Modelling a Quality-Oriented Evaluation System for Primary Health Care Teams in Portugal

José Pedro Gama de Sousa Rodrigues
Master in Data Science
Department of Computer Science
Faculty of Sciences of the University of Porto
2025

Modelling a Quality-Oriented Evaluation System for Primary Health Care Teams in Portugal

José Pedro Gama de Sousa Rodrigues

Dissertation carried out as part of the Master in Data Science

Department of Computer Science

2025

Supervisor
Paulo Santos, Associate Professor, Faculty of Medicine of the University of Porto

Co-supervisor
Inês Dutra, Assistant Professor, Faculty of Sciences of the University of Porto

Dedicated to my family

Acknowledgements

I want to give my deepest appreciation to my supervisors, Prof. Ines and Prof. Paulo. Throughout this last year, they were the main reason I was able to develop this arduous project into a dissertation that gives me confidence and reassurance about my capabilities as a Data scientist. I will be forever grateful for their time and commitment in correcting my mistakes, answering my doubts, and encouraging me to perform to the best of my abilities.

I am eternally grateful to my family. Everything I have accomplished, and will in the future, is thanks to the opportunities, love, and overwhelming support that they have given me throughout my life.

Finally, I want to thank Maria for being the person I endlessly depend on. Thank you for being my motivation and inspiration, guiding me to believe in myself and being the reason I strive to improve every day.

Resumo

Introdução: A monitorização dos Cuidados de Saúde Primários (CSP) é essencial para a adaptação aos desafios e às constantes mudanças na área da saúde. Atualmente, o Índice de Desempenho de Equipa (IDE) é o sistema responsável pela avaliação do desempenho de equipas multiprofissionais nas unidades de saúde. Este sistema de avaliação apresenta várias fragilidades, nomeadamente a fraca evidência científica no desenvolvimento da matriz de indicadores responsável pelo seu cálculo. Adicionalmente, a matriz incentiva profissionais de saúde a terem como objetivo o cumprimento de uma lista de tarefas, ao invés da melhoria contínua de qualidade em unidades de saúde. O trabalho desenvolvido nesta dissertação tem como objetivo criar um sistema de avaliação alternativo, baseado em qualidade. Este sistema utiliza *odds ratios* para desenvolver associações entre determinados parâmetros e os Indicadores de Resultado – Indicador 410, 372, 365, 339.

Metodologia: As investigações foram desenvolvidas em R e Python. Foram aplicadas Regressões Lineares e Correlações para investigar as relações existentes entre os indicadores de resultado, o IDE, e as variáveis de 'Estrutura e demográficas'. Foram estudados relatórios da OCDE e diversos artigos para identificar médias internacionais e categorizar os Indicadores de Resultado. Foram analisadas múltiplas metodologias de *Feature Selection*, com o intuito de identificar o grupo de parâmetros mais influente para cada indicador de resultado. As associações *Odds Ratio* foram desenvolvidas através de modelos de regressão logística, exponenciando os coeficientes e os seus intervalos de confiança.

Resultados: A comparação do IDE com os Indicadores de Resultado revelou que o IDE não descreve qualidade de forma correta. Por outro lado, as variáveis de 'Estrutura e demográficas' tiveram uma relação significativa com os indicadores de resultado e foram adicionadas à análise de *Feature Selection*. De todas as metodologias analisadas, os grupos de variáveis com a seleção mais consistente e influente foram obtidos através do *Stepwise Backwards*, *LassoCV* e *SHAP*. As associações *Odds Ratio* mostraram que os modelos mais confiáveis foram produzidos com os Indicadores de Resultado 410 e 365. As associações também reforçaram a importância da monitorização proativa de doentes com problemas crónicos, de modo a reduzir a intensidade de utilização da rede de urgência hospitalar.

Conclusão: A dissertação demonstrou a necessidade de uma revisão total de todos os indicadores atualmente ativos. Adicionalmente, é importante incentivar profissionais a

aprofundar os conhecimentos sobre como registar os indicadores de uma forma adequada. O sistema desenvolvido nesta dissertação tem como objetivo reformular o IDE, criando um modelo baseado em qualidade. Este modelo permitirá orientar a distribuição de recursos, promovendo uma melhoria contínua das unidades de saúde, beneficiando a nossa comunidade.

Palavras-chave: [Cuidados de Saúde Prímarios, IDE, Qualidade, Indicadores de Resultado, *Odds Ratio*, Unidade de Saúde]

Abstract

Introduction: The monitoring and evaluation of PHC is essential to adapt and overcome any challenges that come with an ever-changing landscape in healthcare. Currently, the IDE system is responsible for the performance assessment of PHC in Portugal. This system has several issues, among them the lack of credible scientific evidence when creating the indicator matrix. The IDE also placed a strong emphasis on policing work rather than on continuous improvement, a deviation from the original mission and spirit that guided our pioneers who developed the 2005 reform. In this dissertation, the work aims to create an alternative team performance evaluation system that assesses PHC in Portugal in terms of quality. To that end, we developed this evaluation system using odds ratio associations based on the four Outcome Indicators that were being tracked at the time of data collection – Indicators 410, 372, 365, and 339.

Methodology: The investigations were performed using R and Python. MLR and Correlation plots were used to investigate relationships between Outcome Indicators, the IDE, and the 'Structural and Demographic' variables. Several articles and OECD reports were used to track international averages, which would act as thresholds in the categorisation of the Outcome Indicators. Various Feature Selection Methods were performed to gather the most consistent set of features. Odds ratio associations were created by developing a Logistic Regression model and exponentiating the coefficients and their confidence intervals.

Results: The comparison between IDE scores and Outcome Indicators showed that IDE does not reflect quality accurately. The 'Structural and Demographic' variables analysis showed that they were significant to the Outcome Indicators and should be included in the Feature Selection process. The Feature Selection section revealed that the best performing and most consistently stable sets came from Stepwise Backwards Feature Selection, LassoCV, and SHAP. The Odds Ratio models showed that the Outcome Indicators with the best performance were Indicators 410 and 365. In addition, these models showed associations that reinforced the importance of proactive monitoring in detecting issues before emergencies could occur.

Conclusion: The dissertation conveyed the necessity for a thorough review of all the current active indicators. In addition, there should be an increase in healthcare professionals' awareness of correctly registering the indicators' information. The evaluation system proposed in this dissertation aims to restructure the IDE, shifting the

current PHC perspective in Portugal from a checklist mentality to one based on quality assessment. This model would guide healthcare units towards allocating resources that reinforce continuous improvement for the betterment of our community.

Keywords: [Primary Health Care, IDE, Quality, Outcome Indicator, Odds Ratio, Healthcare Units]

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ACES	HEALTH CENTRE CLUSTERS (AGRUPAMENTOS DE CENTROS DE SAÚDE
AIC	AKAIKE INFORMATION CRITERION
ARS	ADMINISTRAÇÃO REGIONAL DE SAÚDE
BIC	BAYESIAN INFORMATION CRITERION
CHF	CHRONIC HEART FAILURE
CMZ	US CENTRES FOR MEDICARE AND MEDICAID SERVICES
COPD	CHRONIC OBSTRUCTIVE PULMONARY DISEASE
DF	DEGREES OF FREEDOM
DHCPR	DUTCH HEALTH CARE PERFORMANCE REPORT
DM	DIABETES MELLITUS
FL	FIXED LENGTH
FS	FEATURE SELECTION
HCU	HEALTHCARE UNIT
ICS	SOCIAL DEMOGRAPHIC CONTEXT INDEX (INDICE DE CONTEXTO SOCIODEMOGRÁFICO)
IDE	TEAM PERFORMANCE INDEX (ÍNDICE DE DESEMPENHO DE EQUIPA)
IGZ	HEALTHCARE INSPECTORATE
LOGREG	LOGISTIC REGRESSION
LRT	LIKELIHOOD RATIO TEST
LVT	LISBOA VALE DO TEJO
MLR	MULTIPLE LINEAR REGRESSION
NHS	BRITISH NATIONAL HEALTH SERVICE
NICE	NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE
NZA	DUTCH HEALTHCARE AUTHORITY
OECD	ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT
OR	ODDS RATIO
P4P	PAY FOR PERFORMANCE
PHC	PRIMARY HEALTH CARE
QOF	QUALITY AND OUTCOME FRAMEWORK
R^2	R-SQUARED

RF	RANDOM FOREST
RRE	REGIME REMUNERATÓRIO EXPERIMENTAL
RSS	RESIDUAL SUM OF SQUARES
SE	STANDARD ERROR
SHAP	SHAPLEY ADDITIVE EXPLANATIONS
SIARS	ARS INFORMATION SERVICE (SISTEMA DE INFORMAÇÃO DA ARS)
SIM	SINDICATO INDEPENDENTE DOS MÉDICOS
SNS	SERVIÇO NACIONAL DE SAUDE
TSS	TOTAL SUM OF SQUARES
UCSP	UNIDADES DE CUIDADOS DE SAÚDE PERSONALIZADOS
UF	FUNCTIONAL UNITS (UNIDADES FUNCIONAIS)
USF	UNIDADE DE SAÚDE FAMILIAR
VL	VARIABLE LENGTH
WHO	WORLD HEALTH ORGANISATION
XAI	EXPLAINABLE A

1. Introduction

Primary Health Care (PHC) is not only the main focus of a country's health care system, but also an essential component of the social and economic development of its community. This is due to it being the first point of contact between the national health system and its population. The PHC of a nation has the objective of ensuring the highest possible level of health care, accounting for the continuous shifts in practices, while striving to deliver this care as early as possible [1][3][9].

The development of Primary Health Care in Portugal began with the creation of Health centres in 1971 through Government Decree No. 413/71 [2][5]. These first-generation health centres faced strong opposition due to political influences and a prevailing belief that care should be provided only to individuals with sufficient financial resources [2][7]. Following the revolution on April 25, 1974, and the establishment of democracy in Portugal, healthcare accessibility became a national right for all citizens residing in the country. This led to the development of Portugal's public healthcare system, *Serviço Nacional de Saúde* (SNS), enacted in 1979 [2]. In 1996 and 1999, the *Alfa* project and the *Regime Remuneratório Experimental* (RRE) would lay the groundwork for the breakthrough event that occurred in 2005, the PHC reform [4][13][14]. At the time, Portugal was still showing significant improvements and was a top performer in healthcare indicators compared to its international counterparts. Despite this, there was still a belief in continuous improvement and that conditions could be better. In particular, regarding the negative opinion on organisational services and on the accessibility of the health care units to the population [14]. This 2005 reform was responsible for the PHC organisational overhaul. Leading to the eventual implementations of Family Health Units (*Unidades de Saúde Familiar* - USF) and Health Centre Clusters (*Agrupamentos de Centros de Saúde* - ACeS), established in Decree-Law No. 298/07 [6]. And later, the creation of several Functional Unit types (*Unidades Funcionais* – UF), including the *Unidades de Cuidados de Saúde Personalizados* (UCSP), based on the more traditional models [15].

Currently, the SNS is distributed between multiple health care units with different purposes and structures: the reformed, modernised USF and the more traditional UCSP. UCSPs are health care units whose teams are composed of doctors, nurses, and administrators who provide personalised care, ensuring accessibility, continuity, and comprehensiveness. The UCSP healthcare units follow the older, more traditional structure, either due to not having the accessibility to develop USF units or due to their own preference. This makes them more constrained by not having autonomy regarding

management and objective planning. Similar to UCSP, the USF model also has the objective of providing continuous and accessible personalised care to citizens, but is more organizationally autonomous. A USF unit is divided into three structural models, each having a different degree of autonomy, financial structure, and incentive distribution. Model A is defined by teams still at a maturing phase, focusing on the learning and improvement of the team's performance and cooperation. The public administration sector sets the performance incentives, salaries, and regulations. USF Model B is recommended for teams with a greater organisational maturity, which is reflected through a more demanding level of performance targets. Model B personnel have a more performance-driven compensation, based on conditions stated in Chapter 6 of Decree-Law No. 298/2007 [6]. Model C is defined as a more experimental model aimed at supplementing any shortcomings in SNS. However, due to administrative obstructions and a lack of regulations, Model C units are still currently in a regulatory phase for implementation [2][3][10].

The 2005 reform also consolidated the assessment of performance as the main avenue for PHC development. Influenced by the Quality and Outcome Framework (QOF) from the British National Health Service (NHS), the reform states its accountability approach as an important point of reference for improving healthcare in Portugal. This approach requires transparency in the processes and outcomes of its units, creating a mechanism that could translate the quality of a unit through the evaluation of indicators [13][14]. According to the World Health Organisation (WHO), the monitoring and evaluation of PHC is essential to adapt and overcome any challenges that come with an ever-changing landscape in healthcare [1]. The monitoring of these indicators also enables specialists to determine if implementations are working correctly. The development of USF led to the implementation of a Pay for Performance (P4P) measure, which utilised a matrix of indicators to monitor quality and allocate incentives [10]. This methodology was used to create the current team performance evaluation system, *Índice de Desempenho de Equipa* (IDE), established in the Decree-Law No. 103/2023 and Ordinance No. 411-A/2023 [18]. The Pay for Performance evaluation system was designed to improve the quality of care by establishing incentives based on the accomplishment of specific performance metrics. There is a belief that the P4P model increases the motivation of professionals to improve their commitment and practice by focusing on specified parameters that improve the quality of care and patient-doctor relationships. On the other hand, it can induce ethical problems. According to a study related to the consequences of a P4P in Burkina Faso, the model led to an increase in registration manipulation. It

also introduced issues related to funnelling focus solely on problems referenced in the matrix while neglecting less profitable needs [8][11].

Regarding Portugal's IDE system, there are issues related to the development of the indicator matrix table. There is a lack of significant scientific background in the development of the indicators and the different goals specified in the matrix. Additionally, there is a shortage of meaningful indicator diversity, limiting the matrix's capability to demonstrate whether a unit represents good practice. The matrix also exhibits organisational rigidity, highlighting the lack of foresight related to sociodemographic and regional differences. This is due to every condition specified in the matrix being established solely through the contribution of the Governmental bodies responsible for Public Administration, finances, and health. This created a separation of intent between incentive distribution and healthcare quality, with many believing the matrix is more guided to policing work rather than improving healthcare. There is also growing sentiment that this system contributes to a checklist mentality in Portugal, where the healthcare professionals are only incentivised to focus on reaching the goals present in the indicator matrix. The consensus is that the reform has stagnated throughout the years, favouring the current system of funnelling the focus and checking tasks off a list, rather than the original mission and spirit that guided the motivation that originated the 2005 reform: a system that was developed for the betterment of the community and its people, through the continuous growth and development of healthcare quality [2][3][7][10][12][15][18].

In this dissertation, the work aims to create an alternative team performance evaluation system that assesses PHC in Portugal. This system strives to address the current missteps and the different objectives that need to be considered in developing an evaluation tool that acts in accordance with the principles of our healthcare system and the revolutionary pioneers who have shaped it. This alternative quality-oriented system evaluates team performance through Outcome Indicators.

The distinction between indicator type (Processes, Structure, Outcome) is not currently acknowledged in the IDE, which creates misinterpretations and leads to the inclusion of outcome variables in the matrix [18]. Outcome Indicators are representations of the healthcare quality and are influenced by the conditions and various processes within the unit. Determining the right indicators to represent the true quality of a unit is crucial for the evaluation system, since it also defines the direction in which the system should improve. To that effect, we developed this model using the four Outcome Indicators that were under observation at the time of data collection, using them as measuring tools and

as the main source of indication for priorities and associations. Below are the four Outcome Indicators used in the dissertation [16]:

Indicator 410

SIARS CODE: 2018.410.01 FL

Name: Adjusted annual rate for frequent use of hospital emergency services.

Objective: Monitor the intensity of use of the hospital emergency network by users enrolled in primary health care. The formula calculates the proportion of PHC service users who have four or more episodes requiring urgent healthcare in the last 12 months.

Table 1. Formulation, unit of measurement, state of indicator and output for Indicator 410 [16].

Formula	Unit	Output	State
AA/BB x 100	Per 100	Adjusted Rate	Ongoing

Indicator 372

SIARS CODE: 2017.372.01

Name: Rate of Hospital admissions due to falls with femoral neck fracture

Objective: Monitor the impact of the actions of the different PHC healthcare units on reducing falls in the population aged 75 and over. The formula calculates the proportion of service users aged 75 or above who had hospital admissions due to falls resulting in a femoral neck fracture in the last 12 months

Table 2. Formulation, unit of measurement, state of indicator and output for Indicator 372 [16].

Formula	Unit	Output	State
AA/BB x 100,000	Per 100,000	Admission Rate (per 100,000 registered users per year)	Ongoing

.

Indicator 365

SIARS CODE: 2017.365.01

Name: Rate of avoidable hospitalisations in the adult population (adjusted for a standard population)

Objective: Monitor the effectiveness of care provided by PHC to patients with asthma, COPD, pneumonia, CHF, angina pectoris, hypertension, and diabetes, in their symptomatic control, and in the prevention of complications and exacerbations, using the metric "hospital admission rate". The formula calculates the ratio of avoidable hospitalisations for every 100,000 adult service users.

Table 3. Formulation, unit of measurement, state of indicator and output for Indicator 365 [16].

Formula	Unit	Output	State
AA/BB x 100,000	Per 100,000	Admission Rate (per 100,000 residents per year)	Ongoing

Indicator 339

SIARS CODE: 2018.339.01 FL

Name: Annual adjusted rate of hospital emergency episodes

Objective: Monitor the intensity of use of the hospital emergency network by service users enrolled in primary health care. The formula calculates the annual rate of hospital emergency episodes per 100 service users, adjusted for the standard population distribution by sex and age.

Table 4. Formulation, unit of measurement, state of indicator and output for Indicator 339 [16].

Formula	Unit	Output	State
AA/BB x 100	Per 100	Adjusted Rate	Ongoing

The use of numerous Outcome Indicators leads to a more complex interpretation than the current 0-100 score for the IDE. Nevertheless, this evaluation system allows for a more mathematical and credible approach, using real data and uncovering the true relationships between the indicators and the outcomes, better representing the quality of a unit. This also helps narrow down the indicators to their most impactful for each

outcome, as opposed to the current combination of indicators based on the opinions of different governmental bodies. International benchmarking is also performed through international averages identified through the Organisation for Economic Cooperation and Development (OECD) and other international articles. This allows for the differentiation of the best-performing units from a global perspective. The associations between the indicators and the best-performing units for each Outcome Indicator are then reflected through an Odds Ratio table. To avoid the funnelling of focus on specific parameters and to have the most up-to-date associations possible that discriminate the peak units from the rest, the process of Feature selection and odds ratio development should be performed annually.

This evaluation system is guided by the original sentiment of reforming our healthcare system, emphasising a greater focus on improving healthcare units, considering different expert opinionated pieces and international counterparts as references. In addition to a new evaluation system, this dissertation also questions the extent to which the IDE score reflects the inherent quality of a unit, and whether different structural and demographic conditions of a unit should be included in the evaluation system.

2. Basic Concepts

In this chapter, the foundations and theoretical concepts of the work developed in the dissertation are summarised. These include: the foundation of the IDE and its indicator matrix; the mathematical background for Multiple Linear Regression and Logistic Regression; Different Feature Selection methods that were used in the dissertation, including SHAP thresholding, Wald, Stepwise Backwards, Lasso Regression, Random Forest importance, and Permutation Importance; The Classification performance metrics and the Jaccard Similarity.

2.1. IDE

The IDE is a team performance evaluation system developed in the Decree-Law No. 103/2023 and Ordinance No. 411-A/2023 [18]. It is a 0-100 score with the objective of monitoring and incentive allocation for UCSP and USF unit types. The IDE is based on the work carried out by the multiprofessional team, taking into account citizen accessibility, health and disease management, prescription qualification, and health care integration. The calculation of the IDE score is based on a table found in Ordinance 411-A (Additional Information, Table A1), and has several components impacting the score: Indicators, Importance Weight, Expected Values, and Acceptable Variance, as described next.

- **Indicators:**

The indicators address a specificity about a unit. These values are used for a healthcare unit's monitoring and contractualization objectives, and are the building blocks that generate the IDE score. Each available indicator has its own objective and reasoning, codification system, calculation formula, unit of measurement, and expected value.

The ARS Information Service (SIARS) code is used for identification purposes in the SNS database. It is made up of three sets of numbers separated by a period. The first set corresponds to the year the indicator was introduced, the second constitutes the sequential number where the indicator was introduced in the database, and the final number indicates the version of the specific indicator. The code is followed by an alphabetical sign that explains the measurement period for the indicator, Fixed length (FL – Annual/ Semiannual) or Variable Length (VL – Monthly). As an example, the code **2018.410.01 FL** corresponds to the year 2018 for Indicator 410 version 01 with fixed length.

The formulation of each indicator usually contains the same patterns when creating its metric. They are generated through a formula containing numerators and denominators based on specific conditions for the indicator being upheld (AA/BB). Despite this, every variable can belong to a different unit of measure, as stated below:

- Indices – Range from 0 to 1
- Proportions and Ratios – Range from 0 to 100
- Costs, Expenses, and Rates – Theoretical range from 0 to infinity
- Score values – Range from 0 to 2

Each ID also informs the user about the current state of the indicator, since monitoring/ contractualization indicators can become outdated and are discontinued. These indicators are also designated into three main components related to the state of a healthcare unit: Structural, Process, and Outcome Indicators.

Of the 508 indicators available in the SNS database, only 129 were being tracked and collected during 2023; the rest of the indicators were either discontinued or still in development.

In addition, the table found in Ordinance 411-A states that only 43 should be accounted for in the calculation of the IDE scores. This number was narrowed down using three general assumptions [16]:

- The indicators were independent among themselves;
- Had an ID database page that had a clear description and simple formulation;
- Had a history of at least 2 years under observation

- **Importance Weight:**

Each of the 43 indicators accounted for in the table had an assigned importance weight. These importance weights add up to 100 and enable a clear understanding of which indicators a health care unit should prioritise and how the effort should be distributed among them.

- **Expected Values:**

The expected values were calculated based on the distribution of values in different units at a national level, without incorporating national or internationally established quality standards. They represent an assumption by the administration about what results are expected to be obtained in each assessment indicator.

- **Acceptable Variance:**

The acceptable variance refers to the deviations from expected values that were technically admissible to constitute the health care unit as conducting a good practice for that indicator [18].

2.2. Multiple Linear Regression

Multiple Linear Regression (MLR) is a linear model that showcases the values of a dependent variable explained by a combination of multiple independent variables. Geometrically, Simple Linear Regression creates a linear formula to explain the relationship between 2 variables (Fig.1). MLR also creates a linear formula, but in an n-dimensional space of X_n independent variables (Fig.2), due to a target variable needing more than one predictor variable to best explain their behaviour [19][20].

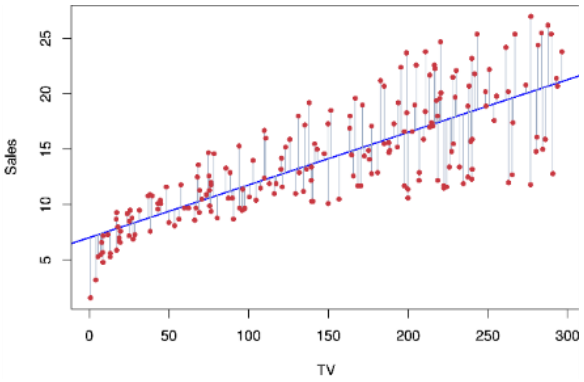

Figure 1. Simple Linear Regression between Sales and TV [20]

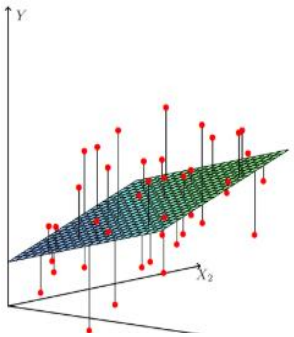

Figure 2. 2-Dimensional hyperplane of an MLR [19]

The MLR formula has n observations and k predictors. It consists of Y as the dependent variable, β_0 and β_j as the regression coefficients, x_{ij} as the value of the independent variable at observation i and predictor j, and the model error (ϵ_i):

$$Y = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_{i,} \qquad i = 1, \ldots, n$$

The Regression Coefficient β_0, also known as the intercept, is the value of the target when all the variables are equal to 0. The rest of the regression coefficients can be estimated in a variety of ways. The MLR models were mainly performed in RStudio with the lm() function, using the least squares method to minimise the residual sum of squares (RSS) and represent the line with the best fit [19][20].

$$L = \sum_{i=1}^{n} \in_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2$$

L is minimised with respect to β and must satisfy

$$\left. \frac{\partial L}{\partial \beta_j} \right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) = 0$$

The RSS is the cost function of the model's line and calculates the sum of squared differences between the real value and the predicted results calculated by the model. It allows the assessment of how the line fits the model.

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The MLR model has several assumptions that are relied upon for more accurate results and a lower probability of Type I/Type II errors. The first assumption is that variables are assumed to be normally distributed, since non-normality can affect significance tests and alter relationships. There is also an assumption of linearity between independent and dependent variables, resulting in the possibility of underestimating non-linear relationships. The observations are also assumed to be independent of each other. The last assumption is homoscedasticity, meaning the errors have the same variance among them; the absence of homoscedasticity may imply the possibility of a need to include additional dependent variables [22][23].

To evaluate an MLR model using R, several conditions are observed. Among them are the variation of the residuals, the F-statistic, the p-value of the linear model, the R^2, and the coefficient analysis. The summary output from a linear model made in RStudio showcases the main evaluation topics to address in a linear regression model. This also checks theoretical hypothesis tests that are usually performed when addressing an MLR model, such as understanding relationships among the variables and looking for the occurrence of non-zero coefficients.

The summary of the linear model below is one that was investigated in Chapter 5.1 (Fig.14). The coefficients of the lm() model are then displayed. This feature mainly demonstrates the estimate, standard error (SE), t-value, and P-value for each independent variable. The estimate column is the β coefficient that is estimated through the least squares method explained in the paragraphs above [24][25].

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.79262    0.02637  30.056  < 2e-16 ***
Ind_365     -0.10172    0.04721  -2.154 0.031498 *
Ind_339     -0.86968    0.24551  -3.542 0.000419 ***
Ind_410      0.66148    0.26314   2.514 0.012129 *
Ind_372     -0.21988    0.05957  -3.691 0.000238 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The standard error measures how precisely the model estimates the unknown coefficient value. The lower the Standard error, the better and more precise the coefficient is. The standard error of a coefficient is calculated using the variance of the error(σ^2), the independent variable at observation i (x_i), and the mean ($\overline{x}$) [24][25].

$$SE\left(\hat{\beta}\right)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The t value calculates the number of standard deviations a coefficient is from 0. A t-statistic above 1 is an indication that the null hypothesis could be rejected and there is a high probability of a relationship between the dependent and independent variables. The probability of observing a value equal to or greater than t is calculated through the p-value. A smaller p-value generally means that it is unlikely that any perceived association between dependent and independent variables occurs by accident. The threshold for a p-value that demonstrates association is usually around 0.05 (5%). The significant codes are markers to easily interpret the extent of the association between an independent variable and the dependent variable [24][25].

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

The R-squared (R^2) of a model measures the proportion of variance in the dependent variable explained by the independent variables, assigning a value between 0 and 1. It can be defined as the Multiple R^2 or Adjusted R^2. For MLR, Adjusted R^2 is the statistic that is usually focused on since it accounts for the general increase in R^2 caused by a higher number of parameters [24][25].

$$R^2 = 1 - \frac{RSS}{TSS}$$

The p-value for the whole model indicates whether there is a statistically significant relationship between the dependent and the set of independent variables. A p-value larger than 0.05 suggests that there is no significant evidence of at least one non-zero coefficient.

```
Residual standard error: 0.2065 on 838 degrees of freedom
Multiple R-squared:  0.07383,   Adjusted R-squared:  0.06941
F-statistic:  16.7 on 4 and 838 DF,  p-value: 3.53e-13
```

An MLR model can be used for several different purposes. It allows an understanding of whether there are relationships between variables, to what extent specific features are important, and how well these variables explain the target output [26].

2.3. Logistic Regression

A Logistic Regression (LogReg) model tries to address a problem by modelling a dichotomous outcome linearly. In other words, while a linear regression model predicts a continuous variable y, a logistic regression model predicts a categorical y target variable. Since the problem is categorical, the formula demonstrates the relationship between variables through a probability rather than modelling the problem as is. This is designated through the log odds (logit) of an event occurring [24][26][27].

$$\log\left(\frac{P(Y_i = 1 \mid x_{i1}, \ldots, x_{ik})}{1 - (Y_i = 1 \mid x_{i1}, \ldots, x_{ik})}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

$$p(x_i) = P(Y_i = 1 \mid x_{i1}, \ldots, x_{ik}) = \frac{\exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij})}$$

This linear function represents an example of a sigmoid function, a core formula in Binary Logistic Regression. The sigmoid forces the probabilities and maps the real values into one between 0 and 1, creating an S-curve (Fig. 3). In the case of LogReg, the distribution of target values can be made through a threshold, set by the model (usually 0.5), forcing a dichotomous outcome depending on its value. Geometrically, a multiple Logistic Regression also fits in an n-dimensional space plane, but rather than creating a straight line, it creates a curve (Fig.4) [24][26][28][29].

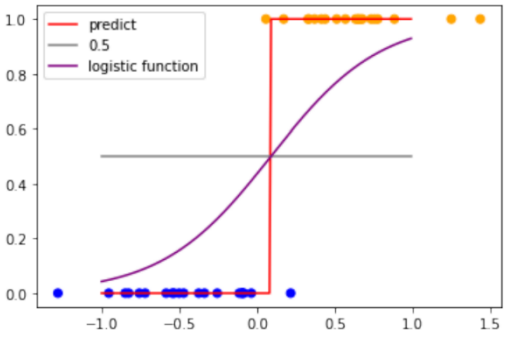

Figure 3. Sigmoid Function of a Simple Logistic Regression

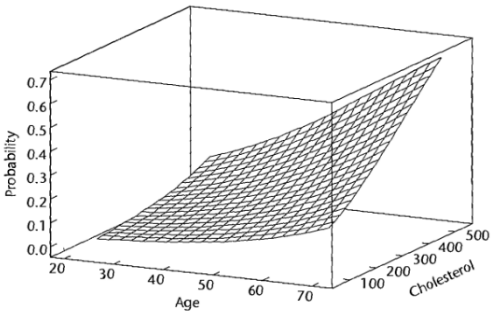

Figure 4. Multiple Logistic Regression, 2-dimensional hyperplane [28]

In terms of assumptions, there are some parallels with MLR, among them the independence of observations and the linearity between the dependent and independent variables, although only the log odds are required to be linearly related. A LogReg model also assumes no multicollinearity among independent variables and benefits from datasets containing larger sample sizes [29].

Since the traditional least squares approach used in MLR would be impossible due to its dichotomous nature, the β coefficients for Logistic regression are estimated using the Maximum Likelihood approach. The maximum likelihood approach uses the likelihood function to estimate the β coefficients that assign a higher probability for events that occur (outcome 1) and the inverse for events that do not occur (outcome 0) [26][24][29].

$$L(\beta_0, \ldots, \beta_k) = \prod_{i=1}^{n} [p(x_i)^{y_i} \, (1 - p(x_i))^{1-y_i}]$$

To evaluate the strength of a Logistic Regression through variation explained is a difficult task, since the standard R^2 statistic and the sum of squares concept from Linear Regression do not apply. This led to several statisticians creating different pseudo-R^2 formulations that have their own strengths and weaknesses, the most recognised one being the McFadden pseudo-R^2. The loglikelihood of a fitted model is represented using ℓ_1 and the loglikelihood of a null model with ℓ_0.

$$McFadden's\ Pseudo\ R^2 = 1 - \frac{\ell_1}{\ell_0}$$

The AIC and BIC scores attribute a value of 0-1 representing the balance between complexity and fit. Both scores measure the amount of information loss when

approximated to reality, with the lower values indicating the better AIC/BIC scores. Both models use a penalisation term that is relevant to the number of features present in the model, with BIC having a stricter penalty.

$$AIC = 2k - 2\ell_1$$

$$BIC = k\ln(n) - 2\ell_1$$

The Likelihood Ratio test is a hypothesis test that compares the goodness of fit between two nested statistical models, the fitted model and the reduced/null model. The LRT follows a chi-squared distribution, with the LRT function as the test statistic and the degrees of freedom being the difference in parameters between the models. The p-value is then identified, with values below 0.05 meaning that the more complex model has a better fit.

$$LRT = 2 * \ell_1 - \ell_{reduced}$$

$$df = k_1 - k_{reduced}$$

The use of log odds can make the interpretability of the estimated Logistic Regression coefficients confusing. To solve this problem, several medical research articles use odds ratios to understand the association between specific features and the binary target outcomes. Odds ratio quantifies the strength and direction of an association between a dependent and independent variable by comparing the odds of an event occurring with its counterpart [30][31][32].

Odds Ratio = Odds of event occurring/Odds of event NOT occurring

OR = 1 No association between the independent and outcome variable

OR > 1 Higher value in the independent variable **increases** outcome odds

OR < 1 Higher value in the independent variable **decreases** outcome odds

A high value odds ratio does not inherently mean that a variable is significant. The use of confidence intervals, p-values, and feature importance evaluations is needed to determine their significance. In terms of coefficients, since they are the change in log-odds of the outcome, the odds ratio of that specific variable is equal to the exponential of that coefficient [30][31].

$$OR_j = \exp(\beta_j) = e^{\beta_j}$$

EX: OR = Odds of heart disease for non-exercisers / Odds of heart disease for exercisers

= 0.25 / 0.11 ≈ 2.27

Non-exercise was positively associated with heart disease, with approximately 2.27 times the odds compared to people who exercise regularly [33].

2.4. Feature Selection

High-dimensional datasets can have problems with predicting a model accurately. One of the main reasons is that models with a high amount of features can be prone to overfitting. Overfitting occurs when a model is overly complex, causing it to train the data in excessive detail, capturing noise (irrelevant information) and negatively impacting the model's performance. Feature Selection methods are used to reduce dimensionality and the possibility of overfitting the model. There are three main types of Feature Selection: Filter, Wrapper, and Embedded methods. The Filter FS type highlights features based only on the intrinsic properties of the data. It evaluates importance separately from the classifier and ignores feature dependencies, but is computationally simple and fast (Ex, Wald). Wrapper FS type combines a model's hypothesis search with creating a subset. It is a search algorithm that evaluates various subsets of features and highlights the best-performing subset. This FS type is computationally expensive and is more prone to overfitting, but is tailored for the classification problem (Ex, Stepwise Regression). Embedded FS types are classifiers that have in-built feature selection. This FS type is not only tailored for the classification problem but also less computationally expensive than the wrapper (Ex, Lasso Regression) [83].

2.4.1. SHAP

SHAP (SHapley Additive exPlanations) is an Explainable AI (XAI) framework developed by Su-In Lee and Scott M. Lundberg in 2017. The XAI utilises SHAP values to help data scientists interpret and understand any machine learning model and its predictions [34][35]. SHAP is not necessarily a Feature Selection method. It is an explanatory model that can be modified to a feature reduction method by adding a threshold to the contribution value.

SHAP values are based on a game-theoretic approach and assign an importance value to a feature by calculating its contribution. It calculates the average contribution of player (feature) j in a game with k players (features), considering all possible permutations of subsets being formed. The weight in the formula ensures that every situation is accounted for according to how often it can occur. The marginal contributions indicate the extra value the player j contributes when joining subset S (can be positive or negative) [34][36][37].

Average Weight Marginal contributions

$$\emptyset_i(k,v) = \frac{1}{k!}\sum_{S \subseteq k/\{j\}} |S|!\,(k - |S| - 1)!\,[v(S \cup \{j\} - v(S)]$$

The SHAP values have four properties that help in the interpretation of the models. They are **symmetric**; players who contribute the same get the same SHAP value. They are **efficient**; the sum of each player's SHAP value is equal to the whole contribution made by the group. Have **Linearity**; the SHAP values of features with different model values can be summed linearly. Have **Missingness**; an individual that does not contribute has a SHAP value of 0, making it robust to missing data and irrelevant features.

The SHAP contributions and feature importance levels can be visualised through different plots that are referenced in the SHAP document. For both Binary Classification and Regression problems, the default summary plot is the beeswarm (Fig. 5). The beeswarm plot uses the SHAP value as the x-axis and features a colour gradient bar that highlights the difference between high (Red) and low (Blue) values in each analysed feature. The density of the points also gives us an understanding of the distribution of the data. The plot is separated by a line at SHAP value 0. For Binary classification, it is the threshold that distinguishes probabilities of being labelled as 0 (left) from being

labelled as 1 (right). For Regression models, it determines a score of confidence in predictions (0 to 1).

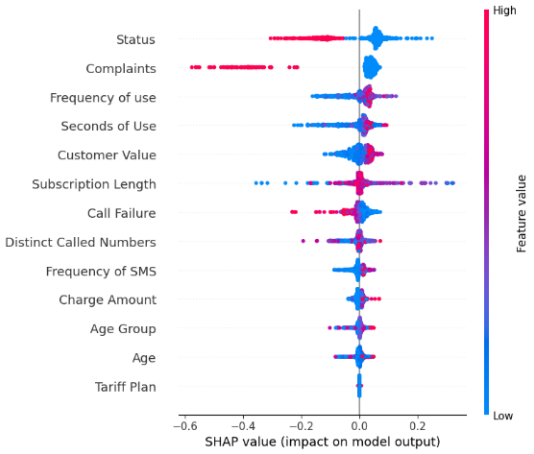

Figure 5. Example of a Beeswarm plot in SHAP Document [34][35]

2.4.2. Wald

The Wald Feature Selection method works by developing a model with every independent feature available. As stated previously, both the Linear Regression and Logistic Regression models contain p-values specific to each feature, allowing us to understand which features are significant and influence the model. Wald uses these significances and suggests creating a different model only with the variables that contained a p-value of 0.05 or lower. Some statisticians oppose the use of a p-value due to the belief that there is over-reliance on the understanding of statistical significance. One proposal is to reduce the p-value threshold to 0.005, but this would create its own issues, such as significantly increasing computational expenses and sample sizes of the model [24][26][38][39].

2.4.3. Stepwise Backwards

The Backwards Step Wise Feature Selection method reduces dimensionality by creating multiple models, starting with every feature available and removing one at a time. The scoring metric can be decided beforehand, with the most usual option being AIC. The model iteratively removes the least impactful feature until no feature that can be removed enhances the model. The stepwise method is a viable choice due to its fast automation and easy implementation. Despite this, it is considered a controversial subject due to several limitations [24][26][40]:

- The assumptions can be violated

- Susceptible to ignoring relevant features due to its automation
- Has a dependency on the order of the features being selected to remove, causing biases and inconsistencies

2.4.4. Lasso

Lasso is the primary regression method preferred for feature reduction. This is due to Lasso reducing the coefficients of less important features to 0, whereas its counterpart, Ridge Regression, only reduces coefficients towards 0. The nonzero coefficients of the Lasso model are then annotated, and a new model with fewer independent features is created.

Lasso Regression function minimises the sum of squares of residuals from Ordinary Least Squares, it also adds a regression penalty lambda (λ) that controls the extent to which the shrinkage is applied to its coefficients. This formula is represented in Fig. 6, the Lasso constraint is indicated as the Blue Diamond, while the RSS corresponds to the red curves. The minimum point is where the constraint and the function meet, which usually forces some coefficients to 0 (in this case, $\beta 1$).

$$\arg\min_{\beta_0,\beta}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{k}\left|\beta_j\right|\right\}$$

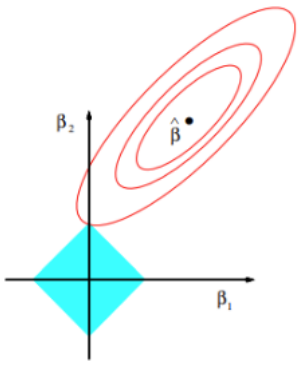

Figure 6. Visual Representation of Lasso Regression (red curve is the function, blue diamond is the constraint). [24]

2.4.5. Random Forest Importance

Random forest is an ensemble technique that averages the responses of Decision tree models differentiated by the random sampling of the dataset. This reduces overfitting and creates more accurate predictions that are less susceptible to outliers. The

predictions for Classification problems are based on majority voting (the label with the most label predictions is the RF model result) [24][26].

A useful feature in RF is the ability to calculate the average importance of each variable. This allowed different statisticians to develop a new method for reducing dimensionality in large datasets. It calculates the importance of a feature as the mean and standard deviation of the accumulation of impurity decreased within each tree. The impurity is calculated using the Gini impurity formula. D stands for the dataset with T classes, and p_c is the probability of a sample belonging to class c. The importances of each class are compared, with the lowest value being the selected split and purer option. This feature importance method comes with a severe limitation, as it can be biased for features with high cardinality [42][43].

$$Gini(D) = 1 - \sum_{c=1}^{T} p_c^2$$

2.4.6. Permutation Importance

Permutation importance is a model-agnostic feature importance technique. The technique uses random shuffling of a single feature and calculates its contribution by measuring the degrading effect on the score. It is usually an alternative to the Gini importance technique due to not being influenced by the cardinality of a model [42][44].

2.5. Model Evaluation

2.5.1. Classification Performance Evaluation

The evaluation of classification evaluation systems allows an understanding of how well a model fits the data and gives the ability to compare different models. One tool that is paramount to these evaluations is the Confusion Matrix (Table 5). The Confusion Matrix is a table for binary classification that compares correct decisions from incorrect ones and is composed of four terms that explain each option of a classification result:

- True Positives (TP): A correctly predicted positive outcome
- True Negatives (TN): A correctly predicted negative outcome
- False Positive (FP): A negative outcome that is incorrectly predicted as a positive, can be known as a Type I error
- False Negative (FN): A positive outcome that is incorrectly predicted as a negative, can be known as a Type II error

Table 5. Confusion Matrix [87]

		Predicted class		
		yes	*no*	Total
Actual class	*yes*	*TP*	*FN*	*P*
	no	*FP*	*TN*	*N*
	Total	P'	N'	$P+N$

These different terms are key building blocks for our evaluation measures, mainly the F1-score. The F1-score, also known as the harmonic mean, combines the precision and recall scores and gives equal weight to each measure. Precision measures the exactness of a model; it calculates the percentage of correctly predicted positive outcomes out of all positive predictions. Recall measures the completeness of a model; it calculates the percentage of true positive outcomes that were correctly predicted [87][88].

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

2.5.2. Jaccard Similarity

Jaccard similarity measures the similarity between two sets. This similarity score quantifies the overlap between two sets, making it particularly useful for stability analysis, due to its simplicity and length-resistant formulation. The Jaccard similarity measure is a 0-1 score that compares the instances where the two sets (A and B) were identical with the number of unique symbols in both sets (Fig. 7)[17][90].

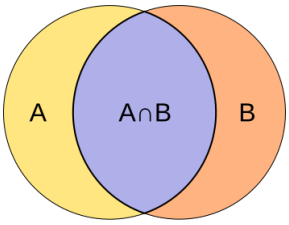

Figure 7. The intersection of two sets in a Venn diagram [90]

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

3. Related Work

In this chapter, different international team performance evaluation systems are analysed and compared with the IDE. In addition, various articles that include an odds ratio analysis in epidemiology or as an evaluation technique were explored.

3.1. International Evaluation Systems

The search for an ideal evaluation system for the performance of multiprofessional teams in health care has been thoroughly investigated worldwide. Internationally, different methodologies are explored, each with its advantages and disadvantages. There are two main strategies a country adopts to monitor the performance of a healthcare unit. Some countries opt for a system that is implemented nationwide, allowing governmental entities to compare performances and instances between units. Other countries go for a more segmented approach through region-specific systems or private frameworks. The subsequent OECD countries contain health care systems that are consistently ranked high in performance.

The United Kingdom's NHS and its Quality and Outcomes Framework (QOF) explore indicators and their effect on performance. It distinguishes itself from the previously referenced Portuguese IDE through its point system. This method works by highlighting different indicators and allocating a certain number of points according to their importance, with a maximum value of 635 points available. Points are attributed depending on performance, with incentives being distributed based on point accumulation [45][46]. An interesting feature in the QOF assessment is the annual refresh of the indicators. An independent institute, the National Institute for Health and Care Excellence (NICE), is responsible for reviewing the indicators and determining whether they are feasible for the system [47]. The development of indicators and their importance in the system also involves a team of Stakeholders, including organisations responsible for representing both patients and carers [48]. For consistency in indicator calculation and their specificities, the NHS developed conditions necessary to be upheld through a document labelled Business Rules [49]. Despite this, the issues between both the IDE and QOF systems remain the same, due to the strong effect on incentive allocation, which encourages professionals to achieve higher scores by focusing solely on these indicators. Medical professionals pursue the development of performance through a checklist mentality rather than by focusing on continuous care [50][2]. The long-term impacts on the quality of care after the establishment of the QOF were investigated through a systematic review conducted by JL Forbes, Lindsay, et al. In the

article, the authors concluded that there was limited evidence that the QOF indicated an increase in the quality of healthcare in a unit, and alternative methods of rewarding good practice could be more appropriate [51].

In Norway, the National Healthcare Quality Indicators have a greater focus on monitoring and improving healthcare rather than distributing funds. This is evident through the lack of a singular fixed scale that determines the quality of a unit, opting for the showcase of all 150 indicator scores. This evaluation system uses benchmarking from previously published indicator scores to show the progression of healthcare performance in the country. Similarly to the indicators calculated in Portugal, they are distributed among three different measurements: Structural Indicators, Process Indicators, and Outcome Indicators. These are then continually published, allowing comparison between them. The lack of a universal scale comes with its benefits and challenges. The system allows for complete transparency in every aspect of a unit, enabling a more personalised focus on particular indicators and their sectors. On the other hand, the total transparency of the indicators can lead to an overload of data that can influence a person to misunderstand their quality. The analysis requires specialised expertise to interpret the data due to its nuanced nature and complex interrelationships between indicators [52][53].

The monitoring of health care performance in the Netherlands is a joint contribution between the Dutch Healthcare Authority (NZA) and the Healthcare Inspectorate (IGZ). The NZA is more focused on cost-related issues in healthcare, nurturing efficiency through a more competitive outlook among healthcare units. The IGZ is responsible for monitoring the safety, quality, and accessibility of healthcare for the patients by imposing fines and penalties. The evaluation of the system comes every four years through the publication of the Dutch Health Care Performance Report (DHCPR). This report contains 125 indicators detailing the quality, accessibility, and costs of the Dutch Healthcare system. Similarly to Norway, it does not have a singular fixed scale quantifying the inherent quality of a unit. Despite this, the DHCPR compares its results using international benchmarking, comparing its performance with other countries [54][55].

Table 6. International systems and their liabilities

Country	Nationwide	System	Liabilities
Portugal	Yes	0-100 score	Checklist encouragement

United Kingdom	Yes	635 Max Point System	Checklist encouragement
Norway	Yes	National Benchmarking	Information Overload
Netherlands	Yes	International Benchmarking	Information Overload

The single, fixed-scale performance evaluation system, which generates a composite score by associating different performance metrics, is not a common evaluation tool used in most countries. This is mainly due to the risk previously stated, where units may divert their focus solely to the specific indicators established in the evaluation system. Other countries opt for a bigger emphasis on continuous internal improvement through benchmarking. The evaluation system also leads to an oversimplified view of the unit's quality, lacking the nuance that the benchmarking system provides, where theoretically, units with the same quality score could have polar opposite indicator scores.

3.2. Odds Ratio Systems

Odds ratio is a common measure used in the medical field to determine the strength of association between an event and its exposure to a variable [56]. The use of the odds ratio as a measuring tool is widely accepted in the medical field and epidemiology.

There are numerous instances of odds ratios being used in the medical field. Some articles investigate the association between a disease and an alleged exposure link. An example of this instance can be found in Remen, T. et al.’s article, where the authors used this metric as a measuring tool to investigate the risk relationship between lung cancer and several smoking metrics through a case-control study conducted in Montreal [57]. Odds ratios can also be used in studies that investigate associations of several risk factors and whether they are suitable for prediction. An example of this instance is in Zhou, Xu et. al.'s article “Multivariate logistic regression analysis of risk factors for birth defects: a study from population-based surveillance data”, published in 2024. The authors conducted a Multiple Logistic Regression analysis with data from 2014 to 2020 from the population-based Birth Defects Surveillance System in Hunan Province, China. This article used adjusted odds ratios as a measuring tool to analyse the strength of the associations and identify the most prevalent risks that contribute to higher birth defects in China [58]. In addition, the odds ratio metric can be used to improve accessibility. In Silva et al.'s article, a cross-sectional study was conducted to examine adolescent

perspectives on healthcare services. In this publication, odds ratios identified a link between an adolescent's value of physician characteristics and the perception of a doctor's skill, highlighting a key component that could foster a better environment [107].

The use of odds ratios as a good metric to evaluate teams' performance in healthcare systems can be advantageous due to their easy interpretability of the risk factors. In 2010, Moore, Lynne, et.al, evaluated the performance of trauma centres in Quebec, ranking the hospitals according to odds ratio after hierarchical Logistic Regression [59]. In a similar article, authors Peter Austin and Douglas Lee estimated the net population benefit through improvements in Hospital standards. The odds ratio measured the effect the variables had on the 30-day Mortality rate and 1-year mortality rate [60]. In March 2021, Nisha Kurian et al. published an article that predicted the overall quality star ratings of Hospitals in the USA. Using the US Centres for Medicare and Medicaid Services (CMS) star ratings provided, the authors created a simple ordinal logistic regression model and stepwise variable selection for different hospitals that were ineligible to have a star rating according to CMS. The Feature Selection method allowed the variables to be reduced from 57 to 20 performance measures, each with a significant effect on the star rating systems. The odds ratios were provided for a more intuitive assessment of the important variables and the hospital's star rating [61].

4. Materials and Methods

In this chapter, all materials used in the dissertation are referenced and explained, including the dataset, different subsets, code languages, libraries, packages, and imports. In addition, all experiments performed were described, including the categorisation of healthcare units based on International Benchmarking, and the IDE Quality, 'Structural and Demographic' variables, Feature Selection, and Odds Ratio analysis.

4.1. BD_BI_CSP_2023 Dataset & Pre-Processing

The BD_BI_CSP_2023 dataset was the focus of the dissertation. It contained 156 columns and 924 observations representing each HCU in the dataset. 128 of the columns contained all the collected indicators in 2023. The indicators in the dataset were primarily categorised into the different units of measure specified in Basic Concepts sub-chapter 1 (Indices, Proportions, Ratios, Costs, Expenses, Rates, and Score values). The remaining features contained structural components that are either necessary to differentiate between health care units (ID, HCU name, etc.) or specific conditions of that unit (No. of Doctors, etc.). These particular features were named 'Structural and Demographic' variables. The description of each column and its purpose is specified in Table A2 (Additional Information).

The pre-processing of the dataset mainly consisted of data cleaning and data integration. The first step involved data profiling, which was primarily conducted through a thorough review of each column, including its data type, objective, and theoretical range of available data (Additional Information, Table A2). All non-indicator variables were continuous, except for:

- 'LUTS_II', responsible for identifying the healthcare administrative region (Norte, LVT, Algarve, Alentejo, Centro)
- 'Tipo Unidade', responsible for identifying whether the healthcare unit was a USF-A, USF-B, or UCSP
- 'ULS', the 39 facilities responsible for integrating and managing the different health care units

The process of data cleaning began by checking the distributions of each variable to identify any severe outliers that contradicted the previously established theoretical ranges. The next step was to look at the missing values displayed per column, where variables that contained more than 30% of their observations missing were removed. Additionally, variables that induced redundancy, perfect separation due to high

correlation, or problems with convergence in the modelling of logistic regression were also removed. The dropped variables were 'ETC outros profissionais saúde', 'IDG IPDA', 'ETC Assistentes técnicos', 'ETC enfermeiros', 'ETC médicos', 'ETC internos', 'ETC outros profissionais', 'Inscritos com MF', '2017.346.01_FL', '2017.347.01_FL', '2017.348.01_FL', '2017.349.01_FL', '2022.434.01_FL', '2022.441.01_FL', '2022.442.01_FL', '2022.443.01_FL', '2022.444.01_FL', '2022.445.01_FL'. After being confident in the variables that were useful for the dissertation, all observations with missing values were removed. The imputation of data in the missing values was not considered adequate for this project, since it could muddle the development of the model's true behaviour and alter the associations we wanted to investigate. For the Feature Selection investigation and the Odds Ratio development, the four Outcome Indicators were transformed into categories through international benchmarking discussed in Chapters 4.4 and 5.2. The cleaned dataset contained 134 variables and 769 observations.

4.2. Materials

This dissertation was conducted using both Python and R programming languages [62][71]. The libraries "tidyr", "ggplot2", "dplyr", "corrplot", "glmnt", "pscl", "data.table" were used in R studio for the pre-processing of the data, the IDE, and the 'Structural and Demographic' analysis [63][64][65][66][67][68][69][70]. The cleaned dataset and normalised datasets were then exported using the "readxl" library for the methodologies explored in Google Colab [72]. Quality of life imports "pandas", "numpy", "statistics", and "Collections" – "Counter" were run for data frame establishments and averages/standard deviations calculations [73][74][75]. The open-source library sklearn was heavily relied upon for imports that allowed for stratified splits of the data and several models, these imports included: "preprocessing"; "model_selection" through "StratifiedKFold", "GridSearchCV", and "train_test_split"; "linear_model through "LogisticRegressionCV", "LogisticRegression", and "LassoCV"; "metrics" through "f1_score", "accuracy_score", "recall_score", "precision_score"; "inspection" through "permutation_importance"; and "utils" through "resample" [42]. For Statistical inference and additional development of Logistic Regression models open-source library "scipy" with import "stats" was used [76].

4.3. How well do IDE scores reflect quality

This IDE Analysis started with the creation of a subset from the BD_BI_CSP_2023_DATASET and the addition of a new variable, the IDE Scores for the year 2023. This data was obtained through the SNS BICSP website [81]. This subset contained 769 observations and 9 variables ('Nome UF', 'ULS', 'LUTS_II', 'Tipo

Unidade', 'Ind_365', 'Ind_372', 'Ind_339', 'Ind_410', 'IDE'). The dataset was normalised using Min-Max Normalisation. Several investigations were performed to evaluate the relationship between the IDE and the Outcome Indicators:

- A box plot diagram comparing the distribution of values between the IDE score and the four Outcome Indicators;
- Four correlation plots, each analysing a Healthcare unit's IDE with its respective Outcome Indicator (Colour coordinated and with a smooth line to represent the trend). The magnitude of the correlation between the IDE and each Outcome Indicator was also identified through the cor() function;
- An MLR with the IDE as the dependent variable and the four Outcome Indicators as independent variables;
- A deeper look into the relationship between these variables was also conducted through box plots of the IDE and the four Outcome Indicators after filtering the subset per region or unit type.

4.4. International Benchmarking

Through the lens of a continuously improving Primary Health Care, the use of international benchmarking was followed. The international average of the four Outcome Indicators was used as our threshold in the development of the dichotomous variables for the Feature Selection and Odds Ratio Analysis.

For Indicators 339 and 365, both international averages were found in the OECD's Health at a Glance 2023 report. Health at a Glance is an annual report that compares key health care indicators across partner countries, including accessibility, quality of care, resources, and emerging risk factors [77]. For Indicator 372, a 2023 study authored by Sing, Chor-Wing, et.al., examined the incidence of hip fractures, treatment, and all-cause mortality due to these fractures at an international level. This article provided possible actions to reduce the burden of primary health care on these hip fracture incidents, and allowed us to identify the global average for our benchmark [78]. For Indicator 410, an article that included an average of several countries was not available. Despite this, two studies investigated the same indicator, one containing information from Italy and the other from Sweden, each with different average scores [79][80]. The quartiles for Portugal's Primary Healthcare Outcome Indicators were also identified to compare all scores (Table 7).

Table 7. PHC quartile scores and International Benchmark for Outcome Indicators 339, 410, 365, 372

	Portugal; PHC 2023			International Benchmark
Indicator	P25	P50	P75	Average
2018.339.01_FL	45.38	52.95	63.02	27.00
2018.410.01_FL	2.34	2.92	3.62	2.90; 4.00
2017.365.01_FL	427.53	529.58	647.90	463.00
2017.372.01_FL	650.46	836.98	1063.83	247.00

These international averages were used in assembling the category labels, also referred to as classes, for the Outcome Indicators. Each Outcome Indicator was transformed into a categorical variable where all values below the international average were labelled 1, while the rest were labelled 0. Portugal's PHC 1st quartile scores (P25) replaced unfeasible benchmarks, created by problems with distribution. This came with the aim of continuous growth, where eventually the P25 thresholds for the category labels would be replaced by the international averages. The balance of distributions between labels and the exact thresholds used for the categories in the Feature Selection and Odds Ratio analyses are found in Chapter 5.2.

4.5. Structural and Demographic Analysis

A study on the relationship between the 'Structural and Demographic' variables of the dataset with our Outcome Indicators enabled the understanding of their general effect on the quality of a healthcare unit. This aimed to determine whether these variables are significant enough to be incorporated into the various Feature Selection methods and subsequently into our Odds Ratio models.

The first part consisted of a national analysis, created by selecting a subset from the cleaned BD_BI_CSP_2023_DATASET that contained only the 'Structural and Demographic' features, as well as the non-categorised Outcome Indicators. This subset contained the same number of observations (769), and 23 variables ('Nome UF', 'ULS', 'LUTS_II', 'Tipo Unidade', 'densidade populacional', 'Idade US', 'Índice de contexto sociodemográfico', 'Assistentes técnicos (n)', 'Enfermeiros (n)', 'Inscritos (n)', 'Médicos (n)', 'Médicos internos (n)', 'Outros profissionais (n)', 'Médicos ponderados (n)', 'Prevalência_DM', 'Proporção idosos', 'Unidades ponderadas', 'Taxa de desemprego', 'taxa abandono escolar', 'Ind_365', 'Ind_372', 'Ind_339', 'Ind_410'). After creating the subset, a min-max normalisation was performed on the dataset to enhance the Multiple

Linear Regression model's performance. The national analysis mainly consisted of one MLR per Outcome Indicator, where an Outcome Indicator was the dependent variable and the 'Structural and Demographic' features were the independent variables. The Adjusted R^2, the model's P-value, and the 5 Significant Variables with the lowest p-values were annotated in a table. In addition, a correlation plot between the four Outcome Indicators and the remaining features was generated.

A deeper evaluation of the relationship between the Outcome Indicators and the 'Structural and Demographic' features was conducted through an analysis by region or unit type. An additional eight subsets were created, all containing the same number of variables as the previous subset but filtered by their corresponding region or unit type.

- Alentejo subset: 23 variables, 46 observations
- Algarve subset: 23 variables, 35 observations
- Centro subset: 23 variables, 150 observations
- LVT subset: 23 variables, 211 observations
- Norte subset: 23 variables, 327 observations
- UCSP subset: 23 variables, 212 observations
- USF-A subset: 23 variables, 227 observations
- USF-B subset: 23 variables, 330 observations

The creation of subsets filtered by both region and unit type was unfeasible due to a lack of observations in certain subsets, since it is recommended for models to have at least 100 observations for complete convergence and greater stability [27]. A total of 32 MLR models were created, one per Outcome Indicator. Both the model's p-value and its R^2 were annotated to get a greater understanding of the effect of these features.

4.6. Feature Selection Analysis

The Feature Selection analysis enabled the narrowing of features into the most important and influential selection that best discriminates between the two labels. This is important in order not to add unnecessary noise to the models, which could lead to less accurate coefficients and a skewed view of the associations. All the Feature Selection methods and evaluations were applied separately for each Outcome Indicator (Ind 410, Ind 372, Ind 365, Ind 339).

4.6.1. Feature Selection Methods

This batch of FS methods was a mix of the three Feature Selection types: Filter methods with Wald thresholding; Wrapper methods with Stepwise Backwards selection, Logistic Regression Permutation importance, and Random Forest Classifier Permutation

importance; and Embedded methods with Lasso Regression, SHAP thresholding, and Random Forest Classifier Gini importance. Since many of the studies for this chapter needed both Scikit-learn and statsmodels open source libraries for statistical inference and modelling, the logistic regression models that were created were constructed in a way that got the closest models possible:

```
#Scikitlearn Version
model = LogisticRegression(penalty=None, solver = 'newton-cg', max_iter = 2000 ,fit_intercept=True)

#Statsmodel Version
model = sm.Logit(y, sm.add_constant(X)).fit(disp = 0)
```

Each Technique required different methodologies to compute the best selection of features:

- **SHAP Thresholding**

This FS Method consisted of performing stratified Cross-Validation, where for each fold, a SHAP explainer would return the SHAP values and store them in a list. The average of the absolute SHAP values per feature would then be stored in a dataframe, where each feature would have its designated mean absolute SHAP value. The summary Beeswarm plot was also displayed to get a better understanding of the associations occurring in the model.

```
skf = StratifiedKFold(n_splits=10, shuffle=True, random_state=1)
shap_values_list = []
X_test_list = []

for train_index, test_index in skf.split(X, y):
    X_train, X_test = X.iloc[train_index], X.iloc[test_index]
    y_train, y_test = y.iloc[train_index], y.iloc[test_index]


    model = LogisticRegression(penalty=None, solver = 'newton-cg', max_iter = 2000 ,fit_intercept=True)
    model.fit(X_train, y_train)

    explainer = shap.Explainer(model, X_train)
    shap_values = explainer(X_test)


    shap_values_list.append(shap_values.values)
    X_test_list.append(X_test)

shap_values_array = np.vstack(shap_values_list)
X_test_array = pd.concat(X_test_list)
mean_shap_values = np.abs(shap_values_array).mean(axis=0)
shap_df_410 = pd.DataFrame({"Feature": X.columns, "Mean |SHAP|": mean_shap_values})
shap_df_410 = shap_df_410.sort_values(by="Mean |SHAP|", ascending=False)

shap.summary_plot(shap_values_array, X_test_array, max_display = 10)
```

To determine the best set of features, a series of thresholds on the data frame was applied (0, 0.1, 0.2, …, 0.8, 0.9, 1). Sets were developed based only on the features with

mean absolute SHAP values above the threshold, as they would represent the most influential features for the model. The set with the highest F1-score was designated as the **SHAP Feature Selection Set** in the evaluation investigations stated in Section 4.6.2.

- **Logistic Regression Permutation Importance**

Due to being too computationally expensive, a cross-validation was unfeasible in this particular Feature Selection method. The alternative was to create a Train Test Split and use the Scikit-learn library “inspection” to import the permutation importance formula.

```
model = LogisticRegression(penalty=None, solver = 'newton-cg', fit_intercept=True, max_iter = 2000)
model.fit(X_train, y_train)
r = permutation_importance(model, X_test, y_test,
                           n_repeats=30,
                           random_state=0)
imp = r.importances_mean

permutation_importance_df410 = pd.DataFrame({
    'Feature': X_train.columns,
    'Importance': imp
})

permutation_importance_df410 = permutation_importance_df410.sort_values(by='Importance', ascending=False)

print(permutation_importance_df410)
```

The number of repeats used for permutation importance was 30, the recommended amount in the scikit learn documentation [42]. The mean importances of each feature were then stored and ordered in a dataframe. The permutation importance averages had a much lower margin than the SHAP contribution values. This meant the thresholding window needed to be lowered. Sets were created based on the thresholds ranging from 0.05 to 0 (0.05, 0.01, 0.005, 0.001, 0). For each set, the features with mean importances above the threshold were stored. The set with the highest F1-score was designated as the **Permutation Importance Feature Selection Set** in the evaluation investigations stated in Section 4.6.2.

- **Random Forest**

Using Random Forest to assess the importance of features in a model enables the investigation of non-linear relationships between the features and the Outcome Indicators. To prevent overfitting and lower the likelihood of poor performance, a grid_search cross-validation was performed, evaluating the best Max_depth Parameter (6, 7, 8) for the RFC model of each Outcome Indicator.

```
cv = StratifiedKFold(n_splits=5)
param_grid = {
    'max_depth': [6, 7, 8],
    }

grid_search = GridSearchCV(
    estimator= RFC(n_estimators=100, random_state=0),
    param_grid=param_grid,
    scoring='f1',
    cv=cv
)
grid_search.fit(X_train, y_train)
```

- **Gini Importance**

The development of Gini importances is a feature available in the Random Forest Classifier import through feature_importances_. The Gini importance values per feature were stored in a dataframe. Sets were created based on thresholds ranging from 0.05 to 0 (0.05, 0.01, 0.005, 0.001, 0). For each set, the features with a Gini importance above the threshold were stored. The set with the highest F1-score was designated as the **RFC Gini importance Feature Selection Set** in the evaluation investigations stated in Section 4.6.2.

```
clf = RFC(n_estimators=100, max_depth = 7, random_state=0)
clf.fit(X_train, y_train)
importance = clf.feature_importances_
feature_importance_dfRFG410 = pd.DataFrame({
    'Feature': X_train.columns,
    'Importance': importance
})
```

- **Permutation Importance**

The Random Forest Classifier Permutation Importance number of repeats was identical to the one using Logistic Regression as an estimator, 30. Sets were created based on thresholds ranging from 0.05 to 0 (0.05, 0.01, 0.005, 0.001, 0). For each set, the features with a Permutation importance above the threshold were stored. The set with the highest F1-score was designated as the **RFC Permutation importance Feature Selection Set** in the evaluation investigations stated in Section 4.6.2.

```
clf = RandomForestClassifier(n_estimators=100, max_depth = 7, random_state=0)
clf.fit(X_train, y_train)
r = permutation_importance(clf, X_test, y_test,
                           n_repeats=30,
                           random_state=0)
imp = r.importances_mean

permutation_importance_dfRFCP = pd.DataFrame({
    'Feature': X_train.columns,
    'Importance': imp
})

permutation_importance_df = permutation_importance_dfRFCP.sort_values(by='Importance', ascending=False)

print(permutation_importance_df)
```

➢ Wald thresholding

For the Wald threshold method, the statsmodel library was used to get the p-values of each feature in a full Logistic Regression model. All features with p-values below the consensus 0.05 were labelled as significant. This group of features was designated as **the Wald Feature Selection Set** in the evaluation investigations stated in Section 4.6.2.

```
X_full410 = sm.add_constant(X)
full_model410 = sm.Logit(y, X_full410).fit(disp = 0)
sig_var410 = full_model410.pvalues[full_model410.pvalues < 0.05].index
```

➢ Lasso Regression

The Lasso Regression was the only Logistic Regression model produced differently from the one established previously. The Lasso Regression model used the penalty 'l1', the solver 'liblinear', and the default alpha (C = 1), representing a default Lasso Regression formula. The import SelectFromModel and the transform function enabled the Selection of the variables that had non-zero coefficients. This group of features was designated as the **Lasso Feature Selection Set** in the evaluation investigations stated in Section 4.6.2.

```
log = LogisticRegression(penalty='l1', solver='liblinear')
log.fit(X, y)
model = SelectFromModel(log, prefit=True)
X_Lasso410 = model.transform(X)
```

➢ Lasso CV

Lasso CV used 5 cross-validation folds to tune the C hyperparameter, using the F1-score statistic as the scoring metric. The higher the C value, the more lenient the penalty and the smaller the sparsity (percentage of zero coefficients). To alleviate the computational cost of the model, a fixed grid of penalties was used, ranging from 0.01 to 4 (0.01, 0.1, 0.5, 2, 4). After fine-tuning the C hyperparameter and creating the model, the significant variables were highlighted and stored as the **LassoCV Feature Selection Set**. This set was later used in the evaluation investigations stated in Section 4.6.2.

```
log_cv_model = LogisticRegressionCV(penalty='l1', solver='liblinear', cv=5,
                                    random_state=42, scoring='f1', max_iter=2000,
                                    Cs=[0.01, 0.1, 0.5, 2, 4])
log_cv_model.fit(X, y)

selector = SelectFromModel(log_cv_model, prefit=True, threshold=1e-5)

selected_mask = selector.get_support()
selected_feature_names_lasso_cv = X.columns[selected_mask]

X_Lasso410_cv = X[selected_feature_names_lasso_cv]
```

- **Stepwise Backwards**

Stepwise Backwards Feature Selection is not a feature readily available in a library/import in Python. For this reason, the collection of features was performed in R Studio. This FS method was performed using the step function in R, using “AIC” as the criterion. After the variables were identified, they were copied into the Google Colab notebook and were stored as the **Stepwise Backwards Feature Selection Set.** This set was later used in used in the evaluation investigations stated in Section 4.6.2.

```
logreg410 <- glm(Ind_410B ~., family = binomial(link='logit'), data = cat410)
bmInd_410B <- step(logreg410,
                   direction = "backward", criterion = "AIC", trace = 0)
summary(bmInd_410B)
```

4.6.2. Feature Selection Evaluation

The amount of evaluation techniques investigated in this section was to address the possible weaknesses of each technique and to identify the best Feature Selection methods specific to this dataset. The evaluation metrics used were F1- score, AIC, BIC, Likelihood Ratio Test, Jaccard Similarity, and Percentage Coverage test.

- **F1-Score**

To evaluate the F1-scores of each Outcome Indicator’s FS set, a stratified Cross-validation was performed where the subset was clarified as the independent variable and the Outcome Indicator as the dependent variable. The average F1-scores and their standard deviation for each set were recorded.

- **AIC & BIC**

The Logit function from “statsmodel” allows the development of the Logistic Regression model using only the group of features established in the Feature Selection section. The AIC and BIC scores of a model were accessed after the creation of the model, giving us a better insight into the balance between the complexity and the fit of a FS set and its Outcome Indicator.

```
X_SHAP410c = sm.add_constant(X_SHAP410)
model_SHAP410 = sm.Logit(y, X_SHAP410c).fit(disp=0)
print("AIC for Set FS with SHAP for Ind 410: ", model_SHAP410.aic)
print("BIC for Set FS with SHAP for Ind 410: ", model_SHAP410.bic)
```

- **Likelihood Ratio Test**

The Statsmodel library enabled the retrieval of the log-likelihoods of a logistic model, laying the groundwork for the development of the LRT function. There were two different

Likelihood ratio tests performed per Feature Selection set: The set's model was compared with a full model (no feature selection) and a null model (only the intercept present). The p-values of the LRT function would indicate whether the null hypothesis could or could not be rejected. P-values below the 0.05 threshold suggested that a more complex model provided a significantly better fit, and the null hypothesis could be rejected with confidence. For the two likelihood ratio tests, a positively viewed result was different depending on whether the FS set model was being compared with a more complex full model or with the null model. When compared with the full model, a favourable outcome had a p-value above 0.05, which indicated that the FS set model was not significantly worse than the full model, and the addition of the remaining features did not improve the fit. When compared with a null model, a favourable outcome had a p-value below 0.05, revealing that the FS set model was significantly better than the null model. For visualisation purposes and to provide less confusion, a table was created where a positive LRT outcome for that test was visualised with a checkmark (☑) and a negative test with a cross (☒).

```
ll_SHAP410 = model_SHAP410.llf
ll_full410 = full_model410.llf
lr_stat = 2 * (ll_full410 - ll_SHAP410)
df = full_model410.df_model - model_SHAP410.df_model
p_value = stats.chi2.sf(lr_stat, df)

print(f"Likelihood-Ratio χ² = {lr_stat:.2f}, df = {df}, p = {p_value:.4g}")
```

- **Stability Measure (Jaccard Similarity & Percentage of Coverage)**

The stability of a group of features was vital to understanding whether the sets of features selected were trustworthy. To evaluate the stability of a group, two tests were performed: Jaccard Similarity and Percentage Coverage. The main functions developed for this test were a bootstrap function, a Jaccard similarity function, a percentage coverage function, and a Feature Selection function relative to the FS set that was being accessed.

The bootstrap function resampled the data with replacement, introducing variability while using only the original observations.

```
def bootstrap(X, y, random_state=None):
    data = pd.concat([X, y], axis=1)
    outcome_name = y.name
    boot_sample = resample(data, replace=True, n_samples=len(data), random_state=random_state)
    X_boot = boot_sample.drop(columns=outcome_name)
    y_boot = boot_sample[outcome_name]
    return X_boot, y_boot
```

The Feature Selection Set function created a list of the most significant features that were present in each bootstrap. The function used the same methodology previously

implemented for each FS set and the same threshold per Outcome Indicator provided, to create the most similar conditions possible (Chapter 4.6.1).

```
def run_shap_selection(X_boot, y_boot, threshold):
    model = LogisticRegression(penalty=None, solver='newton-cg',
        max_iter=2000,fit_intercept=True)
    model.fit(X_boot, y_boot)

    explainer = shap.Explainer(model, X_boot)
    shap_values = explainer.shap_values(X_boot)

    mean_abs_shap = np.mean(np.abs(shap_values), axis=0)

    selected_features = X_boot.columns[mean_abs_shap > threshold]

    return selected_features.tolist()
```

The Jaccard function transformed a list of variables into a set and compared the intersected variables between the two sets, with all the possible outcomes.

```
def jaccard_similarity(list1, list2):
    s1 = set(list1)
    s2 = set(list2)
    return float(len(s1.intersection(s2)) / len(s1.union(s2)))
```

The percentage coverage function covered a dataframe, counting all the features that satisfied two conditions: The feature must be present in at least 70% of all the bootstraps, and the feature must be present in a list provided (the list containing the chosen set of features for the FS Method). A percentage of the Number of features of the FS set that were present in more than 70% of the bootstraps was then revealed.

```
def percentage_function(Sel, df, threshold=0.7):
    Count = 0
    for Col in Sel.columns:
        row = df[df['Feature'] == Col]
        if not row.empty and row['Selection_Frequency'].iloc[0] > threshold:
            Count += 1

    percentage = (Count / len(Sel.columns)) * 100
    return percentage
```

After creating the functions, due to bootstrapping being too computationally expensive, only 10 bootstraps were run. For each bootstrap, the Feature Selection function would select the most important features, which would then be catalogued in a list. In addition, a Jaccard Score between the original FS set and the bootstrapped list was calculated. After finishing all the bootstraps, a dataframe was created where every feature in every bootstrap set was counted and ordered. To get the stability scores, an average Jaccard

score between all the different bootstrap sets and the FS set was calculated, and the percentage of stable features from the original group was revealed.

```
n_bootstraps = 10
bootstrap_shap_lists = []
jaccard_list = []
for i in range(n_bootstraps):
  X_boot, y_boot = bootstrap(X, y, random_state=i)
  selected_features = run_shap_selection(X_boot, y_boot, threshold = 0.5)
  bootstrap_shap_lists.append(selected_features)
  jaccard_list.append(jaccard_similarity(selected_features, X_SHAP410))

all_shap_features = [feature for run in bootstrap_shap_lists for feature in run]
feature_counts = Counter(all_shap_features)

shap_stability_df = pd.DataFrame(
    feature_counts.items(), columns=['Feature', 'Selection_Count']
)
shap_stability_df['Selection_Frequency'] = shap_stability_df['Selection_Count'] / n_bootstraps
shap_stability_df = shap_stability_df.sort_values(
    by='Selection_Frequency', ascending=False
).reset_index(drop=True)

print("\n--- Percentage Coverage ---")
print(percentage_function(X_SHAP410, shap_stability_df))
print("\n--- Jaccard Similarity ---")
avg_jaccard = sum(jaccard_list) / len(jaccard_list)
print(f"Average Jaccard Similarity: {avg_jaccard: .3f}")
```

4.7. Odds Ratio Analysis

After evaluating each performing Feature Selection set, the sets with the best performance and stability for each Outcome Indicator were combined using the pandas.concat() function. To avoid potential bias from the different Feature Selection methods, a voting system was implemented in which only the features present in a majority of the sets were retained as important. After establishing the most consistent and trustworthy group of features for all Outcome Indicators, a logistic regression model using the statsmodels library was developed. The odds ratios were obtained by exponentiating the coefficients and their corresponding confidence intervals. The McFadden pseudo R^2 scores were also extracted for comparison among the different Outcome Indicator models.

5. Results and Discussion

5.1. How well do IDE scores reflect quality

The availability of the 4 Outcome Indicators in our database enabled an assessment of the current IDE system's ability to reflect the quality of a healthcare unit.

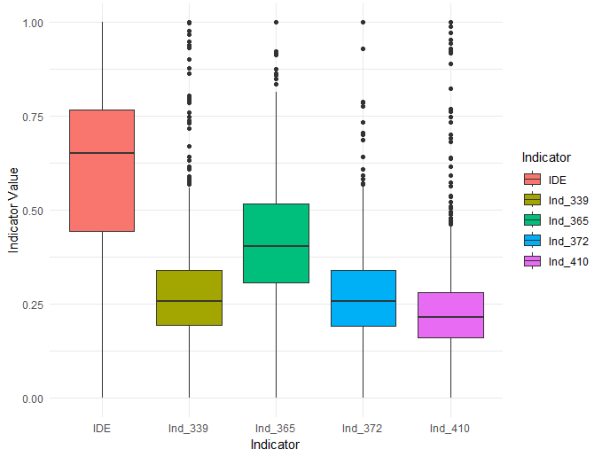

Figure 8. Boxplot diagram of IDE, Ind 339, Ind 365, Ind 372, Ind 410

The boxplots (Fig.8) show the range of distributions for these variables and give us a sense of the general scores for most healthcare units in Portugal. The intrinsic difference between the IDE scores and the Outcome Indicators lies in their evaluation of good performance. The IDE score ranges from 0 to 100 and demonstrates a standard evaluation method, where higher values mean better-performing units. The Outcome Indicators represent real data gathered throughout the year, and since they depict rates of instances in the healthcare unit, lower scores indicate better performance. Understanding that the IDE and Outcome Indicators should mirror each other, the boxplots above demonstrate a slight tendency in that direction.

The correlation plots feature a linear, smooth line that explains the trend across the comparing variables. Looking at the Correlation plot (Fig. 9), for all four Outcome Indicators, higher IDE scores generally corresponded to lower indicator values. Supporting the position that the IDE score is a decent reflection of quality. However, the magnitude of the correlation was also identified (Table 8). Consistent with the trend identified in the plot, the correlation of the IDE had the same direction between all Outcome Indicators. Despite this, the strength of the magnitude of these correlations was weak, ranging from -0.13 to -0.20, suggesting that the IDE does not strongly reflect the behaviour of these indicators.

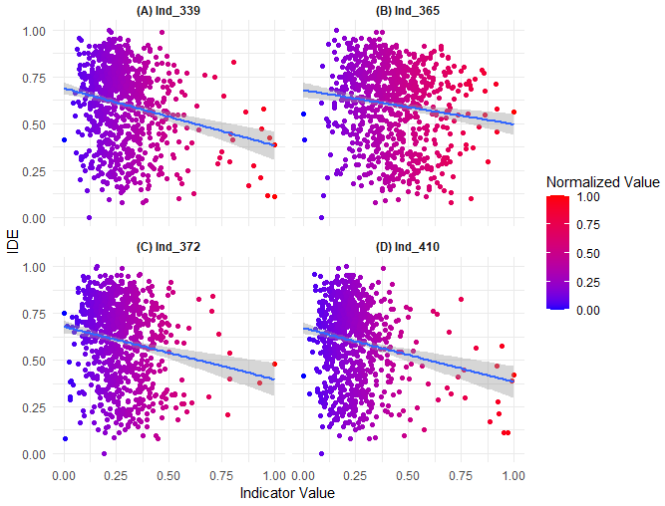

Figure 9. Correlation plots of IDE and (A) Ind 339, (B) Ind 365, (C) Ind 372, (D) Ind 410

Table 8. Magnitude of correlation between IDE and Outcome Indicators

	Ind 410	Ind 372	Ind 365	Ind 339
IDE	- 0.18	- 0.16	- 0.13	- 0.20

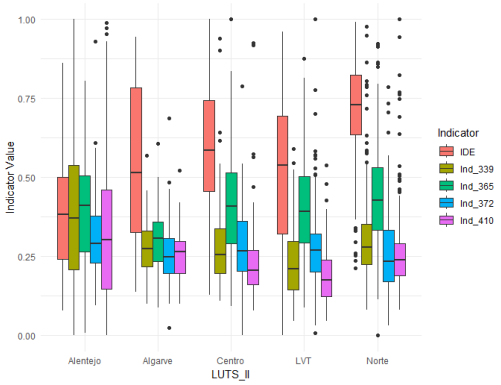

Figure 10. Boxplot diagram of Normalised variables sorted by regions

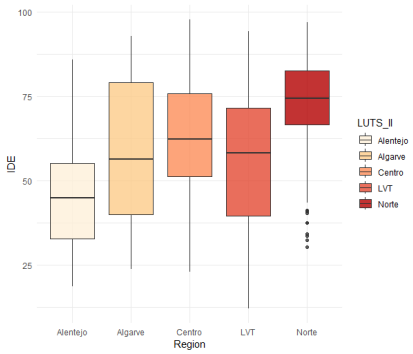

Figure 11. Boxplot diagram of IDE sorted by different regions

Looking at the Normalised comparison between the Outcome Indicators and the IDE scores, a different picture seemed to emerge (Fig.10). Alentejo had a similar distribution of scores between all variables, regardless of being an IDE or an indicator. This implied that lower IDE scores in Alentejo correlated with good performance. Other regions, such as the Algarve, Centro, and LVT, also had overlapping distributions between these variables. Suggesting that the IDE does not illustrate the behaviour of the Outcome Indicators as well as they seem to in the previous figures (Figs. 8 and 9). Regionally,

there was a significant distinction in IDE scores, especially between Norte and Alentejo (Fig.11).

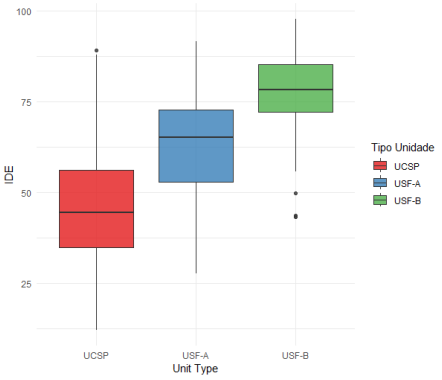

Figure 12. Boxplot diagram of IDE sorted by unit types

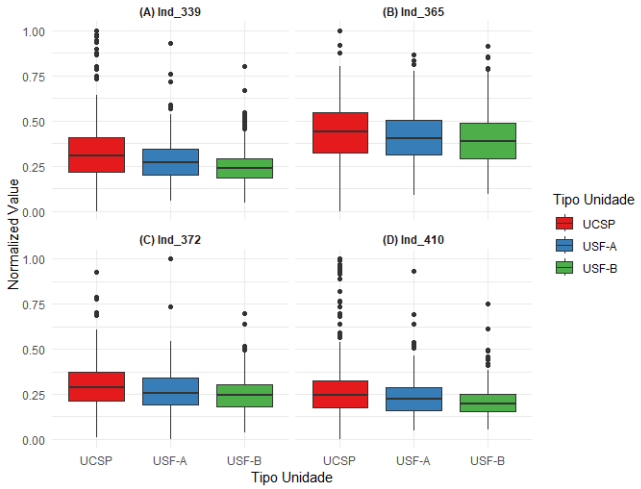

Figure 13. Boxplot diagrams of (A) Ind 339, (B) Ind 365, (C) Ind 372, (D) Ind 410

The distribution between unit types showed that for every variable, USF-B had generally better scores (Figs. 12 and 13). In addition, the differences in distributions of the unit types shown in the IDE score were significantly more embellished than in the indicator results. These vast differences in IDE scores are explained by the composition of the matrix of indicators responsible for their scoring. For example, in the indicator matrix, Indicators 341 and 354 are two financial-economic indicators that have a combined importance level of 18% and account for the prescription costs incurred by healthcare units. In October 2023, the expenses specific to Indicator 354 for unit type model B were averaged at 173€, with their counterparts' expenses being significantly higher. Disregarding this fact, the table gave a consistent expected cost for this indicator of 133€, not accounting for the inflation of expenses throughout the year. These detrimental expected values led to units of models USF-A and UCSP being less likely to achieve a good score for these indicators, widening a gap in IDE scores that is not reflected in the Outcome Indicators [82][83].

The MLR investigation helped understand how well the Outcome Indicators predict the IDE score and, therefore, explain their relationship (Fig. 14). The F-statistic being larger than one and a small p-value suggested that the model was statistically significant, with at least one indicator related to IDE. Looking at the residuals, most were between -0.15 and 0.15, which showed relatively good accuracy when predicting the dependent variable. Despite this, we know from the boxplot diagrams that the distribution of all IDE

scores was close (0.50 and 0.75), which alters our perception of the model since some predictions can be caused by its close distribution. A key feature of the model that expands on this consideration was the low Adjusted R^2, 6.9%. A low R^2 value suggests the model is poorly explained by the indicators, and other unaddressed features are more important. Another observation pertains to the estimated coefficients for each Outcome Indicator and how their estimated values differ from our theoretical understanding. Lower Outcome Indicator values should be designated as high IDE scores, meaning better quality of a health care service. The estimated coefficients, despite all being significant, did not reflect this statement. Indicator 410 behaved directly opposite to the knowledge that the Outcome Indicators were negatively correlated to the IDE score. This MLR model provided us with an understanding that, despite having slightly positive prediction accuracy and a good F statistic, the IDE was a flawed method of attributing quality.

```
Call:
lm(formula = IDE ~ Ind_365 + Ind_339 + Ind_410 + Ind_372, data = NormIDE)

Residuals:
     Min       1Q   Median       3Q      Max
-0.69630 -0.15267  0.04161  0.15481  0.44547

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.79262    0.02637  30.056  < 2e-16 ***
Ind_365     -0.10172    0.04721  -2.154 0.031498 *
Ind_339     -0.86968    0.24551  -3.542 0.000419 ***
Ind_410      0.66148    0.26314   2.514 0.012129 *
Ind_372     -0.21988    0.05957  -3.691 0.000238 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2065 on 838 degrees of freedom
Multiple R-squared:  0.07383,   Adjusted R-squared:  0.06941
F-statistic:  16.7 on 4 and 838 DF,  p-value: 3.53e-13
```

Figure 14. Summary() of MLR model with IDE as dependent variable and Ind 365, Ind 339, Ind 410, and Ind 372 as independent variables

The analysis of the IDE Scores, as calculated by the table in Ordinance 411-A/2023 (Additional Information, Table A1), reflects how well this scoring system evaluates the multiprofessional team performance of a healthcare unit. Comparing this scoring system with the four Outcome Indicators showed us that, on a general basis, the IDE seemed to represent quality marginally. However, a deeper analysis demonstrated how a different number of flawed choices when developing the table affected the results. The choices of allocation of importance levels and expected values with little to no mathematical background led to unit types such as USF-A and UCSP being at a disadvantage in scoring a high IDE. The decision of indicators used in the 411-A table is also debatable, specifically using one of the Outcome Indicators - Indicator 365 (Additional Information, Table A1). Using an Outcome Indicator in the table for evaluating efficiency is unethical and leads to a muddying of the intentions of a healthcare professional. This is because

it conflates indicators representing outcomes with the performance of healthcare professionals. Professionals can not directly control these Outcome Indicators. They assess a particularity of a unit that can only be improved through the improvement of its surrounding conditions and processes. The inclusion of Indicator 365 in the table can also lead to misleading conclusions in our analysis. Any correlation between the IDE and Indicator 365 may be inflated and demonstrate unreliable results. These different discussions showcase the limitations of the current IDE score and emphasise the need for a different system with a greater focus on continuous improvement [82][7].

5.2. International Benchmarking

The balance of the distribution of the labels for category classification had a clear effect on the results of the models. An imbalanced outcome target can lead to misleading accuracy scores due to a bias towards the majority class [84]. According to Google's Machine Learning Crash course, the minimum percentage of data that belongs to the minority class while maintaining a relatively balanced distribution is 20%; with anything below it being considered moderate to extremely imbalanced [85]. Looking at the distribution of classes for the four Outcome Indicators with the international averages as a threshold, we can see that Indicators 339 and 372 had extreme imbalance (Table 9). This ensured that model development using these thresholds was unfeasible. This distribution demonstrates the rarity of units displaying scores below the international average for these indicators. The balance in the distribution of both articles relevant to Indicator 410 was also analysed [79][80]. Comparing the two averages, two different analyses of our healthcare system can be made. In general, both distributions displayed a positive indication for Primary Health Care units for this indicator. The article based on the results in Italy had more healthcare units below the average than its counterpart. This demonstrated that almost half of the units in Portugal had better scores than the average displayed for this indicator [79]. Evaluating the PHC units using the average from Sweden shows us a significantly more optimistic outlook, with more than 80% of the units in the dataset having lower scores than the average [80].

Table 9. Distribution of category labels according to different benchmark thresholds (thresholds identified inside brackets) and percentage of coverage of label 1 (Better performance)

Indicator	Above Threshold	Below Threshold	Percentage (%)
Ind 339 (27.00)[77]	769	0	0 %
Ind 410 (2.90)[79]	394	375	48.80 %

Ind 410 (4.00)[80]	131	638	83.00 %
Ind 365 (463.00)[77]	516	253	32.90 %
Ind 372 (247.00)[78]	762	7	0.90 %

After researching the international averages and analysing the distributions of the units, the thresholds for our models were established (Table 10). For Indicator 410, the threshold that was the most suitable for our classification problem was from the article authored by Furia, G. et al., developed in Italy, with an average of 2.9 [79]. This article contained more observations, giving us a more realistic outlook on the international conditions for this indicator. It also displayed a more balanced distribution for our labels, since a percentage coverage of more than 80% could lead to problems with the development of the model. The choice for a more pessimistic outlook also aligns with our goal of striving for continuous improvement in health care units. Indicator 365 had a relatively balanced distribution (32.9%), meaning the international threshold of 463 could be used for the categorisation of this Outcome Indicator. For Indicators 339 and 372, the development of the model through the international benchmark was not implemented due to its impact on performance. To substitute these values, the lower quartile (P25) of the PHC scores was used as the threshold (Table 7). This allowed the creation of a model that highlights the most important variables, their strength, and direction. Using the P25 as a threshold came with the understanding that focusing on the parameters established later, along with consistently refreshing the significant variables, should lead to continuous improvement. The intuition being that the subsequent improvement of the indicator scores will result in values that eventually become low enough to compare with the international average.

Table 10. Thresholds opted for model development, and the percentage of coverage of label 1 (Better performance for the indicator)

Indicator	Threshold	Percentage (%)
Ind 410	2.90	48.80 %
Ind 372	650.46	24.00 %
Ind 365	463.00	32.90 %
Ind 339	45.38	24.50 %

5.3. Structural and Demographic Analysis

Table 11. R^2 and P-value scores for the 4 MLR models using the Outcome Indicators as dependent variables and 'Structural and Demographic' variables as independent variables

	Ind 410	**Ind 372**	**Ind 365**	**Ind 339**
R^2	0.19	0.10	0.12	0.22
P-value	<2.20e-16	6.94e-13	<2.20e-16	<2.20e-16
Sig. Variable 1	Proporção Idosos	Enfermeiros (n)	Prevalencia_DM	Taxa de desemprego
Sig. Variable 2	Ind.Contexto Sociodemográfico	Ind. Contexto Sociodemográfico	Ind. Contexto Sociodemográfico	Ind. Contexto sociodemográfico
Sig. Variable 3	Unidades ponderadas	Tax. Desemprego	Idade US	Enfermeiros (n)
Sig. Variable 4	Tax. Desemprego	Tax. Abondono Escolar	Unidades Ponderadas	Proporção idosos
Sig. Variable 5	Médicos Ponderados (n)	Unidades ponderadas	Médicos ponderados (n)	Médicos ponderados (n)

The national 'Structural and Demographic' features analysis showed that the percentages of variance explained by the variables ranged from 10% to 22%. According to Gupta, Avi, et al., a good threshold for defining a good R^2 score varies depending on the context of the study being conducted, but the consensus is around 15%. Using this as a principle, we can see that two of the Outcome Indicators (Ind 399 & 410) had an R^2 value above this threshold. Likewise, the p-value of the model should be below the 0.05 threshold, allowing for the distinction of statistical significance. According to Table 11, all the multiple linear regression models had p-values below the 0.05 threshold, further supporting the belief that the 'Structural and Demographic' models were significant [89]. Analysing the significance level of each variable in the models, we can see that the 'Indice de Contexto Sociodemográfico' (ICS), also known as the Social Demographic context index, was a relevant feature for all four Outcome Indicators. The ICS is a calculated variable that allows for the contextualization of the efficiency and economic analysis of a specific healthcare unit [16]. This index is a 0-100 score that is calculated by giving importance weights to 4 different attributes of a location (Unemployment rate, School dropout rate, Proportion of Elderly, and Population Density), with higher scores meaning fewer desirable conditions. For Indicators 339, 372, and 410, the coefficient was around 0.36, indicating a relatively high correlation between higher Social Demographic context scores and worse quality for those indicators. Opposite to the other models' behaviour, Indicator 365 showed a negative correlation with a coefficient of -0.27. The ICS variable for each model had a p-value below the 0.05 threshold. Despite this, the Social Demographic context index for Indicator 365 had a slightly higher p-value

and standard error relative to the other Outcome Indicator models. This suggests that the variable is less statistically robust for this indicator compared to its counterparts.

Table 12. R^2 and P-value scores for the MLR model after filtering based on region, using the Outcome Indicators as dependent variables and 'Structural and Demographic' variables as independent variables.

ARS Region		Ind 410	Ind 372	Ind 365	Ind 339
Alentejo	R^2	0.19	- 3.54 e-03	0.36	0.20
	P-value	0.11	0.49	0.01	0.10
Algarve	R^2	-0.18	0.42	0.45	-0.21
	P-value	0.80	0.02	0.02	0.83
Centro	R^2	0.10	0.15	0.20	0.12
	P-value	0.01	1.10 e-03	4.78e-05	5.26 e-03
LVT	R^2	0.28	0.07	0.20	0.33
	P-value	4.67 e-11	0.02	4.83e-07	1.39 e-13
Norte	R^2	0.20	0.10	0.06	0.24
	P-value	1.20 e-11	2.47 e-05	3.33 e-03	1.05 e-14

The filtering of units by region allowed us to have a deeper understanding of the effect the 'Structural and Demographic' variables had on the Outcome Indicators (Table 12). The first observation visible was the effect of a low number of observations in the different subsets. The regions of Alentejo and Algarve had the most volatile percentages of variability explained in all the subsets analysed, having both the highest and the lowest R^2 scores. Small samples can lead to inflated or deflated R^2 values depending on how the data fits the model [89]. This is also highlighted in the p-value scores, where more than half of the models explored for these regions had a p-value above the 0.05 threshold. A p-value above 0.05 indicates that we cannot presume that the variables have coefficients different from 0. Looking at the more stable regions, we can see that most models followed the national interpretation of 10% to 25% of the variation explained due to these variables.

Table 13. R^2 and P-value scores for the MLR model after filtering based on unit type, using the Outcome Indicators as dependent variables and 'Structural and Demographic' variables as independent variables.

Unit Type		Ind 410	Ind 372	Ind 365	Ind 339
UCSP	R^2	0.08	0.09	0.05	0.08
	P-value	5.98 e-03	2.68 e-03	0.04	5.26 e-03
USF - A	R^2	0.15	0.07	0.10	0.19
	P-value	2.09 e-05	7.46 e-03	9.61 e-04	2.10 e-07
USF - B	R^2	0.22	0.05	0.19	0.24
	P-value	3.29 e-13	0.01	2.67 e-11	8.49 e-15

The separation of the dataset by unit type allowed for more balanced and stable models of the Outcome Indicators, since all the different subsets had more than 200 observations. All models evaluated had p-values below the threshold of 0.05. Despite this, the differences in unit types were evident in these models (Table 13). Looking at the R^2 values, the quality of USF unit types seemed to be more influenced by the structural and demographic landscape of their surroundings than the UCSPs. Table 13 also displayed the clear hierarchy of effect from these features. USF-B generally had higher R^2 values in all models, the outlier being Indicator 372, where UCSP had more of the percentage variability explained.

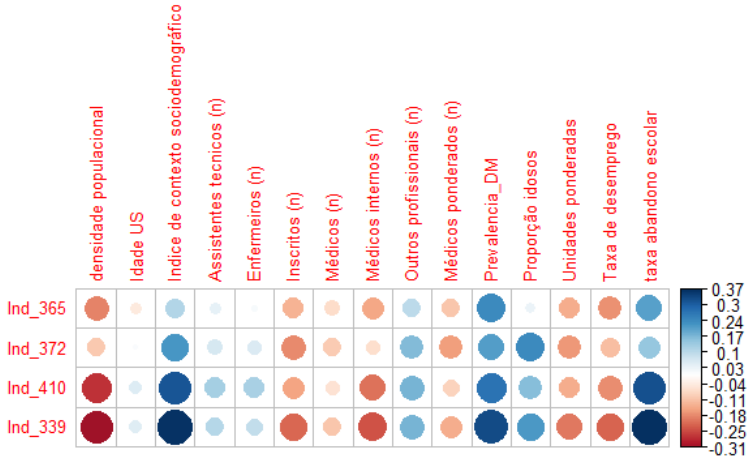

Figure 15. Correlations between outcome indicators and 'Structural and Demographic' variables.

The correlation analysis between the Outcome Indicators and the 'Structural and Demographic' features also displays attributes of the relationship between these variables. As expected, through the previous R^2 scores, the colour gradient limits in Fig. 15 show that the correlations between these variables were moderately weak (0.37, -0.31). In addition, the correlation plot revealed that the Outcome Indicators had the same correlated direction for all variables. This demonstrates that the Outcome Indicators showed similar distributions and corroborates the intuition that they display the actual performance of a unit. In terms of specific correlations, we can observe an increase in variables such as School dropout rate, Proportion of the Elderly, and Social Demographic context, correlating these variables with worse-performing healthcare units. On the other hand, population density, number of registered service users, and higher unemployment rates seemed to be more present in better-performing units. Having unemployment with a negative association may lead us to incorrectly assume that a higher unemployment rate results in better-performing models. However, correlation does not mean causality.

Having variables with the same direction despite having a theoretical contradiction underlines the extent to which these variables affect the Outcome Indicators. A deeper analysis of this particularity reveals how the population density of a location influences the underlying associations of these variables. According to the EU official website, in 2023, rural areas had a lower unemployment rate (5.9%) than urban areas (7.3%) [91]. A higher population density characterises urban cities with better investment opportunities and more attractive employment prospects for more experienced doctors and nurses. This could be a factor that can explain the association between the Outcome Indicator scores and these variables [16].

The combination of 'Structural and Demographic' variables, despite showing moderately weak correlations, contained R^2 values and specific variables that displayed significant influence on the Outcome Indicators. This investigation revealed the importance of being aware of these variables and their inclusion in our Feature Selection investigation.

5.4. Feature Selection Analysis

The evaluation scores for each of the indicators' Feature Selection sets models were annotated in the tables below.

Table 14. Number of variables in Sets after Feature Selection

FS Method	Ind 410	Ind 372	Ind 365	Ind 339
Null Model	0	0	0	0
Full model	129	129	129	129
SHAP	37	40	26	60
PI	67	55	32	45
RFC G.I	21	25	26	21
RFC P.I	73	47	104	18
Wald	20	7	14	35
S.W Backwards	45	36	39	62
Lasso	49	41	55	55
Lasso (CV)	71	92	89	96

Table 15. Average F1-Scores and standard deviation for each Feature Selection Set across the different Outcome Indicators

FS Method	**Ind 410**	**Ind 372**	**Ind 365**	**Ind 339**
Null Model	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Full Model	0.73 ± 0.08	0.28 ± 0.07	0.46 ± 0.08	0.64 ± 0.07
SHAP	0.79 ± 0.07	0.25 ± 0.10	0.44 ± 0.12	0.68 ± 0.08
PI	0.77 ± 0.05	0.22 ± 0.09	0.50 ± 0.11	0.63 ± 0.12
RFC G.I	0.76 ± 0.05	0.14 ± 0.12	0.37 ± 0.10	0.62 ± 0.14
RFC P.I	0.72 ± 0.04	0.13 ± 0.07	0.47 ± 0.09	0.47± 0.09
Wald	0.79 ± 0.07	0.00 ± 0.00	0.35 ± 0.08	0.68 ± 0.09
S.W. Backwards	0.80 ± 0.05	0.32 ± 0.11	0.53 ± 0.08	0.73 ± 0.06
Lasso	0.79 ± 0.06	0.27 ± 0.12	0.51 ± 0.09	0.71 ± 0.10
Lasso CV	0.76 ± 0.06	0.26 ± 0.13	0.50 ± 0.10	0.68 ± 0.08

Table 16. AIC Scores for each Feature Selection Set across the different Outcome Indicators

FS Method	Ind 410	Ind 372	Ind 365	Ind 339
Null Model	1067.59	859.65	976.27	859.66
Full Model	764.60	912.10	956.90	519.48
SHAP	690.53	809.68	896.10	520.31
PI	720.94	833.79	863.44	559.27
RFC G.I	750.13	812.08	912.10	588.86
RFC P.I	870.07	848.55	939.21	724.59
WALD	677.55	834.31	900.27	491.91
S.W. Backwards	644.37	768.56	820.54	421.70
Lasso	701.58	792.72	861.04	492.83
Lasso CV	695.06	851.57	896.51	481.22

Table 17. BIC Scores for each Feature Selection Set across the different Outcome Indicators

FS Method	Ind 410	Ind 372	Ind 365	Ind 339
Null Model	1072.24	864.30	980.92	864.30
Full Model	1368.46	1515.96	1560.76	1123.35

SHAP	867.05	1000.13	1021.52	803.66
PI	1036.81	1093.92	1016.73	772.95
RFC G.I	852.33	932.85	1037.52	691.05
RFC P.I	1213.81	1071.51	1426.95	812.85
WALD	775.10	871.47	969.95	659.14
S.W. Backwards	858.05	940.43	1006.34	714.34
Lasso	933.84	987.81	1121.16	752.96
Lasso CV	1029.51	1283.56	1314.57	931.80

Table 18. Likelihood ratio test compared with Null and Full models for each Feature Selection Set across the different outcome indicators

FS	Ind 410		Ind 372		Ind 365		Ind 339	
	Null Model	Full Model	Null Model	Full Model	Null Model	Full Model	Null Model	Full Model
SHAP	☑	☑	☑	☑	☑	☒	☑	☒
P.I	☑	☑	☑	☑	☑	☑	☑	☒
RFC G.I	☑	☒	☑	☑	☑	☒	☑	☒
RFC P.I	☑	☒	☑	☒	☑	☑	☑	☒
WALD	☑	☑	☑	☒	☑	☒	☑	☒
S. W	☑	☑	☑	☑	☑	☑	☑	☑
Lasso	☑	☑	☑	☑	☑	☑	☑	☒
Lasso CV	☑	☑	☑	☑	☑	☑	☑	☑

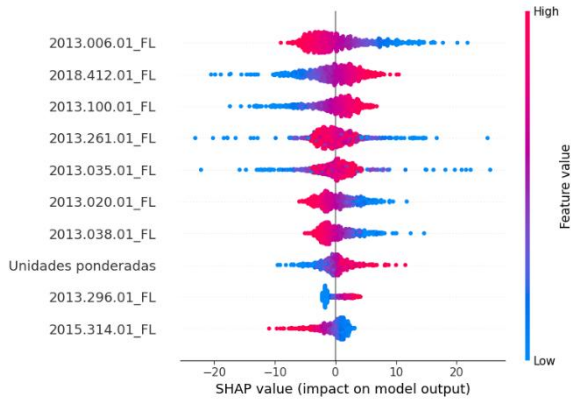

Figure 16. SHAP summary plot for Indicator 410

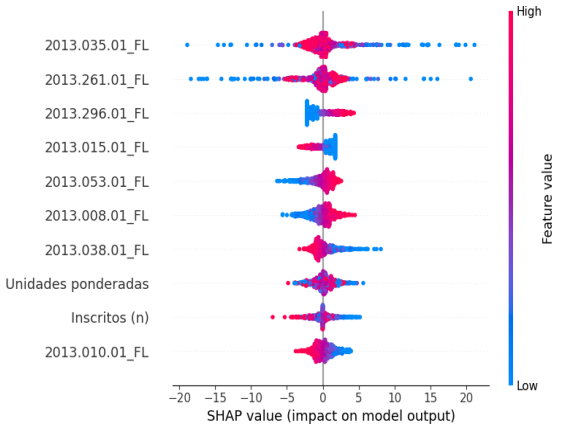

Figure 17. SHAP summary plot for Indicator 372

Indicator 410 was the Outcome Indicator with the highest F1-scores and displayed consistently improved results compared to its Full and Null model counterparts. The AIC and BIC Scores also displayed lower values throughout most Feature Selection sets against the base models. The Stepwise Backwards FS set displayed the best metrics in most evaluation analyses, with the lowest AIC score, and the highest F1-score among

the sets investigated. The model distinguished itself as having the greatest capability for representing the generalisation of its outcome variables, among all analysed. Wald thresholding also displayed positive results, exhibiting the lowest BIC value, demonstrating a greater representation of the significant features in a stricter evaluation system [92]. The worst-performing FS set was a non-linear Feature Selection method, with RFC permutation importance having the highest AIC/BIC scores and the lowest F1-scores. In addition, the non-linear FS sets (RFC Permutation and Gini Importance) were the only models with a negative LRT test result, suggesting the full model was a better fit and still contained relevant information. Regarding specific variables, parameters 2013.006.01_FL and 2018.412.01_FL were consistently present in all Feature Selection sets and valued substantially high in the methods that sorted features by importance/ degree of contribution. Looking at Figure 16, the two features displayed opposite directions in their relationship with Indicator 410. Specifically, Indicator 006 associated lower values with better-performing health care units for outcome 410, whereas Indicator 412 implied that higher values were more associated with label one. Parameters 2013.020.01_FL, 2015.314.01_FL, and 2013.038.01_FL also had an impact on the model, albeit to a lesser extent and displayed similar behaviour to Indicator 006. In terms of their relevance among the different Feature Selection methods, all three indicators were present in all linear FS methods. Features 2013.100.01_FL, 'Unidades ponderadas', and 2013.296.01 displayed similar relationships with the dependent variable, where higher values had stronger associations with better-performing health care units. Of the three, Indicator 100 was the only feature highlighted in all linear FS methods. Indicator 296 and 'Unidades ponderadas' were absent after Wald thresholding, having p-values of 0.109 and 0.567, respectively. Features 2013.261.01_FL and 2013.035.01_FL were the most unique features addressed by SHAP. Both were considered highly impactful, with only Indicators 412, 006 and 100 having a mean SHAP value above them. On the other hand, both indicators were absent in all other Feature reduction models. According to the SHAP plot in Fig. 16, both indicators were the only parameters that suggested primarily low values to affect the prediction of a model, be it positive or negative, while high values had minimal impact.

Indicator 372 was unanimously the Outcome Indicator with the lowest F1-scores of all the outcome variables analysed. The lower scores may suggest that the features do not have an underlying relationship with Indicator 372. One factor to consider is the distinction of Outcome Indicator 372. While the other Outcome Indicators displayed a more general outcome related to hospitalisation events, Indicator 372 was primarily focused on a particular injury (hip fractures). Although this indicator can hint at the

capabilities of a unit addressing residents with chronic pain, its specification means that theoretically a lower number of the indicators investigated should affect the outcome. Hence, a higher difficulty in narrowing down significant features. Considering the other evaluation scores, the AIC scores showed consistently lower values compared to the full model, but similar scores to the null model. This implied a similar fit between the different Feature Selection sets and the intercept-only model. Despite this, both Stepwise Backwards and the Lasso Feature Selection sets displayed lower values, suggesting that some selected features could help understand the associations that differentiate the labels in Indicator 372. On the other hand, the sets displayed persistently higher BIC scores than the null model. In terms of individual parameter importance, there was no specific standout parameter in all the sets discussed. Regardless, Indicators 2013.296.01_FL and 2013.010.01_FL were the variables most consistently present in the different FS sets. Examining Figure 17, both parameters had a low amount of impact on the SHAP model's prediction. The SHAP plot pattern indicated that high values for parameter 296 had a positive association with label 1, while the opposite effect occurred on Indicator 010. Indicators 296 was distinguished as the most impactful parameter in the LogReg Permutation importance analysis, with an importance of 0.11. However, a comparison between the permutation importances on the other Outcome Indicators' models implies a low degree of effect on Indicator 372, since an important variable usually ranged from 0.20 to 0.30.

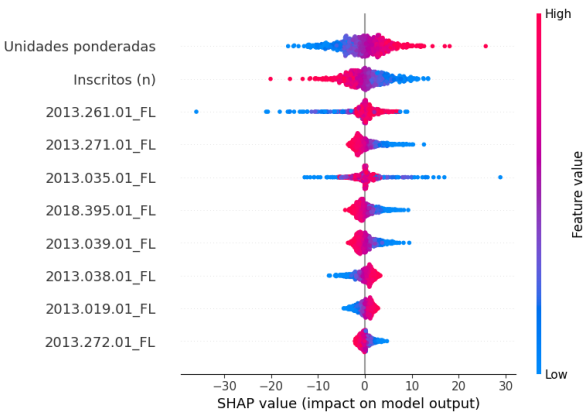

Figure 18. SHAP Summary plot for Indicator 365

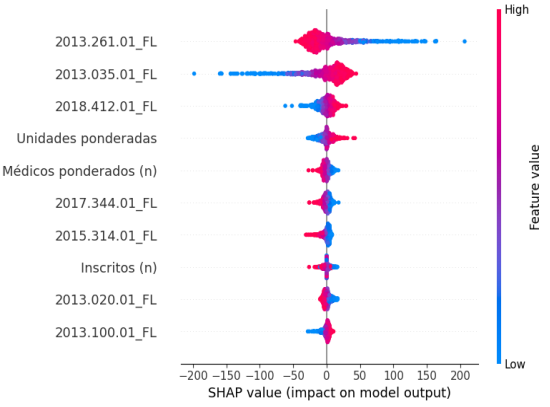

Figure 19. SHAP summary plot for Indicator 339

Indicator 365 had slightly more positive F1-scores, with all models having a range between 0.45 and 0.55. The outlier was the FS set based on Wald thresholding, which had a distinctly lower average, indicating that the mere 14 features were insufficient in explaining the difference in the data. Likewise, the AIC score for this particular set was one of the highest among the different FS methodologies. The other FS sets with higher

AIC scores were the non-linear RFC FS methods. RFC permutation importance, despite having positive Likelihood ratio tests, still developed AIC scores close enough to the Null and Full model, suggesting no significant improvement. The BIC scores show that while the different Feature Selection methods improve the fit when compared with the full model, the null model produces BIC scores lower than the FS sets. Features ‘Unidades ponderadas’ and ‘Inscritos (n)’ were consistently highlighted as important in most Feature Selection methods. This is evident in Figure 18, where these features appeared to have almost double the effect in contribution compared to the rest. ‘Unidades ponderadas’ identified a positive association between higher values and label 1. As expected, ‘Inscritos (n)’ showed that a lower number of registered service users for the unit was associated with lower rates of avoidable hospitalisations. Indicators '2013.039.01_FL' and '2018.395.01_FL' were the only parameters that were present after all the different FS selection methods. The SHAP plot indicates that both these features had very similar distributions, both having a similar impact on the model and suggesting an association between lower values in these features and better performance.

Indicator 339 had positive F1-scores on most Feature Selection sets, displaying higher prediction ability than the full model. On the other hand, the non-linear FS method, RFC Permutation importance, developed models with a low prediction ability. This lack of generalisation and performance is also evident in a significantly higher AIC score, demonstrating a lower capability in balancing the complexity of the data and its fit [92][93]. In terms of the LRT test, both the stepwise regression and Lasso CV selected features were the only models with positive scores. The BIC tests displayed a clear improvement when compared with the full parameter model. Contrarily, when compared with the null model, the models showed no improvement with Lasso CV, and mild improvement to significant improvement in the remaining FS sets. Looking at the most important features and their associations, a clear parallel between Indicators 339 and 410 is evident. Similar variables such as 2018.412.01_FL and '2015.314.01_FL were highlighted as important while showing identical association directions, albeit with lesser impact on the model.

Examining the different tables (Tables 14 - 17), the tendency of the result scores for each Outcome Indicator appeared similar, regardless of the Feature Selection set. As expected, due to their scoring criteria, Stepwise Backwards FS had the lowest AIC scores. Likewise, in the remaining evaluation metrics analysed, it displayed the most promising results. This is surprising, given the numerous setbacks associated with this

Feature Selection method and its controversial status [95][96][97][98]. Main concerns with Stepwise Backwards selection include how the feature reduction technique may overfit the training data, overemphasising the importance of some features due to bias. There is also the removal of parameters that may be significant due to its computing-focused framework, often leading to surprising results through a drop in performance when evaluated with cross-validation [95][96][97][98]. To address these issues, we used expert recommendations in ways to restrain its limitations and take advantage of its perks, such as the use of cross-validation, the gathering of subsets of features through other methods for comparison, and the evaluation with multiple metrics (F1-score, AIC, BIC, LRT, Stability) [99]. Default Lasso Regression, SHAP, and Permutation Importance were also standout Feature reduction methods that displayed positive results. Despite their different methodologies in providing feature importance, the sets often highlighted the same core 5-10 parameters as the most influential. On the other hand, the non-linear FS methods showed more disappointing performances. This can be attributed to how RFC and LogReg differ when handling high variance in the parameters. LogReg can ignore redundant fluctuations and noise. RFC, on the other hand, has a bigger tendency to overfit due to these variations [100]. Another reason, RFC performance could be worse, is simply due to the relationship between the dependent variables and the parameters being mostly linear. This is mainly hinted at through the different permutation importance scores for each variable. Using Logistic regression as the estimator, multiple features had an importance of around 0.20 to 0.30. Conversely, Random Forest had significantly lower importance values, with none of the parameters exceeding an importance of 0.08. Many articles also suggest an alternative method for distributing feature importance by focusing on the parameters referenced by an expert opinion [101][102]. However, this methodology led to the issues currently observed in the IDE system, including an overemphasis on specific features and the misguided use of Outcome Indicators in the 411-A indicator matrix [102][3][2][7]. This dissertation opted for the reliability of a mathematical background and Machine learning applications to discover the associations and the different levels of importance.

The Likelihood ratio test comparing the different subsets of parameters with the null model suggested a more positive outlook, with all the subsets having a better fit (Table 18). In contrast, some AIC/BIC scores, in particular for Indicators 372 and 365, suggested no improvement by the addition of the different features. Despite both being used for comparing nested models, they provide different information. Likelihood ratio tests helped provide information on whether the more complex model was significantly better due to the additional information [103]. AIC provided a numeric value to

understand the efficiency of a model by finding the best trade-off between fit and complexity [92][93]. These differences in objectives and penalisation techniques can often lead to conflicting results.

Table 19. Average Jaccard Similarity between the Feature Selection Sets and the highlighted features per bootstrap

FS Method	Ind 410	Ind 372	Ind 365	Ind 339
SHAP	0.47	0.45	0.45	0.47
PI	0.55	0.46	0.49	0.35
RFC G.I	0.52	0.24	0.42	0.55
RFC P.I	0.39	0.34	0.65	0.13
Wald	0.37	0.16	0.27	---
S.W Backwards	0.40	0.29	0.37	0.43
Lasso	0.54	0.45	0.51	0.59
Lasso CV	0.58	0.65	0.65	0.70

Table 20. Percentage of features per Feature Selection Set that were present in more than 70% of their bootstraps

FS Method	Ind 410	Ind 372	Ind 365	Ind 339
SHAP	72.97	72.50	65.38	95.00
PI	47.76	63.64	43.75	91.11
RFC G.I	66.67	24.00	30.77	52.38
RFC P.I	13.69	80.85	57.69	22.22
Wald	65.00	42.86	35.71	---
S.W Backwards	64.44	44.44	66.67	46.77
Lasso	46.94	39.02	41.81	47.27
Lasso CV	50.70	58.70	68.54	66.67

The average Jaccard similarity of features displayed that Lasso CV had the highest consistency among the different bootstraps and Outcome Indicator evaluations (Table 19). On the other hand, Wald's method was clearly unstable. This FS method had the lowest scores compared to all the other FS sets. In particular, Indicator 339 did not have a score due to a syntax error in a bootstrap, revealing perfect separation. This suggests that Indicator 339 had overly optimistic values in the previous evaluation metrics.

The percentage coverage analysis revealed different interpretations of the Feature Selection sets (Table 20). SHAP had the most consistent results, where all Outcome Indicators had more than 65% of the subset present in more than 7 of the 10 bootstraps. The differences in results on the two stability tests highlight the differing philosophies in the Feature Selection methods. Lasso Regression reduces complexity by shrinking feature coefficients to 0. Correlated variables could incentivise Lasso to declare one of them redundant. Across the 10 bootstraps, different correlated variables could be classified as important, lowering the percentage of coverage. The SHAP's marginal contribution approach is used mainly as an explanatory tool. This means all features that have a contribution effect are highlighted, regardless of the correlation between them. The high percentage coverage scores spotlight the consistency of attributed high contributions (SHAP values) to the specified features throughout the different bootstraps in SHAP. Contrastingly, the influence of the bootstrap effect on the data may have resulted in the fixed threshold impacting the Jaccard similarity scores, either causing an influx or a reduction in the number of impactful features.

After evaluating all the Feature Selection sets, the groups chosen to produce the final cluster of features for the Odds Ratio analysis were Stepwise Backwards FS, Lasso Regression CV, and SHAP thresholding. Stepwise Backwards consistently had the best performing AIC, BIC, LRT and F1-scores. LassoCV and SHAP had the most consistently stable selection of features after the bootstrap evaluation. This combination of features enables the collection of the most reliable core of features while excluding the biases of each Feature Selection method.

5.5. Odds Ratio Analysis

Table 21. Odds Ratio Table of Associations for Outcome Indicators 410, 372, 365, 339

Variable	Ind 410	Ind 372	Ind 365	Ind 339
2013.001.01_FL				**0.96 (0.93 - 0.99)**
2013.002.01_FL	0.98 (0.88 - 1.09)	0.92 (0.83 - 1.02)		**0.79 (0.67 - 0.93)**
2013.005.01_FL	**0.97 (0.95 - 0.98)**			**0.97 (0.94 – 1.00)**
2013.006.01_FL	**0.69 (0.57 - 0.84)**	**1.13 (1.00 - 1.28)**		0.95 (0.71 - 1.26)
2013.008.01_FL		1.02 (0.97 - 1.07)		
2013.009.01_FL				**0.95 (0.91 - 0.99)**
2013.010.01_FL		**0.95 (0.91 - 0.99)**		
2013.011.01_FL	0.98 (0.93 - 1.02)			0.97 (0.91 - 1.03)

2013.014.01_FL	**1.03 (1.00 - 1.06)**			1.05 (0.99 - 1.10)
2013.015.01_FL	0.97 (0.93 - 1.01)	0.97 (0.94 - 1.01)		
2013.016.01_FL				0.98 (0.94 - 1.01)
2013.017.01_FL	1.00 (0.98 - 1.02)	1.00 (0.98 - 1.02)		1.01 (0.97 - 1.04)
2013.018.01_FL	**1.06 (1.02 - 1.11)**		1.04 (0.99 - 1.09)	
2013.019.01_FL			1.03 (0.99 - 1.06)	
2013.020.01_FL	**0.91 (0.84 - 0.98)**	0.97 (0.94 - 1.00)		**0.87 (0.77 - 0.98)**
2013.023.01_FL			**1.04 (1.01 - 1.07)**	
2013.030.01_FL			1.03 (0.99 - 1.07)	
2013.031.01_FL		1.04 (1.00 - 1.07)	1.02 (0.99 - 1.05)	**1.10 (1.03 - 1.16)**
2013.032.01_FL		**0.97 (0.95 - 0.98)**		
2013.034.01_FL		1.03 (0.99 - 1.07)		**1.11 (1.05 - 1.18)**
2013.035.01_FL			1.15 (0.56 - 2.35)	1.86 (0.64 - 5.35)
2013.036.01_FL		**0.97 (0.95 - 0.99)**		**0.93 (0.89 - 0.97)**
2013.038.01_FL	**0.93 (0.90 - 0.97)**	0.95 (0.90 - 1.00)	1.03 (0.97 - 1.09)	
2013.039.01_FL		**1.08 (1.00 - 1.17)**	0.88 (0.80 - 0.96**)**	
2013.046.01_FL	**0.96 (0.93 - 0.99)**			
2013.049.01_FL	**1.04 (1.02 - 1.05)**			**1.03 (1.00 - 1.06)**
2013.053.01_FL		**1.11 (1.01 - 1.23)**		1.06 (0.92 - 1.22)
2013.054.01_FL		**0.97 (0.95 - 1.00)**		0.97 (0.93 - 1.01)
2013.063.01_FL		0.97 (0.93 - 1.01)	0.97 (0.93 - 1.01)	**0.92 (0.86 - 0.99)**
2013.091.01_FL			**1.05 (1.01 - 1.10)**	
2013.094.01_FL	0.95 (0.9 - 1)	**0.92 (0.87 - 0.98)**	0.97 (0.92 - 1.02)	0.99 (0.94 - 1.04)
2013.097.01_FL				**0.94 (0.90 - 0.98)**
2013.098.01_FL			0.95 (0.91 – 1.00)	
2013.099.01_FL	**0.95 (0.91 - 0.99)**			0.97 (0.89 - 1.07)
2013.100.01_FL	**1.46 (1.20 - 1.77)**		**1.17 (1.09 - 1.25)**	1.33 (0.99 - 1.78)
2013.261.01_FL			0.89 (0.44 - 1.82)	0.55 (0.19 - 1.60)
2013.262.01_FL	0.98 (0.95 - 1.00)			
2013.267.01_FL		7.46 (0.01 - 6692.38)		
2013.269.01_FL			0.20 (0.01 - 2.99)	
2013.270.01_FL	>0.01 (>0.01 - 0.12)			>0.01 (>0.01 – 1.86)
2013.274.01_FL		**0.97 (0.96 - 0.99)**		**0.97 (0.94 - 0.99)**
2013.275.01_FL		1.01 (1.00 - 1.03)		0.98 (0.96 - 1.01)
2013.276.01_FL			**1.05 (1.01 - 1.10)**	1.08 (1.00 - 1.17)
2013.277.01_FL			**0.95 (0.93 - 0.98)**	1.01 (0.96 - 1.07)
2013.278.01_FL	0.93 (0.87 - 1.00)		**0.89 (0.83 - 0.95)**	
2013.294.01_FL			**1.00 (1.00 - 1.00)**	
2013.295.01_FL				1.01 (0.98 - 1.05)

2013.296.01_FL	**1.05 (1.00 - 1.10)**	1.04 (1.00 - 1.08)	1.00 (0.99 - 1.01)	1.00 (0.99 - 1.02)
2013.297.01_FL			**1.08 (1.03 - 1.14)**	**1.31 (1.04 - 1.65)**
2013.300.01_FL			**69.79 (4.18 - 1164.97)**	
2013.413.01_FL			**0.99 (0.97 – 1.00)**	
2015.306.01_FL	0.94 (0.88 - 1.00)	1.04 (0.99 - 1.09)		
2015.307.01_FL		1.02 (1.00 - 1.05)		
2015.308.01_FL				**1.05 (1.01 - 1.10)**
2015.309.01_FL				**0.95 (0.91 - 0.99)**
2015.310.01_FL	**327.26 (9.40 - 11399.46)**			**367.96 (1.41 - 96113.36)**
2015.311.01_FL		6.49 (0.46 - 92.08)		
2015.312.01_FL			**142.42 (12.64 - 1604.08)**	
2015.314.01_FL	**0.91 (0.86 - 0.96)**		0.95 (0.90 - 1.01)	**0.89 (0.82 - 0.97)**
2015.315.01_FL		1.03 (0.99 - 1.06)		
2015.316.01_FL	1.05 (0.97 - 1.13)		0.97 (0.94 – 1.00)	**1.14 (1.02 - 1.27)**
2017.257.01_FL	**1.20 (1.05 - 1.37)**			
2017.331.01_FL		8.32 (0.63 - 110.39)	**26.36 (3.07 - 226.53)**	1.08 (0.01 - 159.67)
2017.335.01_FL				0.96 (0.92 - 1)
2017.342.01_FL			0.98 (0.95 - 1.01)	
2017.344.01_FL	**0.88 (0.84 - 0.92)**			**0.73 (0.68 - 0.79)**
2017.350.01_FL			1.00 (0.99 - 1.00)	1.00 (1.00 - 1.01)
2017.353.01_FL		1.00 (0.99 - 1.02)		
2017.354.01_FL	1.02 (1.00 - 1.05)			1.01 (0.98 - 1.05)
2017.378.01_FL				0.67 (0.43 - 1.04)
2017.379.01_FL	2.10 (0.87 - 5.09)	0.16 (0.02 - 1.34)		**19.87 (6.61 - 59.79)**
2017.381.01_FL		0.97 (0.94 - 1.01)		0.97 (0.93 - 1.03)
2017.382.01_FL			1.09 (0.97 - 1.22)	1.1 (0.92 - 1.31)
2017.384.01_FL	**0.94 (0.89 - 0.98)**		0.98 (0.95 - 1.01)	**0.94 (0.88 - 0.99)**
2017.389.01_FL			0.70 (0.48 - 1.04)	
2017.390.01_FL	**0.37 (0.19 - 0.72)**			
2017.391.01_FL				**0.35 (0.17 - 0.72)**
2017.392.01_FL	**0.31 (0.18 - 0.56)**	**1.79 (1.04 - 3.08)**		
2017.393.01_FL	**3.32 (1.79 - 6.16)**			**7.63 (3.13 - 18.6)**
2018.395.01_FL		0.91 (0.82 - 1.01)	**0.93 (0.89 - 0.97)**	0.89 (0.76 - 1.05)
2018.397.01_FL	1.02 (1.00 - 1.05)			**0.95 (0.92 - 0.99)**
2018.398.01_FL	**0.99 (0.98 - 1.00)**	1.01 (1.00 - 1.02)		
2018.404.01_FL			**1.01 (1.00 - 1.01)**	1.00 (1.00 - 1.01)

2018.405.01_FL	**1.24 (1.17 - 1.30)**		**1.05 (1.03 - 1.08)**	**1.24 (1.17 - 1.32)**
2018.409.01_FL	**1.31 (1.13 - 1.52)**			0.69 (0.40 - 1.20)
2018.412.01_FL	**1.33 (1.27 - 1.40)**	1.02 (1.00 - 1.05)		**1.59 (1.45 - 1.75)**
2019.414.01_FL				10.76 (0.05 - 2505.38)
2019.427.01_FL				1.02 (0.99 - 1.05)
2019.428.01_FL				0.38 (0.14 - 1.04)
2019.430.01_FL~	1.91 (0.97 - 3.75)			**3.03 (1.04 - 8.86)**
2021.439.01_FL	**1.08 (1.05 - 1.12)**	1.18 (0.65 - 2.16)		**1.12 (1.07 - 1.17)**
"Enfermeiros (n)"	0.87 (0.75 - 1.01)	0.87 (0.73 - 1.03)	**0.79 (0.70 - 0.90)**	0.87 (0.69 - 1.10)
"Idade US"	0.95 (0.90 - 1.01)		1.05 (1.00 - 1.10)	
"Indice de contexto sociodemográfico"		**0.88 (0.80 - 0.97)**		0.94 (0.88 - 1.01)
"Inscritos (n)"		1.00 (1.00 - 1.00)	**1.00 (1.00 - 1.00)**	1.00 (1.00 - 1.00)
"Médicos (n)"		1.19 (1.00 - 1.43)		1.18 (0.90 - 1.55)
"Médicos internos (n)"	**1.23 (1.09 - 1.39)**		1.09 (0.99 - 1.21)	
"Médicos ponderados (n)"	**0.64 (0.44 - 0.92)**			**0.35 (0.20 - 0.63)**
"Outros profissionais (n)"		0.82 (0.67 - 1.01)		
"Prevalencia_DM"			**0.64 (0.54 - 0.76)**	**0.68 (0.51 - 0.91)**
"Proporção idosos"				0.95 (0.80 - 1.11)
"Taxa abandono escolar"		1.07 (1.00 - 1.15)		
"Taxa de desemprego"	1.14 (0.91 - 1.42)	**1.49 (1.04 - 2.12)**	**1.44 (1.17 - 1.78)**	
"Unidades ponderadas"	**1.00(1.00 - 1.00)**	1.00(1.00 - 1.00)	**1.00(1.00 - 1.00)**	1.00 (1.00 - 1.00)
Pseudo R^2	**0.4699**	**0.1844**	**0.2372**	**0.5839**

The McFadden pseudo R^2 can often give a pessimistic outlook on the general improvement of a model. Due to the formulation of the metric, it is stated that values between 0.2 and 0.4 are considered very good fits and an indication of a significantly improved model [104]. Using this reference, we should take away an understanding that Indicators 410, 365 and 339 have models with good pseudo R^2 scores. Interestingly, Indicator 339 had a score above 0.5, an unlikely value. This may indicate overfitting or perfect separation. Additionally, Indicator 372 contained a McFadden R^2 score below the literature threshold and was the Outcome Indicator with the lowest amount of significant

features in the model, with only 11 (Table 21). The low pseudo R^2 score indicated that the additional features only improved the model marginally, suggesting a lack of influential features in the BD_BICSP_DATASET specific to this Outcome Indicator. This demonstrates a need to develop more up-to-date and relevant indicators that can help visualise and assess Outcome Indicator 372. The lack of definitive, significant features specific to this indicator may lead us to collect unreliable associations, overestimating the wrong features. Associations that, on raw data, may reveal one particularity can be confounded by redundant features and the addition of noise, leading to less reliable odds ratios [27]. In addition, other associations seemed to indicate a connection, but further investigation revealed that they should not interact.

Of the statistically significant variables in the Outcome Indicator 410's model, features 2013.100.01_FL (OR = 1.46 (95% C.I.: 1.20 - 1.77)), 2018.405.01_FL (OR = 1.24 (95% C.I.: 1.17 - 1.30)), 2018.409.01_FL (OR = 1.31 (95% C.I.: 1.13 - 1.52)), 2018.412.01_FL (OR = 1.33 (95% C.I.: 1.27 - 1.40)), and "Medicos internos (n)" (OR = 1.23 (95% C.I.: 1.09 - 1.39)) had a positive association. Additionally, features 2015.310.01_FL (OR = 327.26 (95% C.I.: 9.40 - 11399.46)) and 2017.393.01_FL (OR = 3.32 (95% C.I.: 1.79 - 6.16)) had the strongest positive associations with this outcome variable. Contrarily, variables 2017.344.01_FL (OR = 0.88 (95% C.I.: 0.84 - 0.92)), 2013.006.01_FL (OR = 0.69 (95% C.I.: 0.57 - 0.84)), "Médicos ponderados (n)"(OR = 0.64 (95% C.I.: 0.44 - 0.92)), 2017.390.01_FL (OR = 0.37 (95%C.I.: 0.19 - 0.72)), and 2017.392.01_FL (OR = 0.31 (95% C.I.: 0.18 - 0.56)) were negatively associated with Outcome 410.

Indicator 365 had strong positive associations with 2013.100.01_FL (OR = 1.17 (95% C.I.: 1.09 - 1.25)), 2013.300.01_FL (OR = 69.79 (95% C.I.: 4.18 - 1164.97)), 2015.312.01_FL (OR = 142.42 (95% C.I.: 12.64 - 1604.08)), 2017.331.01_FL (OR = 26.36 (95% C.I.: 3.07 - 226.53)). The Prevalence of Diabetes (OR = 0.64 (95% C.I.: 0.54 - 0.76)) was the feature with the strongest negative association with Indicator 365.

Indicator 339 had strong positive associations with 2013.297.01_FL (OR = 1.31 (95% C.I.: 1.04 - 1.65)), 2015.310.01_FL (OR = 367.96 (95% C.I.: 1.41 - 96113.36)), 2015.316.01_FL (OR = 1.14 (95% C.I.: 1.02 - 1.27)), 2013.100.01_FL(OR = 1.33 (95% C.I.: 0.99 - 1.78)), 2017.379.01_FL (OR = 19.87 (95% C.I.: 6.61 - 59.79)), 2017.393.01_FL (OR = 7.63 (95% C.I.: 3.13 - 18.6)), 2018.405.01_FL (OR = 1.24 (95% C.I.: 1.17 - 1.32)), 2018.412.01_FL (OR = 1.59 (95% C.I.: 1.45 - 1.75)), 2019.430.01_FL (OR = 3.03 (95% C.I.: 1.04 - 8.86)). The strong negative associations included 2013.002.01_FL (OR = 0.79 (95% C.I.: 0.67 - 0.93)), 2013.020.01_FL (OR = 0.87 (95% C.I.: 0.77 - 0.98)), 2017.344.01_FL (OR = 0.73 (95% C.I.: 0.68 - 0.79)), 2017.391.01_FL

(OR = 0.35 (95% C.I.: 0.17 - 0.72)), “Médicos ponderados (n)” (OR = 0.35 (95% C.I.: 0.2 - 0.63)), and “Prevalencia_DM” (OR = 0.68 (95% C.I.: 0.51 - 0.91)).

Examining Outcome Indicators 410, 365 and 339, we can see that most of the strong positive associations were related to the continuous monitoring of different service users and their diseases. The Odds Ratios also revealed that an improvement in accessibility to healthcare units other than where the service user is registered can have a substantial impact on these Outcome Indicators. A systematic review conducted in 2015 explored numerous studies and different intervention methods to reduce the number of frequent emergency service visits. In this review, similarly to our Odds Ratio table, two of the most consistent intervention approaches were case management and information sharing. After studying 17 different articles, the author highlights the importance of managing varying service users' conditions, stating that proactive monitoring and follow-ups were essential in helping detect issues before emergencies could occur [105]. Indicator 2017.393.01_FL assigns a score to each healthcare unit's capabilities in the internal training of the different multidisciplinary teams. This feature was very significant for both Outcome Indicators 410 and 339. This highlights the importance of well-trained professionals in reducing the number of emergency episodes in a healthcare unit.

Indicators 2013.006.01_FL and 2013.002.01_FL describe the overall utilisation rates of medical consultations. According to our models, there was a negative association between these indicators and the Outcome Indicators 410 and 339. This meant that an increase in overall consultations usually indicated a higher level of intensity related to frequent hospital emergencies. The prevalence of diabetes was also a significant feature for both Outcome Indicators 339 and 365. This underscores the importance of understanding the healthcare environment and regional conditions, as it shows that units with a higher prevalence of diabetes require more targeted resources to help prevent avoidable hospitalisations and emergency admissions. According to a study authored by A. Ramalho et al. [106], there is an inefficiency in the resources used to address patients with diabetes. Through the assessment of 54 different ACeS health groups, only 13 reached their established efficiency frontier. The study further suggests that the better management of human resources could improve these scores by around 50%. This is supported by our Outcome Indicator models, which place a strong emphasis on supporting programs focused on the monitoring and management of chronic diseases, including diabetes.

For Outcome Indicator 372, significantly fewer features had both substantial and significant associations in the model. Of these, Social Demographic Index (OR = 0.88

(95% C.I.: 0.80 - 0.97)) and Indicator 392 (OR = 1.79 (95% C.I.: 1.04 - 3.08)) were the standout features. The relationship between the Social Demographic index and outcome 372 was evident, considering how this feature was calculated and influenced by the proportion of elderly individuals in the healthcare system. Lower Social Demographic scores meant more desirable regional conditions, which were consistent with the odds ratio of this parameter [16]. A lower Social Demographic index score was correlated with a lower amount of elderly individuals, which meant a lower probability of having issues related to service users above the age of 75 developing femoral neck fractures. Indicator 392 calculated the consistency with which prevention strategies were implemented. The original purpose for developing Outcome Indicator 372 was to monitor actions implemented by HCU to reduce falls in the elderly population. The odds ratio for this particular feature highlighted this objective, since it showcased a positive association between implementing more strategies and reducing neck fracture hospital admission rates.

Looking at Table 21, many of the significant features had an exposure close to one, suggesting no overall impact on the model. Despite this, the confidence intervals of these variables gave us further insight into whether they were essential to an Outcome Indicator. A variable with confidence intervals crossing one was a key giveaway on whether the associations of these variables were significant. An example of this was in Indicator 410's model with feature 2013.011.01_FL (OR = 0.98 (95% C.I.: 0.93 - 1.02)), which demonstrated that this feature was insignificant to outcome 410. On the other hand, several features were significant, with both the odds ratios and the confidence intervals marginally above 1.00. At first glance, the impact of these features may seem inconsequential, but they still displayed a slight association. The analysis of a continuous variable's odds ratio shows the effect of a one-unit increase in the exposure event (label 1) [27]. One example being Indicator 2013.014.01_FL (OR: 1.03 (95% C.I.: 1.00 - 1.06), where an increase of one unit in the proportion of at least one consultation of a newborn in 28 days increased the probability of having a low rate of frequent emergency users by three per cent. These variables have a moderate effect and can incrementally impact the Outcome Indicator.

Other variables had odds ratio values that seemed more impactful for the outcome, but had p-values above 0.05 and confidence intervals crossing one. An example of this instance was feature 2013.035.01_FL with OR = 1.86 (95% C.I.: 0.64 – 5.35) for Outcome Indicator 339. This table also displayed variables with considerable odds ratio values, such as Indicator's 2015.310_FL's OR = 327.26 (95% C.I.: 9.40 - 11399.46).

These variables typically had a significant value below 0.05, despite having the widest confidence interval margins in the table. The wide intervals showed us that while these variables still had strong associations and their direction was clear, there was still some uncertainty around their exact odds ratio value.

One factor that can significantly affect the model is the issue of multicollinearity. Reinforcing Miguel Melo and Jaime Correia de Sousa, many of the indicators had a questionable scientific basis [12]. In addition, several indicators display overlapping information, either due to a lack of commitment when logging the data by professionals or due to oversight in their development. This caused problems with the modelling of the data, as it would introduce correlated variables, adding noise and affecting the logistic regression's calculated coefficients. The indicators are the most essential component in this evaluation method. The representation of all relevant processes and structural components of the healthcare unit is crucial in uncovering the most consistent and accurate associations possible. This enables us to understand where and how implementations should be made for continuous improvement in healthcare quality.

6. Conclusion

Throughout the ever-changing landscape of Primary Health Care developments, a shift has occurred in Portugal. The introduction of the Pay for Performance evaluation metric, IDE, marks a particular landmark that represents a departure from the goals and commitments made by our pioneers, who strove for continuous improvement in healthcare quality and accessibility. This IDE system simplifies the complexity and nuance of a healthcare unit's quality, relying on an arbitrary 0 to 100 scoring system that cannot accurately evaluate the unit's performance. This was clear through our IDE analysis of its reflection of quality. The comparison between the IDE scores and the different Outcome Indicator values showed only a marginal relationship in the MLR and the correlation plots. Additionally, the focus on each healthcare unit by region or unit type revealed extreme variability in this relationship, particularly in Alentejo, where the correlation suggested better quality healthcare units were more consistent with lower IDE scores. This scoring system also incentivised the funnelling of healthcare professionals' focus on the few indicators that were classified as necessary. This is especially concerning, considering the lack of credible scientific evidence on specific indicators and on the different components of the IDE's indicator matrix.

This dissertation utilised the four Outcome Indicators that were observed in the BD_BICSP_2023 dataset as references, recognising their capability in representing quality. In addition, categories were created using the international averages from these Outcome Indicators, which helped spotlight the best-performing features from the rest. This developed issues in the modelling, mainly due to the lack of sufficiently balanced distributions between the two different category labels. To counteract this, the unfeasible benchmarks were substituted with the lower quartile value scores of the year. This approach was derived from the intuition that by committing to lowering these quality Outcome Indicator scores, the international benchmark could be used in the future. All active indicators, as well as 'Structural and Demographic' variables, were gathered, and numerous Feature Selection techniques were performed, where the best performing and most consistently stable sets came from Stepwise Backwards Feature Selection, LassoCV, and SHAP. The sets enabled the collection of the best group of features for each Outcome Indicator as well as the development of the odds ratio associations and their respective confidence intervals. Among the four different models, the Outcome Indicators that performed best were Indicators 410 and 365. Indicator 339 developed an overly optimistic model that relied on multicollinearity and was overfitted. Indicator 372 had the worst performance, suggesting a need for a different collection of indicators that

could better impact the outcome. Nevertheless, the Odds Ratio models still showed interesting associations. In particular, the impact that case management and communication can have on the intensity of emergency services. These associations reinforced the importance of proactive monitoring in detecting issues before emergencies could occur.

For future developments, allocating more time and resources could help mitigate the limitations imposed by computational costs. These include expanding the various Feature Selection methods and incorporating cross-validation into the different permutation importance and stepwise regression analyses. Since the dataset used in this dissertation is from 2023, a more recent collection of data can help narrow the features and their associations to more current issues. Additionally, more awareness needs to be placed on accurately logging information and monitoring the healthcare unit's performance. Missing or incorrect values are detrimental to the evaluation system, as they hinder the development of models by either classifying the unit's information as unfeasible for model production or by misleading the models and providing incorrect associations. A greater amount of data could also provide flexibility to the model by enabling the creation of more personalised odds ratio associations, either by separating the units by region or unit type. Finally, a complete and thorough review of all the current active indicators is necessary. This helps remove redundant features that either contain overlapping information or provide similar details, resulting in high correlation. This indicator review could also help identify the particularities of a unit that are missing and may be crucial in locating the variables that have the most impact on the Outcome Indicators, such as Indicator 372.

In conclusion, the monitoring of performance is paramount for healthcare teams to identify weaknesses and capitalise on their strengths. To that end, the restructuring of the current IDE evaluation system is crucial for our healthcare units to contribute to the best of their abilities. The evaluation system proposed in this dissertation aims to shift the current PHC perspective in Portugal from a checklist mentality and policing work to one based on quality assessment, guiding healthcare units towards allocating resources that reinforce continuous and sustainable improvement for the betterment of our community.

References

[1] World Health Organization & United Nations Children's Fund (UNICEF). (2018). *A vision for primary health care in the 21st century: Towards universal health coverage and the Sustainable Development Goals*. World Health Organization. https://iris.who.int/bitstream/handle/10665/328065/WHO-HIS-SDS-2018.15-eng.pdf?sequence=1&isAllowed=y

[2] Silva, C. F., Beirão, D., & Santos, P. (2021). 40 anos de desenvolvimento dos Cuidados de Saúde Primários em Portugal. In *International handbook for the advancement of public health policies–Health Policy, Planning and Management* (pp. 30-48). Porto, Portugal: Publicações ESS.

[3] Santos, P. (2024). A saúde em Portugal no equilíbrio entre a acessibilidade e a qualidade. *Revista Portuguesa de Medicina Geral e Familiar*, *40*(5), 428-9.https://doi.org/10.32385/rpmgf.v40i5.14173

[4] Magalhães, Nuno. (2013). *Unidades de Saúde Familiar: O seu papel na reforma dos Cuidados de Saúde Primários.* Faculdade de Medicina da Universidade de Coimbra.

[5] Ministerio da Saude e Assistencia. (1971). *Decreto-Lei n.º 413/71*. Diário do Governo n.º 228/1971, Série I de 1971-09-27, páginas 1406 - 1434. https://diariodarepublica.pt/dr/detalhe/decreto-lei/413-1971-632738

[6] Ministerio da Saude. (2007). *Decret-Lei n.º 298/2007*. Diário da República n.º 161/2007, Série I de 2007-08-22, páginas 5587 - 5596. https://diariodarepublica.pt/dr/detalhe/decreto-lei/298-2007-640665

[7] De Sousa, Jaime. (2019). Celebrando 20 anos do Regime Remuneratório Experimental em Cuidados de Saúde Primários: uma reflexão pessoal sobre um percurso único. *Revista Portuguesa de Medicina Geral e Familiar*, *35*(5), 340-4. DOI: 10.32385/rpmgf.v35i5.12619

[8] Turcotte-Tremblay, A. M., Gali Gali, I. A., & Ridde, V. (2022). An Exploration of the Unintended Consequences of Performance-Based Financing in 6 Primary Healthcare Facilities in Burkina Faso. *International journal of health policy and management*, *11*(2), 145–159. https://doi.org/10.34172/ijhpm.2020.83

[9] World Health Organization & United Nations Children's Fund. (1978). *Declaration of Alma-Ata*. International Conference on Primary Health Care. https://www.who.int/teams/social-determinants-of-health/declaration-of-alma-ata

[10] Reis, I., Envia, G., & Santos, P. (2022). Impact of the primary care residents on the productivity of the ambulatory health centres in Portugal: a cross-sectional study. *BMC Medical Education*, *22*(1), 465. https://doi.org/10.1186/s12909-022-03528-y

[11] Petersen, L. A., Woodard, L. D., Urech, T., Daw, C., & Sookanan, S. (2006). Does pay-for-performance improve the quality of health care?. *Annals of internal medicine*, *145*(4), 265–272. https://doi.org/10.7326/0003-4819-145-4-200608150-00006

[12] Melo, M., De Sousa, J. (2011). Os indicadores de desempenho contratualizados com as USF: Um ponto da situação no actual momento da reforma. *Revista Portuguesa De Medicina Geral E Familiar*, *27*(1), 28-34. https://doi.org/10.32385/rpmgf.v27i1.10818

[13] Biscaia, A. R., & Heleno, L. C. (2017). Primary Health Care Reform in Portugal: Portuguese, modern and innovative. A Reforma dos Cuidados de Saúde Primários em Portugal: portuguesa, moderna e inovadora. *Ciencia & saude coletiva*, *22*(3), 701–712. https://doi.org/10.1590/1413-81232017223.33152016

[14] Biscaia, A. (2006). A reforma dos cuidados de saúde primários e a reforma do pensamento. *Revista Portuguesa De Medicina Geral E Familiar*, *22*(1), 67-79. https://doi.org/10.32385/rpmgf.v22i1.10211

[15] Silva, M. (2023). Os cuidados de saúde primários em Portugal: uma história de mais de 50 anos – a perspetiva de um futuro médico de família . *Revista Portuguesa De Medicina Geral E Familiar*, *39*(4), 355-9. https://doi.org/10.32385/rpmgf.v39i4.13732

[16] Ministério da Saúde. (2017). *Bilhete de Identidade dos Indicadores dos Cuidados de Saúde Primários para o Ano de 2017*. Administração Central do Sistema de Saúde, IP.

[17] Nogueira, S., Sechidis, K., & Brown, G. (2018). On the stability of feature selection algorithms. *Journal of Machine Learning Research*, *18*(174), 1-54

[18] Finanças e Saúde. (2023). *Portaria n.º 411-A/2023*. Diário da República n.º 234/2023, 1º Suplemento, Série I de 2023-12-05, páginas 2 – 11 . https://diariodarepublica.pt/dr/detalhe/portaria/411-a-2023-225265452

[19] T. Hastie, R. Tibshirani, and J. Friedman. (*2001*). The Elements of Statistical learning. *Springer Series in Statistics Springer New York Inc., New York, NY, USA,* https://www.sas.upenn.edu/~fdiebold/NoHesitations/BookAdvanced.pdf

[20] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

[21] Montgomery, D. C., & Runger, G. C. (2019). *Applied statistics and probability for engineers*. John wiley & sons.

[22] Osborne, J. W., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical assessment, research, and evaluation*, *8*(1). https://www.researchgate.net/publication/234616195_Four_Assumptions_of_Multiple_Regression_That_Researchers_Should_Always_Test

[23] Frost, Jim. (2019) *Regression Analysis: An Intuitive Guide for Using and Interpreting Linear Models*. Statistics by Jim Publishing

[24] James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2021). *An Introduction to Statistical Learning with Applications in R* (Second Edition). Springer. https://www.stat.berkeley.edu/~rabbee/s154/ISLR_First_Printing.pdf

[25] *R Programming for Statistics*. (2024). (n.p.): EduGorilla Publication.

[26] T. Hastie, R. Tibshirani, and J. Friedman. (*2001*). The Elements of Statistical learning. *Springer Series in Statistics Springer New York Inc., New York, NY, USA,* https://www.sas.upenn.edu/~fdiebold/NoHesitations/BookAdvanced.pdf

[27] Hosmer, D. W., Hosmer, D. W., Lemeshow, S. (2004). *Applied Logistic Regression*. Alemanha: Wiley.

[28] Neter, J., Wasserman, W., Kutner, M. H. (1989). *Applied Linear Regression Models*. Taiwan: Irwin.

[29] Meyers, L. S., Gamst, G., Guarino, A. (2006). *Applied Multivariate Research: Design and Interpretatio*n. Índia: SAGE Publications

[30] Szumilas M. (2010). Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry = Journal de l'Academie canadienne de psychiatrie de l'enfant et de l'adolescent*, *19*(3), 227–229.

[31] UCLA: Statistical Consulting Group. (n.d). FAQ: *How do I interpret odds ratios in logistic regression?*.https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/

[32] Grimes, D. A., & Schulz, K. F. (2008). Making sense of odds and odds ratios. *Obstetrics & Gynecology*, *111*(2 Part 1), 423-426. DOI: 10.1097/01.AOG.0000297304.32187.5d

[33] GeeksforGeeks. (2024). *How to interpret odds ratios in logistic regression*. GeeksforGeeks.

[34] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

[35] Awan, A. A. (2023). An introduction to SHAP values and machine learning interpretability. *ww. datacamp. com/tutorial, June*, *28*.

[36] Roth, A. E. (Ed.). (1988). *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press. doi:10.1017/CBO9780511528446. ISBN 0-521-36177-X.

[37] O'Sullivan, Conor. (2022). From Shapley to SHAP – Understanding the Math. *Medium*. https://medium.com/data-science/from-shapley-to-shap-understanding-the-math-e7155414213b

[38] Pal, Saurabh & Aggraawl, Ritu. (2021). P-Value Feature Selection Technique for Prediction of Student Performance. *International Journal of Research Publication and Reviews*. www.ijrpr.com ISSN 2582-7421.

[39] Di Leo, G., & Sardanelli, F. (2020). Statistical significance: p value, 0.05 threshold, and applications to radiomics—reasons for a conservative approach. *European radiology experimental*, *4*(1), 18. https://doi.org/10.1186/s41747-020-0145-y

[40] Flom, P. L., & Cassell, D. L. (2007). Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. In *NorthEast SAS Users Group Inc 20th Annual Conference* (Vol. 11).

[41] Doosa, Ganesh (2021). The Mathematical background of Lasso and Ridge Regression. *Medium*. https://medium.com/codex/mathematical-background-of-lasso-and-ridge-regression-23b74737c817

[42] Pedregosa, F. et al., (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

[43] Saad, Mohammed. (2021). Detailed Explanation of Random Feature importance Bias. *Medium.* https://medium.com/@eng.mohammed.saad.18/detailed-explanation-of-random-forests-features-importance-bias-8755d26ac3bc

[44]Breiman, Leo. (2001). "Random Forests." *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

[45]Department of Health. (2004). *Quality and Outcomes Framework (QOF).*

https://www.health-ni.gov.uk/topics/quality-and-outcomes-framework-qof

[46]NHS England. (2024). *Quality and Outcomes Framework (QOF) 2023-2024 results*. https://qof.digital.nhs.uk/

[47]NHS Old School Surgery. (n.d.). *How we perform QOF*. https://www.oldschoolsurgery.org.uk/policies/how-we-perform-qof/

[48] National Institute for Health and Care Excellence (NICE). (2012). *About NICE guidance. How NICE clinical guidelines are developed: An overview for stakeholders, the public and the NHS*. https://www.nice.org.uk/process/pmg6/resources/how-nice-clinical-guidelines-are-developed-an-overview-for-stakeholders-the-public-and-the-nhs-2549708893/chapter/about-nice-guidance

[49] National Institute for Health and Care Excellence (NICE). (2014). *Nice indicator process guide*. https://www.nice.org.uk/process/pmg44

[50] NHS Digital. (2018). *Business Rules Supporting Information*.

[51] Forbes, L. J., Marchand, C., Doran, T., & Peckham, S. (2017). The role of the Quality and Outcomes Framework in the care of long-term conditions: a systematic review. *The British journal of general practice : the journal of the Royal College of General Practitioners*, *67*(664), e775–e784. https://doi.org/10.3399/bjgp17X693077

[52] Helsedirektoratet. (2022). *National Healthcare Quality Indicators*. https://www.helsedirektoratet.no/english/national-health-care-quality-indicators

[53] Rostad, H. M., Burrell, L. V., Skinner, M. S., Hellesø, R., & Sogstad, M. K. R. (2023). Quality of municipal long-term Care in Different Models of care: A cross-sectional study from Norway. *Health Services Insights*, *16*, 11786329231185537. https://doi.org/10.1177/11786329231185537

[54] Government of the Netherlands. (n.d.). *Quality of Healthcare – Monitoring and quality requirements*. https://www.government.nl/topics/quality-of-healthcare/monitoring-and-quality-requirements/monitoring-the-quality-of-healthcare

[55] van den Berg, M. J., Kringos, D. S., Marks, L. K., & Klazinga, N. S. (2014). The Dutch Health Care Performance Report: seven years of health care performance

assessment in the Netherlands. *Health research policy and systems*, *12*, 1. https://doi.org/10.1186/1478-4505-12-1

[56] Remen, T., Pintos, J., Abrahamowicz, M., & Siemiatycki, J. (2018). Risk of lung cancer in relation to various metrics of smoking history: a case-control study in Montreal. *BMC cancer*, *18*(1), 1275. https://doi.org/10.1186/s12885-018-5144-5

[57] Szumilas M. (2010). Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry = Journal de l'Academie canadienne de psychiatrie de l'enfant et de l'adolescent*, *19*(3), 227–229.

[58] Zhou, X., He, J., Wang, A., Hua, X., Li, T., Shu, C., & Fang, J. (2024). Multivariate logistic regression analysis of risk factors for birth defects: a study from population-based surveillance data. *BMC public health*, *24*(1), 1037. https://doi.org/10.1186/s12889-024-18420-1

[59] Moore, L., Hanley, J. A., Turgeon, A. F., & Lavoie, A. (2010). Evaluating the performance of trauma centers: hierarchical modeling should be used. *The Journal of trauma*, *69*(5), 1132–1137. https://doi.org/10.1097/TA.0b013e3181cc8449

[60] Austin, P. C., & Lee, D. S. (2020). Estimating the Net Benefit of Improvements in Hospital Performance: G-Computation With Hierarchical Regression Models. *Medical care*, *58*(7), 651–657. https://doi.org/10.1097/MLR.0000000000001312

[61] Kurian, N., Maid, J., Mitra, S., Rhyne, L., Korvink, M., & Gunn, L. H. (2021). Predicting Hospital Overall Quality Star Ratings in the USA. *Healthcare*, *9*(4), 486. https://doi.org/10.3390/healthcare9040486

[62] R Core Team. (2021). *R: A language and environment for statistical computing* (Version 4.x). R Foundation for Statistical Computing. https://www.R-project.org/

[63] Wickham, H., Vaughan, D., & Girlich, M. (2025). *tidyr: Tidy Messy Data* (R package version 1.3.1). R Foundation for Statistical Computing. https://tidyr.tidyverse.org

[64] Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

[65] Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2025). *dplyr: A Grammar of Data Manipulation* (R package version 1.1.4). R Foundation for Statistical Computing. https://dplyr.tidyverse.org dplyr.tidyverse.org

[66] Wei, T., & Simko, V. (2024). *corrplot: Visualization of a Correlation Matrix* (R package version 0.95). https://github.com/taiyun/corrplot

[67] Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01

[68] Jackman, S. (2024). *pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory* (R package version 1.5.9). University of Sydney. https://github.com/atahk/pscl/

[69] Barrett, T., Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., Hocking, T., Schwendinger, B., & Krylov, I. (2025). *data.table: Extension of 'data.frame'* (R package version 1.17.8). https://CRAN.R-project.org/package=data.table

[71] Python Software Foundation. (2023). *Python (Version 3.11)*. https://www.python.org

[70] Posit team (2025). RStudio: Integrated Development Environment for R. Posit Software, PBC, Boston, MA. URL http://www.posit.co/.

[72] Google. (2024). Google Colaboratory. from https://colab.research.google.com/

[73] McKinney, W. (2010). Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference* (pp. 56–61).

[74] The pandas development team. (2020). *pandas-dev/pandas: pandas* (2.3.2) .Zenodo. https://doi.org/10.5281/zenodo.3509134

[75] Harris, C. R., Millman, et.al. (2020). Array programming with NumPy. *Nature, 585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

[76] Virtanen, P., Gommers, R., Oliphant, T. E., et.al. SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods, 17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

[77] OECD. (2023). Health at a Glance 2023: OECD indicators. *OECD Publishing*, Paris. https://doi.org/10.1787/7a7afb35-en

[78] Sing, C. W., Lin, T. C., Bartholomew, S., Bell, J. S., Bennett, C., Beyene, K., Bosco-Levy, P., Bradbury, B. D., Chan, A. H. Y., Chandran, M., Cooper, C., de Ridder, M., Doyon, C. Y., Droz-Perroteau, C., Ganesan, G., Hartikainen, S., Ilomaki, J., Jeong, H. E., Kiel, D. P., Kubota, K., … Wong, I. C. K. (2023). Global Epidemiology of Hip Fractures: Secular Trends in Incidence Rate, Post-Fracture Treatment, and All-Cause Mortality. *Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research*, *38*(8), 1064–1075. https://doi.org/10.1002/jbmr.4821

[79] Furia, G., Vinci, A., Colamesta, V., Papini, P., Grossi, A., Cammalleri, V., Chierchini, P., Maurici, M., Damiani, G., & De Vito, C. (2023). Appropriateness of frequent use of emergency departments: A retrospective analysis in Rome, Italy. *Frontiers in public health*, *11*, 1150511. https://doi.org/10.3389/fpubh.2023.1150511

[80] Hansagi, H., Olsson, M., Sjöberg, S., Tomson, Y., & Göransson, S. (2001). Frequent use of the hospital emergency department is indicative of high use of other health care services. *Annals of emergency medicine*, *37*(6), 561–567. https://doi.org/10.1067/mem.2001.111762

[81] Ministério da Saúde. (n.d.). Mediana IDE (UCSP). *BI-CSP*. https://bicsp.min-saude.pt/pt/contratualizacao/ide/Paginas/default.aspx

[82] Sindicato Independente dos Médicos. (2023). *Portaria que Regula o IDE das USF. SIM*.https://www.simedicos.pt/pt/noticias/5352/portaria-que-regula-o-ide-das-usf/

[83] Entidade Reguladora da Saúde. (2024). *Cuidados de saúde Primários – qualidade e eficiência nas UCSP e USF*. https://www.ers.pt/pt/atividade/regulacao-economica/selecionar/estudos/lista-de-estudos/estudo-cuidados-de-saude-primarios-qualidade-e-eficiencia-nas-ucsp-e-usf

[84] Sydow, David. (2024). Stay Balanced — The Importance of a Balanced Dataset in ML. *Medium.* https://medium.com/@dsydow/the-importance-of-a-balanced-dataset-in-ml-classification-algorithms-3a55dff768f7

[85] Google Developers. (n.d.). Imbalanced datasets. *Machine Learning Crash Course*. https://developers.google.com/machine-learning/crash-course/overfitting/imbalanced-datasets

[86] Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics (Oxford, England)*, *23*(19), 2507–2517. https://doi.org/10.1093/bioinformatics/btm344

[87] Han, Kamber & Pei. (2006). Data Mining: Concepts and Techniques. 2nd ed. *The Morgan Kaufman Series in Data Managment*.

[88] Kundu, Rohit. (2022). F1 Score in Machine Learning: Intro & Calculation. *V7labs*. https://www.v7labs.com/blog/f1-score-guide

[89] Gupta, A., Stead, T. S., & Ganti, L. (2024). Determining a meaningful R-squared value in clinical medicine. *Academic Medicine & Surgery*. https://academic-med-

surg.scholasticahq.com/article/125154-determining-a-meaningful-r-squared-value-in-clinical-medicine

[90] Kulla, S. (2006). *Intersection of sets A and B*. Wikimedia commons.https://commons.wikimedia.org/wiki/File:Intersection_of_sets_A_and_B.svg

[91] Eurostat. (2024). Urban-rural Europe – labour market. https://ec.europa.eu/eurostat/statistics-explained/SEPDF/cache/112338.pdf

[92] Chakrabarti, A., & Ghosh, J. K. (2011). AIC, BIC and recent advances in model selection. *Philosophy of statistics*, 583-605.

[93] Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological methods*, *17*(2), 228.

[94] Wang, L., Han, M., Li, X., Zhang, N., & Cheng, H. (2021). Review of classification methods on unbalanced data sets. *Ieee Access*, *9*, 64606-64628.

[95] Huberty, C. J. (1989). Problems with stepwise methods-better alternatives. *Advances in social science methodology*, *1*, 43-70.

[96] Snyder, P., Thompson, B., & Snyder, P. (1991). Three reasons why stepwise regression methods should not be used by researchers. *Advances in educational research: Substantive findings, methodological developments*, *1*, 99-105.

[97] Thompson, B. (1989). Why won't stepwise methods die?. *Measurement and evaluation in counseling and development*, *21*(4), 146-148.

[98] Wang, K., & Chen, Z. (2016). Stepwise regression and all possible subsets regression in education. *Electronic International Journal of Education, Arts, and Science (EIJEAS)*, *2*(Special Issue), 60-81.

[99] Ruengvirayudh, P., & Brooks, G. P. (2016). Comparing stepwise regression models to the best-subsets models, or, the art of stepwise. *General linear model journal*, *42*(1), 1-14.

[100] Kirasich, K., Smith, T., & Sadler, B. (2018). Random forest vs logistic regression: binary classification for heterogeneous datasets. *SMU Data Science Review*, *1*(3), 9.[11]

[101] Bagherzadeh-Khiabani, F., Ramezankhani, A., Azizi, F., Hadaegh, F., Steyerberg, E. W., & Khalili, D. (2016). A tutorial on variable selection for clinical prediction models:

feature selection methods in data mining could improve the results. *Journal of clinical epidemiology*, *71*, 76-85.

[102] Walter, S., & Tiemeier, H. (2009). Variable selection: current practice in epidemiological studies. *European journal of epidemiology*, *24*(12), 733-736.

[103] Lewis, F., Butler, A., & Gilbert, L. (2011). A unified approach to model selection using the likelihood ratio test. *Methods in ecology and evolution*, *2*(2), 155-162.

[104]Hensher, D. A., & Stopher, P. R. (2021). Behavioural travel modelling. In *Behavioural travel modelling* (pp. 11-52). Routledge.

[105] Soril, L. J., Leggett, L. E., Lorenzetti, D. L., Noseworthy, T. W., & Clement, F. M. (2015). Reducing frequent visits to the emergency department: a systematic review of interventions. *PloS one*, *10*(4), e0123660. https://doi.org/10.1371/journal.pone.0123660

[106] Ramalho, A., Souza, J., Castro, P., Lobo, M., Santos, P., & Freitas, A. (2021). Portuguese primary healthcare and prevention quality indicators for diabetes mellitus–a data envelopment analysis. *International journal of health policy and management*, *11*(9), 1725 doi: 10.34172/ijhpm.2021.76.

[107] Silva, C. F., Barreira, V., Beirão, D., Sá, L., & Santos, P. (2025). The perspective of socially vulnerable youth on healthcare services: a Portuguese cross-sectional study. *Archives of Public Health*, *83*(1), 202.

[108] Lim, S. W., Jin, J. Z., Jin, L., Jin, J., & Li, C. (2015). Role of dipeptidyl peptidase-4 inhibitors in new-onset diabetes after transplantation. *The Korean journal of internal medicine*, *30*(6), 759–770. https://doi.org/10.3904/kjim.2015.30.6.759

[109] Direção-Geral da Saúde. (2019). *Abordagem terapêutica farmacológica na diabetes mellitus tipo 2 no adulto* (Norma n.º 025/2011 de 28/12/2011 atualizada a 30/06/2019). https://normas.dgs.min-saude.pt/wp-content/uploads/2019/09/abordagem-terapeutica-farmacologica-na-diabetes-mellitus-tipo-2-no-adulto.pdf

[110] Direção-Geral da Saúde. (2019). *Diagnóstico e rastreio laboratorial da infeção pelo vírus da imunodeficiência humana (VIH)* (Norma n.º 015/2013 de 24/09/2013 atualizada a 30/06/2019). https://normas.dgs.min-saude.pt/wp-content/uploads/2019/09/diagnostico-e-rastreio-laboratorial-da-infecao-pelo-virus-da-imunodeficiencia-humana-vih.pdf

[111] Direção-Geral da Saúde. (2019). *Exames ecográficos na gravidez de baixo risco* (Norma n.º 021/2011 de 28/12/2011 atualizada a 30/06/2019). https://normas.dgs.min-saude.pt/wp-content/uploads/2019/09/exames-ecograficos-na-gravidez-de-baixo-risco.pdf

[112] Direção-Geral da Saúde. (2011). *Exames laboratoriais na gravidez de baixo risco* (Norma nº 037/2011, atualizada a 20/12/2013) [PDF]. Ordem dos Médicos/Reprodução da DGS. https://ordemdosmedicos.pt/files/pdfs/vI2V-Exames_Gravidez_Baixo_risco_37_2011.pdf

[113] Direção-Geral da Saúde. (2011). *Processo Assistencial Integrado da Diabetes Mellitus tipo 2* (Norma n.º 005/2011). https://www.dgs.pt/directrizes-da-dgs/normas-e-circulares-normativas/norma-n-0052011-de-21012011-pdf.aspx

[114] Direção-Geral da Saúde. (2011). *Prescrição de Exames Laboratoriais para Avaliação de Dislipidemias no Adulto* (Norma n.º 066/2011). https://normas.dgs.min-saude.pt/wp-content/uploads/2019/09/prescricao-de-exames-laboratoriais-para-avaliacao-de-dislipidemias-no-adulto.pdf

[115] Direção-Geral da Saúde. (2011). *Abordagem Terapêutica da Hipertensão Arterial* (Norma n.º 026/2011). https://normas.dgs.min-saude.pt/wp-content/uploads/2019/09/abordagem-terapeutica-da-hipertensao-arterial.pdf

[116] Saúde. (2018). *Portaria n.º 126/2018*. Diário da República n.º 88/2018, Série I de 2018-05-08, páginas 2039 – 2044. https://diariodarepublica.pt/dr/detalhe/portaria/126-2018-115235763

[117] Direção-Geral da Saúde. (2011). *Prescrição de Exames Laboratoriais para Avaliação de Dislipidemias no Adulto* (Norma n.º 066/2011). https://normas.dgs.min-saude.pt/wp-content/uploads/2019/09/prescricao-de-exames-laboratoriais-para-avaliacao-de-dislipidemias-no-adulto.pdf

[118] Saúde - Gabinete do Secretário de Estado Adjunto e da Saúde. (2021). *Despacho n.º 9390/2021*. Diário da República n.º 187/2021, Série II de 2021-09-24, páginas 96 – 103. https://diariodarepublica.pt/dr/detalhe/despacho/9390-2021-171891094

Additional Information

Table A 1. Indicator matrix with importance weight, Expected values, and Acceptable variance that calculates the IDE. Found in Ordinance 411-A/2023 [18]

N.º Indicador	**Ponderação**	**Valores Esperados**	**Variações Aceitáveis**
8	4.0	[60;100]	[38;60[U]100;100]
294	4.0	[500;1500]	[150;500[U]1500;1500]
330	5.0	[0.82;2]	[0.7;0.82[U]2;2]
331	5.0	[0.76;2]	[0.6;0.76[U]2;2]
335	2.0	[85;100]	[80;85[U]100;100]
11	1.2	[91;100]	[85;91[U]100;100]
34	1.2	[72;100]	[55;72[U]100;100]
45	1.2	[60;100]	[37;60[U]100;100]
46	1.2	[65;100]	[45;65[U]100;100]
54	1.2	[60;100]	[40;60[U]100;100]
63	1.2	[80;100]	[65;80[U]100;100]
95	1.2	[95;100]	[90;95[U]100;100]
98	1.2	[93;100]	[80;93[U]100;100]

N.º Indicador	Ponderação	Valores Esperados	Variações Aceitáveis
269	1.2	[0.87;1]	[0.7;0.87[U]1;1]
295	1.2	[77;100]	[51;77[U]100;100]
302	1.2	[0.93;1]	[0.82;0.93[U]1;1]
308	1.2	[80;100]	[60;80[U]100;100]
310	1.2	[0.79;1]	[0.62;0.79[U]1;1]
311	1.2	[0.54;1]	[0.4;0.54[U]1;1]
312	1.2	[0.43;1]	[0.3;0.43[U]1;1]
384	1.2	[90;100]	[80;90[U]100;100]
397	1.2	[25;100]	[15;25[U]100;100]
404	1.2	[60;10000]	[20;60[U]10000;10000]
435	1.2	[66;100]	[62;66[U]100;100]
409	1.2	[91.5;100]	[89;91.5[U]100;100]
20	1.5	[67;100]	[45;67[U]100;100]
23	1.5	[80;100]	[60;80[U]100;100]
36	1.5	[75;100]	[60;75[U]100;100]
37	1.5	[87;100]	[70;87[U]100;100]

N.º Indicador	Ponderação	Valores Esperados	Variações Aceitáveis
39	1.5	[70;100]	[50;70[U]100;100]
49	1.5	[60;100]	[30;60[U]100;100]
261	1.5	[87;100]	[70;87[U]100;100]
274	1.5	[82;100]	[65;82[U]100;100]
275	1.5	[70;100]	[50;70[U]100;100]
314	1.5	[0;15]	[0;0[U]15;28]
315	1.5	[48;100]	[33;48[U]100;100]
380	1.5	[81;100]	[74;81[U]100;100]
436	1.5	[70;100]	[35;70[U]100;100]
437	1.5	[49;100]	[35;49[U]100;100]
341	8.0	[0;133]	[0;0[U]133;163]
354	10.0	[0;47]	[0;0[U]47;57]
365	7.0	[0;480]	[0;0[U]480;620]
412	10.0	[60;85]	[40;60[U]85;95]

Table A 2. Description of BD_BICSP_2023 Dataset, highlighted columns are the Outcome Indicators [16]

ID	Description	Objective
2013.001.01_FL	1 - Proportion of consultations carried out by the respective family doctor	Overview of accessibility between the family doctor and HCU service users. Understand the capability of doctor exchange in a health care unit.
2013.002.01_FL	2 – Overall utilization rate of medical consultations	Assess access to medical consultations by the registered population.
2013.003.01_FL	3 – Home medical consultation rate per 1000 service users	Monitor productivity related to the performance of home consultations
2013.005.01_FL	5 - Proportion of consultations carried out by the respective family nurse	Monitor Service users' access to their respective family nurse and the capacity for nurses in the health unit to substitute for each other
2013.006.01_FL	6 - Overall utilization rate of medical consultations in the last 3 years	Assess access to medical consultations by the registered population in the last 3 years
2013.008.01_FL	8 - Rate of use of family planned (PF) consultations (doctor or nurse)	Monitor the use of reproductive health and family planning (FP) consultations by women of childbearing age (MIF)
2013.009.01_FL	9 - Utilization rate of family planning nursing consultations	Monitor the use of reproductive health and family planning (FP) *nursing* consultations by women of childbearing age (MIF)
2013.010.01_FL	10 - Utilization rate of family planning (PF) medical consultations	Monitor the use of reproductive health and family planning (FP) consultations by women of childbearing age (MIF)
2013.011.01_FL	11 - Proportion of pregnant women with their first pregnancy monitoring medical appointment, carried out in the 1st trimester	Monitoring of the Maternal Health Surveillance Program
2013.014.01_FL	14 - Proportion of newborns with at least one medical surveillance consultation carried out within 28 days of being born	Monitor early surveillance of newborns
2013.015.01_FL	15 - Proportion of newborns with home nursing consultations carried out up to the 15th day of being born	Monitor newborn care
2013.016.01_FL	16 - Proportion of children with at least 6 child health surveillance medical appointments in the first year of life	Monitoring the Child Health Program - 1st Year of Life.
2013.017.01_FL	17 - Proportion of children with at least 3 child health surveillance medical appointments in the 2nd year of life	Acompanhamento do Programa de Saúde Infantil - 2º ano de vida
2013.018.01_FL	18 - Proportion of service users with high blood pressure, with at least one BMI record in the last 12 months	Monitor service users with high blood pressure (body mass index – BMI)
2013.019.01_FL	19 - Proportion of service users with high blood pressure, with	Monitor service users with high blood pressure (blood pressure result)

	blood pressure recorded every six months	
2013.020.01_FL	20 - Proportion of service users with high blood pressure, under 65 years of age, with blood pressure below 150/90 mmHg	Monitor service users with high blood pressure (blood pressure result)
2013.023.01_FL	23 -Proportion of service users with high blood pressure (without cardiovascular disease or diabetes), with cardiovascular risk determined in the last 3 years	Monitoring the hypertension program (cardiovascular risk)
2013.030.01_FL	30 – Proportion of elderly or chronic disease with the flu vaccine prescribed or administered in the last 12 months	Adult health program monitoring (Flu vaccine)
2013.031.01_FL	31 - Proportion of children aged 7 with weight and height recorded in the range [5; 7[years]	Monitor the child and youth health program (Weight and height recording)
2013.032.01_FL	32 - Proportion of 14-year-olds with weight and height recorded in the range [11; 14[years	Monitor the child and youth health program (Weight and height recording)
2013.034.01_FL	34 - Proportion of obese service users aged 14 years or older who underwent obesity surveillance consultation in the last 2 years	Monitor the child and youth health program (Obese surveillance consultation)
2013.035.01_FL	35 - Proportion of service users with diabetes with at least one foot examination recorded in the last year	Monitor the diabetes program (Foot exam performance)
2013.036.01_FL	36 - Proportion of service users with diabetes, with a record of therapeutic regimen management (3 items) in the last year	Monitor the diabetes program (Therapeutic regimen management)
2013.037.01_FL	37 - Proportion of service users with diabetes who had a diabetes surveillance nursing consultation in the last year	Monitor the diabetes program. (Nursing surveillance consultation)
2013.038.01_FL	38 - Proportion of service users with diabetes, with at least 2 HbA1c in the last year, provided they cover both semesters	Monitor the diabetes program (HbA1c result recording)
2013.039.01_FL	39 - Proportion of service users with diabetes, with the last HbA1c record equal to or less than 8.0%	Monitor the diabetes program (HbA1c result recording)
2013.045.01_FL	45 - Proportion of women aged 25-60 years who have had cervical cancer screening	Monitor the oncological screening program (screening and early detection of cervical cancer)
2013.046.01_FL	46 - Proportion of service users aged between [50; 75[years, with colon and rectal cancer screening performed	Monitor the cancer screening program. Parameter (screening and early detection of colon and rectal cancer)

2013.049.01_FL	49 - Proportion of patients with COPD with at least one FeV1 assessment record in the last 3 years	Monitor the respiratory disease follow-up program. Parameter (FeV1 in Chronic obstructive pulmonary disease (COPD) patients)
2013.053.01_FL	53 - Proportion of service users aged 14 or over, with quantification of alcohol consumption, recorded in the last 3 years	Monitor the youth and adult health program (Alcohol consumption record)
2013.054.01_FL	54 - Proportion of service users aged 14 or over with the problem of "excessive alcohol consumption" who received at least one related consultation in the last 3 years	Monitor the youth and adult health program (consultation related to excessive alcohol consumption)
2013.057.01_FL	57 - proportion of newborns with early diagnosis (TSHPKU) carried out up to the sixth day of being born	Monitor the child health program (Early diagnosis)
2013.059.01_FL	59 - Proportion of 2-year-old children with weight and height recorded in the last year	Monitor the child health program, 2nd year of life (weight and height record)
2013.063.01_FL	63 - Proportion of children aged 7 years with a medical check-up carried out between [5; 7[years] and PNV fully complied with by their 7th birthday	Monitoring the child health program (7-year-old cohort) PNV – National Vaccination Program
2013.091.01_FL	91 - Proportion of service users with diabetes, aged under 65 years, with the last HbA1c record equal to or less than 6.5%	Monitor the diabetes program (HbA1c result)
2013.093.01_FL	93 - Proportion of children aged 2 with PNV fulfilled or in progress at the indicator's reference date	Monitor compliance with the PNV among users with active registration - 2-year cohort.
2013.094.01_FL	94 - Proportion of children aged 7 with PNV fulfilled or in progress at the indicator's reference date	Monitor compliance with the PNV among users with active registration - 7-year cohort.
2013.095.01_FL	95 - Proportion of children aged 14 with PNV fulfilled or in progress at the indicator's reference date	Monitor compliance with the PNV among users with active registration - 14-year cohort.
2013.097.01_FL	97 - Proportion of service users with diabetes, with microalbuminuria in the last year	Monitor the diabetes program. (microalbuminuria result recording)
2013.098.01_FL	98 - Proportion of service users aged 25 or over who have an up-to-date tetanus vaccine	Monitor compliance with the PNV among users with active registration (tetanus vaccine)
2013.099.01_FL	99 - Overall utilization rate of nursing consultations in the last 3 years	Assess access to nursing consultations by the registered population
2013.100.01_FL	100 - Overall utilization rate of medical or nursing consultations in the last 3 years	Assess access to medical or nursing consultations by the registered population

2013.261.01_FL	261 - Proportion of service users with diabetes, with assessment of the risk of foot ulceration in the last year	Monitor the diabetes program. (perform examination and assess the risk of foot ulceration)
2013.262.01_FL	262 - Proportion of service users with a risk assessment for type 2 diabetes recorded in the last 3 years	Monitor the diabetes program. (Type 2 diabetes risk record)
2013.267.01_FL	267 – Index: Rate of adequate monitoring of family planning in women of childbearing age	Monitor the family planning program (FP)
2013.269.01_FL	269 – Index: Adequate monitoring index in child health, 2nd year of life	Monitor the child health program, 2nd year of life
2013.270.01_FL	270 – Index: Adequate monitoring index in maternal health	Monitor the maternal health program
2013.271.01_FL	271 – Index: Adequate monitoring index in users with Diabetes Mellitus	Monitor the diabetes program
2013.272.01_FL	272 – Index: Adequate monitoring index for service users with high blood pressure	Monitor the hypertension program
2013.274.01_FL	274 - Proportion of patients with type 2 diabetes and indication for insulin therapy, undergoing appropriate therapy	Monitor the effectiveness of the application of DGS standard 052/2011 [2]
2013.275.01_FL	275 - Proportion of patients with a new diagnosis of type 2 diabetes who initiate therapy with metformin monotherapy	Monitor the effectiveness of pharmacological therapy for type 2 Diabetes Mellitus (metformin therapy)
2013.276.01_FL	276 - Ratio between the sum of DDD prescribed in DPP-4 inhibitors and the sum of DDD prescribed in non-insulin antidiabetics, in patients with type 2 Diabetes Mellitus	Monitor the effectiveness of pharmacological therapy for type 2 Diabetes Mellitus DDD – Defined Daily Dose DPP-4 inhibitors - a class of oral medications used to manage type 2 diabetes [108]
2013.277.01_FL	277 - Proportion of users aged 14 or over with smoking habits who received a smoking-related consultation in the last year	Monitor the youth and adult health program (smoking-related consultation)
2013.278.01_FL	278 - Proportion of prescription drug packages of therapeutic classes with generics	Monitor the frequency of prescription of drugs from therapeutic classes with generics
2013.294.01_FL	294 – Rate of Home nursing consultation per 1,000 enrolled elderly individuals	Monitor the productivity of nursing teams when carrying out home nursing for elderly people
2013.295.01_FL	295 - Proportion of postpartum women with 5 or more maternal health nursing consultations during pregnancy and with a postpartum review consultation	Monitoring of the Maternal Health Surveillance Program, an area of access to maternal health nursing consultations during pregnancy and the postpartum period

2013.296.01_FL	296 - Proportion of households of postpartum women or newborns who received nursing home visits	Monitoring of maternal health surveillance and child and youth health surveillance programs
2013.297.01_FL	297 - Proportion of users aged 65 or over, without long-term prescription of anxiolytics, sedatives or hypnotics, in the period under analysis	Monitor the mental health program (Prescription of anxiolytics, sedatives, and hypnotics in the elderly)
2013.300.01_FL	300 - Average number of physiatry consultation prescriptions per user	Monitor MCDT prescription program (physiatry prescription)
2013.302.01_FL	302 – Index: Adequate monitoring index in child health, 1st year of life	Monitor the child health program, 1st year of life
2013.413.01_FL	413 - Proportion of patients undergoing insulin therapy among those enrolled with type 2 diabetes and HbA1c greater than 9% for 6 + months	Monitor the effectiveness of the application of DGS standard 052/2011 [109]
2015.306.01_FL	306 - Proportion of service users consulted in the last 12 months and without HIV/AIDS screening who underwent it in that period	Monitoring compliance with DGS clinical guidance standard 58/2011 [110]
2015.307.01_FL	307 - Proportion of pregnant women who underwent an ultrasound examination during the 1st trimester of pregnancy	Monitoring compliance with DGS clinical guidance standard 23/2011 [111]
2015.308.01_FL	308 - Proportion of pregnant women who underwent an ultrasound examination during the 2nd trimester of pregnancy	Monitoring compliance with DGS clinical guidance standard 23/2011 [111]
2015.309.01_FL	309 - Proportion of pregnant women who underwent an ultrasound examination during the 3rd trimester of pregnancy	Monitoring compliance with DGS clinical guidance standard 23/2011 [111]
2015.310.01_FL	310 – Index: Rate of laboratory tests performed in the 1st trimester of pregnancy	Monitoring compliance with DGS clinical guidance standard 37/2011 [112]
2015.311.01_FL	311 – Index: Rate of laboratory tests performed in the 2nd trimester of pregnancy	Monitoring compliance with DGS clinical guidance standard 37/2011 [112]
2015.312.01_FL	312 - Index: Rate of laboratory tests performed in the 3rd trimester of pregnancy	Monitoring compliance with DGS clinical guidance standard 37/2011 [112]
2015.314.01_FL	314 - Proportion of service users with diabetes with a last blood pressure reading above 140/90 mmHg	Monitor compliance with DGS standard no. 005/2011 (Prevention and Assessment of Diabetic Nephropathy) [113]
2015.315.01_FL	315 - Proportion of service users with diabetes with LDL cholesterol levels below 100 mg/dL in the last 12 months	Monitor compliance with DGS standard no. 066/2011 on (Prescription of Laboratory Tests for the Assessment of Dyslipidemia in Adults) [114]

2015.316.01_FL	316 - Proportion of users with arterial hypertension, aged between [18; 65[years, with blood pressure below 140/90 mmHg	Monitor compliance with DGS standard no. 026/2011 on (Therapeutic Approach to Arterial Hypertension) [115]
2017.255.01_FL	255 - Proportion of quinolones among billed antibiotics (packaging, to registered users)	Monitor drug prescriptions in terms of the quantity of quinolones prescribed
2017.257.01_FL	257 - Proportion of cephalosporins among billed antibiotics (packaging, to registered users)	Monitor drug prescriptions in terms of the quantity of cephalosporins prescribed
2017.259.01_FL	259 - Proportion of COX-2 inhibitors among nonsteroidal anti-inflammatory drugs billed (DDD, to registered users)	Monitor drug prescription in the quantity of COX-2 inhibitor prescribed
2017.330.01_FL	330 – Index: Utilization rate scaled to the estimated annual need for medical consultations	Monitor the frequency with which necessary medical consultations based on sociodemographic and morbidity criteria are carried out
2017.331.01_FL	331 – Index: Utilization rate scaled to the estimated annual need for nursing consultations	Monitor the frequency with which necessary nursing consultations based on sociodemographic and morbidity criteria are carried out
2017.335.01_FL	335 - Proportion of remote consultations with a prescription issued in the first 3 working days after the respective request	Monitor guaranteed maximum response times (MRRT), applied to "medication renewal in case of chronic illness" requested by users through remote consultations, per current legislation
2017.341.01_FL	341 - Average expenditure (PVP) of prescribed and reimbursed medicines, per standard registered user	Monitor expenditure on reimbursed prescription drugs, carried out in the context of the activities of primary health care functional units
2017.342.01_FL	342 - Proportion of medical appointments initiated by users scheduled in less than 15 working days	Monitor guaranteed maximum response times (GRTs) for medical consultations
2017.344.01_FL	344 - Proportion of medical consultations carried out on the same day the appointment was registered	Monitor the proportion of medical consultations carried out on the day the appointment is registered
2017.345.01_FL	345 - Proportion of nursing consultations carried out on the day of the appointment	Monitor the proportion of nursing consultations carried out on the day of the appointment
2017.346.01_FL	346 - Proportion of consultations carried out in the interval [8; 11[hours (Q1)	Monitor the time distribution of in-person consultations
2017.347.01_FL	347 - Proportion of consultations carried out in the interval [11; 14[h (Q2)	Monitor the time distribution of in-person consultations
2017.348.01_FL	348 - Proportion of consultations carried out in the interval [14; 17[(Q3)	Monitor the time distribution of in-person consultations
2017.349.01_FL	349 - Proportion of consultations carried out in the interval [17; 20]h (Q4)	Monitor the time distribution of in-person consultations

2017.350.01_FL	350 - Cost of therapy for patients with Diabetes Mellitus	Monitor the cost of treating Diabetes Mellitus
2017.351.01_FL	351 - Cost of therapy for patients with controlled Diabetes Mellitus	Monitor the efficiency of treating Diabetes Mellitus
2017.352.01_FL	352 - Cost of therapy for patients with high blood pressure (HTA)	Monitor the cost of treating high blood pressure
2017.353.01_FL	353 - Cost of therapy for patients with controlled high blood pressure (HTA)	Monitor the efficiency of treating high blood pressure
2017.354.01_FL	354 - Average expenditure, based on the price of prescribed CDTMs, per standard registered user	Monitor expenditure on prescribed complementary diagnostic and therapeutic methods (CDTMs), carried out in the context of the activities of primary health care functional units [116]
2017.365.01_FL	365 - Rate of avoidable hospitalizations in the adult population (adjusted for a standard population)	Monitor the effectiveness of care provided by PHC to patients with asthma, COPD, pneumonia, CHF, angina pectoris, hypertension, and diabetes, in their symptomatic control, in the prevention of complications and exacerbations, (hospital admission rate)
2017.372.01_FL	372 – Rate of Hospital admissions due to falls with femoral neck fracture	Monitor the impact of the actions of the different PHC UFs on reducing falls in the population aged 75 and over
2017.378.01_FL	378 - Proportion of PVP billed for medicines to users registered in PHC, prescribed by the family doctor himself in the context of private practice	Monitor the frequency of medication prescriptions in the context of private clinical practice carried out by family doctors to service users registered in the prescribing physician's file
2017.379.01_FL	379 - Proportion of PVP billed for medicines to users registered in the PHC, prescribed by the family doctor in a functional unit or service different from the unit where the service users are registered	Monitor the frequency of medication prescriptions made by family doctors carried out in a functional unit or service other than the unit where the users are registered
2017.380.01_FL	380 - Proportion of adult service users with clinical records showing the existence of asthma, COPD or chronic bronchitis, with a diagnosis recorded in the problem list	Monitor the frequency of potential errors in coding health problems in the problem list (classified by ICPC-2)
2017.381.01_FL	381 - Proportion of adult service users with clinical records showing the existence of depression or anxiety, with a diagnosis recorded in the problem list	Monitor the frequency of potential errors in coding health problems in the problem list (classified by ICPC-2)
2017.382.01_FL	382 - Proportion of adult service users with clinical records showing the existence of Diabetes Mellitus, with a diagnosis recorded in the list of problems	Monitor the frequency of potential errors in coding health problems in the problem list (classified by ICPC-2)

2017.383.01_FL	383 - Proportion of adult service users with clinical records showing the existence of Arterial Hypertension, with a diagnosis recorded in the list of problems	Monitor the frequency of potential errors in coding health problems in the problem list (classified by ICPC-2)
2017.384.01_FL	384 - Proportion of newborns whose mother has a pregnancy record	Monitor the frequency of potential errors in recording "pregnancies" in the electronic clinical record
2017.389.01_FL	389 - Score for evaluating the "assistance services" dimension	Assess the professional team's commitment to providing care activities in common ACES services.
2017.390.01_FL	390 - Score for assessing the dimension "non-assistance services - clinical governance activities of ACES"	Assess the team's commitment to providing non-care services related to ACES clinical governance
2017.391.01_FL	391 - Score for evaluating the "continuous improvement of quality in the area of access" dimension	--
2017.392.01_FL	392 - IDS of the "Continuous Quality Improvement Programs and Integrated Care Processes" dimension	--
2017.393.01_FL	393 - Score for evaluating the "training of the multidisciplinary team" dimension of the "internal training" subarea	--
2017.394.01_FL	394 - Score for evaluating the "training of interns and students" dimension of the "internal training" subarea	--
2018.339.01_FL	339 – Annual adjusted rate of hospital emergency episodes	Monitor the intensity of use of the hospital emergency network by service users enrolled in primary health care
2018.395.01_FL	395 - Proportion of service users aged 15 or over, with quantification of smoking habits in the last 3 years	Monitor records of tobacco use by healthcare professionals in primary care units. Support clinical governance processes that can lead to the prevention of tobacco use initiation and promote smoking cessation.
2018.397.01_FL	397 - Proportion of smokers aged 15 or over who received counseling intervention, based on a brief approach, in the last year	Monitor the frequency with which primary care health services intervene in promoting smoking cessation
2018.398.01_FL	398 - Proportion of pregnant smokers aged 15 years or over who received counseling intervention, based on a brief approach, in the first trimester of pregnancy	Monitor the frequency with which primary care health services intervene in promoting smoking cessation
2018.404.01_FL	404 - Annual incidence of people who have completed 12 months of smoking abstinence, based on the registered population aged 15 or over	Monitor the efficacy and effectiveness of health professionals' interventions to promote tobacco abstinence

2018.405.01_FL	405 - Proportion of medical consultations carried out on the day the appointment was registered, in a functional unit other than the one in which the user is registered	Measure the accessibility of users to the registration functional unit, with the respective result varying in inverse proportion to the level of accessibility provided
2018.409.01_FL	409 - Proportion of patients without a long-term prescription for anxiolytics, sedatives, or hypnotics, adjusted for a standard population	Monitor the mental health program. (Long-term prescription of anxiolytics, sedatives, and hypnotics)
2018.410.01_FL	410 - Adjusted annual rate of frequent or very frequent users of hospital emergency services	Monitor the intensity of hospital emergency network use by patients enrolled in primary care (Frequent and very frequent users)
2018.412.01_FL	412 - Proportion of "acute illness" medical consultations carried out in the user's UF of registration, counting in the denominator the sum of emergency episodes of registered users with the acute illness consultations carried out in the ACES of registration	Measure the accessibility of users to the registration functional unit, with the respective result varying in direct proportion to the level of accessibility provided
2019.414.01_FL	414 – Index: Adequate monitoring rate in the treatment of patients with pressure ulcers	Monitor pressure ulcer treatment follow-up across four dimensions: Teamwork; Continuity of care; Results/resolution; Efficiency
2019.415.01_FL	415 – Index: Adequate monitoring rate in the treatment of patients with venous ulcers	Monitorizar o acompanhamento no tratamento de úlceras venosas, em 5 dimensões: Adequação do processo de cuidados; Trabalho em equipa; Continuidade de cuidados; Resultados / resolutividade; Eficiência
2019.427.01_FL	427 - Proportion of pregnant women who delivered by cesarean section	Monitor the impact of USF/UCSP actions in reducing the cesarean section rate
2019.428.01_FL	428 - Score for assessing the "user safety" dimension	Assess the degree of implementation of continuous quality improvement processes aimed at user safety, under the terms of the National Plan for Patient Safety DGS, published by Order No. 9390/2021 [117]
2019.430.01_FL	430 - Score for assessing the "citizen participation" dimension	Assess the extent to which the functional unit is organized by placing the citizen/user at the center of the system
2021.436.01_FL	436 - Proportion of patients with COPD aged 40 or over who underwent a COPD surveillance consultation in the last year	Monitor the respiratory disease program. (COPD surveillance consultation)
2021.437.01_FL	437 - Proportion of users with asthma aged 18 or over who underwent asthma surveillance consultation in the last year	Monitorizar o programa doenças respiratórias. (consulta de vigilância de asma em adultos)
2021.439.01_FL	439 - Proportion of COPD patients over 6 months of age who have had the flu vaccine prescribed or administered in the last 12 months	Respiratory Disease Program Monitoring: (Flu Vaccine)

2022.434.01_FL	434 – Index: Adequate Monitoring Index in Family Planning - V2	Monitor the family planning program [6]
2022.441.01_FL	441 – Index: Adequate Maternal Health Monitoring Index - V2	Monitor the Maternal Health program
2022.442.01_FL	442 – Index: Adequate monitoring index in Child Health in the 1st year of life - V2	Monitor the Child Health Program
2022.443.01_FL	443 – Index: Adequate monitoring index in Child Health in the 2nd year of life - V2	Monitor the Child Health Program
2022.444.01_FL	444 – Index: Adequate monitoring index in users with Diabetes Mellitus - V2	Monitor the Diabetes Mellitus program
2022.445.01_FL	445 – Index: Adequate monitoring index for patients with high blood pressure - V2	Monitor the Arterial Hypertension program
Nome UF	Name of the Health Care Unit	--
ULS	Local Health Unit	--
LUTS_II	Administrative region	Norte, Lisboa Vale e Tejo, Alentejo, Algarve, Centro
Tipo Unidade	Type of Health care unit	USF- A, USF – B, UCSP
IDG IPDA	Overall performance index (after IPDA)	--
densidade populacional	population density	--
ETC Assistentes técnicos	Equivalent to full-time technical assistants	--
ETC enfermeiros	Equivalent to full-time nurses	--
ETC médicos	Equivalent to full-time doctors	--
ETC internos	Equivalent to full-time interns	--
ETC outros profissionais	Equivalent to full-time other professionals	--
ETC outros profissionais saúde	Equivalent to full-time other health professionals	--
Idade US	Health Care unit Age	--
Indice de contexto sociodemográfico	Social Demographic context index	Determining a representative sample of HCU so that these can be used to perform efficiency analyses "equalizing the context" or econometric analyses; Encourage health professionals to choose vacancies in places with a "worse context"; Increase the amount paid for each weighted unit of users in locations with a "worse context"
Assistentes tecnicos (n)	Number of technical assistants	--
Enfermeiros (n)	Number of Nurses	--
Inscritos (n)	Number of registered service users	--
Inscritos com MF	Number of registered service users with Family doctor	--
Médicos (n)	Number of Doctors	--
Médicos internos (n)	Number of Intern doctors	--
Outros profissionais (n)	Number of Other professionals	--
Médicos ponderados (n)	Number of weighted doctors	--

Prevalencia_DM	Prevalence of Diabetes Mellitus	--
Proporção idosos	Elderly proportion	--
Unidades ponderadas	Weighted Units	--
Taxa de desemprego	Unemployment rate	--
taxa abandono escolar	school dropout rate	Proportion of adults without the 9th year of schooling