do you rate the relevance of the following factors in predicting the turnaround time for exam reports? (Rate each item from 1 to 4, where 1 = Not at All Relevant and 4 = Extremely Relevant, or N/A) *

	1	2	3	4	N/A
Unit where the exam was performed	☐	☐	☐	☐	☐
Origin of the patient visit (e.g. emergency, outpatient clinic)	☐	☐	☐	☐	☐
Exam room availability	☐	☐	☐	☐	☐
Technical problems in the information system: Some platforms may be more or less efficient.	☐	☐	☐	☐	☐
Availability of speech recognition and transcription tools: this tool helps in producing these reports and can therefore impact delivery time.	☐	☐	☐	☐	☐
Availability of artificial intelligence tools (for example, image processing): these tools assist in producing these reports and can therefore impact delivery time.	☐	☐	☐	☐	☐

Figure A.9: Table made for evaluating resources category. Here, both reporting-prescribing and reporting physicians evaluate the relevance of resources concerning their work.

As a prescribing physician, with regard to resources, how do you rate the relevance of the following factors in predicting the turnaround time for exam reports? (Rate each item from 1 to 4, where 1 = Not at All Relevant and 4 = Extremely Relevant, or N/A) *

	1	2	3	4	N/A
Unit where the exam was performed	☐	☐	☐	☐	☐
Origin of the patient visit (e.g. emergency, outpatient clinic)	☐	☐	☐	☐	☐
Exam room availability	☐	☐	☐	☐	☐
Technical problems in the information system: Some platforms may be more or less efficient.	☐	☐	☐	☐	☐

Figure A.10: Table made for evaluating resources category. Here, prescribing physicians evaluate the relevance of resources concerning their work.

Appendix B

Group-Weighted Questionnaire Results

Table B.1: Quantitative analysis of the importance physicians attributed to factors related to the reporting physician.

Variable	Weight Average: RP Physicians	Weight Average: R Physicians	WA
RepPhysicianSpecialization	3.8333	4	3.875
RepPhysicianAvailability	4	4	4
RepPhysicianAge	3.3333	4	3.5
RepPhysicianPendingReports	3.6667	3	3.5
RepPhysicianRequestTime	3	3	3

Table B.2: Quantitative analysis of the importance physicians attributed to factors related to the exam.

Variable	Weight Average: RP Physicians	Weight Average: P Physicians	WA
ExamPriority	4	3.875	3.9
ExamSpecialty	3.1667	3.2083	3.2
ExamType	3.5	3.6087	3.587
ExamBodyPart	3.5	2.8333	2.9667
ExamElementsToAnalyze	3.1667	3.4583	3.4
ExamComplexity	3.3333	3.4783	3.4493

Table B.3: Quantitative analysis of the importance physicians attributed to factors related to the patient.

Variable	Weight Average: RP Physicians	Weight Average: P Physicians	Weight Average: R Physicians	WA
PatientAge	3.5	2.5217	4	2.7976
PatientGender	2.6	1.8696	3	2.0772
PatientExamHistory	3.1667	3.0455	4	3.1278
PatientNextAppointment	3.5	3.5217	3.5	3.5163
PatientClinicalStatus	3.8333	3.8636	4	3.8665

Table B.4: Quantitative analysis of the importance physicians attributed to factors related to resources.

Variable	Weight Average: RP Physicians	Weight Average: P Physicians	Weight Average: R Physicians	WA
ResourceUnit	3.8333	2.9091	3.5	3.1193
ResourceRequestOrigin	3.8333	3.75	4	3.7813
ResourceExamRoomAvailability	3.3333	3.0909	4	3.1932
ResourceTechnicalIssues	4	3.4091	4	3.5568
ResourceVoiceRecognitionTools	3.6	-	4	3.7
ResourceAI_Tools	2.8	-	3	2.85

Table B.5: Quantitative analysis of the importance physicians attributed to factors related to time.

Variable	Weight Average: RP Physicians	Weight Average: R Physicians	WA
TemporalHospitalDemand	2.8333	4	3.125
TemporalReportShift	3	4	3.25
TemporalDayOfWeek	3.8333	4	3.875
TemporalMonth	3	4	3.25

Appendix C

Medical Acts - Complete List

Table C.1: Counts of medical acts of Non X-ray radiological data of Site 1, before categorical outlier removal.

Medical Act	Count	Percentage
Brain MRI (Cranial-Encephalic)	17491	6.9616%
Brain CT	15970	6.3562%
Chest CT	12346	4.9139%
Chest-Abdominal-Pelvic CT	12263	4.8808%
Abdominal Ultrasound	11780	4.6886%
Renal-Bladder Ultrasound	9552	3.8018%
Bilateral Breast Ultrasound	9242	3.6784%
Abdominal-Pelvic CT	8792	3.4993%
Lumbar Spine MRI	7783	3.0977%
Cervical-Thoracic-Abdominal-Pelvic CT	5838	2.3236%
Mammography 4 Views	5805	2.3105%
High Resolution Chest CT	5795	2.3065%
Thyroid Ultrasound	5686	2.2631%
Abdominal-Pelvic CT Angiography	5177	2.0605%
Cervical Ultrasound	4967	1.9769%
Cervical Spine MRI	4640	1.8468%
Abdominal and Reno-Pelvic Ultrasound	4563	1.8161%
Chest-Abdominal CT	3993	1.5893%
Soft Tissue Ultrasound	3639	1.4484%
Upper Abdominal MRI	3321	1.3218%
Thoracic CT Angiography	3294	1.3110%
Axillary Ultrasound	3031	1.2064%
Pelvic MRI	2865	1.1403%
MR Cholangiopancreatography	2799	1.1140%
Lumbar Spine CT	2381	0.9477%
Abdominal and Renal Ultrasound	1764	0.7021%
Thoracic Spine MRI	1734	0.6902%
Breast MRI	1696	0.6750%
Abdominal-Pelvic Ultrasound	1599	0.6364%
Right Knee MRI	1517	0.6038%
Bladder + Prostatic Ultrasound	1473	0.5863%
Paranasal Sinuses CT	1425	0.5672%
SA Trunks + Circle of Willis CT Angiography	1414	0.5628%
Left Knee MRI	1399	0.5568%
Ultrasound Guided Core Biopsy	1388	0.5524%
Renal Transplant Ultrasound	1343	0.5345%
Ultrasound Guided Neck Biopsy	1305	0.5194%
Cardiac CT Angiography	1305	0.5194%
Abdominal MR Angiography	1299	0.5170%
Liver MRI	1215	0.4836%

Cardiac MRI	1190	0.4736%
Cervical Spine CT	1185	0.4716%
Prostatic Ultrasound	1176	0.4681%
Abdominal CT Angiography	1158	0.4609%
Cerebral MR Angiography	1151	0.4581%
Pituitary MRI	1096	0.4362%
Right Shoulder MRI	1095	0.4358%
Renal Ultrasound	1091	0.4342%
Prostate MRI	1043	0.4151%
CT Guided Thoracic Biopsy	1002	0.3988%
Upper Abdominal CT	933	0.3713%
Hip Ultrasound	926	0.3686%
Face CT	895	0.3562%
Joint Ultrasound	878	0.3495%
Orbital MRI	866	0.3447%
Cervical-Thoracic CT	854	0.3399%
Left Shoulder MRI	788	0.3136%
Pre-operative Marking Mammography	786	0.3128%
Scrotal Ultrasound	757	0.3013%
Thoracic Spine CT	753	0.2997%
Entero-CT	751	0.2989%
Thoracic-Abdominal CT Angiography	714	0.2842%
Central Nervous System Ultrasound (Pediatric)	710	0.2826%
Ear MRI (Internal Auditory Canals)	702	0.2794%
Pre-operative Localization with Clip (Titanium)	682	0.2714%
Renal Transplant Doppler Ultrasound	665	0.2647%
Encephalic, Cervical and Facial Angiography	661	0.2631%
Cervical CT	654	0.2603%
Ear CT	652	0.2595%
TAVI CT	614	0.2444%
Pelvic CT	606	0.2412%
Sacroiliac Joint MRI	604	0.2404%
Renal Doppler Ultrasound	549	0.2185%
Cervical-Thoracic-Abdominal CT	538	0.2141%
Inguinal Ultrasound	531	0.2113%
Lower Limb CT Angiography	525	0.2090%
Paranasal Sinuses CT	482	0.1918%
Ultrasound Guided Abdominal Biopsy	462	0.1839%
Left Foot MRI	459	0.1827%
Renal CT	448	0.1783%
TMJ MRI (Temporomandibular Joint)	429	0.1707%
Right Foot MRI	426	0.1696%
Pelvic Ultrasound	421	0.1676%
Brain MRI + MR Angiography	395	0.1572%
Unilateral Mammography	391	0.1556%
Abdominal + Pelvic MRI	388	0.1544%
Ultrasound Guided Abdominal Drainage	386	0.1536%
Pelvic CT	378	0.1504%
Right Hip MRI	358	0.1425%
Face MRI	349	0.1389%
Neck CT	344	0.1369%
Pelvis MRI	344	0.1369%
Pelvic CT Angiography	342	0.1361%
Thoracic Ultrasound	338	0.1345%
Neck MRI	336	0.1337%
Adrenal CT	331	0.1317%
Renal MRI	329	0.1309%
Left Hip MRI	305	0.1214%
Carotid Doppler Ultrasound	302	0.1202%
Left Ankle MRI	301	0.1198%
Right Hand MRI	285	0.1134%

Orbital CT	283	0.1126%
Right Ankle MRI	282	0.1122%
Screening Mammography	278	0.1106%
Abdominal Doppler Ultrasound	269	0.1071%
Hip Joint CT	263	0.1047%
Ultrasound Guided Core Biopsy	260	0.1035%
Salivary Gland Ultrasound	252	0.1003%
Neck CT Angiography	251	0.0999%
Adrenal MRI	250	0.0995%
Penile Doppler Ultrasound	238	0.0947%
Sacrum MRI	232	0.0923%
Pelvic MRI - Defecography	228	0.0907%
Pre-operative Localization with Clip (Titanium)	223	0.0888%
Right Shoulder CT	210	0.0836%
Right Wrist MRI	206	0.0820%
Brain MRI + Spectroscopy	205	0.0816%
Prostatic Ultrasound	203	0.0808%
Chest MRI	200	0.0796%
Spinal Cord MRI (Cervical+Thoracic)	199	0.0792%
Coronary CT - Calcium Score	197	0.0784%
Left Hand MRI	197	0.0784%
Portal Vein Doppler Ultrasound	188	0.0748%
Left Knee CT	188	0.0748%
Stereotactic Guided Vacuum Biopsy	187	0.0744%
Entero MRI	183	0.0728%
Right Knee CT	180	0.0716%
Brain Ultrasound	173	0.0689%
Left Shoulder CT	168	0.0669%
Paranasal Sinuses MRI	166	0.0661%
Left Wrist MRI	166	0.0661%
Magnetic Resonance with Tractography	164	0.0653%
Cerebral Stereotactic CT	163	0.0649%
Ultrasound Guided Drainage	161	0.0641%
Breast Ultrasound	157	0.0625%
Cervical Spinal Cord MRI	152	0.0605%
Aortic MR Angiography	148	0.0589%
Left Thigh CT	143	0.0569%
Sacrococcygeal CT	132	0.0525%
Right Thigh CT	130	0.0517%
Brain MRI with CSF Flow Study	130	0.0517%
Pancreas MRI	127	0.0505%
3D Facial CT	124	0.0494%
Knee CT	124	0.0494%
Head and Neck Embolization/Chemoembolization	124	0.0494%
Parotid MRI	123	0.0490%
Dentascan	121	0.0482%
CT Guided Abdominal Biopsy	120	0.0478%
Abdominal and Pelvic Embolization	117	0.0466%
Right Wrist CT	116	0.0462%
Thoracic Spinal Cord MRI	114	0.0454%
Green Pathway Protocol Cerebral CT	112	0.0446%
Bronchial Arteriography with Embolization	112	0.0446%
CT Guided Musculoskeletal System Biopsy	112	0.0446%
Right Leg MRI	110	0.0438%
Supra-aortic Trunks MR Angiography	110	0.0438%
Spinal Cord Ultrasound	109	0.0434%
Voiding Urosonography with Contrast	108	0.0430%
Lumbar Myelo MRI	107	0.0426%
Cervico-encephalic Angiography by Direct Puncture	107	0.0426%
Ultrasound Guided Thoracic Drainage	106	0.0422%
Left Leg MRI	104	0.0414%

Cervical Myelo MRI	101	0.0402%
Pharynx CT	100	0.0398%
Central Catheter Placement	100	0.0398%
Pharynx MRI	100	0.0398%
Right Ankle CT	99	0.0394%
Ultrasound Guided Thoracic Biopsy	99	0.0394%
Thigh MRI	98	0.0390%
Thoracic Myelo MRI	97	0.0386%
Left Ankle CT	95	0.0378%
Right Elbow MRI	94	0.0374%
Right Thigh MRI	92	0.0366%
Testicular Doppler Ultrasound	92	0.0366%
Left Wrist CT	90	0.0358%
Left Thigh MRI	90	0.0358%
Upper Abdominal Ultrasound	86	0.0342%
Unilateral Mammography - 2 Projections	86	0.0342%
Left Elbow MRI	84	0.0334%
Left Foot CT	78	0.0310%
Second Reading of External Mammography	77	0.0306%
Ultrasound Guided Head Biopsy	77	0.0306%
Left Leg CT	75	0.0299%
CT Guided Abdominal Drainage	74	0.0295%
Right Leg CT	73	0.0291%
CT Guided Abdominal Thermoablation	73	0.0291%
Instillation of Pharmacological Agents for Therapy	72	0.0287%
Sacroiliac Joint CT	72	0.0287%
Cranio-cervical Junction MRI with CSF Flow Study	71	0.0283%
Breast Tomosynthesis	71	0.0283%
Percutaneous Sclerosis of Venous Malformation	71	0.0283%
Right Foot CT	70	0.0279%
Hip MRI	69	0.0275%
Whole Body MRI	68	0.0271%
Cervical Doppler Ultrasound	67	0.0267%
Bilateral Mammography - 4 Projections	66	0.0263%
Neck MR Angiography	62	0.0247%
Ultrasound Guided Lower Limb Drainage	62	0.0247%
Muscle Ultrasound	61	0.0243%
Cerebral Venous CT	61	0.0243%
Barbotage of Calcific Tendinitis	59	0.0235%
Virtual Colonoscopy by CT	59	0.0235%
Ultrasound Guided Upper Limb Biopsy	59	0.0235%
Osteoarticular Infiltration of Therapeutic Agents	58	0.0231%
Ultrasound Guided Lower Limb Biopsy	58	0.0231%
Ultrasound Guided Pelvic Biopsy	56	0.0223%
Right Hand CT	56	0.0223%
Left Arm CT	55	0.0219%
Right Arm MRI	55	0.0219%
Abdominal and Pelvic Chemoembolization	54	0.0215%
CT Guided Pelvic Drainage	52	0.0207%
Left Forefoot MRI	51	0.0203%
Right Forefoot MRI	51	0.0203%
Gynecological Ultrasound	50	0.0199%
Leg MRI	49	0.0195%
Brachial Plexus MRI	48	0.0191%
Scrotal Doppler Ultrasound	47	0.0187%
Adrenal Ultrasound	47	0.0187%
Mouth MRI	47	0.0187%
Maxillofacial MRI	47	0.0187%
Catheter Removal	46	0.0183%
Cervico-occipital Junction CT	45	0.0179%
Right Arm CT	45	0.0179%

Left Arm MRI	45	0.0179%
Contrast Enhanced Ultrasound (CEUS)	44	0.0175%
Ultrasound Guided Upper Limb Drainage	43	0.0171%
Lower Abdominal MRI	42	0.0167%
Thoracic MR Angiography	42	0.0167%
CT Guided Pelvic Biopsy	42	0.0167%
Right Shoulder MR Arthrography	40	0.0159%
Left Hand CT	37	0.0147%
Mechanical Thrombectomy	37	0.0147%
Left Shoulder MR Arthrography	36	0.0143%
Ultrasound Guided Musculoskeletal System Biopsy	36	0.0143%
Ultrasound Guided Pelvic Drainage	35	0.0139%
Ultrasound Guided Musculoskeletal System Drainage	35	0.0139%
Left Elbow CT	35	0.0139%
Tongue MRI	34	0.0135%
Soft Tissue Ultrasound	34	0.0135%
Brachial Plexus MRI	34	0.0135%
Right Elbow CT	34	0.0135%
Second Reading of External Breast Ultrasound	33	0.0131%
Right Lower Limb MRI	33	0.0131%
PTC - Percutaneous Transhepatic Cholangiography	32	0.0127%
Rib MRI	32	0.0127%
Visceral Doppler Ultrasound (Triplex)	31	0.0123%
Left Forearm MRI	31	0.0123%
Visceral Doppler Ultrasound (Triplex)	31	0.0123%
Right Forearm MRI	30	0.0119%
Cervico-occipital Junction MRI	30	0.0119%
Lumbosacral Plexus MRI	30	0.0119%
Left Forearm CT	29	0.0115%
Sternum MRI	28	0.0111%
Abdominal and Pelvic Radioembolization	28	0.0111%
Thigh CT	27	0.0107%
Fluoroscopy Guided Abdominal Drainage	27	0.0107%
Right Forearm CT	26	0.0103%
Lower Limb MR Angiography	26	0.0103%
Left Hip Joint MRI	26	0.0103%
Head and Neck Vascular Prosthesis Placement	25	0.0100%
Transvaginal Gynecological Ultrasound	25	0.0100%
Left Lower Limb MRI	25	0.0100%
Mammography Biopsy Collection	25	0.0100%
Aspiration Biopsy Ultrasound	25	0.0100%
Left Femur MRI	24	0.0096%
Joint Capsule Hydrodistension	24	0.0096%
Fetal Neuro MRI (CNS)	23	0.0092%
CT Guided Thoracic Drainage	23	0.0092%
Petrous Bone MRI	21	0.0084%
Brain MRI with Functional Study	21	0.0084%
TMJ CT (Temporomandibular Joints)	21	0.0084%
Ultrasound Guided Aspiration Biopsy	21	0.0084%
Percutaneous Gastrostomy / Gastrojejunostomy	21	0.0084%
MRI Guided Breast Biopsy	21	0.0084%
Ultrasound Guided Axillary Lymph Node Aspiration Biopsy	21	0.0084%
Thoracic-Abdominal-Pelvic MRI	20	0.0080%
Face CT Angiography	20	0.0080%
Percutaneous Sclerosis of Lymphatic Malformation	20	0.0080%
External, Mixed and Internal Percutaneous Biliary Drainage	19	0.0076%
Ultrasound Guided Abdominal Drainage	19	0.0076%
Right Hip Joint MRI	18	0.0072%
Ultrasound Guided Vacuum Biopsy	18	0.0072%
Larynx CT	18	0.0072%
Right Knee MR Arthrography	18	0.0072%

Ultrasound Guided Abdominal Biopsy	18	0.0072%
Oropharynx MRI	17	0.0068%
Ankle CT	16	0.0064%
Shoulder MRI	15	0.0060%
Bronchial Artery Angiography	15	0.0060%
Percutaneous Cholecystostomy	15	0.0060%
Musculoskeletal System Embolization	15	0.0060%
Second Reading of External Breast MRI	15	0.0060%
Mediastinum MRI	15	0.0060%
Hand CT	14	0.0056%
Petrosal Sinuses	14	0.0056%
Ultrasound Guided Neck Drainage	13	0.0052%
Pituitary CT	13	0.0052%
Abdominal Percutaneous Radiofrequency Thermoablation	13	0.0052%
CT Guided Musculoskeletal System Drainage	13	0.0052%
Foot CT	13	0.0052%
Right Upper Limb MRI	13	0.0052%
Leg CT	12	0.0048%
Internal Auditory Canal/Pontine Cisterns MRI	12	0.0048%
Spinal Cord MR Angiography	12	0.0048%
Thoracic Embolization	11	0.0044%
Arm MRI	11	0.0044%
Larynx MRI	11	0.0044%
Thoracic + Abdominal MRI	11	0.0044%
Mastoid CT	11	0.0044%
Hip MR Arthrography	11	0.0044%
Right Femur MRI	11	0.0044%
Percutaneous Nephrostomy	10	0.0040%
Adrenal Vein Blood Collection for Dosage	10	0.0040%
Lower Limb MRI	10	0.0040%
T.A.B. Transthoracic Aspiration Biopsy	9	0.0036%
Ultrasound Guided Supraclavicular Lymph Node Aspiration Biopsy	9	0.0036%
Left Breast MRI	9	0.0036%
Left Upper Limb MR Angiography	9	0.0036%
Left Knee MR Arthrography	9	0.0036%
Abdominal and Pelvic Percutaneous Drainage	8	0.0032%
Upper Limb MR Angiography	8	0.0032%
Lymphangiography	8	0.0032%
Lumbar Spine Discography	8	0.0032%
Foot MRI	8	0.0032%
Right Breast MRI	8	0.0032%
Fluoroscopy Guided Thoracic Biopsy	8	0.0032%
Right Upper Limb MR Angiography	8	0.0032%
Brain Perfusion CT	8	0.0032%
Fluoroscopy Guided Thoracic Drainage	8	0.0032%
Wrist CT	7	0.0028%
Percutaneous Cholangiography	7	0.0028%
Macroradiography	7	0.0028%
Pulley Release	7	0.0028%
CT Guided Neck Biopsy	7	0.0028%
Abdominal and Pelvic Biliary or Digestive Prosthesis Placement	7	0.0028%
Cavum CT	6	0.0024%
Cervical Spine MRI with CSF Flow Study	6	0.0024%
Fetal MRI	6	0.0024%
Petrous Bone CT	6	0.0024%
Drain Review and/or Replacement	6	0.0024%
Stereotactic Guided Core Biopsy (Mammography)	6	0.0024%
Left Upper Limb MRI	6	0.0024%
Hip Joint MRI	6	0.0024%
Femur MRI	6	0.0024%
Spinal Angiography	5	0.0020%

Iliac MRI	5	0.0020%
Ankle MRI	5	0.0020%
Fluoroscopy Guided Pelvic Drainage	5	0.0020%
Left Humerus MRI	5	0.0020%
Spleen MRI	5	0.0020%
Bronchial Arteriography	5	0.0020%
Ultrasound Guided Left Breast Node 1 Core Biopsy	5	0.0020%
Soft Tissue Doppler Ultrasound	5	0.0020%
Surgical Specimen Mammography	5	0.0020%
Fluoroscopy Guided Lower Limb Biopsy	4	0.0016%
Fluoroscopy Guided Transjugular Liver Biopsy	4	0.0016%
CT Guided Thoracic Thermoablation	4	0.0016%
Breast Macroradiography	4	0.0016%
TIPS Placement	4	0.0016%
MRI, Supplementary MR Angiography with Gadolinium	4	0.0016%
Manubrium MRI	4	0.0016%
Right Humerus MRI	4	0.0016%
Upper Limb MRI	4	0.0016%
Percutaneous Cholangioplasty (Benign Stenosis of Biliary Tract)	3	0.0012%
Sella Turcica CT	3	0.0012%
Celiac Plexus Neurolytic Block	3	0.0012%
Fluoroscopy Guided Pelvic Biopsy	3	0.0012%
Fluoroscopy Guided Musculoskeletal System Biopsy	3	0.0012%
Vertebro-medullary CT Angiography	3	0.0012%
Ultrasound Guided Neck Biopsy	3	0.0012%
MRI Angiography of Carotids	3	0.0012%
Percutaneous Sclerosis of Abdominal Cystic Lesions	3	0.0012%
Parathyroid MRI	3	0.0012%
Hand MRI	3	0.0012%
Breast Ultrasound	3	0.0012%
Knee MRI	3	0.0012%
Thoracic Spine Myelography	3	0.0012%
Abdominal and Pelvic Percutaneous Chemoablation	2	0.0008%
Directed Puncture Echotomography	2	0.0008%
Pre-operative Localization with Ultrasound-guided Carbon	2	0.0008%
Excretory Tract Ultrasound	2	0.0008%
Intra-articular Infiltration	2	0.0008%
Thymus MRI	2	0.0008%
Soft Tissue Ultrasound	2	0.0008%
Percutaneous Radiofrequency Thermoablation of Muscle	2	0.0008%
CT Arthrography	2	0.0008%
Fluoroscopy-guided Neck Drainage	2	0.0008%
Vascular MRI	2	0.0008%
Elbow MRI	2	0.0008%
Lumbar Spine Myelography	2	0.0008%
Doppler Ultrasound of Hepatic Vessels	2	0.0008%
Radioscopic Support for Catheterization	2	0.0008%
Ultrasound-guided Head Drainage	2	0.0008%
CT Myelography	2	0.0008%
Pre-operative Localization with Metallic Reference	2	0.0008%
Fluoroscopy-guided Lower Limb Drainage	2	0.0008%
Forearm MRI	2	0.0008%
CT Limbs (each anatomical segment)	1	0.0004%
CT - Intervention Support	1	0.0004%
Pre-operative Localization with Carbon by Stereotaxy	1	0.0004%
Ultrasound-guided Biopsy of Thorax and Breast	1	0.0004%
Ultrasound-guided Drainage of Thorax and Breast	1	0.0004%
Fluoroscopy-guided Drainage of Musculoskeletal System	1	0.0004%
Renal and Suprarenal Ultrasound	1	0.0004%
Ultrasound of 1 Territory	1	0.0004%
Angiography of Cavernous Bodies	1	0.0004%

Ultrasound-guided Microbiopsy of Left Breast Node 2	1	0.0004%
Brain MRI + CAIS	1	0.0004%
Abdominal MRI	1	0.0004%
Dry-needling for Epicondylitis	1	0.0004%
Ultrasound-guided Microbiopsy of Right Breast	1	0.0004%
Suprapubic Pelvic Ultrasound	1	0.0004%
Fluoroscopy-guided Abdominal Biopsy	1	0.0004%
Pelvic MRI	1	0.0004%
Drain Extraction	1	0.0004%
Dry-needling for Plantar Fascia	1	0.0004%
MRI Support for Intervention Procedures	1	0.0004%
TIPS Placement (Abdomen and Pelvis)	1	0.0004%
Cervical Spine Myelography	1	0.0004%
Fluoroscopy-guided Neck Biopsy	1	0.0004%
Fluoroscopy-guided Upper Limb Biopsy	1	0.0004%
Hepatic Catheterization	1	0.0004%
Venous Doppler Ultrasound of Lower Limbs	1	0.0004%
Central Catheter Placement for Hemodialysis	1	0.0004%
Super-selective Angiography	1	0.0004%
Thyroid MRI	1	0.0004%
Perioperative Ultrasound	1	0.0004%
Mammography	1	0.0004%
Right Wrist MR Arthrography	1	0.0004%
CT-guided Drainage - Abdomen	1	0.0004%
Measurement of Hepatic Portions	1	0.0004%
Thyroid Ultrasound	1	0.0004%
Ultrasound-guided Pelvic Drainage	1	0.0004%
MRI Intervention	1	0.0004%
Peripheral CT Angiography	1	0.0004%
Fluoroscopy-guided Head Biopsy	1	0.0004%
Post-operative Cholangiography (Radiology Only)	1	0.0004%
Wrist MRI	1	0.0004%

Table C.2: Counts of medical acts of Non X-ray radiological data of Site 2, before categorical outlier removal.

Medical Act	Count	Percentage
3D post-processing or sequence with reconstruction	45033	15.2567%
CT, intravenous contrast supplement	24890	8.4325%
MRI, post-processing	13462	4.5608%
MRI, additional diffusion study	10817	3.6647%
Chest CT	10190	3.4523%
Upper abdominal ultrasound	10016	3.3933%
MRI, contrast supplement	9556	3.2375%
Cranioencephalic CT	9541	3.2324%
Cranioencephalic MRI	9452	3.2022%
CT of thorax, upper abdomen and pelvis	8528	2.8892%
Renal and adrenal ultrasound	7403	2.5081%
Breast ultrasound	7039	2.3847%
MRI of lumbar and sacral spine	5268	1.7847%
CT of upper abdomen and pelvis	4970	1.6838%
Mammography, two views per breast	4375	1.4822%
Bladder ultrasound via suprapubic approach	4325	1.4653%
MRI, additional cortical mapping	4236	1.4351%
CT of the chest and upper abdomen	3913	1.3257%
High-resolution chest CT	3351	1.1353%
Bone densitometry of lumbar spine and femoral neck	3131	1.0607%
Tomosynthesis	3037	1.0289%
MRI of cervical spine	3028	1.0259%
Pelvic ultrasound via suprapubic approach	2724	0.9229%

CT, additional Angio CT	2593	0.8785%
Soft tissue ultrasound	2567	0.8697%
CT of upper abdomen	2403	0.8141%
Positron tomography - Whole Body Study	2374	0.8043%
CT of paranasal sinuses	2166	0.7338%
Thyroid ultrasound	2120	0.7182%
Post-void ultrasound with residual volume calculation	2066	0.6999%
MRI, additional non-contrast Angio-MRI (TOF)	1942	0.6579%
CT of urinary system (Uro-CT)	1747	0.5919%
CT of neck, thorax, upper abdomen and pelvis	1655	0.5607%
Cerebral Angio-CT	1638	0.5549%
MRI of right knee	1504	0.5095%
MRI of left knee	1491	0.5051%
Whole body bone scintigraphy	1409	0.4774%
MRI, Cholangiography (MRCP) - includes Abdominal MRI	1383	0.4685%
Selective arteriography	1296	0.4391%
Angio CT post-processing	1294	0.4384%
MRI of upper abdomen	1277	0.4326%
Histology puncture	1238	0.4194%
Pelvic MRI	1199	0.4062%
CT of lumbar spine	1195	0.4049%
Angio-CT of supra-aortic trunks	1177	0.3988%
CT of neck and thorax	1157	0.3920%
MRI of thoracic spine	1084	0.3672%
Prostate MRI	1057	0.3581%
Histology puncture - Thorax and Breast	1051	0.3561%
Abdominal Doppler, each viscera or vascular territory	1051	0.3561%
Breast MRI	990	0.3354%
MRI, additional associated with examination execution	979	0.3317%
Marking of subclinical lesion by clip	954	0.3232%
CT, support for intervention procedures	895	0.3032%
CT of neck (soft tissues)	870	0.2947%
Pelvic CT	866	0.2934%
CT of both lower limbs	837	0.2836%
Ultrasound, support for intervention procedures	794	0.2690%
Ultrasound of left hip	762	0.2582%
Ultrasound of right hip	749	0.2538%
CT of ears	739	0.2504%
Ultrasound of neck (soft tissues)	723	0.2449%
MRI, additional perfusion study	714	0.2419%
Doppler of neck vessels	628	0.2128%
MRI of right shoulder	612	0.2073%
Vascular access for angiographic procedure	602	0.2040%
MRI of pituitary	592	0.2006%
Cardiac morpho-functional MRI	553	0.1874%
Ultrasound, additional associated with examination execution	530	0.1796%
MRI of ears	484	0.1640%
CT of cervical spine	468	0.1586%
Wrist bone densitometry	466	0.1579%
Maxillofacial CT	456	0.1545%
Cytology puncture - Neck	451	0.1528%
Percutaneous drainage - Abdomen and Pelvis	451	0.1528%
Ultrasound support in Imaging service	434	0.1470%
Ultrasound of right shoulder	430	0.1457%
Supraselective arteriography	416	0.1409%
Examination performed with anesthesiology support	406	0.1375%
Scrotal ultrasound	398	0.1348%
Ultrasound of left shoulder	377	0.1277%
MRI of left shoulder	376	0.1274%
CT of thoracic spine	371	0.1257%
Joint post-processing	371	0.1257%

MRI of neck	364	0.1233%
Pulmonary ventilation/inhalation scintigraphy	353	0.1196%
Pulmonary perfusion scintigraphy	352	0.1193%
Cardiac MRI with myocardial perfusion study	344	0.1165%
MRI of right ankle	332	0.1125%
Lymphoscintigraphy for sentinel lymph node detection	330	0.1118%
MRI of right hip	329	0.1115%
MRI of left hip	324	0.1098%
MRI of left ankle	322	0.1091%
MRI of sacroiliac joints	311	0.1054%
Percutaneous closure of arterial access	304	0.1030%
Vascular access for angiographic procedure	302	0.1023%
Thyroid scintigraphy	290	0.0982%
MRI of pelvis	280	0.0949%
Lower jaw CT dental/implantology	279	0.0945%
MRI, Enterography	273	0.0925%
Ultrasound of salivary glands	268	0.0908%
Histology puncture - Abdomen and Pelvis	267	0.0905%
MRI of orbits	267	0.0905%
CT of sacro-coccygeal spine	263	0.0891%
Exam Not Performed with Time Spent in Room	257	0.0871%
Ultrasound of right hand	245	0.0830%
Doppler of venous sector of left lower limb	234	0.0793%
CT of pelvis	232	0.0786%
CT of right knee	232	0.0786%
Upper jaw CT dental/implantology	229	0.0776%
Arterial embolization	226	0.0766%
Gastrostomy button placement	225	0.0762%
Doppler of venous sector of right lower limb	217	0.0735%
CT of left knee	213	0.0722%
Esophageal transit	212	0.0718%
Liver biopsy	208	0.0705%
Ultrasound of left hand	206	0.0698%
Cerebral Tomography With 123I - Ioflupane	206	0.0698%
Mammography, each additional view	206	0.0698%
Iodine-131 therapy for hyperthyroidism	204	0.0691%
Joint access by image control, for drainage	203	0.0688%
CT of pharynx	203	0.0688%
CT, additional associated with examination execution	191	0.0647%
CT of right hip	187	0.0634%
CT of left hip	184	0.0623%
CT, oral contrast	184	0.0623%
Histology puncture - Neck	182	0.0617%
Dentascan post-processing	175	0.0593%
MRI of left forefoot	173	0.0586%
MRI of right forefoot	170	0.0576%
Ultrasound of left knee	166	0.0562%
Renogram with MAG 3	161	0.0545%
PET - Positron tomography with 68 Ga-Peptides	155	0.0525%
Ultrasound of right knee	151	0.0512%
MRI, Defecography	151	0.0512%
Ultrasound of right wrist	150	0.0508%
Myocardial perfusion scintigraphy at stress/strain	150	0.0508%
CT of right thigh	148	0.0501%
Bone densitometry of lumbar spine	147	0.0498%
Myocardial perfusion scintigraphy at rest	145	0.0491%
Prostate and seminal vesicles ultrasound, via transrectal	142	0.0481%
MRI of TMJ	142	0.0481%
MRI, Cholangiography (MRCP)	140	0.0474%
Marking of subclinical lesion by harpoon	138	0.0468%
MRI of right hand	138	0.0468%

MRI of right wrist	136	0.0461%
Nerve root or nerve plexus block under image guidance	129	0.0437%
Vacuum biopsy	127	0.0430%
Cytology puncture	126	0.0427%
Ultrasound of left wrist	125	0.0423%
MRI of paranasal sinuses	124	0.0420%
Indiscriminate CT	122	0.0413%
MRI, cerebral spectroscopy study	119	0.0403%
CT of left ankle	119	0.0403%
Percutaneous drainage - Thorax	118	0.0400%
Percutaneous thermoablation procedure	117	0.0396%
Stent placement in arterial access territory	117	0.0396%
PTHC - Percutaneous transhepatic cholangiography	116	0.0393%
MRI of face	115	0.0390%
Ultrasound of right elbow	115	0.0390%
Ultrasound of right forefoot	111	0.0376%
Parathyroid scintigraphy	110	0.0373%
Ultrasound of right ankle	110	0.0373%
Three-phase bone scintigraphy	110	0.0373%
Doppler of arterial sector of right lower limb	109	0.0369%
Renal scintigraphy with 99mTc - DMSA	108	0.0366%
Drainage of musculoskeletal collections by image control	108	0.0366%
MRI of left wrist	108	0.0366%
Mammography of surgical specimen	108	0.0366%
Post-processing metal artifact removal	107	0.0363%
CT of right ankle	106	0.0359%
Doppler of arterial sector of left lower limb	102	0.0346%
Selective arteriography of supra-aortic trunks	100	0.0339%
Ultrasound of left forefoot	100	0.0339%
CT of left thigh	99	0.0335%
MRI of left hand	98	0.0332%
Selective angioplasty	95	0.0322%
Renal MRI	94	0.0318%
CT of left leg	94	0.0318%
CT of right shoulder	93	0.0315%
Cardiac scintigraphy Tc99m-DPD	93	0.0315%
Ultrasound of left ankle	92	0.0312%
MRI of left leg	91	0.0308%
MRI of right thigh	90	0.0305%
CT of left shoulder	88	0.0298%
Ultrasound of left elbow	88	0.0298%
Thoracic ultrasound	87	0.0295%
Gynecological ultrasound via endocavitary (intravaginal)	85	0.0288%
MRI of left thigh	84	0.0285%
PET with 68 GA-PSMA	83	0.0281%
CT of orbits	82	0.0278%
CT, Colonography (virtual colonoscopy)	74	0.0251%
Implantation of device for central venous access	73	0.0247%
MRI of right leg	72	0.0244%
Bone densitometry of femoral neck	72	0.0244%
Joint CT, each joint	71	0.0241%
Fetal cerebral MRI	71	0.0241%
MRI of cranio-vertebral junction	71	0.0241%
CT of right leg	71	0.0241%
Cranioencephalic Angio-MRI	70	0.0237%
CT, support for other procedures	70	0.0237%
MRI of thorax	68	0.0230%
MRI of left elbow	67	0.0227%
Retrograde cystography with vesicoureteral reflux study	66	0.0224%
Patient with ICD	65	0.0220%
Biliary stent placement or recanalization	64	0.0217%

Cholangiography by external drain (Kehr)	64	0.0217%
CT for measurements	64	0.0217%
CT of left forefoot	64	0.0217%
Contrast-enhanced mammography	63	0.0213%
CT of right forefoot	59	0.0200%
CT of right wrist	58	0.0196%
Cerebral Venous-CT	58	0.0196%
Cytology puncture - Thorax	58	0.0196%
CT, rectal contrast	58	0.0196%
Time spent, contrast nephropathy prophylaxis	57	0.0193%
Central angiographic venous study	57	0.0193%
CT of left wrist	55	0.0186%
Breast cyst drainage or cystography, each	54	0.0183%
Additional contrast supplement	53	0.0180%
Supraselective angioplasty	52	0.0176%
Long-term CVC placement	51	0.0173%
Swallowing and/or esophagus study with cineradiography	50	0.0169%
MRI of right elbow	50	0.0169%
Nasogastric tube placement	49	0.0166%
CT of lower limbs (each anatomical segment)	49	0.0166%
Positron tomography - Brain study with F-FDG	46	0.0156%
Retrograde and voiding urethrography	45	0.0152%
CT of sacroiliac joints	45	0.0152%
CT of right arm	43	0.0146%
Cytology puncture - Musculoskeletal	42	0.0142%
Cytology puncture - Abdomen and Pelvis	42	0.0142%
Vessel Wall Study	42	0.0142%
Arteriography of aortic arch, supra-aortic trunks	41	0.0139%
CT, Enterography	39	0.0132%
Arteriography of right lower limb	39	0.0132%
Percutaneous endoscopic gastrostomy	39	0.0132%
Penile Eco-Doppler	38	0.0129%
CT of left arm	37	0.0125%
Doppler of venous sector of right upper limb	34	0.0115%
Gastro-duodenal transit double contrast (with or without small intestine)	34	0.0115%
Gastro-duodenal transit single contrast (with or without small intestine)	32	0.0108%
CT of mandible	31	0.0105%
MRI of right arm	31	0.0105%
Doppler of other sector (e.g., soft tissue mass)	31	0.0105%
Adrenal MRI	30	0.0102%
Feed and Wrap Technique	29	0.0098%
Lymphoscintigraphy	28	0.0095%
MRI of left arm	28	0.0095%
CT of cranio-vertebral junction	27	0.0091%
CT of right elbow	26	0.0088%
Angioplasty in venous territory	26	0.0088%
Salivary gland scintigraphy	26	0.0088%
MRI of right forearm	26	0.0088%
CT of right hand	26	0.0088%
CT of left forearm	26	0.0088%
MRI, additional Arthrography	25	0.0085%
CT of right forearm	25	0.0085%
MRI with dynamic CSF flow study	24	0.0081%
Intra-arterial mechanical thrombectomy	23	0.0078%
CT of left hand	23	0.0078%
Ultrasound, additional associated with examination execution	23	0.0078%
MRI of left forearm	22	0.0075%
Abdominal aortography	21	0.0071%
Arteriography of left lower limb	21	0.0071%
CT of left elbow	20	0.0068%
Transjugular liver biopsy	20	0.0068%

Joint ultrasound	19	0.0064%
CT examination complement	18	0.0061%
Indirect cystoscintigraphy	18	0.0061%
Doppler of venous sector of left upper limb	17	0.0058%
CT, perfusion study	17	0.0058%
MRI of sternoclavicular joints	17	0.0058%
Iliac arteriography	15	0.0051%
CT for sacroiliac measurements	15	0.0051%
Bronchial arteriography	15	0.0051%
CT of TMJ	14	0.0047%
Prevention – Interventional Neuroradiology	14	0.0047%
MRI, additional tractography	14	0.0047%
Doppler (additional to any ultrasound examination)	14	0.0047%
PET — Positron tomography with 18F-Choline	14	0.0047%
Phlebography of right upper limb	14	0.0047%
Portal pressure assessment by hepatic vein catheterization	13	0.0044%
Hepatic Angio-MRI	13	0.0044%
Angio-MRI of supra-aortic trunks	13	0.0044%
Import of external examination	13	0.0044%
Portal embolization	13	0.0044%
Catheter placement (internal code - not export)	13	0.0044%
Histology puncture - Musculoskeletal	13	0.0044%
Doppler of arterial sector of right upper limb	12	0.0041%
TIPS	12	0.0041%
Percutaneous drainage - Musculoskeletal	12	0.0041%
Drainage of abscesses and hematomas	12	0.0041%
Ascending urethrography with post-void study	11	0.0037%
CT of upper limb (each anatomical segment)	11	0.0037%
Fetal MRI of central nervous system	10	0.0034%
Stent placement in supraselective venous territory	10	0.0034%
Doppler of subclavian veins	10	0.0034%
Hepatic radioembolization	10	0.0034%
Scrotal Doppler	10	0.0034%
Doppler of arterial sector of left upper limb	10	0.0034%
Arthrography of right shoulder	9	0.0030%
Pelvic Doppler	9	0.0030%
Internal jugular phlebography	9	0.0030%
Abdominal aortography with selective catheterization	9	0.0030%
Orthopantomography	9	0.0030%
CT of sternoclavicular joints	9	0.0030%
MRI for measurements	9	0.0030%
MRI, support for intervention procedures	8	0.0027%
Arteriography of lower limbs including iliac region	8	0.0027%
Embolization of spermatic vein or ovarian vein	8	0.0027%
Percutaneous Aspiration puncture of cyst or cavity	8	0.0027%
MRI examination complement	8	0.0027%
Vascular access study	8	0.0027%
Angio-MRI of lower limbs	8	0.0027%
PET — Positron tomography with 18F-DOPA	7	0.0024%
Penile ultrasound	7	0.0024%
Phlebography of pelvic veins	7	0.0024%
Arteriography of right upper limb	7	0.0024%
Small intestine transit (by ingestion)	7	0.0024%
MRI of thorax, upper abdomen and pelvis	7	0.0024%
Positron Tomography - Cardiac Study with F-FDG	7	0.0024%
Hepatic venography by selective catheterization	7	0.0024%
Percutaneous pyelography (imaging support added)	7	0.0024%
Whole Body Scintigraphy with 123 MIBG	7	0.0024%
Cystography, with established route, supervision	7	0.0024%
Pharmacological Intervention - Diuretic	7	0.0024%
Thoracic aortography	7	0.0024%

Arthrocentesis under image control	7	0.0024%
Ultrasound of spinal canal and contents	6	0.0020%
Barium enema	6	0.0020%
Angio-MRI of left upper limb	6	0.0020%
Arthrography of left hip	6	0.0020%
Ultrasound of sternoclavicular joints	6	0.0020%
Unspecified Scintigraphy of Any Organ	6	0.0020%
Gastric Emptying Study	6	0.0020%
Pelvic embolization (non-uterine and non-prostatic)	6	0.0020%
Ultrasound of sacroiliac joints	6	0.0020%
Prevention – Interventional Radiology	6	0.0020%
MRI of posterior fossa	6	0.0020%
Additional for contrast administration	6	0.0020%
Percutaneous cholangiography (PTC)	6	0.0020%
Descending pyeloureterography, established route	5	0.0017%
Lumbar myelography, with contrast	5	0.0017%
CT sella turcica	5	0.0017%
Direct percutaneous portography	5	0.0017%
Partial bone scintigraphy	5	0.0017%
Arthrography of right hip	5	0.0017%
Vascular access study for hemodialysis	5	0.0017%
Meckel's diverticulum search	5	0.0017%
Pulmonary arteriography	4	0.0014%
Direct cystoscintigraphy	4	0.0014%
Cisternography with CT support for other procedures	4	0.0014%
Selective phlebography of left spermatic, unilateral	4	0.0014%
Adrenal cortex scintigraphy	4	0.0014%
Percutaneous cementoplasty	4	0.0014%
Pleural drainage	4	0.0014%
Arteriography of genital arteries	4	0.0014%
Percutaneous drainage - Neck	4	0.0014%
Phlebography of left upper limb	3	0.0010%
Prostatic embolization	3	0.0010%
Left renal phlebography	3	0.0010%
Angio-MRI of right upper limb	3	0.0010%
Aorto-iliac Angio-MRI	3	0.0010%
Arthrography of right knee	3	0.0010%
Uterine embolization	3	0.0010%
Myelography and cervical Myelo-CT	3	0.0010%
Renal Angio-MRI	3	0.0010%
Hypogastric arteriography	2	0.0007%
Central venous access revision	2	0.0007%
Cardiac Scintigraphy with 123I-MIBG / with study	2	0.0007%
Cystography, with established route (support added)	2	0.0007%
Whole body scintigraphy with 131I	2	0.0007%
Hepatic suprahepatic – angioplasty	2	0.0007%
Arthrography of right wrist	2	0.0007%
Arthrography of left wrist	2	0.0007%
Mechanical thrombectomy in venous territory	2	0.0007%
Selective phlebography of right spermatic, unilateral	2	0.0007%
Whole body scintigraphy with 131IMIBG	2	0.0007%
Nephrostomy catheter replacement (support added)	2	0.0007%
Endovascular treatment of vascular malformation	2	0.0007%
Arteriography of left upper limb	2	0.0007%
Direct trans-hepatic portography	2	0.0007%
Liver scintigraphy with marked red blood cells	2	0.0007%
Embolization of varices after TIPS placement	1	0.0003%
MRI, additional pharmacological stress test	1	0.0003%
Selective arteriography of supra-aortic trunks	1	0.0003%
Whole body scintigraphy with 67Ga (5mCi)	1	0.0003%
MRI of thorax and upper abdomen	1	0.0003%

Gynecological ultrasound via endocavitary	1	0.0003%
CT, additional intrathecal injection (myelo-CT)	1	0.0003%
Hepato-splenic scintigraphy	1	0.0003%
Digestive hemorrhage search	1	0.0003%
Radiological support for ERCP	1	0.0003%
MRI, Enteroclysis	1	0.0003%
Myelography	1	0.0003%
Myelography and thoracic Myelo-CT	1	0.0003%
Right renal phlebography	1	0.0003%
Thoracic Angio-MRI	1	0.0003%
Ultrasound for hydrostatic reduction (pediatrics)	1	0.0003%
Vena cava filter placement	1	0.0003%
Arteriography of thoracic wall (e.g., mammary)	1	0.0003%
Right adrenal phlebography	1	0.0003%
Indirect portography	1	0.0003%
AVM embolization	1	0.0003%
Iliocavography	1	0.0003%
CT, additional volumetric calculation (e.g., liver)	1	0.0003%
Fetal MRI	1	0.0003%
Renogram	1	0.0003%
Myelography and lumbar Myelo-CT	1	0.0003%
Cisternography	1	0.0003%
Arthrography of left shoulder	1	0.0003%
Portal vein – angioplasty/thrombolysis	1	0.0003%
Enema through cutaneous stoma	1	0.0003%
Dynamic cavernosography	1	0.0003%
Synovial biopsy	1	0.0003%
Sialography of right gland	1	0.0003%
Intraoperative ultrasound	1	0.0003%
Fetal thoracic MRI	1	0.0003%
Gastroesophageal reflux search	1	0.0003%

Table C.3: Counts of medical acts of X-ray radiological data of Site 1, before categorical outlier removal.

Medical Act	Count	Percentage
Chest (Frontal)	55717	33.0378%
Chest (Frontal + P)	11327	6.7164%
Pelvis (Frontal)	5787	3.4314%
Pelvis under load (Frontal)	4637	2.7495%
Spine Extra-long (Frontal + P)	3975	2.3570%
Feet under load (Bilateral)	3971	2.3546%
Lower Limbs Extra-long Plate	3615	2.1435%
Hands Bilateral	3278	1.9437%
Lumbosacral Spine - Hyperflexion and Hyperextension	3183	1.8874%
Thoracic + Lumbar Spine (Frontal + P)	3055	1.8115%
Cervical Spine (Frontal + P)	3017	1.7890%
Hip (Right Hip Joint)	2741	1.6253%
Knees Under Load (Bilateral, frontal under load + 30° Lateral)	2727	1.6170%
Hip (Left hip Joint)	2724	1.6152%
Cervical Spine - Hyperflexion and Hyperextension	2508	1.4871%
Lumbar Spine (Frontal + Lateral)	2396	1.4207%
Hips (Hip Joints Bilateral)	2376	1.4089%
Knees Under Load (Frontal + P)	2139	1.2683%
Lumbosacral Spine (Frontal + P)	1951	1.1569%
Patellae (Axial at 90°)	1674	0.9926%
Left Knee (Frontal + Lateral)	1552	0.9203%
Right Knee (Frontal + Lateral)	1526	0.9049%
Chest (Frontal + Right Lateral)	1494	0.8859%
Cervical Spine (Frontal + Lateral + Oblique)	1440	0.8539%
Extra-long Spine (Frontal)	1414	0.8384%

Weight-bearing Lumbar Spine (Frontal + Lateral)	1377	0.8165%
Orthopantomography	1346	0.7981%
Left Wrist	1344	0.7969%
Thoracic Spine (Frontal + Lateral)	1271	0.7536%
Right Shoulder Standard (Frontal + Lamy Lateral)	1255	0.7442%
Abdominal Standing Plain Film	1255	0.7442%
Left Shoulder Standard (Frontal + Lamy Lateral)	1193	0.7074%
Right Ankle	1183	0.7015%
Left Ankle	1159	0.6872%
Right Wrist	1141	0.6766%
Extra-long Spine (Lateral)	926	0.5491%
Right Foot	876	0.5194%
Weight-bearing Knees (2 Views)	860	0.5099%
Right Hand	822	0.4874%
Left Foot	812	0.4815%
Thoraco-abdominal	759	0.4501%
Bone Age - Hand and Wrist	745	0.4418%
Chest (Frontal + Left Lateral + Right Lateral)	739	0.4382%
Left Femur	692	0.4103%
Sacroiliac - 2 Oblique Views	684	0.4056%
Right Elbow	682	0.4044%
Right Femur	668	0.3961%
Right Shoulder Arthrosis (Double Oblique Frontal + Bernargeau Lateral)	622	0.3688%
Left Elbow	591	0.3504%
Left Hand	584	0.3463%
Left Shoulder Arthrosis (Double Oblique Frontal + Bernargeau Lateral)	544	0.3226%
Right Leg	526	0.3119%
Bilateral Weight-bearing Ankles	525	0.3113%
Plain Abdominal Film	471	0.2793%
Left Forearm	464	0.2751%
Sacroiliac Joints - Bilateral View	438	0.2597%
Bilateral Wrists	426	0.2526%
Bilateral Feet (Frontal + Internal Oblique)	423	0.2508%
Left Hand Fingers	392	0.2324%
Right Forearm	387	0.2295%
Esophageal-Gastro-Duodenal Transit	380	0.2253%
Patellae (1 Axial)	379	0.2247%
Left Leg	379	0.2247%
Lumbar Spine (Frontal + Lateral + Obliques)	368	0.2182%
Weight-bearing Lumbosacral Spine (Frontal + Lateral)	367	0.2176%
Right Hand Fingers	360	0.2135%
Bilateral Knees (Frontal + Lateral)	348	0.2063%
Weight-bearing Pelvis (Frontal)	346	0.2052%
Right Shoulder Calcifications (Frontal + Internal Rotation + External Rotation)	339	0.2010%
Renovesical	331	0.1963%
Left Shoulder Calcifications (Frontal + Internal Rotation + External Rotation)	296	0.1755%
Left Arm (Humerus)	270	0.1601%
Bilateral Elbows	265	0.1571%
Cervical Spine (Trans-Oral View)	264	0.1565%
Right Arm (Humerus)	246	0.1459%
Right Clavicle	224	0.1328%
Facial Cephalometry	219	0.1299%
Lumbosacral Spine (Frontal + Lateral + Obliques)	215	0.1275%
Adult Skeleton (See Protocol for Views)	211	0.1251%
Left Clavicle	211	0.1251%
Chest (Frontal in Expiration)	204	0.1210%
Extra-long Spine - Lateral Inclinations	189	0.1121%
Left Knee	180	0.1067%
Right Knee	170	0.1008%
Patellae (Axial 30-60-90°)	165	0.0978%
Videofluoroscopy	150	0.0889%

Esophageal Transit	143	0.0848%
Bilateral Rib Cage	135	0.0800%
Bilateral Frontal Weight-bearing Knees	127	0.0753%
Bilateral Feet (Frontal + External Oblique)	120	0.0712%
Child's Lower Limbs (1 View)	110	0.0652%
Radioscopic Support	103	0.0611%
Left Shoulder Acromioclavicular (Frontal + Zanca)	99	0.0587%
Bilateral Ankles	95	0.0563%
Right Shoulder	94	0.0557%
Gastro-Duodenal Transit with Hiatal Hernia Assessment	88	0.0522%
Left Knee (Monopodal Support Frontal + 30° Lateral)	87	0.0516%
Right Knee (Monopodal Support Frontal + 30° Lateral)	86	0.0510%
Skull (Frontal + Lateral)	82	0.0486%
Left Shoulder	81	0.0480%
Retrograde Urethrocystography with Voiding Phase	81	0.0480%
Right Shoulder Acromioclavicular (Frontal + Zanca)	79	0.0468%
Right Rib Cage	79	0.0468%
Left Calcaneus	77	0.0457%
Left Rib Cage	74	0.0439%
Extra-long Spine - Hyperflexion and Hyperextension	73	0.0433%
Bilateral Clavicles	72	0.0427%
Right Foot Toes	65	0.0385%
Bilateral Legs	63	0.0374%
Right Calcaneus	60	0.0356%
Weight-bearing Knees (Frontal)	58	0.0344%
Drain Cholangiography	56	0.0332%
Sacral Spine (Frontal + Lateral)	55	0.0326%
Left Foot Toes	51	0.0302%
Coccygeal Spine (Frontal + Lateral)	47	0.0279%
Left Shoulder Instability (Frontal + Bernargeau + Apical)	45	0.0267%
Right Shoulder Instability (Frontal + Bernargeau + Apical)	43	0.0255%
Bilateral Knees	40	0.0237%
Sternum (Oblique + Lateral)	37	0.0219%
Hysterosalpingography (Radiology)	36	0.0213%
Left Scapula	36	0.0213%
Child Skeleton (See Protocol for Views)	32	0.0190%
Skull (3 Views)	32	0.0190%
Right Scapula	32	0.0190%
Bilateral Shoulders	31	0.0184%
Bilateral Knees 30 Degrees Lateral	31	0.0184%
Sternoclavicular (Joints) Each View	29	0.0172%
Bilateral Forearms	26	0.0154%
Bilateral Knees (30 Degrees Lateral)	25	0.0148%
Patellae (3 Axial)	24	0.0142%
Lumbosacral Spine - Right and Left Lateral Inclinations	16	0.0095%
Barium Enema	16	0.0095%
Long Bones	16	0.0095%
Small Bowel Transit with Loop Visualization	14	0.0083%
Cavum or Rhinopharynx	13	0.0077%
Bilateral Calcanei	12	0.0071%
Weight-bearing Knees (Frontal)	12	0.0071%
Cervical - Soft Tissues	12	0.0071%
Face	11	0.0065%
Bilateral Arms	10	0.0059%
Retrograde Urethrography	10	0.0059%
Gastro-Duodenal Transit	10	0.0059%
Fistulography	7	0.0042%
Bilateral Scapulae	7	0.0042%
Retrograde Urethrography with Cystography	6	0.0036%
Barium Enema via Colostomy	6	0.0036%
Descending Cystography - 3 Views (Obliques + Empty)	6	0.0036%

Ascending Cystography	4	0.0024%
Left Shoulder	3	0.0018%
Percutaneous Pyelography	3	0.0018%
Mandible	3	0.0018%
Chest Tomosynthesis	3	0.0018%
Right Shoulder	3	0.0018%
Paranasal Sinuses - 3 Views	3	0.0018%
Per-drain Cholangiography	2	0.0012%
Left Scaphoid Bone	2	0.0012%
Paranasal Sinuses - 2 Views	2	0.0012%
Sialography	2	0.0012%
Nasal Bones	1	0.0006%
Stump (Amputation)	1	0.0006%
Temporomandibular Joints, Open and Closed Mouth	1	0.0006%
Pharynx (With Contrast)	1	0.0006%
Knee Arthrography	1	0.0006%
Facial Bones - Waters View	1	0.0006%
Shoulder 1 View	1	0.0006%
Abdominal Plain Film, One View	1	0.0006%
Jejun Tube Placement	1	0.0006%
Paranasal Sinuses - Waters View	1	0.0006%
Unilateral Hip, Two Views	1	0.0006%

Table C.4: Counts of medical acts of X-ray radiological data of Site 2, before categorical outlier removal.

Medical Act	Count	Percentage
Chest, one view	39766	22.3126%
Radiology, additional fee for inpatient exam	16038	8.9989%
Radiology, additional fee for pediatric exam	10849	6.0874%
Chest, two views	9673	5.4275%
Pelvis	7690	4.3148%
Radiology, support in operating room	5985	3.3582%
Weight-bearing knees, two views	5198	2.9166%
Lumbar spine, two views	4253	2.3863%
Fluoroscopy control	4244	2.3813%
Full-leg radiograph, extra-long film	3732	2.0940%
Right foot, weight-bearing, two views	3190	1.7899%
Left foot, weight-bearing, two views	3160	1.7731%
Cervical spine, two views	3065	1.7198%
Left knee, two views	2885	1.6188%
Lumbosacral spine, hyperflexion and hyperextension	2873	1.6120%
Right knee, two views	2841	1.5941%
Axial study of patella, one view	2561	1.4370%
Spine radiograph, extra-long film, two views	2323	1.3034%
Left wrist, two views	2248	1.2613%
Right hand, two views	2177	1.2215%
Weight-bearing lumbosacral spine, two views	2177	1.2215%
Weight-bearing pelvis	2121	1.1901%
Right wrist, two views	2093	1.1744%
Left hand, two views	1948	1.0930%
Right shoulder, two views	1891	1.0610%
Cervical spine, hyperflexion and hyperextension	1808	1.0145%
Dorsal spine, two incidences	1782	0.9999%
Simple abdomen, one incidence	1699	0.9533%
Left shoulder, two incidences	1588	0.8910%
Right unilateral hip, two incidences	1248	0.7003%
Left unilateral hip, two incidences	1179	0.6615%
Left ankle, two incidences	1147	0.6436%
Right ankle, two incidences	1135	0.6368%
Right foot, two incidences	1127	0.6324%

Left foot, two incidences	1034	0.5802%
Left elbow, two incidences	1030	0.5779%
Right ankle weight-bearing, two incidences	1011	0.5673%
Left ankle weight-bearing, two incidences	1004	0.5633%
Right elbow, two incidences	933	0.5235%
Right unilateral hip, one incidence	879	0.4932%
Left unilateral hip, one incidence	803	0.4506%
Axial study of patella, three incidences	674	0.3782%
Bone age (left hand and wrist)	619	0.3473%
Bilateral hip, one incidence	598	0.3355%
Right thigh, two incidences	592	0.3322%
Left hand fingers, two incidences	579	0.3249%
Right hand fingers, two incidences	570	0.3198%
Left thigh, two incidences	543	0.3047%
Left leg, two incidences	440	0.2469%
Bilateral hip, two incidences	439	0.2463%
Right leg, two incidences	430	0.2413%
Right shoulder, one incidence	408	0.2289%
Sacroiliac joints, one bilateral incidence	401	0.2250%
Left shoulder, one incidence	335	0.1880%
Right arm, two incidences	313	0.1756%
Left sacroiliac joint, one incidence	309	0.1734%
Right sacroiliac joint, one incidence	304	0.1706%
Left forearm, two incidences	292	0.1638%
Left arm, two incidences	277	0.1554%
Right forearm, two incidences	269	0.1509%
Bilateral hip weight-bearing, one incidence	263	0.1476%
Left clavicle, two incidences	263	0.1476%
Right clavicle, two incidences	231	0.1296%
Lumbar spine, four incidences	226	0.1268%
X-ray of spine, extra-long film, one incidence	214	0.1201%
Pharyngeal cavity, one incidence	208	0.1167%
Foot, two incidences	179	0.1004%
Wrist, two incidences	175	0.0982%
Right calcaneus, two incidences	170	0.0954%
Lateral cephalometric skull X-ray	160	0.0898%
Left calcaneus, two incidences	158	0.0887%
Stereotaxy, support for intervention gestures	145	0.0814%
Simple pelvis X-ray	139	0.0780%
Right foot fingers, two incidences	138	0.0774%
Left unilateral hip weight-bearing, one incidence	136	0.0763%
Ankle, two incidences	134	0.0752%
Shoulder blade, two incidences	133	0.0746%
Right unilateral hip weight-bearing, one incidence	116	0.0651%
Elbow, two incidences	113	0.0634%
Skeleton (child)	111	0.0623%
Left foot fingers, two incidences	100	0.0561%
Acromioclavicular joint, bilateral	96	0.0539%
Radiology, additional information associated with performing contrast examinations on children	95	0.0533%
Hand, two incidences	93	0.0522%
Fingers of the hand, two incidences	89	0.0499%
Sacrum and coccyx, two incidences	80	0.0449%
Shoulder, two incidences	77	0.0432%
Left rib cage, two incidences	77	0.0432%
Cervical spine, four incidences	76	0.0426%
Knee, two incidences	73	0.0410%
Right rib cage, two incidences	69	0.0387%
Unilateral hip, one incidence	67	0.0376%
Right clavicle, one incidence	67	0.0376%
Dorso-lumbar transition, two incidences	66	0.0370%

Simple abdomen, each additional incidence	64	0.0359%
Lumbosacral spine, lateral inclinations	63	0.0353%
Left clavicle, one incidence	61	0.0342%
Lower limbs of child, two incidences	59	0.0331%
Acromioclavicular joint right, each incidence	54	0.0303%
Left shoulder blade, two incidences	52	0.0292%
Skull, two incidences	46	0.0258%
Right shoulder blade, two incidences	43	0.0241%
Acromioclavicular joint left, each incidence	42	0.0236%
Sacral spine, one incidence	30	0.0168%
Thigh, two incidences	24	0.0135%
Cervical spine, lateral inclinations	21	0.0118%
Skeleton (adult) (includes the following incidences)	21	0.0118%
Lumbosacral hinge, two incidences	21	0.0118%
Left rib cage, one incidence	20	0.0112%
Right shoulder blade, one incidence	20	0.0112%
Right rib cage, one incidence	19	0.0107%
Foot fingers, two incidences	19	0.0107%
Right sternoclavicular joint, two incidences	19	0.0107%
Lower limbs, each additional incidence	18	0.0101%
Cervicothoracic transition, two incidences	18	0.0101%
Left shoulder blade, one incidence	17	0.0095%
Calcaneus, two incidences	13	0.0073%
Craniovertebral hinge, two incidences	11	0.0062%
Leg, two incidences	11	0.0062%
Left sternoclavicular joint, two incidences	11	0.0062%
Neck soft tissues, two incidences	9	0.0050%
Sternum, two incidences	9	0.0050%
Skull, one incidence	9	0.0050%
Facial bones, two incidences	8	0.0045%
Jaw, one incidence	7	0.0039%
Rib cage, two incidences	7	0.0039%
Sternum, one incidence	6	0.0034%
Forearm, two incidences	6	0.0034%
Lower limb radiometry by segment	5	0.0028%
Neck soft tissues, one incidence	5	0.0028%
Right upper limb of child, minimum two incidences	4	0.0022%
Arm, two incidences	4	0.0022%
Left upper limb of child, minimum two incidences	4	0.0022%
Radiology, additional iodinated contrast for urological examination	4	0.0022%
Radiography, support in Radiology service	2	0.0011%
Upper limbs, each additional incidence	2	0.0011%
Orbit, two incidences	2	0.0011%
Sacroiliac joints, one unilateral incidence	2	0.0011%
Radiography of surgical specimen	2	0.0011%
Paranasal sinuses, two incidences	2	0.0011%
Facial bones, one incidence	1	0.0006%
Atlas and axis, two incidences	1	0.0006%
Temporomandibular joint, open and closed mouth	1	0.0006%
Right internal auditory canal, two incidences	1	0.0006%
Sella turcica	1	0.0006%
Lumbar discography	1	0.0006%
Radiology, additional iodinated contrast for gynecological examination	1	0.0006%
Left internal auditory canal, two incidences	1	0.0006%
Pelvimetry	1	0.0006%