FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Development of a trip purpose estimation model

João Fernandes



Mestrado em Engenharia de Software

Supervisor: Prof. Marta Campos Ferreira

Second Supervisor: Prof. Teresa Galvão Dias

October 20, 2022

Development of a trip purpose estimation model

João Fernandes

Mestrado em Engenharia de Software

Approved in oral examination by the committee:

President: Prof. Nuno Flores
Referee: Prof. Rui Gomes
Referee: Prof. Marta Campos Ferreira

October 20, 2022

Resumo

Os Sistemas de Bilhética Electrónica (AFC) são importantes não apenas para facilitar o uso dos transportes públicos, mas também para a colecta de dados que são importantes para a melhoria desse sistema. A estimação do próposito da viagem é extremamente útil para que os provedores de serviço possam saber o padrão de mobilidade dos cidadãos e que atividades estes realizam entre viagem e para que assim possam desenvolver politicas e adaptar a sua oferta consoante os casos. Este estudo tem como objetivo o desenvolvimento de um modelo de estimação do propósito da viagem utilizando em conjunto os dados do sistema de bilhética eletrónica e os dados dos pontos de interesse (POIs) da área metropolitana do Porto (PMA).

Para alcançar a estimação, esta pesquisa apresenta os métodos para estudar os dados de AFC e POIs e como prepará-los para obter conhecimento. Os dados AFC são carregados, analisados e preparados, utlizando os processos do CRoss Industry Standard Process for Data Mining (CRISP-DM), para obter informação apartir dos dados, identificando estudantes, trabalhadores, turistas e reformados e para aumentar a performance de análise ao utilizar os dados. De seguida, os dados dos POIs são trabalhados da mesma maneira, quando o final dataset dos POIs é atingido essa base de dados é utilizada em conjunto com uma tabela de estações, utilizando longitude, latitude e distancia para caracterizar uma paragem pela frequencia relativa dos POIs que se encontram a 400m da mesma. Em seguida, esta tabela é cruzada com os dados da tabela AFC, juntamente com pressupostos analisando a hora de validação, duração entre viagens e tipo de utilizador para criar um sistema baseado em regras para identificar o propósito de uma viagem. No fim, apresenta o desenvolvimento dos modelos de estimação, os seus resultados e a discussão dos mesmos.

Os melhores resultados entre: *Decision Tree; Deep Learning; Naïve Bayes; Neural Networks*; e *Random Forests* foi *Decision Tree* com *bagging*. No entanto, melhores resultados poderiam ter sido atingidos se esta pesquisa tivesse um inquérito para suportar as suas estimações e se o ambiente computacional tivesse sido mais poderoso. O objetivo principal que era usar os dados AFC com a base de dados dos POIs da PMA foi alcançado.

Keywords: Objetivo da viagem, estimativa, inteligência artificial, sistemas AFC, mineração de dados, dados de POI, previsão do objetivo da viagem, estimativa do objetivo da viagem, sistema baseado em regras

Abstract

Automated Fare Collection (AFC) Systems are important to not only facilitate the usage of the public transportation system but also to collect vital data to improve said system. Estimating trip purpose is very useful for service providers to know the mobility profile of citizens and the activities they perform between trips and to develop policies and adapt the offer accordingly. This study aims to use the Andante AFC data with the POIs Data from the Porto Metropolitan Area (PMA) to develop a model to estimate trip purposes.

In order to achieve the estimation, this research presents the methods to study the AFC and POIs data and how to prepare them to obtain knowledge. The AFC Data is loaded, analysed and prepared, using CRoss Industry Standard Process for Data Mining (CRISP-DM) processes, to obtain information from the data, identifying students, tourists, retirees and regular users, and to achieve the best performance when using the dataset; Next, the POIs Dataset is worked in the same manner, after achieving the final dataset with the POI information, this database is crossed with a stations table using latitude, longitude and distance to characterize a stop by the POIs relative frequency 400 meters around them. Next, this final POI Dataset is crossed with the AFC Data, and with assumptions analysing the date of the validation, the time between trips and the type of user, in order to create a rule-based system to find the purpose of a trip. In the end, it presents the development of the estimation models, their results and the discussion of said results.

The best result among Decision Tree; Deep Learning; Naïve Bayes; Neural Networks; and Random Forests was the Decision Tree Model with bagging. However, better performance of the models could be achieved if the estimation had a trip survey to support the estimations upon and if the computational environment was more powerful. The main objective, which was to use AFC Data along with the PMAs POIs dataset to estimate trip purpose was achieved.

Keywords: Trip Purpose, Estimation, Artificial Intelligence, AFC Systems, Data Mining, POI Data, Trip Purpose Prediction, Trip Purpose Estimation, Rule-Based

Acknowledgements

É difícil de acreditar que esta longa fase da minha vida, a vida académica, está a chegar ao fim.

Primeiramente, gostaria de agradecer às minhas orientadoras, Marta Campos Ferreira e Teresa Galvão Dias por todo o seu apoio incondicional, paciência e confiança que depositaram em mim durante toda a dissertação.

Ao meu irmão, pai e mãe cuja paciência, apoio e amor me ajudaram durante os anos e continuarão a ajudar-me. Amo-vos para sempre.

Ao Flávio Teixeira, Pedro Sousa e à Diana Mourão, que tenho o privilégio de poder chamar de amigos, obrigado por me ouvirem e apoiarem-me durante os momentos mais difíceis.

Por último, a todos os meus outros amigos e família que eu não mencionei mas que também me são muito importantes.

Aos meus bisavós João e Alice.

João Fernandes

"On ne voit bien qu'avec le cœur.
L'essentiel est invisible pour les yeux."

Antoine de Saint-Exupéry, *Le Petit Prince*

Contents

1 Introduction **1**
- 1.1 Problem Context . . . 1
- 1.2 Motivation . . . 1
- 1.3 Objectives . . . 2
- 1.4 Document Structure . . . 2

2 Literature Review **3**
- 2.1 Introduction . . . 3
  - 2.1.1 Finding the revision papers . . . 3
- 2.2 Automated Fare Collection Systems . . . 4
  - 2.2.1 Usage of the AFC Data . . . 4
  - 2.2.2 Estimation of alighting point . . . 4
  - 2.2.3 The Data . . . 5
  - 2.2.4 Trip Purpose . . . 7
- 2.3 Discussion/Conclusion . . . 11

3 Context **12**
- 3.1 The Porto Metropolitan Area . . . 12
- 3.2 The Public Transport System of the PMA . . . 13

4 Methodology **16**
- 4.1 The Data . . . 16
  - 4.1.1 The CRISP-DM Process . . . 16
  - 4.1.2 AFC Data . . . 18
  - 4.1.3 POIs Data . . . 22
- 4.2 Rule-Based System . . . 30
  - 4.2.1 Health Care and Shopping . . . 31
  - 4.2.2 Work trips . . . 31
  - 4.2.3 Education trips . . . 33
  - 4.2.4 Home trips . . . 34
  - 4.2.5 Food trips . . . 37
  - 4.2.6 Tourists . . . 38
- 4.3 Final Dataset . . . 38
- 4.4 Technologies . . . 39

5 Results and Discussions **41**
- 5.1 The Estimation Models . . . 41
  - 5.1.1 Decision Trees . . . 43

5.1.2 Deep Learning 45
5.1.3 Naïve Bayes 45
5.1.4 Neural Networks 45
5.1.5 Comparisions 46

6 Conclusion and Future Work **48**
6.1 Conclusion 48
6.2 Future Work 48

References **50**

List of Figures

2.1 Data Mining Process [38] . . . 6
2.2 Alsger A et al. - Logical inference framework . . . 10
2.3 Alsger A et al. - Structure of the model development. . . . 10

3.1 PMA Map [5] . . . 12
3.2 Andante Zoning [5] . . . 14

4.1 CRISP-DM Phases [21] . . . 18
4.2 Feature Engineering Main Components [4] . . . 20
4.3 Passenger Characterization Diagram . . . 21
4.4 Education . . . 27
4.5 Food . . . 27
4.6 Healthcare . . . 28
4.7 Leisure . . . 28
4.8 Lifestyle . . . 28
4.9 Lodging . . . 28
4.10 Religion . . . 29
4.11 Services . . . 29
4.12 Shopping . . . 29
4.13 Transportation . . . 29
4.14 Worker's Validations Bars - October 2019 - 30 min periods . . . 32
4.15 Worker's Validations Line - October 2019 . . . 32
4.16 Student's Validations Bars - October, 2019 - 30 min periods . . . 33
4.17 Student's Validations Line - October, 2019 . . . 34
4.18 Without 'LastTrip' Validations Line - October, 2019 . . . 35
4.19 Without 'LastTrip' Validations Bars - October, 2019 . . . 35
4.20 Validations Line - October, 2019 . . . 36
4.21 Validations Bars - October, 2019 . . . 36

5.1 RapidMiner Main Phase - Common Phase . . . 42
5.2 RapidMiner Validation Phase - Random Forest Example . . . 42
5.3 Example Decision Tree using the Data - Zoomed . . . 44
5.4 Neural Network Example,[29] . . . 46

List of Tables

2.1 AFC Systems 5

3.1 Monthly Discount Requeriments 15

4.1 Original Dataset Description 19
4.2 POIs Dataset Attributes 22
4.3 POIs Categories 23
4.4 Count of POIs in PMA 23
4.5 Example - Stations Dataset 24
4.6 POIs Station Database 26
4.7 POIs Station Database - Example 26
4.8 Frequency Relative of POIs in Stations - Example 26
4.9 Classified Stations - Example 27
4.10 Variables of the final dataset 39

5.1 Variables fed into model generation 41
5.2 Confusion Matrix for Decision Tree Model with Bagging 43
5.3 Model Comparison Table - Accuracy Results 47

1 Discount 54
2 Tariff Code 54
3 Confusion Matrix for Deep Learning 10 epochs 55
4 Confusion Matrix for Deep Learning 200 epochs 55
5 Confusion Matrix for Deep Learning 500 epochs 55
6 Confusion Matrix for Decision Tree Model 56
7 Confusion Matrix for Decision Tree Model with Bagging 56
8 Confusion Matrix for Naïve Bayes 56
9 Confusion Matrix for Neural Networks 56
10 Confusion Matrix for Random Forest 100 trees 56
11 Confusion Matrix for Random Forest 200 trees 57

Listings

4.1 Code that matches POIs around a station 25
4.2 Code that sets 'purpose' variable as "hospitals" 31
4.3 Code that sets 'purpose' variable as "shopping" 31
4.4 Code that sets 'purpose' variable as "work" 32
4.5 Code that sets 'purpose' variable as "education" 34
4.6 Code that sets 'purpose' variable as "home" for students 37
4.7 Code that sets 'purpose' variable as "home" for workers 37
4.8 Code that sets 'purpose' variable as "food" for students 37
4.9 Code that sets 'purpose' variable as "food" for workers 37
4.10 Code that sets 'purpose' variable as "food" for workers 38
4.11 Code that sets 'purpose' variable as "lodging" for tourists 38
4.12 Code that sets 'purpose' variable as "lodging" for tourists if 'EntranceStation' is the airport 38

Abbreviations

AFC	Automated Fare Collection
CBD	Central Business District
CNIL	Commission nationale de l'informatique et des libertés
CP	Caminhos de Ferro Portugueses, EP / Comboios de Portugal
CRISP-DM	Cross Industry Standard Process for Data Mining
CSV	Comma-separated values
CTA	Chicago Transit Authority
EU	European Union
FEUP	Faculdade de Engenharia da Universidade do Porto
GDPR	General Data Protection Regulation
Geo Location	Geographic Location
GPS	Global Positioning System
GTFS	General Transit Feed Specification
GUI	Graphical User Interface
HMM	Hidden Markov Model
HTTP	Hypertext Transfer Protocol
HTS	Household Travel Surveys
IBM	International Business Machines Corporation
IDs	Identifications
O-D	Origin-Destination
PMA	Porto Metropolitan Area
POI	Point of Interest
POIs	Points of Interest
SPAM	Sending and Posting Advertisement in Mass
STCP	Sociedade de Transportes Colectivos do Porto
TIP	Transportes Intermodais do Porto
UTC	União de Transportes dos Carvalhos

Chapter 1

Introduction

1.1 Problem Context

Since the 1990s Automated Fare Collection (AFC) Systems have become more common in public transportation systems around the world. These systems were initially thought of with the intent to facilitate payment by the use of smart cards which eliminated cash from the travel process. However, these systems also collected data, which over the years have been used in several kinds of research that had a great impact on studies in the area of public transportation. The information obtained from this data can even be more powerful if we cross it with other datasets such as the Points-of-Interest dataset.

The Porto Metropolitan Area (PMA), which is the area of study of this dissertation, has a vast public transport system which has Andante as its AFC System. Andante, which was implemented in 2002, is an entrance-only AFC System which works with contactless fare cards that can be used on validators on stops, for light rail and trains, or inside each land bus.

This thesis uses the data from the Andante AFC database together with the data of Porto's Points-of-Interest (POIs) to build a rule-based system to identify trip purpose to then be used to develop an estimation model.

1.2 Motivation

Estimation of trip purpose is usually based on expensive and time-consuming national travel surveys. Estimating the purpose of the trip is important so that service providers can know the mobility profile of citizens and the activities they perform between trips to develop policies and adapt their offers accordingly.

The Andante Network has never done any survey, thus is important to use alternative data sources, like AFC Data alongside the POIs Dataset, to do this estimation.

1.3 Objectives

This dissertation aims to use the AFC Data from the Andante Network alongside the POIs Dataset from the PMA to:

- Estimate the passengers' trip destination;
- Calculate the relative frequency of POIs around a station to estimate its type;
- and use the above to reach the grand final objective Estimate the Trip Purpose.

To achieve these objectives first it is important to analyse the AFC Data and extract information from it and remove the unnecessary data. It is in this step that, regular users, students, retirees, and tourists are identified; the stations where they alight and the trip duration.

Secondly, the POIs Dataset is worked on. This dataset is then crossed with a stations table, using latitude, longitude and distance, and depending on what POIs are around a certain station that is how, that station, is going to be classified.

Finally, by crossing the characterized stations with the worked AFC Data taking into account trip characteristics, data analysis and what is known about the Portuguese Lifestyle, a rule-based system characterizes the purpose of the trip.

1.4 Document Structure

First a brief insight into the current state of the field is given in Chapter 2. Then Chapter 3 presents a brief introduction to the PMA and its public transport system. The next chapter describes all the methodology and processes with the data. The fifth chapter presents the development of the estimation models and results. The ending chapter concludes the dissertation and what steps should be taken to improve even further.

Chapter 2

Literature Review

In this chapter, a revision of previous works which utilize data from Automated Fare Collection Systems is presented to the readers.

2.1 Introduction

Finding ways to estimate the trip purpose is important because usually they are based on expensive and time-consuming surveys. Nowadays alternative options are being studied by researchers mainly by collecting data from AFC Systems and inferring the purpose by studying and working on that data. This estimation is important so that not only money can be saved on surveys but also that providers can know where to invest in the public infrastructure, like new stations and new routes.

This review will discuss the current state of AFC Data and how to work with that data and will explain that while it was possible to use data from Google Places and social media by referring to other studies, nowadays it is harder to do so. It finishes with a revision of several studies that lean on different methods to achieve the main theme of "Estimating trip purposes". This revision was important so that:

- The current state of estimating trip purposes could be understood;
- Learn some aspects which can help during the process of working with Porto's public transport database;
- To have an idea on what to work on so to improve processes for making it easier for future studies in this area of estimating trip purposes.

2.1.1 Finding the revision papers

The author thinks is also important to mention the way review papers were found and chosen.

The research was done mainly in Scopus by using the query "trip estimation purposes" which were then filtered by selecting only the fields of computer science and engineering. The resulting

papers were then selected by their title. While reading the final papers the author also found other papers by cross-referencing, which has proven to be helpful for the final result.

2.2 Automated Fare Collection Systems

This section discusses concepts and previous works within the scope of the trip purpose estimation area.

Used by many public transport systems around the world an Automated Fare Collection System (AFC) is the collection of components that automate the ticketing system of a public transportation network. AFC systems originated with tokens or paper tickets dispensed by staff or from self-service vending machines. Nowadays these tokens have been in general replaced by cards, and with the evolving technology, some systems already accept contactless bank cards as payment methods [42]. There are already many of these systems implemented worldwide such as the Oyster Card in London, the MetroCard in New York City, Navigo in Paris, OV-Chipkaart in all of the Netherlands and the main subject of this thesis Andante, in the Porto Metropolitan Area.

2.2.1 Usage of the AFC Data

Barry et al, [8], in 2002 enumerated the following advantages of AFC-based methodology

- Large accurate database containing 4 million subway transactions per day,
- Savings of $2 million in data collection and editing costs,
- Available data in a few weeks rather than in a year or more,
- Data continuously updated 24 h per day, and
- Potential for expanding the process to include bus trips.

The last point nowadays is already true, Andante is an example of an AFC system on a bus.

But with such power over the data, all of the AFC data must be treated as sensibly as possible, confidentiality is a must. AFC data is no different from data obtained by other individual data collection systems, as pointed out by Clarke, [11]. In 2003, the *Commission nationale de l'informatique et des libertés* (CNIL) gave the recommendation to treat this data with much care as personal data from French citizens and public authorities of the French Republic were involved, [34].

2.2.2 Estimation of alighting point

When an entity wants to implement an AFC system card validation is one of the most important steps. When building the system, the entity responsible for the public/private transportation network has to choose if they want an entry-only validation system or an entry-exit validation system. For example, table I, shows the methods use in the systems mentioned in the previous section.

	Entry	Exit
Andante	✓	
MetroCard	✓	
Navigo	✓	✓
OV-Chipkaart	✓	✓
Oyster Card	✓	✓

Table 2.1: AFC Systems

Estimating the alighting point of an entry-exit system is far easier as when the passenger disembarks the mean of transportation the passenger is required to do the validation again. Using the entry-only solution raises the problem that the trip's destination will be unknown, but in the case of Porto is the only solution due to the layout of the system, having stations on the street level does not allow exit control. In these cases, entry-only, the alighting point has to be determined so studies were performed: Barry et al, [8], in the same study mentioned above made two key assumptions:

- The exit station of a person's current subway trip is the same as the entry station of the passenger's next subway trip.
- The exit station of a person's last subway trip of the day is the same as the entry station of the passenger's first trip of the day

This thesis used this method as the basis to determine the exit station of the Andante users as the primary objective was not to find the exit station but to what the passenger will do at said location.

Only the trips that have +30min between them will be considered this is due to Ali A et al research, [2], which defines that trips occurring between 30 min of each other should be distinguished as transfer points.

Analysing the data from the CTA (Chicago Transit Authority) system, Zhao et al, [44], reached the conclusion that the maximum walking distance is 400m.

This walking distance was also used in this work.

2.2.3 The Data

2.2.3.1 Data preparation

AFC data will need to be worked on so for that Data Mining techniques will be applied.

Data Mining, [43] , [38], can be defined as the process of extracting and discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. The next figure shows the data mining process:

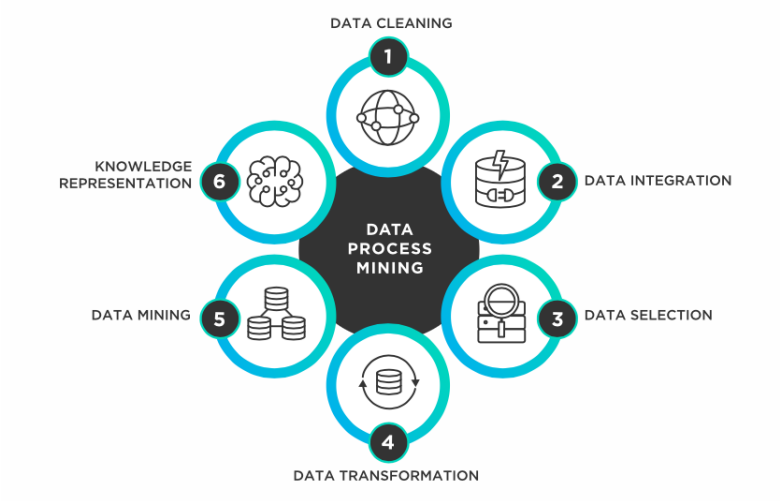


Figure 2.1: Data Mining Process [38]

First, it is important to clean the data, to remove inconsistencies, like repeated values or data where the dates are off (noisy data) and others.

The data produced by the AFC system is so large, due to the number of passengers and the easiness of collecting the information that without the removal of data it would make the process of processing the data a lot slower so as to facilitate and speed it up useless attributes can be removed from the data: redundant data; and irrelevant data to the case study. This will help reduce the dataset to a much lighter version. If needed and possible some information can also be compressed.

After the data is clean it needs to be integrated, and join it with the different sources, for example, joining the metro stations' geo-location with the Google Places data, which has to also be cleaned as much as possible.

Finally, it is important to do transformations on the data, make it more uniform, and transform it to get the best performance to the case in mind. Some examples of data transformation: Normalisation, Aggregation, Discretisation,....

With the data mining process information on the data can be discovered, because data is meaningless without providing information. With the data mining process, information will be extracted, and in the case of this thesis, a heuristic model will be designed to make assumptions on the data, as explained in the next section.

At the end of all these processes, knowledge is obtained.

2.2.3.2 Assumptions on the Data

Some assumptions of trip purposes can be made using AFC data and land usage data, working with temporal and spatial attributes and by leaning on a rule-based approach. This method is later discussed in the next section in a more detailed perspective presenting studies.

But they are only assumptions because as Bagchi and White, [7], mentioned in their work, the absence of trip purpose and activity information of the passenger is an intrinsic limitation of the

data itself, no definite certainty can be obtained. The use of household travel surveys can also be used to enrich the AFC data by connecting the two data sources as studied by Trépanier et al. , [39].

2.2.3.3 Gathering of Data - Google Places

Studying the works of Cui Y et al, [12], and Ermagun et al, [14], it is proven that the gathering of data from Google Places is possible. To accomplish this gathering of data the use of their Application Programming Interface(API) is required. An API, [33], is a set of definitions and protocols for building and integrating application software. It can be referred to as a mediator, as a contract, between an information provider (Google) and an information user (Scholar). This API allows the connection to the Google Places service and download of the data without needing to understand how the service is implemented. The Google Places API returns information about places using HTTP requests, [20]. In this thesis the nearby_search request is used, using the provided alighting station coordinates a cluster around that station is generated with relevant places in its vicinity (600m as defined previously), there is a lot of data that can be retrieved from Google Places, here it will mostly be used the business hours and the type of place (leisure, work, education,..). Unfortunately, to the date of this work, the real-time occupation information is still not provided, [36]. As noted by Cui Y et al, [12], a free Google Place API account had a limit of 150,000 free requests per 24 h period and would only return at most 60 Points of Interest (POI), unfortunately, from 2021 onwards, , Google no longer offers a free tier option.

2.2.3.4 Gathering of Data - Social Media

In the same study Cui Y et al., [12], use Twitter data to get the popularity of a certain place by combining it with the Google Places data. So it's possible to obtain that data from Twitter.

Instagram is not possible to use without creating an additional tool to geotag the place due to Facebook cutting out the ability to pull Instagram user location data via an API, so it is no longer possible to get GPS coordinates from Instagram and only the location names, without the GPS coordinates it will harder to geotag the locations making it hard to use this data. This is mainly due to GDPR rules that came into effect a few years ago.

Due to all of this, Social Media data was not considered for estimating the trip purpose in this work.

2.2.4 Trip Purpose

In this subsection, the various approaches adopted by previous researchers to estimate the trip purpose are presented to the reader.

First, works that are based on rule-based processing approaches are presented.

When using this approach four attributes are usually important and have a big weight on the inference.

- Activity duration
- Departure time
- Frequency
- Card type

Lee and Hickman, [26], using AFC data, Google's general transit feed specification (GTFS), and parcel-level land use data were successful in inferring trip purposes. From the data, they developed a training set of behavioural and heuristic rules to derive trip purpose. Some aspects used to create the rules:

- A transit user (e.g., commuter) visits the same location on multiple days in the same week.
- A transit stop used frequently is near to the transit user's workplace or home.
- A specific activity has a certain understood time duration.
- The start and end times of specific activities may be associated with certain times of the day. In the morning peak periods, a well-recognised transit demand moves from a residential area to the workplace, while in the evening peak periods this demand shifts from the workplace back to a residential area. Meaning work-related trips

Lee and Hickman also conclude that trips made with student cards with exits near schools/universities are education-related trips.

Devillaine F et al, [13], use smart card databases and other available information about land use and user behaviour for the detection and estimation of the location, time, duration, and purpose of activities of public transport users. Their method was applied to Santiago, Chile, and Gatineau, Quebec, Canada. First, they process the databases to estimate boarding and alighting stops and times using methods from other studies. This leads to a full set of stages being obtained, each one with a full set of attributes that characterises each database. Next, they differentiate trips from transfers. In the end, all data is combined to find the purpose, four purpose types are defined: work, study, home, or other. They also set some rules.

For work:

- Card type is adult and,
- Trip before the activity was not last of the day.
  For Canada:
- Activity is longer than 5h,
- Activity is between 2 and 5h long, and the bus stop zone is not a hospital or recreational.
  For Chile:
- Activity is longer than 2h and,

For school:

- Card type is student or underage and,
- Trip before the activity is not last of the day.
  For Canada:
- Activity is longer than 5h and,
- Activity is between 2 and 5 h long, and bus stop zone is education.
  For Chile:
- Activity is longer than 5h.

Home trips are considered the last of the day.

After creating this training set, a Decision Tree based classification technique is conducted with a test set that is used to determine the performance of the model. Later this model is validated with onboard surveys. The set of heuristic and behavioural rules is determined by the Farecard Type, duration of the journey, frequency of the journey, and others.

Lu K et al. [27] using data from the Beijing Metro Transit inferred trip purpose from the passengers alighting and boarding station. For example: if the trip started in a residential area and headed to the Central Business District (CBD) it was considered a work-related trip. They based on the land use and the daily entry flow pattern for each metro station, and processed it using the k-means clustering method and with the results classified the stations into 7 categories: Work, Residential, Working-residential, Transit hub, Railway, Shopping-sightseeing and Rural stations. They estimated the trip purpose using the category of the station. They give an example of how it works in their paper: If a passenger started his trip from a residential station and finished his trip at a working station. Based on the station category, the trip could be labelled as a working trip. This process could efficiently determine the trip's purpose during the day. On stations with the two categories like the "working-residential", they recorded the movements for a week, filtered them, and studied the aspects of it, and then based on a rule-model approach they tagged it as work or residential. Time of day and frequency of trip were important aspects. When validating the model 75% accuracy was obtained.

Ali A et al, [1] used Decision trees and rule-based models, R-Tree and C50, to estimate trip purposes. This was applied to data from Seoul, South Korea. C50 algorithm proved to be the best with 80% accuracy.

Apart from rule-based approaches:

Kusakabe and Asakura [25] in order to supplement absent behavioural attributes in the smart card data, developed a data fusion methodology of AFC data with the person trip survey data with the naïve Bayes probabilistic model. By doing this fusion of data they were able to estimate the trip purposes with a success rate of 86.2% when using the validation set.

Alsger A et al.,[3], combine data from the GoCard data (AFC System from Brisbane, Queensland), Household Travel Surveys (HTS) and land use databases to generate the trip purpose inference model, as shown in Fig. 4.

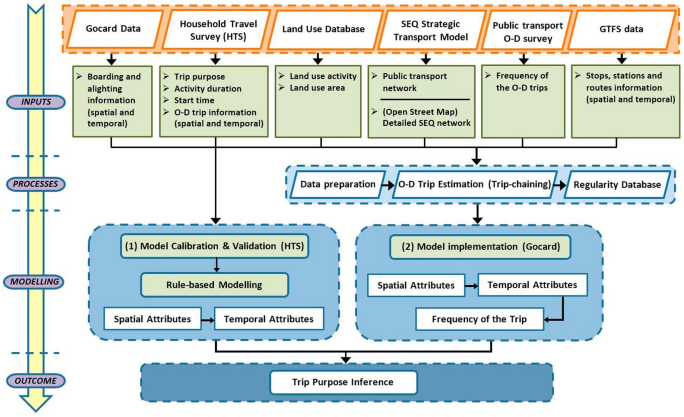


Figure 2.2: Alsger A et al. - Logical inference framework

First, they work on the data (all of the data preparation phases like data cleaning and refining), do some analysis and then integrate the data from the different sources. As the previous picture shows two stages are needed to infer passengers' trip purpose, this is due to GoCard data not including trip purpose information, which, as previously mentioned in the last section, is an intrinsic limitation of AFC data, so the use of the HTS data is needed to calibrate and validate the model, the first stage of the process, otherwise results would be meaningless. The second stage is the application of the model to the GoCard data. The model is built using a rule-based modelling approach, where the probability of a specific trip purpose is chosen relative to the other trip purposes based on the temporal and spatial attributes. The attributes that affect the travel purpose inference are:

- Temporal Attributes (extracted from HTS),
- Spacial Attributes (extracted from land-use database) and,
- Frequency of trip (Public transport O-D surveys).

Fig 5. shows the structure of the model

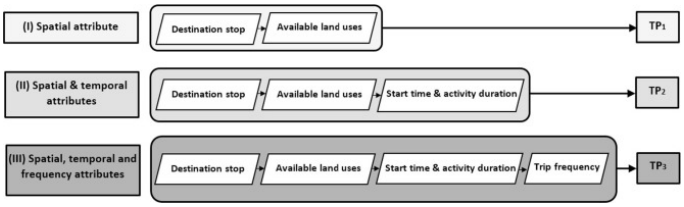


Figure 2.3: Alsger A et al. - Structure of the model development.

This approach has proven to show a strong capability to predict trip purpose at a high level of accuracy.

Chen et al., [10] formed a Hidden Markov Model (HMM) by fusing together a Markov chain and a Gaussian Mixture. The procedure is as follows:

- Trip chained is extracted from the data,
- The number of potential activities is determined,
- Possible number of clusters is determined,
- Feature data are collected for every hidden activity within the sequence of activities for trip chains in the sample,
- Implementation of a solution algorithm for estimating the parameters of an HMM,
- Characterise clusters and map each state with them by investigating the estimated parameters,
- Using the parameters, the most probable sequence of activities is imputed for trip chains.

The data used was from GPS data of Yizhuang district, Beijing, China. This method has an accuracy of 86.43%

Cui Y et al. , [12], in the same study mentioned in the previous section, used a Bayesian neural network to model the trip dependence on each individual's daily trip chain and infer the trip purpose. Their work was to predict the current trip purpose and the next trip purpose. They register that Google Places and Twitter information greatly improve the overall accuracy of prediction of certain activities.

2.3 Discussion/Conclusion

The estimation of the trip purpose has several ways to be achieved. It requires a long and attentive process, from the Data Mining to the final result.

From all the research done a better perspective was obtained on what to do and how to do it. It will be important to work the data and to clean it and process it as good as possible to improve the processing times. Using the data from Google Places is also a must as it improves greatly the estimation of the trip purpose.

Estimating the trip purpose is important as it can help to understand the current state of public transport users and help to know where to invest more service and policy wise.

Chapter 3

Context

This chapter describes the Porto Metropolitan Area (PMA) using POI data extracted from Google Places and also describes the public transportation characteristics using information available on the PMA's website [5].

3.1 The Porto Metropolitan Area

Located on the north coast of Portugal the PMA embraces a geographic area of 17 municipalities, from Póvoa do Varzim to Arouca in its northwestern and southern extremes respectively, spreading around an area of approximately 2.040km^2 with a population rounding up to 1.7M inhabitants.

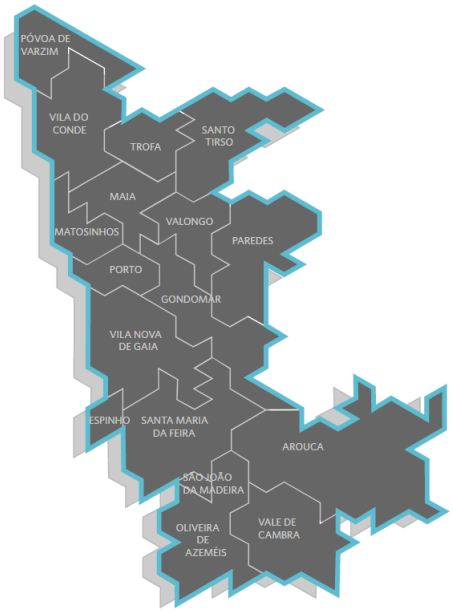


Figure 3.1: PMA Map [5]

3.2 The Public Transport System of the PMA

The PMA is served by different transport companies, each having a role in the different means of transportation:

- Sociedade de Transportes Coletivos do Porto (STCP) – for the main public transport operator;
- Metro do Porto SA (Metro do Porto) – for the light rail network;
- 29 other private bus transport operators, which serve not only the PMA but also have trips to several cities of Portugal outside the PMA and to the other EU Member States;
- Caminhos de Ferro Portugueses, EP (CP) – train network.

On the 20th of December 2002, the three main companies: Metro do Porto, STCP, and CP to allow intermobility in the public transportation systems of the PMA constituted a joint venture called Transportes Intermodais do Porto (TIP), which is responsible for the ticket system Andante.

Andante is an entry-only AFC system, having ticket readers on the entrance of the stations and inside each BUS, its main objective is to allow a common ticketing system between the previously mentioned three companies and all the others which also decided to join, like União de Transportes do Carvalhos (UTC) and Espirito Santo. The Andante network is divided into several zones, Fig. 3.2, the tariff paid and the time of validity for the occasional ticket depends on the number of zones acquired, in case the user needs to travel to a zone to which the user has no valid ticket a new one has to be bought. The Andante 24 ticket is valid for 24 consecutive hours after the 1st validation.

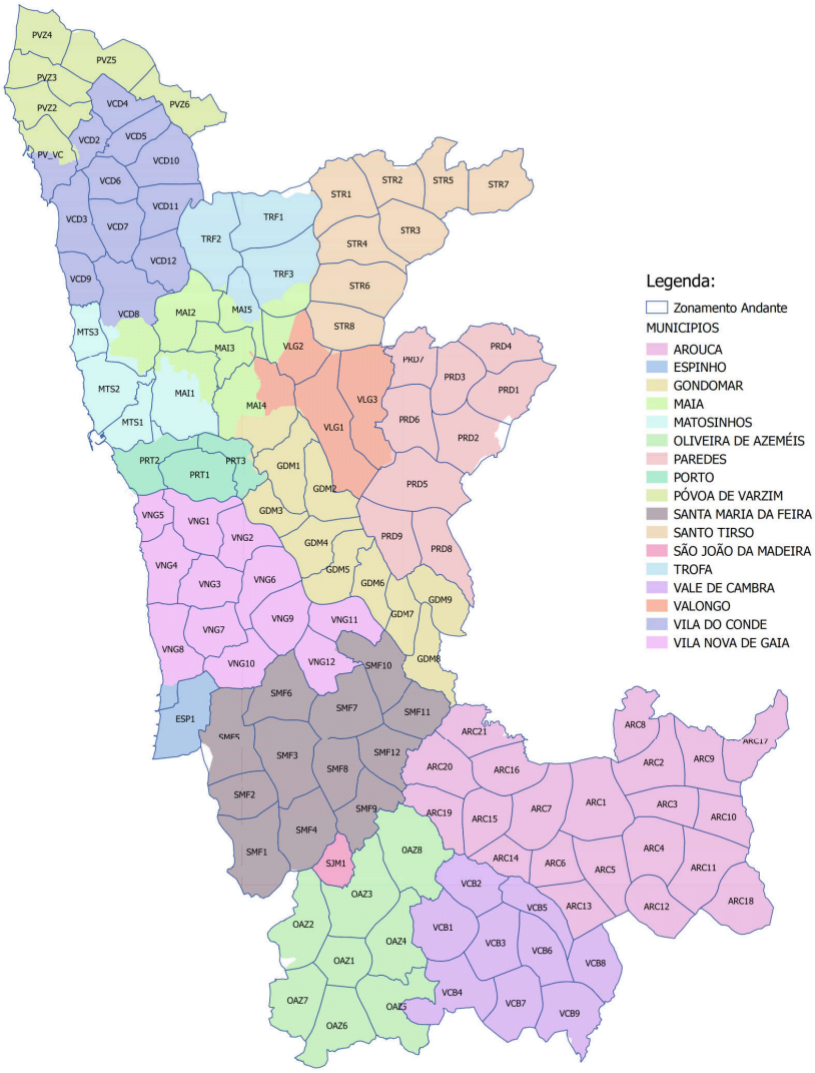


Figure 3.2: Andante Zoning [5]

There are three main types of Andante Cards[6]:

- **Occasional Tickets:**
  - *Titulo de Viagem*: Valid for a set of rings that are counted around the zone where someone has started their journey to the limit of the acquired zones during the purchase of the travel title.
- **Andante 24:**
  - Andante 24, which grants 24 consecutive hours after the first validation to travel inside the PMA.

- **Tour:**
  - Tour 1: Valid for 24h in all the Andante Network.
  - Tour 3: Valid for 72h in all the Andante Network.

  Both "tour" tickets have unlimited validations during the time period and are not rechargeable.

- **Monthly Subscriptions:**
  - 3Z: Valid in 3 contiguous zones selected by the client in the act of purchase in any line and any vehicle in the Andante Network.
  - Municipal: Valid for all zones inside a municipality selected by the client during purchase act in any line and any vehicle in the Andante Network.
  - *Metropolitano*: Valid for all the PMA in any line and any vehicle in the Andante Network.

  All monthly subscriptions are valid for a month and can be obtained with a discount if someone has the requirements to apply to said discount, refer to table 3.1

Category	Requirements
Social+	Aimed at individuals and families with lower economic resources
Retired/Senior	For People Aged 65 or older
Students 4_18	For all students from 4 to 18 years old (inclusive) who do not attend higher education and who do not benefit from school transport assured by the City Councils.
Sub23	For all students up to the age of 23 (inclusive) enrolled in study programmes officially recognised by the Ministry responsible for Higher Education. 24 for Medicine and Architecture Students
Family	Households (minimum 3 members) with tax residence in one of the 17 municipalities of Porto Metropolitan Area.
Sub13	Free for all children from 4 to 12 years old (inclusive) that allows the free use of public transports integrated in the Andante intermodal system throughout the Porto Metropolitan Area.

Table 3.1: Monthly Discount Requeriments

Chapter 4

Methodology

This chapter focuses on explaining the data, and how it was processed. Then it explains the methodological steps used to process the data and how they helped achieve a final solution for the problem in question.

4.1 The Data

In this section all the processes related to the data are explained, the method used and then for each of the datasets an explanation from its raw form to the final data file is given.

4.1.1 The CRISP-DM Process

The CRoss Industry Standard Process for Data Mining (CRISP-DM) [21], is a process model that serves as the base for a data science process. It has six sequential phases:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

Which will be explained in the next pages.

Business Understanding:

To be able to create information from data the first step is to understand what is required to be obtained from the data. This is where the Business Understanding phase is important, to understand the business perspective. A meeting was scheduled at the beginning of the thesis development

process between the student and the supervisors to understand what was required of the estimation model.

Data Understanding:

This is the process where the analyst gets familiar with the data, analysing, identifying quality problems, and getting the first insights.

Regarding the AFC Data this process was done in Tableau, and with the initial raw data loaded first insights on it were taken and quality problems were identified.

The POI data was analysed in Python, but the process was essentially the same as with the AFC Data.

Data Preparation:

In this process of the CRISP-DM, all activities to construct the final data, from cleaning, to feature engineering to the obtention of information are covered.

Modelling:

In this phase of the project, a method for the model is chosen and the model is developed. Using software like RapidMiner can greatly help in this phase by removing the coding step of this phase.

Evaluation:

During this phase, the results and processes are evaluated, and it is determined what to do next if deploying the model or keep doing iteration of it.

Deployment:

Deploying the model for whatever task it is required to perform.

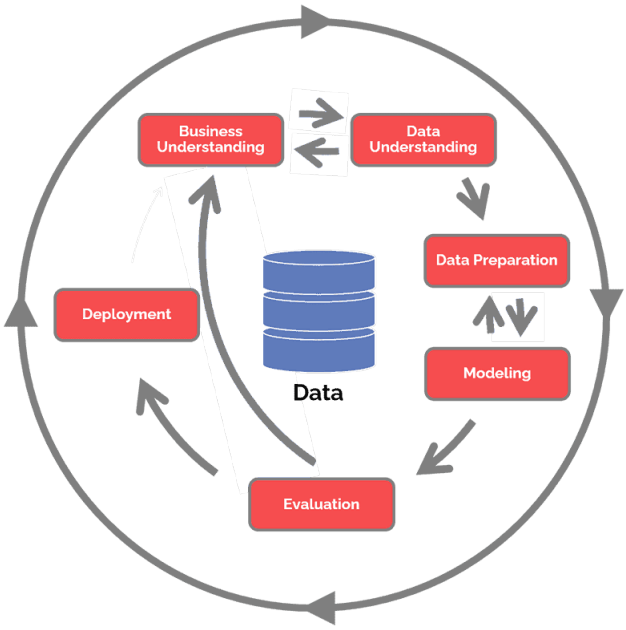


Figure 4.1: CRISP-DM Phases [21]

4.1.2 AFC Data

This project uses data from the Andante Network AFC System, this data is obtained when each passenger validates his/her card at the beginning of the journey. Validators are found at the entrance of each station and inside each BUS. The AFC Data and all its additional support tables, on the annexe, were kindly provided by TIP.

Data Understanding:

The dataset used in this estimation model was from the year 2019 and it initially contained approximately 82 million validations from 4.4 million different cards.

In total, 28 CSV files from two datasets were provided. These files contained in their essence the same data however the Metro and CP (trains) trips were taken from one dataset, the ROD table, and the STCP (BUS) trips from the other, the RR table. This is due to the fact that the BUS trip's "entrance station" information was missing from one of the datasets. When joining the datasets it was also required to use a support table that contained the stations with IDs -> String to convert the information present in the dataset to strings. Each CSV corresponds to one month of the year, May and October are distributed into two files due to the large number of trips that occurred during that period.

This dataset has several recorded attributes, the most important ones can be seen in Table 4.1:

Name	Type	Description
Code	Integer	ID of the Validation Transaction
Card	Hexadecimal	Encrypted ID of the Travel Card
SNValidator	String	ID of the AFC Validation Machine
ProviderCode	Integer	ID Code for the Transportation Service provider
Discount	Integer	ID Code for the Discount Type
TarifCode	Integer	ID Code for the AFC Card Type
OperatorID	Integer	ID Code for the Operator
DataHoraValidacao	DateTime	Time of the Validation
EntryZone	String	Origin Zone
ExitZone	String	Predicted Destination Zone
EntranceStation	String	Origin Station (Station where validation occurred)

Table 4.1: Original Dataset Description

From all these files only the month of October ended up being used, which has about 15.6 million validations out of the 82 million. October was chosen because it was a month where there were already students and workers doing their activities but also a significant number of tourists.

Data Preparation:

During the previous phase while analysing the data it was seen that most information was not useful for the development of the estimation model.

As such in order to improve future performance, save time, and storage space all the unnecessary data was removed.

First and foremost, “SNValidator”, “ProviderCode”, and “Code” were removed as they were unnecessary to the study.

Secondly “EntranceZone” and “ExitZone” columns were also removed due to the PMA zones being so vast and no useful information would come of it to the end purpose.

In the early stages of working with the data the "Card" variable was kept to do feature engineering and find the exit stops but later it was removed from the final dataset because it was wanted for the model to be independent from the passenger.

All duplicate validations were removed and it was made sure that every card that only had 1 validation per day would be removed as that would not give any significant information with Andante being an entrance-only validation system and no prediction of the exit station could be achieved using those records.

After analysis it was noticed that Trindade being mostly a transit station, had 1 million more validations than the second most used stop, was an outlier so it was decided to completely remove it.

During this phase Feature Engineering was performed, which according to Kaggle [24] is:

"Feature Engineering

The most effective way to improve your models"

Feature Engineering [31] is used to make the data better suited to the problem trying to be solved. It is done to improve model performance; reduce computational needs and improve results. Feature Engineering's main components can be seen in Fig 4.2:

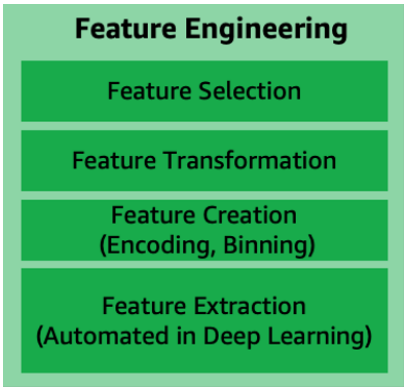


Figure 4.2: Feature Engineering Main Components [4]

Feature Engineering processes in this work were used to:

- Create variables to help create the rules for estimating the purpose of the trip. For example, "DurationofStay" which is the TimeDelta between two validations of the same card on a said day.
- Using the support tables TIP provided the numbered "Station" variable in the RR table was converted to Strings.
- Improve model performance by using Encoding in the variable "ExitStation", which was created during this phase.
- Selecting features such as only the trips with more than one validation per day.

Passenger Categorization:

To categorize the passengers, a rule-based system, Fig. 4.3, that uses the passenger card's 'Tarif-Code' and 'Discount' was developed to determine whether the passenger was a:

- Single User
- Regular User
- Student
- Retired User

- Tourist

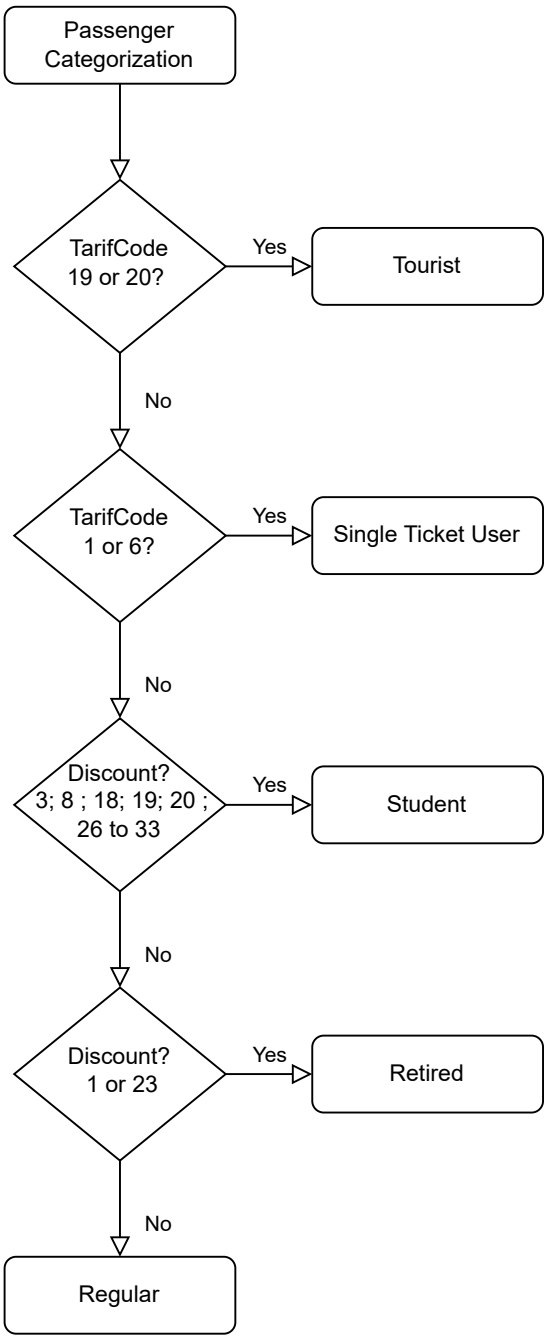


Figure 4.3: Passenger Characterization Diagram

As shown in the decision flowchart, Fig. 4.3, the data from the support data tables provided by TIP, which can be found in the appendix, were used to characterize the passengers.

This step is of most importance to better define the passengers' trip purposes. As, for example, a student is more likely to take an education trip than a retired user.

In the end four additional variables were obtained using feature engineering:

- "IsStudent";
- "IsRegular";
- "IsRetired" and
- "IsTourist".

These variables are only used to assist in finding the estimation of purpose at a later stage of the work and are not fed into the model.

4.1.3 POIs Data

The Points-of-Interest Data used to classify the stations was obtained by João Gonçalves, another student of FEUP, which he obtained while doing his thesis: "Understanding urban mobility patterns of tourists through data extraction", [19], this data was provided by the professors supervising.

Data Understanding:

The POIs dataset comes in a CSV file and it contains around 300 thousand POIs surrounding Porto. This dataset has the following attributes :

Name	Type
Name	String
Type	String
Longitude	Integer
Latitude	Integer

Table 4.2: POIs Dataset Attributes

Data Preparation:

Removing unnecessary information is important, so in this phase, the "Name" column was removed, maintaining only "Category", "Latitude" and "longitude" columns.

Secondly all the data that was wrongly formatted was removed.

Thirdly all duplicates or more were removed.

Lastly, all POIs with only the tag "establishment" were also removed since "establishment" is not specific enough to take any information from it.

In the end, 97 categories were observed, which were then grouped into 10 main categories:

Main Category	Includes
Education	Libraries, Schools and Universities.
Food	Bakeries, Cafés and Restaurants.
Healthcare	Dentists, Doctors, Physiotherapists, Hospitals and Pharmacies.
Leisure and Culture	Bars, Cinemas, Clubs, Museums, Parks, Stadiums, ...
Lifestyle and Beauty	Gyms, Haircare, Spas, ...
Lodging	Motels, Hotels, ...
Religion	Churches, Cemeteries and Temples.
Service	Car Rental, Courthouse, Embassies, Fire Stations, Insurance agencies, ...
Shopping	Grocery Stores, Shopping Centres, ...
Transportation	Airports, Taxi Stands, Transit Stations and Parking.

Table 4.3: POIs Categories

The final count of the POIs in the PMA, after all the data removal phases, was of 43903:

tipo	total
shopping	11103
food	9475
service	8932
healthcare	3215
transportation	2368
lifestyle_and_beauty	2364
lodging	1762
education	1693
leisure_and_culture	1670
religion	1321

Table 4.4: Count of POIs in PMA

Defining Stations by POIs

Due to the nature of this work, it was of most importance to characterize the stations using POI Data. To do so two different datasets were used: the station dataset provided by the supervisors, Table 4.5; and the previously described POI Dataset, table 4.2.

Provider	Code	ExtCode	IsActive	Name	ShortName	Latitude	Longitude
STCP	ANT2	ANT2	1	ANTAS (L. CIDADÃO)		41,16492702	-8,589191697
STCP	FMGC2	FMGC2	1	FORMIGA CEMITÉRIO		41,20955566	-8,542263813
STCP	BFAL1	BFAL1	1	BR. FALCÃO		41,15862168	-8,575365586
STCP	C24A4	C24A4	1	CAMPO 24 AGOSTO		41,14892532	-8,598545718
STCP	CMPO	CMPO	1	CAMPO		41,18301311	-8,469202054
STCP	TLHO1	TLHO1	1	TELHEIRO		41,18486023	-8,613143874
STCP	BVE1	BVE1	1	ERMESINDE - B. V.		41,21696773	-8,55143055
STCP	PROS2	PROS2	1	PARAÍSO (METRO)		41,15699309	-8,610080628
STCP	MAKR3	MAKR3	1	MAKRO		41,12036068	-8,626735599
STCP	DISP2	DISP2	1	DISPENSÁRIO		41,12927741	-8,615485597
STCP	CALE	CALE	1	R.CAMPO ALEGRE		41,15313852	-8,628624776
STCP	ZIND4	ZIND4	1	ZONA INDUSTRIAL		41,23263809	-8,665856231
STCP	AMRQ3	AMRQ3	1	ANTÓNIO MARQUES		41,23865852	-8,6135744
STCP	JAS2	JAS2	1	JOAQUIM ALVES DA SILVA (SUL)		41,2163051	-8,692486385
STCP	VGM5	VGM5	1	VASCO DA GAMA		41,24509704	-8,515236578

Table 4.5: Example - Stations Dataset

Later the Haversine Formula 4.1:

$$\alpha = \sin^2\left(\frac{\Delta\phi}{2}\right) + \cos(\phi_1)\cdot\cos(\phi_2)\cdot\sin^2\left(\frac{\Delta\lambda}{2}\right) \tag{4.1}$$

With:

$$c = 2 \cdot \arctan 2\left(\sqrt{\alpha}, \sqrt{1-\alpha}\right) \tag{4.2}$$

and:

$$d = R \cdot c \tag{4.3}$$

Where:

$$\begin{aligned} \phi &= Latitude \\ \lambda &= Longitude \\ R &= 6371 \text{ (Earth's mean radius in kilometers)} \end{aligned}$$

was applied using a Python Script, adapted from GitHub [18], listing 4.1, that uses longitude and latitude to provide in our case stations 400 meters around a POI, which was previously defined and explained in the State of the Art the reason for the 400 meters.

```
1  R = 6373e3  #radius of the Earth
2  for items in stations.index:
3
4      for item in pois_comstations.index:
5
6          rlat1 = stations['Latitude'][items] * (math.pi / 180)
7          rlon1 = stations['Longitude'][items] * (math.pi / 180)
8          rlat2 = pois_comstations['lat'][item] * (math.pi / 180)
9          rlon2 = pois_comstations['lng'][item] * (math.pi / 180)
10
11         lon2_or = pois_comstations['lng'][item]
12         lat2_or = pois_comstations['lat'][item]
13         lon1_sta = stations['Longitude'][items]
14         lat1_sta = stations['Latitude'][items]
15
16         dlon = (pois_comstations['lng'][item] - stations['Longitude'][items]) * (
               math.pi / 180)
17                             #change in coordinates
18         dlat = (pois_comstations['lat'][item] - stations['Latitude'][items]) * (
               math.pi / 180)
19
20         a = (math.sin(dlat/2) * math.sin(dlat/2)) + (math.cos(rlat1) * math.cos(
               rlat2) * (math.sin(dlon/2) * math.sin(dlon/2)))
21         c = 2 * math.atan2(math.sqrt(a), math.sqrt(1-a))
```

```
22          distance = R * c
23
24          if distance<=400:
25              if (pois_comstations['station'][item]) is None:
26                  pois_comstations['station'][item]=[]
27                  (pois_comstations['station'][item]).append(stations['Name'][items])
28              (pois_comstations['station'][item]).append(stations['Name'][items])
```

Listing 4.1: Code that matches POIs around a station

After running this script and making a few transformations a new dataset was generated and it contained the following variables:

Name	**Type**
Tipo	String
lat	Integer
lng	Integer
station	String

Table 4.6: POIs Station Database

tipo	**lat**	**lng**	**station**
food	41.125977	-8.603506	João de Deus
food	41.125977	-8.603506	JOÃO DE DEUS I
food	41.125977	-8.603506	COLEGIO DE GAIA Vi
food	41.125977	-8.603506	Raimundo Carvalho I

Table 4.7: POIs Station Database - Example

Note that as it can be seen in table 4.7 a POI can have more than one station associated too.

Lastly, using this dataset the relative frequency of the POIs in the stations was calculated and the one with the highest value was chosen to define/characterize it.

$$\text{Relative Frequency} : \quad = \quad \frac{\text{Number of POIs from type in Station}}{\text{Total POIs of Type in PMA}} \tag{4.4}$$

station	**education**	**food**	**healthcare**	**leisure_and_culture**	**lifestyle_and_beauty**	**lodging**	**religion**	**service**	**shopping**	**transportation**
14 DE OUTUBRO I	0.001772	0.002639	0.001555	0.0	0.005076	0.000568	0.000757	0.001791	0.003332	0.000422
14 De Outubro - MAFAMUDE	0.002953	0.003061	0.001866	0.0	0.005922	0.000568	0.000757	0.002015	0.003693	0.000422
14 de Outubro	0.002363	0.003061	0.001866	0.0	0.005499	0.000568	0.000757	0.001903	0.003513	0.000422

Table 4.8: Frequency Relative of POIs in Stations - Example

station	type
14 DE OUTUBRO I	lifestyle_and_beauty
14 DE OUTUBRO - MAFAMUDE	lifestyle_and_beauty
14 DE OUTUBRO	lifestyle_and_beauty

Table 4.9: Classified Stations - Example

POIs Distribution in the PMA:

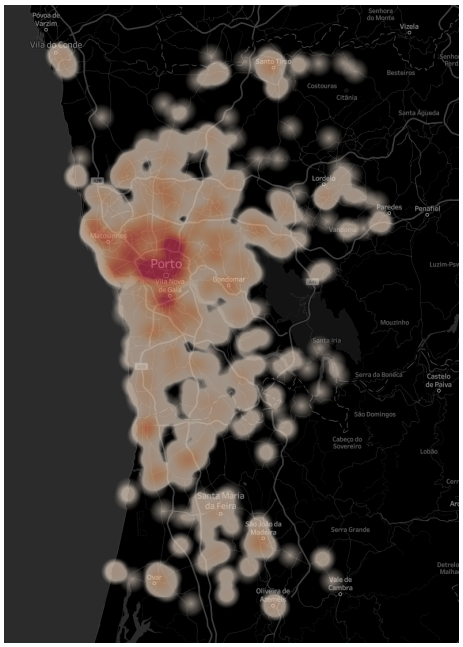


Figure 4.4: Education

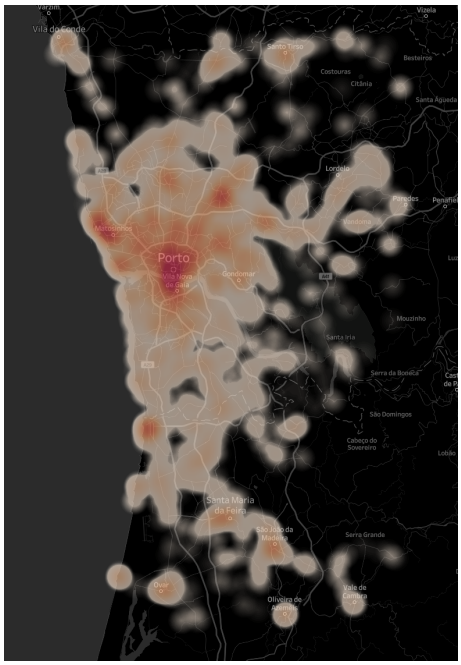


Figure 4.5: Food

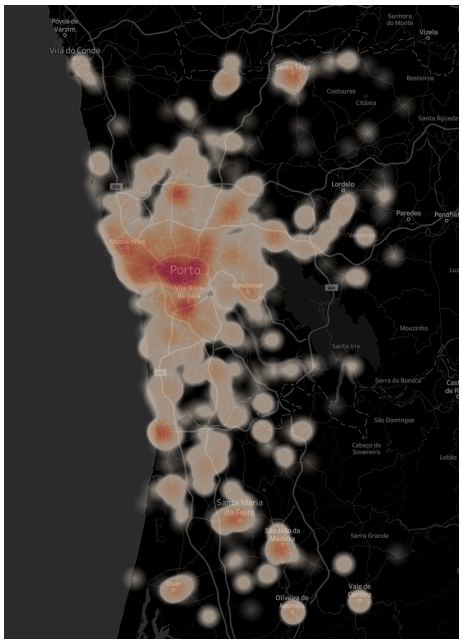


Figure 4.6: Healthcare

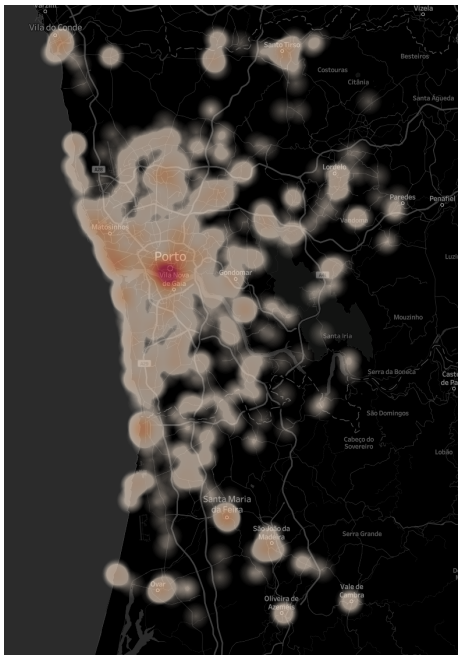


Figure 4.7: Leisure

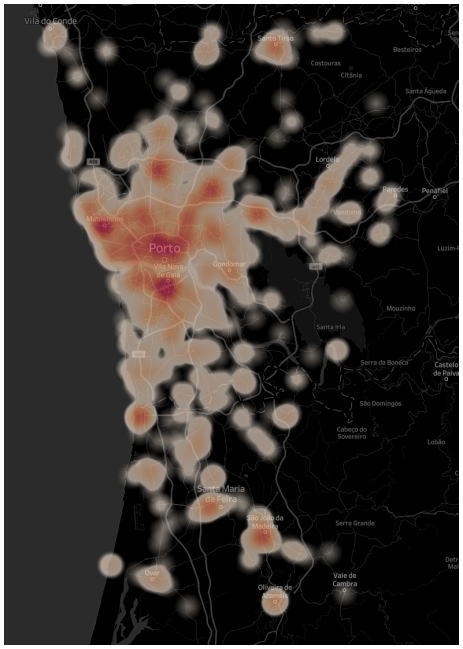


Figure 4.8: Lifestyle

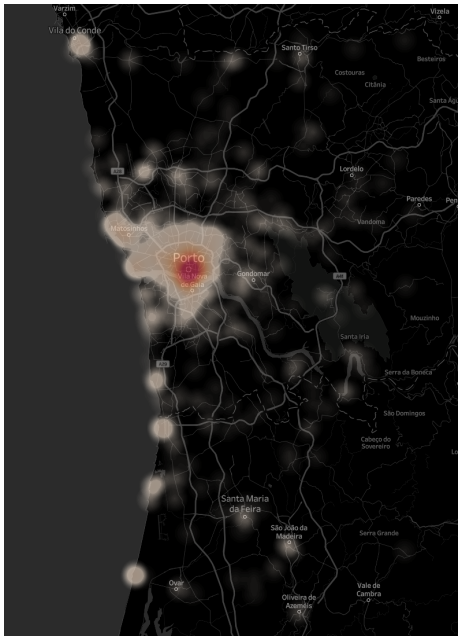


Figure 4.9: Lodging

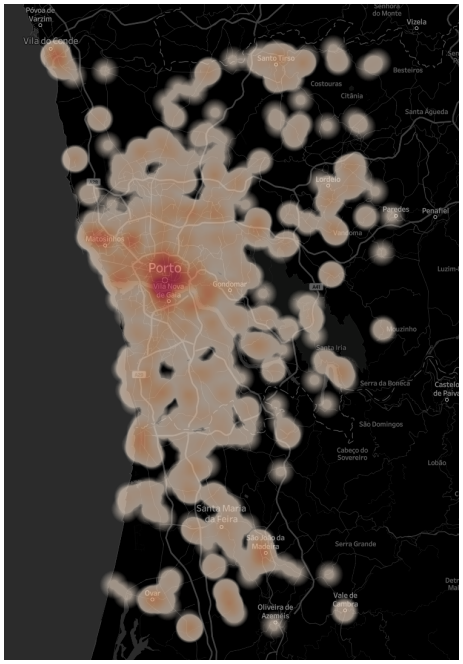


Figure 4.10: Religion

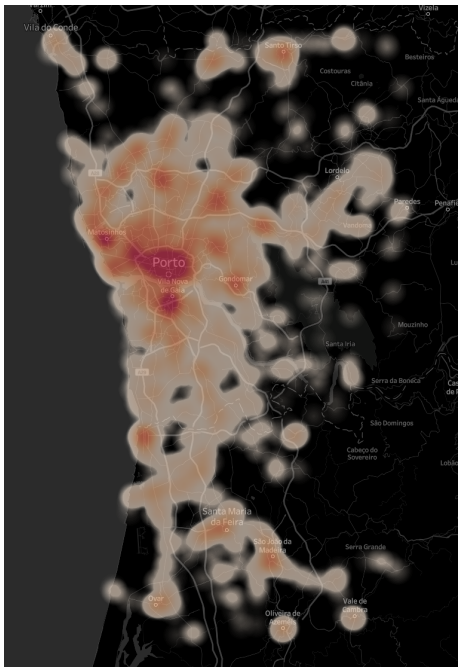


Figure 4.11: Services

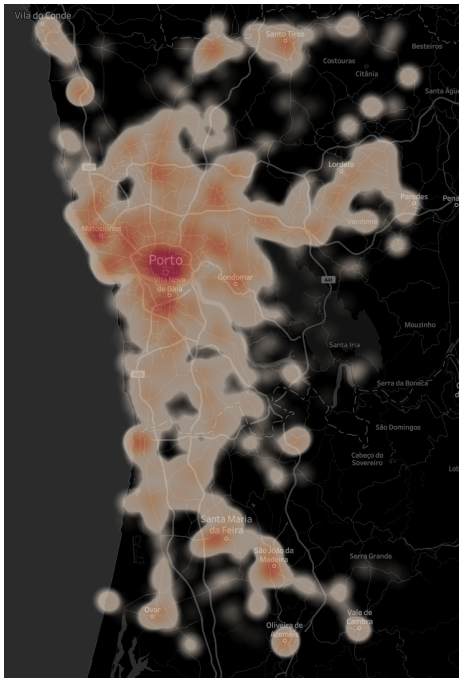


Figure 4.12: Shopping

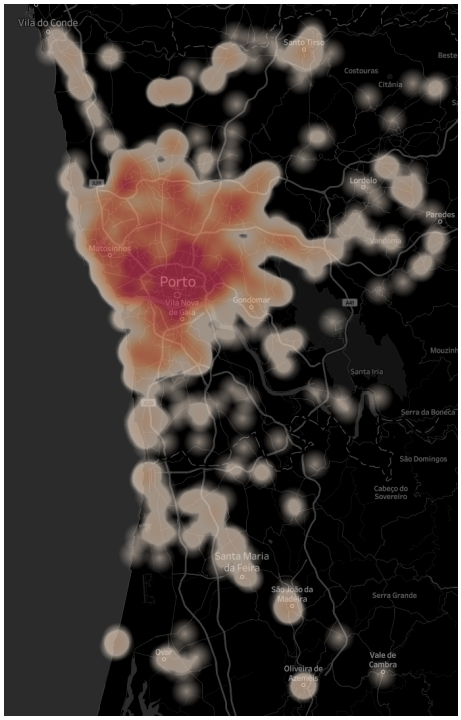


Figure 4.13: Transportation

All map data comes from Mapbox:OpenStreetMaps.

The previous Figures 4.4 to 4.13 shows the density of the POIs on the PMA by type.

- Fig. 4.4 shows us the distribution of the Education POIs, as it can be seen they are distributed all around the Porto Area with the biggest concentration being in the big cities of the PMA;
- Fig. 4.5, the restaurants and cafés are mostly distributed in the city centre of Porto, with other hotspots in areas like Espinho and Matosinhos, areas well known for their seafood restaurants;
- Fig. 4.6 Hospitals are located in big populated areas;
- Fig. 4.7 Leisure activities, having the biggest concentration in Porto city centre, can be found all around the PMA;
- Fig. 4.8 Lifestyle activities have a very similar concentration as the leisure ones;
- Fig. 4.9 The city centre of Porto contains the most lodging places, the area around the city centre also has them but to a less extent;
- Fig. 4.10 Churches and Cemeteries are scattered around the PMA. The Porto historical centre has the most concentration of these places;
- Fig. 4.11 Most services are located yet again in the centre of Porto with others being scattered around the PMA;
- Fig. 4.12 Shopping has a similar distribution as the services;
- Fig. 4.13 The transportation map shows a lot of POIs because it includes every station of every type in the PMA. Porto City centre and its surrounding areas have the most concentration of transportation.

4.2 Rule-Based System

The nature of this work required the creation of a rule-based system based on assumptions to achieve the end goal of estimating the trip purpose. These assumptions were sustained on the support tables provided; the study of the AFC Data; using other studies that were analysed during the Literature Review; and the reality of the Portuguese Lifestyle.

The following sections are ordered by the script running order.

The initial classification to every trip which was not the 'LastTrip' ('LastTrip' is always the last **'ExitStation'** of the day), due to the fact that Andante's AFC System being entrance-only, utilized the logic referred in the previous section of using the relative frequency to define the stations. Every other followed this logic:

4.2.1 Health Care and Shopping

Trips are classified as healthcare if the user has alighted in a station whose name includes 'HOSP' in its name or as alighted in the 'IPO' station.

```
VAL_res['purpose']=np.where((VAL_res.ExitStation.str.contains('HOSP')) | (VAL_res['
    ExitStation'] =='IPO') & (VAL_res['TravelTime'] < dtime.timedelta(0,0,0,0,0,4))
    ,'healthcare',VAL_res['purpose'])
```

Listing 4.2: Code that sets 'purpose' variable as "hospitals"

Trips are classified as shopping if the user has alighted in a station whose name includes 'SHOP' in its name.

```
VAL_res['purpose']=np.where(VAL_res.ExitStation.str.contains('SHOP'),'shopping',
    VAL_res['purpose'])
```

Listing 4.3: Code that sets 'purpose' variable as "shopping"

4.2.2 Work trips

Work-related trips were defined using the following rules:

- User needs to be a regular, not a student, not a retiree and not a tourist;
- The trip needs to be performed during a weekday; (Mon-Fri);
- The duration between trips must be 6 or more hours, normally a workday in Portugal is 8 hours, [37]
- The validation must occur between 6 to 11 in the morning as that is the interval of the peak hours, Figs: 4.14 and 4.15.

As it can be seen by the graphs below there is a peak of validations between 6 in the morning to 11, those are the hours for which if validation has occured and if it has a duration of 6 or more hours it's considered a 'work' trip.

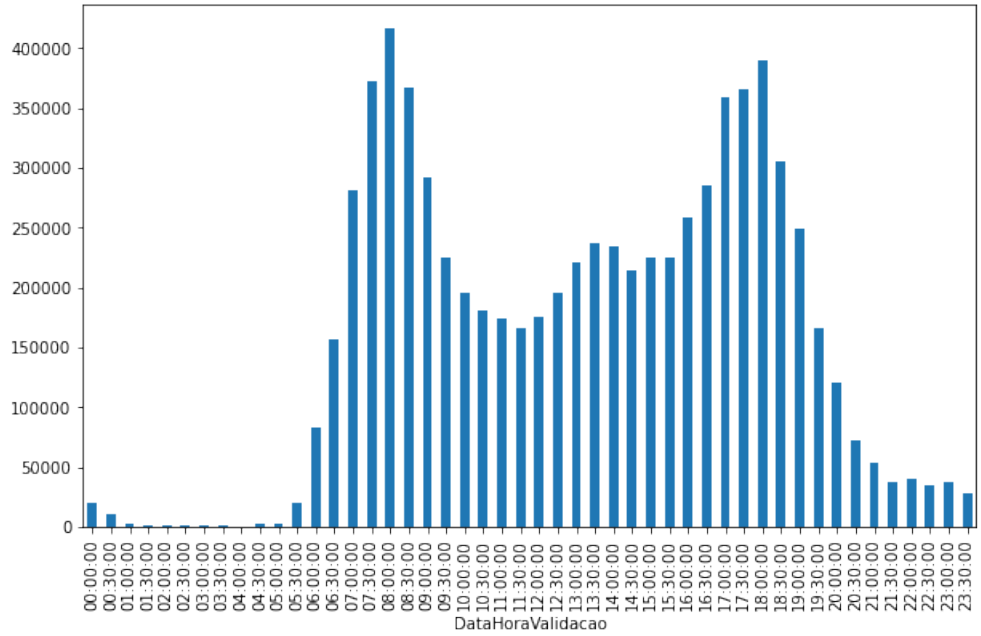


Figure 4.14: Worker's Validations Bars - October 2019 - 30 min periods

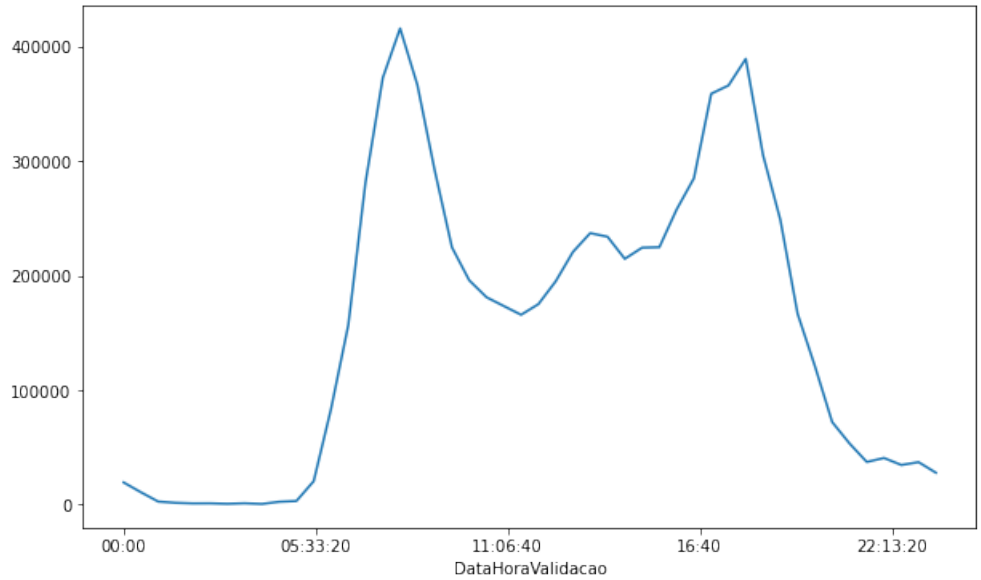


Figure 4.15: Worker's Validations Line - October 2019

```
VAL_res['purpose']=np.where((VAL_res['IsRegular']==1) & (VAL_res['IsStudent']==0) &
    (VAL_res['IsWeekDay'] == True) & (VAL_res['IsRetired'] == 0) & (VAL_res['
   IsTourist'] == 0) & (VAL_res['TravelTime'] > dtime.timedelta(0,0,0,0,0,6)) & ((
   VAL_res.DataHoraValidacao.dt.hour >= 5) & (VAL_res.DataHoraValidacao.dt.hour <=
    11)),'work',VAL_res['purpose'])
```

Listing 4.4: Code that sets 'purpose' variable as "work"

4.2.3 Education trips

Education-related trips were defined using the following rules:

- User needs to be a student, not a tourist and not a retiree;
- The trip needs to be performed during a weekday; (Mon-Fri);
- The duration between trips must be 4 or more hours, taking into account the usual travel time and classes duration in Portugal;
- The validation must occur between 6 in the morning to 11 which are the peak morning hours interval, Figs: 4.16 and 4.17.

As it can be seen by the graphs below there is a peak of validations between 6 in the morning to 11, those are the hours for which if validation has occured and if it has a duration of 4 or more hours it's considered an 'education' trip.

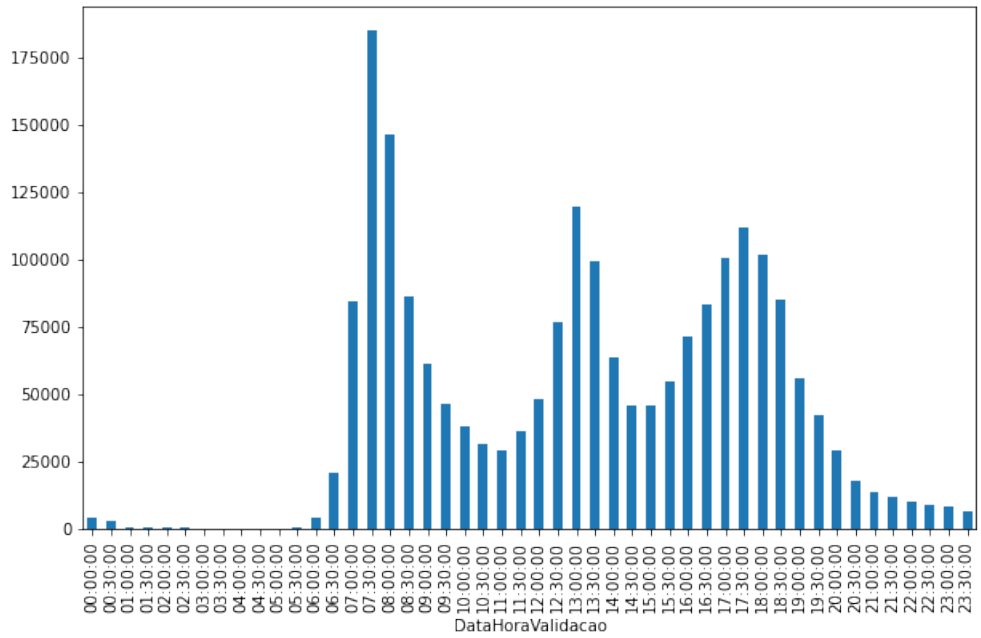


Figure 4.16: Student's Validations Bars - October, 2019 - 30 min periods

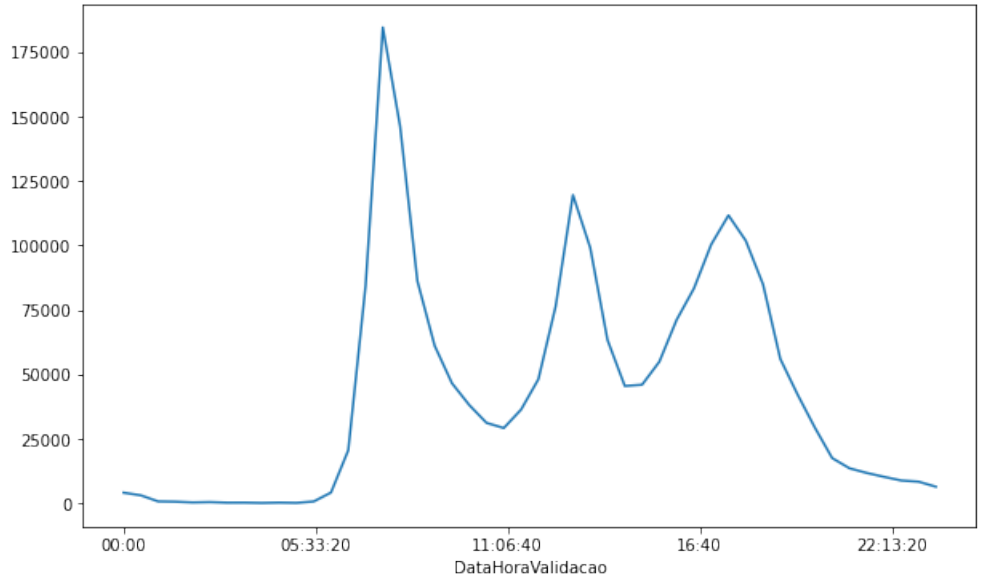


Figure 4.17: Student's Validations Line - October, 2019

```
1 VAL_res['purpose']=np.where((VAL_res['IsStudent']==1) & (VAL_res['IsWeekDay'] ==
      True) & (VAL_res['IsRetired'] == 0) & (VAL_res['IsTourist'] == 0) & (VAL_res['
      TravelTime'] > dtime.timedelta(0,0,0,0,0,4)) & ((VAL_res.DataHoraValidacao.dt.
      hour >= 6) & (VAL_res.DataHoraValidacao.dt.hour <= 11)),'education',VAL_res['
      purpose'])
```

Listing 4.5: Code that sets 'purpose' variable as "education"

4.2.4 Home trips

Home trips are classified as the same for workers and students.

By analysing and comparing the graph: 4.18 to 4.21, it was decided that it would be considered ' Home' Trips every last trip of the day from 16h, where the difference in the line is noticeable, until the end of the day, weekday. On weekends and on every ticket, except tourists, that the **'ExitStation'** is "LastTrip" it is classified as 'home' trip from 18h onwards.

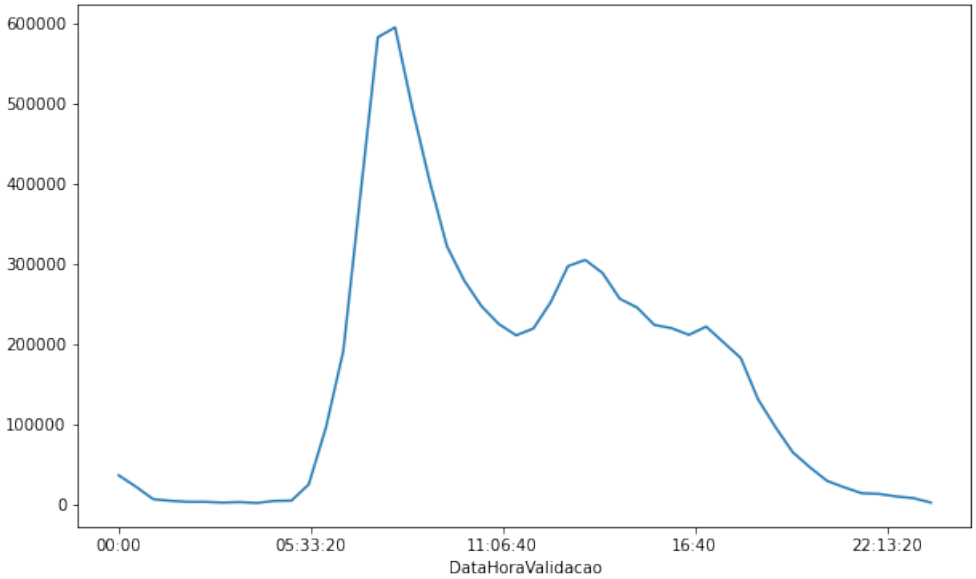


Figure 4.18: Without 'LastTrip' Validations Line - October, 2019

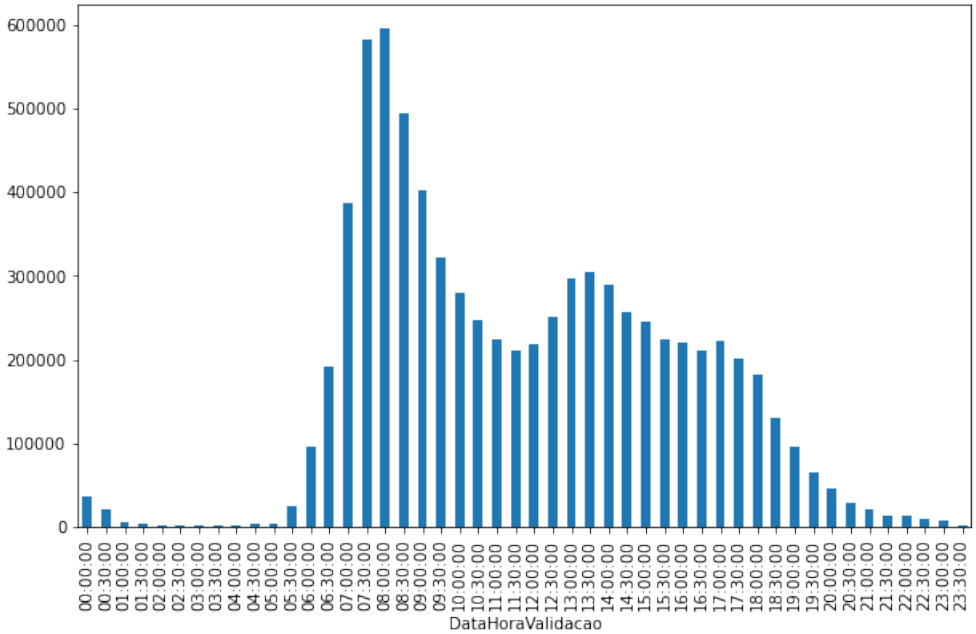


Figure 4.19: Without 'LastTrip' Validations Bars - October, 2019

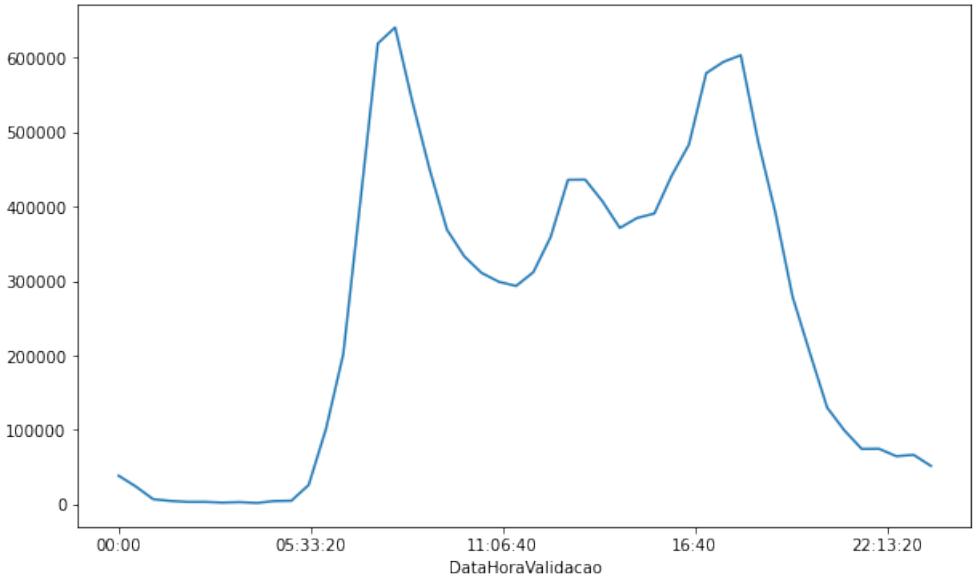


Figure 4.20: Validations Line - October, 2019

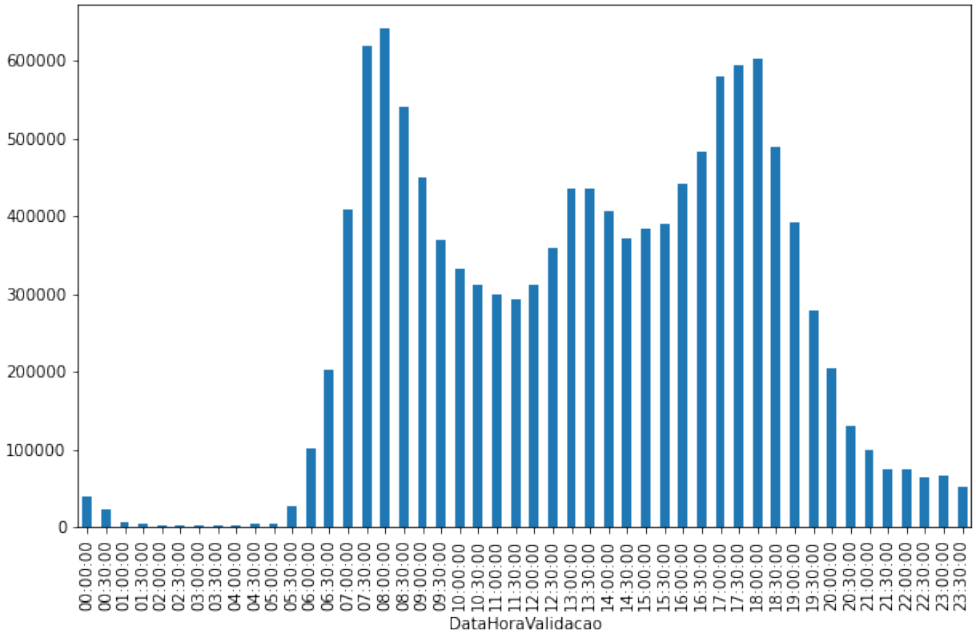


Figure 4.21: Validations Bars - October, 2019

Students Code:

```
VAL_res['purpose']=np.where( (VAL_res['IsStudent']==1) & (VAL_res['IsWeekDay'] ==
    True) & (VAL_res['IsRetired'] == 0) & (VAL_res['IsTourist'] == 0) & ((VAL_res.
    DataHoraValidacao.dt.hour >= 16) & (VAL_res.DataHoraValidacao.dt.hour <= 23)) &
     (VAL_res['ExitStation'] == 'LastTrip'),'home',VAL_res['purpose'])
```

Listing 4.6: Code that sets 'purpose' variable as "home" for students

Workers Code:

```
VAL_res['purpose']=np.where((VAL_res['IsRegular']==1) & (VAL_res['IsStudent']==0) &
     (VAL_res['IsWeekDay'] == True) & (VAL_res['IsRetired'] == 0) & (VAL_res['
    IsTourist'] == 0) & ((VAL_res.DataHoraValidacao.dt.hour >= 16) & (VAL_res.
    DataHoraValidacao.dt.hour <= 23)) & (VAL_res['ExitStation'] == 'LastTrip'),'
    home',VAL_res['purpose'])
```

Listing 4.7: Code that sets 'purpose' variable as "home" for workers

4.2.5 Food trips

Food trips classification follows the same logic for workers, students and tourists.

As seen in graph, 4.20, there is a peak between 12h to 14h, every journey between that peak is classified as a 'food' trip.

Students Code:

```
VAL_res['purpose']=np.where((VAL_res['IsStudent']==1)  & (VAL_res['IsRetired'] ==
    0) & (VAL_res['IsTourist'] == 0) & ((VAL_res.DataHoraValidacao.dt.hour >= 12) &
     (VAL_res.DataHoraValidacao.dt.hour <= 14)),'food',VAL_res['purpose'])
```

Listing 4.8: Code that sets 'purpose' variable as "food" for students

Workers Code:

```
VAL_res['purpose']=np.where((VAL_res['IsRegular']==1) & (VAL_res['IsStudent']==0) &
     (VAL_res['IsRetired'] == 0) & (VAL_res['IsTourist'] == 0) & ((VAL_res.
    DataHoraValidacao.dt.hour >= 12) & (VAL_res.DataHoraValidacao.dt.hour <= 14)),'
    food',VAL_res['purpose'])
```

Listing 4.9: Code that sets 'purpose' variable as "food" for workers

With tourists, trips from 19h to 21h were also considered as' food' related.
Tourists Code:

```
VAL_res['purpose']=np.where((VAL_res['IsTourist'] == 1) & ((VAL_res.
    DataHoraValidacao.dt.hour >= 12) & (VAL_res.DataHoraValidacao.dt.hour <= 14)),'
    food',VAL_res['purpose'])
VAL_res['purpose']=np.where((VAL_res['IsTourist'] == 1) & ((VAL_res.
    DataHoraValidacao.dt.hour >= 19) & (VAL_res.DataHoraValidacao.dt.hour <= 21)),'
    food',VAL_res['purpose'])
```

Listing 4.10: Code that sets 'purpose' variable as "food" for workers

4.2.6 Tourists

Tourists use the same logic as 'home' trips to be classified as 'lodging' except that they do not have time restrictions.

```
VAL_res['purpose']=np.where((VAL_res['IsTourist'] == 1) & (VAL_res['ExitStation']
    == 'LastTrip'),'lodging',VAL_res['purpose'])
```

Listing 4.11: Code that sets 'purpose' variable as "lodging" for tourists

If the tourists' entrance station is the airport it is also classified as lodging:

```
VAL_res['purpose']=np.where((VAL_res['IsTourist'] == 1)  & (VAL_res.EntranceStation
    .str.contains('AERO')) ,'lodging',VAL_res['purpose']))
```

Listing 4.12: Code that sets 'purpose' variable as "lodging" for tourists if 'EntranceStation' is the airport

Every other journey of tourists is classified as "Leisure and Culture" and if the trip occurred between 12h to 14h and 19h to 21h it is classified as 'food' related as previously mentioned.

4.3 Final Dataset

In the end, the final dataset had the following variables with around 13.4 million validations:

Name	Type
Code	Integer
Card	String
Discount	Integer
TarifCode	Integer
DataHoraValidacao	DateTime
EntranceStation	String
ExitStation	String
IsTourist	Integer
IsRegular	Integer
IsStudent	Integer
IsRetired	Integer
TimeinLocation	DeltaTime
IsWeekDay	Integer
Purpose	String

Table 4.10: Variables of the final dataset

4.4 Technologies

To this work data is everything, so it was needed to comprehend the data to make information out of it, without understanding it data is meaningless. The software that was to be chosen had to facilitate the data analysis and simplify it as much as possible, and possess a graphical user interface (GUI). Tableau was chosen for that matter, when information was created from the data it was important to study it more deeply, to do that Python using Panda's library was the elected candidate. RapidMiner was used to develop the estimation model without having to perform the coding phase.

- **Tableau**: Tableau is a data visualisation software developed by Tableau Software that lets you see and understand data easily with its GUI. This eliminates the coding process from the data understanding process which is time-consuming. The current stable Tableau version is: 2022.2

- **Python**: Python is a high-level, general-purpose, interpreted programming language designed by Guido van Rossum in the late 1980s and first released in 1991. It is used by a lot of Data Analysts and is well established in the community due to the vast libraries available for that purpose, one of the most well-known is Pandas, which is the main library used in this work. Python is more powerful than Tableau when advanced data transformations are required, in this project it is used to deeply study the information we created with Tableau and to make the estimation model. The current stable Python version is: 3.10.

- **RapidMiner**: RapidMiner is a data science software which allows us to do data mining tasks without doing the coding. Rapidminer was useful because it removes the coding process of the model generation, even though it's not an extensive coding process it saved time. The current stable RapidMiner version is: 9.10.

Chapter 5

Results and Discussions

This chapter initially will discuss the generation of the estimation model and why it was decided to use that determined classification algorithm. And will end by presenting the results in a summarized table.

5.1 The Estimation Models

The choice of what classification algorithms methods were to be used was based on the research done in the second chapter, for example, Ermagun et al work [14](Decision Trees and others), Lee and Hickman [26], [9],.. and on some later searching on various webpages like JavatPoint [23] and GeeksforGeeks [15] [16].

It also takes into account experience gained taking a class, *Análise de Dados e Engenharia de Software*, while taking this master's degree at FEUP.

To develop the models RapidMiner was used as previously mentioned it speeds up the process by taking away the coding phase.

Every process developed in RapidMiner followed the same logic as seen in Fig: 5.1, only changing the Learning Algorithm used on the part of the validation process seen in Fig: 5.2.

For every learning algorithm learning phase, the following data was selected from the final dataset, Table. 4.10:

Name	Type
Discount	Integer
TarifCode	Integer
DataHoraValidacao	DateTime
ExitStation	Integer
TimeinLocation	Hour
IsWeekDay	Integer
Purpose	String

Table 5.1: Variables fed into model generation

Due to RapidMiner's limitations, 'TravelTime' variable was converted to Hours format again, this was the first phase of the process, and then 'Stations' variable was transformed into a numeral form which improves the performance significantly, this is one process of feature engineering. Lastly, the validation phase which splits the dataset into 70% for training and 30% for testing.

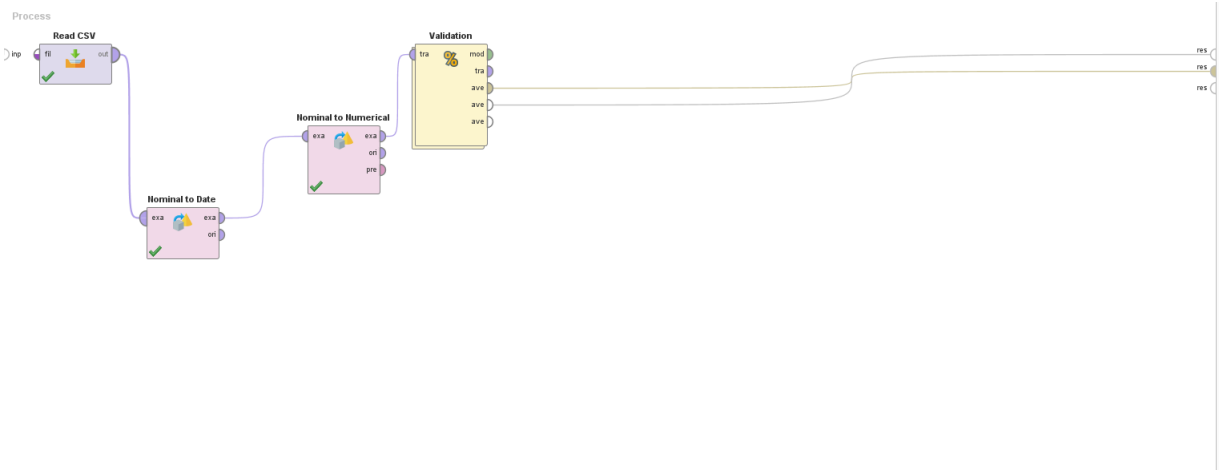


Figure 5.1: RapidMiner Main Phase - Common Phase

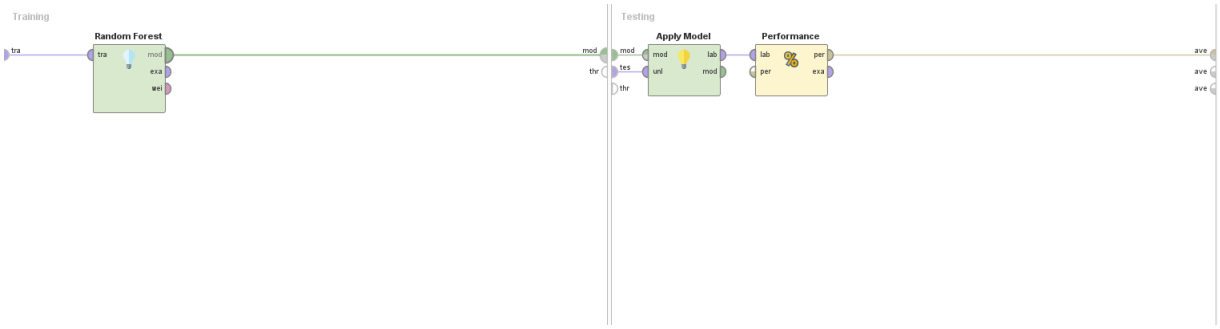


Figure 5.2: RapidMiner Validation Phase - Random Forest Example

Validation

As mentioned above the validation phase separates randomly the data into a training set and a test set. The training set has 70% of the data and it's in that set that the model is going to be taught and developed. The other 30% serves to test the model, the model is applied to the test data and the results will be shown in a confusion matrix.

The following table shows the confusion matrix for Decision Tree Model with Bagging:

	true religion	true home	true lodging	true transportation	true other	true education	true food	true work	true healthcare	true lifestyle_and_beauty	true leisure_and_culture	true shopping	true service	class precision
pred. religion	43031	0	0	0	0	552	7948	9	0	1	0	1	0	**83.49%**
pred. home	0	854914	0	0	205046	0	122108	0	0	0	0	0	0	**72.32%**
pred. lodging	1471	0	238325	12975	0	3714	54385	33	10586	9382	841	1723	627	**71.34%**
pred. transportation	223	0	881	32820	0	577	5565	51	802	758	60	185	50	**78.19%**
pred. other	0	26827	44	0	150843	0	9261	0	0	0	0	0	0	**80.68%**
pred. education	207	0	1863	1212	0	136209	15877	46	1490	557	33	64	17	**86.44%**
pred. food	34933	0	277904	274374	10	81219	416380	3	170406	146960	19738	26230	8307	**28.59%**
pred. work	673	0	4262	3449	0	636	29805	428089	2234	2361	101	431	117	**90.67%**
pred. healthcare	1	0	18	9	0	267	20110	2	88533	4	15	2	2	**81.25%**
pred. lifestyle_and_beauty	2	0	10	13	0	111	7728	6	6	26951	55	0	0	**77.26%**
pred. leisure_and_culture	0	0	0	12	0	1	2452	0	0	0	6770	0	0	**73.31%**
pred. shopping	0	0	1	0	0	0	0	0	0	1	0	3	0	**60.00%**
pred. service	0	0	0	0	0	0	0	0	0	0	0	0	0	**0.00%**
class recall	**53.43%**	**96.96%**	**45.54%**	**10.10%**	**42.38%**	**61.00%**	**60.20%**	**99.96%**	**32.30%**	**14.41%**	**24.52%**	**0.01%**	**0.00%**	

Table 5.2: Confusion Matrix for Decision Tree Model with Bagging

All the confusion matrixes obtained for the developed models of this research can be found in Appendix B.

Accuracy

Accuracy is the percentage of correct predictions and it is calculated by dividing the number of correct predictions over the total cases:

$$\text{Accuracy} = \frac{\sum \text{CorrectPredictions}}{\text{TotalCases}} \tag{5.1}$$

In the case of the table, 7, the accuracy would be:

$$\text{Accuracy} = \frac{2422868}{4035901} = 0.600329 \simeq 60.03\% \tag{5.2}$$

5.1.1 Decision Trees

A Decision Tree,[35], is a supervised learning method used for classification and regression. It creates a model by learning simple rules inferred from the data features, in this case, the rule-based system created previously.

Bagging, [41], is used to improve the stability and accuracy of the classification. It reduces variance and helps to avoid overfitting.

This was the model, recurring to bagging, gave us the best results with **60,03%** of accuracy, see table: 5.3, for comparision and table: 7 for the confusion matrix. The same approach without recurring to bagging gave **53,64%** of accuracy.

Random forests,[9], is an ensemble learning approach, that extends bagging, where a collection of decision trees are built on the training data, one of its important features is that it can handle categorical variables as it is in the case of classification problems. This approach of using Random Forests gave an accuracy of **58,12%** with 100 trees and 200 trees gave an accuracy of **58,25%**.

Usually with Random Forest models increasing the number of trees raises the accuracy, with the data this dissertation was worked upon, as seen in Table. 5.3, 100 to 200 trees did not make much of a difference.

An higher number could not be tried due to hardware limitations that would cause RapidMiner to crash, so it's not possible to know if the results would improve even further with, for example, 300 trees.

This approach, Decision Trees, was the first choice since the branches of the tree represent the outcome of a test on an attribute,[15], which later resembles the rule-based system, as in for example in Fig. 5.3

The Estimation Process

During the generation of the decision tree,[17], the algorithm(decision tree), makes the splitting into branches to achieve a prediction, Fig.5.3, asking questions. This split is done by calculating the entropies of all the attributes and the one with the least entropy is selected for the split.

Here the Gain Ratio criterion, [30], is used. Gain Ration is a variant of the information gain criterion that adjusts the information gain for each Attribute to allow the breadth and uniformity of the attribute values.

Fig. 5.3 shows the tree zoomed into a final prediction zone.

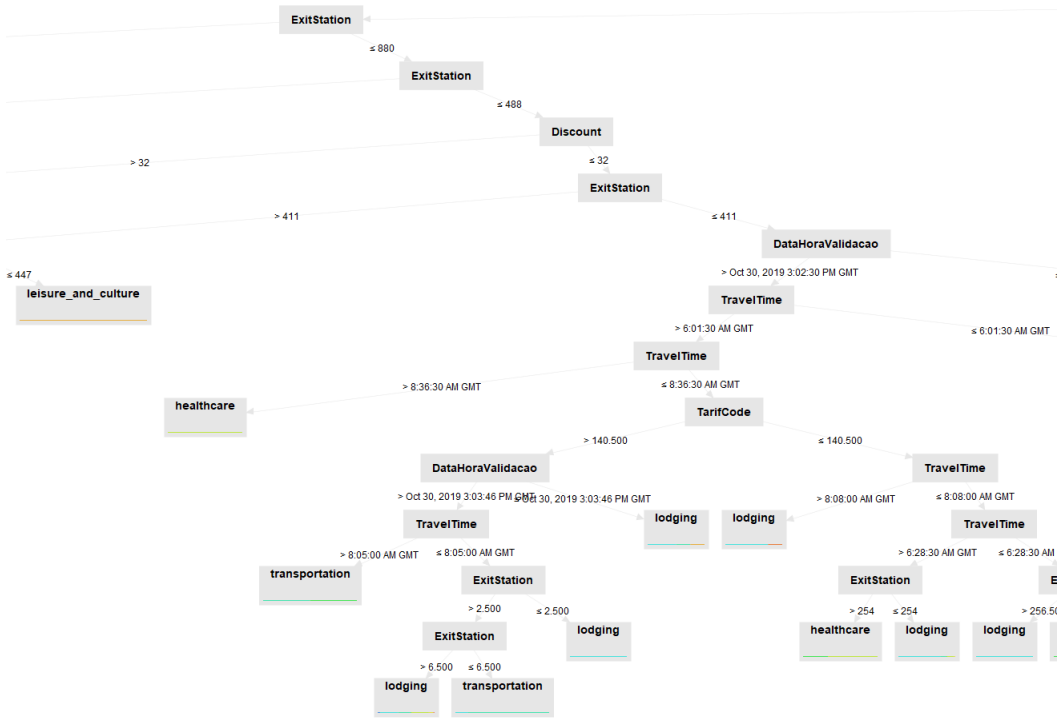


Figure 5.3: Example Decision Tree using the Data - Zoomed

The process is the same for Random Forests and Bagging, with the nuance that in Random Forests,[32] a majority vote is taken after this process and from the number of selected trees it selects the prediction that appears more times, hence "majority vote".

5.1.2 Deep Learning

The Deep Learning algorithm works in a way that teaches computers to learn as humans do, learn by examples. Deep Learning is, for example, behind the technology that teaches cars to drive. Deep Learning is essencially, according to IBM,[22], a neural network with three or more layers.

Deep Learning is an algorithm that requires a large amount of data to perform best.

This approach gave an accuracy of **54.52%**.

5.1.3 Naïve Bayes

The Encyclopedia of Machine Learning,[28], defines Naïve Bayes as a supervised learning algorithm that uses the Bayes Theorem, Eq. 5.3.

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)} \tag{5.3}$$

According to Javatpoint,[23], the Naïve Bayes algorithm is mostly used in text classification, so as this is one of the simplest and most effective classification algorithms and the target variable on the final dataset is text it was also taken into account to generate a model. As mentioned it uses the Bayes Rule so it's a probabilistic classifier -> it predicts on the basis of the probability of an object.

Uses of the Naïve Bayes algorithm:

- Detection of Sending and Posting Advertisement in Mass (SPAM);
- Filtering mature words;
- Article Classification;
- ...;

The use of Naïve Bayes in this type of data proved not to be useful as it gave the lowest results of the model sets, with **52,90%** of accuracy.

5.1.4 Neural Networks

Neural Network, [40], or artificial neural network is a model of neurons inspired by the human brain. It is made up of many neurons that are connected to each other to simulate the synapsis.

They differ from Deep Learning on the basis of their depth and number of hidden layers, [16].

This approach gave an accuracy of **53.92%**.

Uses of the Neural Networks:

- Medical Diagnosis;
- Face Detection;
- Voice Recognition;
- ...;

The Estimation Process

Neural Networks,[29] algorithm have a layered structure, Fig. 5.4. The input layer receives a signal which passes through an arbitrary number of hidden layers and in the end an output layer that makes a decision or prediction about the input. Each layer is made of a set of neurons (the circles in the image), and neurons within individual layers are connected to adjacent layers. A Neural Network is Deep when it has lots of layers.

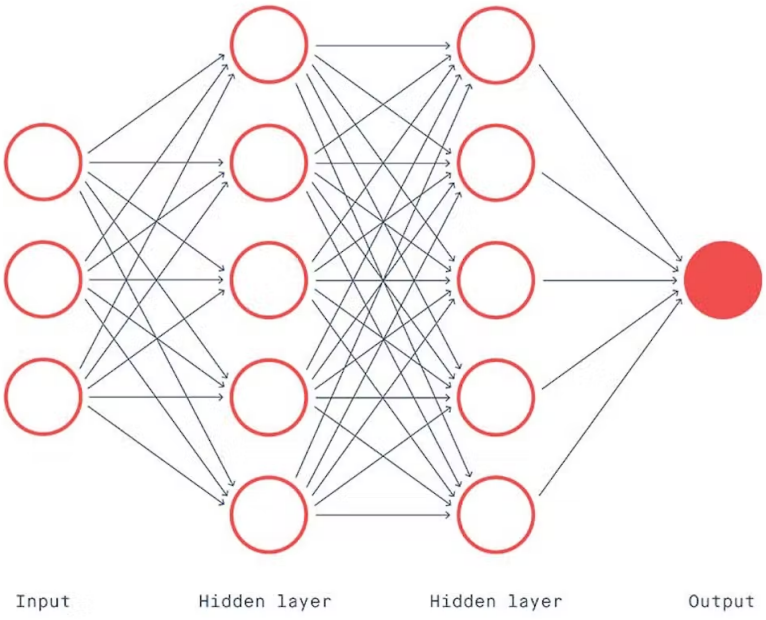


Figure 5.4: Neural Network Example,[29]

5.1.5 Comparisions

The following table presents in a summarised form the performance (accuracy) results of the developed estimation models:

Algorithm	**Accuracy**	**Number of**	**Details**
Deep Learning	54,52%	10 epochs	Default RapidMiner Values
Deep Learning	54.31%	200 epochs	Default RapidMiner Values
Deep Learning	54,44%	500 epochs	Default RapidMiner Values
Decision Tree	53,64%	N/A	Default RapidMiner Values
Decision Tree	**60,03%**	**10 iterations (of Bagging)**	**Bagging with Pre-prunning**
Naïve Bayes	52,90%	N/A	Default RapidMiner Values
Neural Networks	53,92%	200 training cycles	Default RapidMiner Values
Random Forest	58,12%	100 trees	Pre-Prunning
Random Forest	58,25%	200 trees	Pre-Prunning

Table 5.3: Model Comparison Table - Accuracy Results

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This research aimed to develop a model to estimate the trip purpose of users' in the Porto Metropolitan Area (PMA) using and crossing the data from Andante's AFC network with the PMA POIs dataset.

By analysing each user's trip characteristics and the characterization of the alighting station with a lot of work and empirical decisions it was possible to estimate the trip purpose and with that develop the model.

However, it is important to remember that Porto's Andante Network is an entrance-only AFC system so the data would be much more accurate if surveys were done to ensure the rule-based system had theoretical data to fall upon rather than only empirical assumptions on the Portuguese lifestyle and based on studies from other previous estimation works from other countries.

Having the trip purpose estimated the next step was to develop the classification model, so several algorithms that could use polynomial variables as a target were identified and tested.

Finally, from all the models developed, for the data used, a decision tree with bagging was identified as the one with the best results. However, better results on this or in the other models would probably have been able to be achieved if all of the processes were developed in a more powerful computing environment.

Even though results could be better, the main objective which was to estimate trip purpose by combining AFC Data with the POIs source was achieved, and the generated models were able to identify trip purposes in the AFC Data from the Andante Network with a certain level of accuracy. This was never done in the PMA and the data may be useful in the future to apply changes to the policies and offers of the public transports in Porto.

6.2 Future Work

Future works in this area of research could be greatly improved. The first improvement would be being able to have surveys available to have a more theoretical approach rather than an empirical

one to estimate the trip purposes. The method used here works but it would be much more accurate if it had a survey to fall its assumptions upon. Another would be using more powerful hardware so the models could be generated quicker and with more iterations than the ones possible in this research.

References

[1] A. Ali and S. Lee. *Destination and activity estimation*. 2017.

[2] Atizaz Ali, Jooyoung Kim, and Seungjae Lee. Travel behavior analysis using smart card data. *KSCE Journal of Civil Engineering*, 20:1532–1539, 5 2016.

[3] Azalden Alsger, Ahmad Tavassoli, Mahmoud Mesbah, Luis Ferreira, and Mark Hickman. Public transport trip purpose inference using smart card fare data. *Transportation Research Part C: Emerging Technologies*, 87:123–137, 2 2018.

[4] Amazon Web Services. Feature engineering - machine learning lens, 2022. Available at: https://docs.aws.amazon.com/wellarchitected/latest/machine-learning-lens/feature-engineering.html Accessed on 6 August 2022.

[5] AMPorto. Amporto, 2019. Available at: http://portal.amp.pt/pt/noticia/481 Accessed on 6 March 2022.

[6] Andante. Títulos de transporte, n.d. Available at: https://linhandante.com/en/web/tip/titulos-de-transporte , Accessed on 11 March 2022.

[7] M. Bagchi and P. R. White. The potential of public transport smart card data. *Transport Policy*, 12:464–474, 9 2005.

[8] James J. Barry, Robert Newhouser, Adam Rahbee, and Shermeen Sayeda. Origin and destination estimation in new york city with automated fare system data:. *https://doi.org/10.3141/1817-24*, pages 183–187, 1 2002.

[9] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[10] Y. Chen, Z. Jin, and C. Li. Trip purpose prediction based on hidden markov model with gps and land use data. pages 55–59, 2020.

[11] Roger Clarke. Person location and person tracking - technologies, risks and policy implications. *Information Technology People*, 14:206–231, 6 2001.

[12] Yu Cui, Chuishi Meng, Qing He, and Jing Gao. Forecasting current and next trip purpose with social media data and google places. *Transportation Research Part C: Emerging Technologies*, 97:159–174, 12 2018.

[13] Flavio Devillaine, Marcela Munizaga, and Martin Trépanier. Detection of activities of public transport users by analyzing smart card data:. *https://doi.org/10.3141/2276-06*, 2276:48–55, 1 2012.

[14] Alireza Ermagun, Yingling Fan, Julian Wolfson, Gediminas Adomavicius, and Kirti Das. Real-time trip purpose prediction using online location-based search and discovery services. *Transportation Research Part C: Emerging Technologies*, 77:96–112, 4 2017.

[15] GeeksforGeeks. Decision tree. Available at: https://www.geeksforgeeks.org/decision-tree/ Accessed on 22 September 2022.

[16] GeeksforGeeks. Difference between a neural network and a deep learning system, 2022. Available at: https://www.geeksforgeeks.org/difference-between-a-neural-network-and-a-deep-learning-system/ Accessed on 22 September 2022.

[17] GitHub. Pythondatasciencehandbook/05.08-random-forests.ipynb at master · jakevdp/pythondatasciencehandbook · github, 2018. Available at: https://github.com/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/05.08-Random-Forests.ipynb Accessed on 2 September 2022.

[18] GitHub. Github - technicalvillager/distance-bearing-calculation: Calculating distance and bearing between two co-ordinates, 2019. Available at: https://github.com/TechnicalVillager/distance-bearing-calculation Accessed on 20 June 2022.

[19] João Manuel Angélico Gonçalves. Understanding urban mobility patterns of tourists through data extraction, 7 2021.

[20] Google. Overview | places api | google developers, 2022. Available at: https://developers.google.com/maps/documentation/places/web-service/overview Accessed on 24 January 2022.

[21] Nick Hotz. What is crisp dm? - data science process alliance, 2022. Available at: https://www.datascience-pm.com/crisp-dm-2/ Accessed on 26 July 2022.

[22] IBM Cloud Education. What is deep learning?, 2020. Available at: https://www.ibm.com/cloud/learn/deep-learning Accessed on 22 September 2022.

[23] Javatpoint. Naive bayes classifier in machine learning, n.d. Available at: https://www.javatpoint.com/machine-learning-naive-bayes-classifier Accessed on 6 September 2022.

[24] Kaggle. Learn feature engineering tutorials, n.d. Available at: https://www.kaggle.com/learn/feature-engineering Accessed on 6 August 2022.

[25] Takahiko Kusakabe and Yasuo Asakura. Behavioural data mining of transit smart card data: A data fusion approach. *Transportation Research Part C: Emerging Technologies*, 46:179–191, 9 2014.

[26] Sang Gu Lee and Mark Hickman. Trip purpose inference using automated fare collection data. *Public Transport*, 6:1–20, 4 2014.

[27] K. Lu, A. Khani, and B. Han. A trip purpose-based data-driven alighting station choice model using transit smart card data. *Complexity*, 2018, 2018.

[28] Paul Munro, Hannu Toivonen, Geoffrey I. Webb, Wray Buntine, Peter Orbanz, Yee Whye Teh, Pascal Poupart, Claude Sammut, Caude Sammut, Hendrik Blockeel, Dev Rajnarayan, David Wolpert, Wulfram Gerstner, C. David Page, Sriraam Natarajan, and Geoffrey Hinton. *Bayes Rule*. Springer US, 2011.

[29] Peltarion. How does deep learning algorithms work?, n.d. Available at: https://peltarion.com/deep-learning/deep-learning-algorithms Accessed on 6 September 2022.

[30] Tung M Phung. Information gain, gain ratio and gini index, 2021. Available at: https://tungmphung.com/information-gain-gain-ratio-and-gini-index/ Accessed on 6 September 2022.

[31] ProjectPro. 8 feature engineering techniques for machine learning, 2022. Available at: https://www.projectpro.io/article/8-feature-engineering-techniques-for-machine-learning/423 Accessed on 6 August 2022.

[32] RapidMiner. Random forest - rapidminer documentation, n.d. Available at: https://docs.rapidminer.com/9.10/studio/operators/modeling/predictive/trees/parallel_random_forest.html Accessed on 6 September 2022.

[33] Red Hat. What is a rest api?, 2020. Available at: https://www.redhat.com/en/topics/api/what-is-a-rest-api Accessed on 14 January 2022.

[34] République Française. Délibération n° 2003-38 du 16 septembre 2003 portant adoption d'une recommandation relative à la collecte et au traitement d'informations nominatives par les sociétés de transports collectifs dans le cadre d'applications billettiques - légifrance, 9 2003. Available at: https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000245108 Accessed on 14 January 2022.

[35] Scikit Learn. 1.10. decision trees — scikit-learn 1.1.2 documentation, n.d. Available at: https://scikit-learn.org/stable/modules/tree.html Accessed on 22 September 2022.

[36] Stack Overflow. Using google api to tell how busy a place is - stack overflow, 2020. Available at: https://stackoverflow.com/questions/60652129/using-google-api-to-tell-how-busy-a-place-is Accessed on 24 January 2022.

[37] SuperCasa. Portugal no top 10 de países onde se trabalham mais horas - supercasa, 2021. Available at: https://supercasa.pt/noticias/portugal-no-top-1-de-paises-onde-se-trabalham-mais-horas/n513 Accessed on 18 August 2022.

[38] TIBCO Software. What is data mining?, n.d. Available at: https://www.tibco.com/reference-center/what-is-data-mining Accessed on 11 January 2022.

[39] Martin Trepanier, Catherine Morency, and Carl Blanchette. Enhancing household travel surveys using smart card data. 2009.

[40] W3Schools. Neural networks, n.d. Available at: https://www.w3schools.com/ai/ai_neural_networks.asp Accessed on 22 September 2022.

[41] Wikipedia. Bootstrap aggregating. Available at: https://en.wikipedia.org/wiki/Bootstrap_aggregating Accessed on 22 September 2022.

[42] Wikipedia. Automated fare collection - wikipedia, n.d. Available at: https://en.wikipedia.org/wiki/Automated_fare_collection Accessed on 10 January 2022.

[43] Wikipedia. Data mining - wikipedia, n.d. Available at: https://en.wikipedia.org/wiki/Data_mining Accessed on 11 January 2022.

[44] Jinhua Zhao, Adam Rahbee, and Nigel H.M. Wilson. Estimating a rail passenger trip origin-destination matrix using automatic data collection systems. *Computer-Aided Civil and Infrastructure Engineering*, 22:376–387, 7 2007.

Appendix A

Support Tables

Codigo	Descricao
0	Normal
1	Reformado
2	3ª Idade
3	Estudante
4	Menores 13 anos
8	Estudante Mais
18	4_12_escola.tp
19	13_18_escola.tp
20	sub23@superior.tp
21	Social+
22	Social+ A
23	Social+ R
24	Social+ D
26	4_12_escola.tp A
27	13_18_escola.tp A
28	4_12_escola.tp B
29	13_18_escola.tp B
30	sub23_superior.pt A
31	4_12_escola.tp F
32	13_18_escola.tp F
33	sub23_superior.pt F
34	Sub13
35	Família

Table 1: Discount

Codigo	Descricao
1	Título de Viagem
2	Assinatura
6	Andante 24
12	Bilhete Euro - 1 Dia
13	Bilhete Euro - 3 Dias
19	Andante Tour-1
20	Andante Tour-3
140	Assinatura 3Z
141	Assinatura Municipal
142	Assinatura Metropolitana
152	Assinatura AMP/CP

Table 2: Tariff Code

Appendix B

Results Tables

	true religion	true home	true lodging	true transportation	true other	true education	true food	true work	true healthcare	true lifestyle_and_beauty	true leisure_and_culture	true shopping	true service	class precision
pred. religion	3418	0	33	0	0	0	0	0	0	0	0	0	0	**99.04%**
pred. home	552	836636	1860	420	257475	40	147953	0	515	18	0	0	0	**67.17%**
pred. lodging	33452	0	245262	68757	0	14652	94754	286	107161	55562	3479	8404	1919	**38.70%**
pred. transportation	22298	0	83398	141579	0	12402	86542	30	37885	39791	4805	4045	2570	**32.52%**
pred. other	178	45105	433	127	98424	14	18418	0	122	26	0	3	0	**60.44%**
pred. education	726	0	2699	1996	0	154145	26324	4345	1121	1048	120	146	43	**79.99%**
pred. food	18970	0	167301	105607	0	39074	270720	3843	104965	81199	8365	14828	4289	**33.05%**
pred. work	878	0	4819	3929	0	1593	28776	419735	2468	2692	112	453	132	**90.15%**
pred. healthcare	61	0	17419	2222	0	1356	12931	0	19774	6570	1	746	165	**32.29%**
pred. lifestyle_and_beauty	8	0	84	64	0	10	4	0	44	69	3	14	2	**22.85%**
pred. leisure_and_culture	0	0	0	163	0	0	5197	0	2	0	10728	0	0	**66.67%**
pred. shopping	0	0	0	0	0	0	0	0	0	0	0	0	0	**0.00%**
pred. service	0	0	0	0	0	0	0	0	0	0	0	0	0	**0.00%**
class recall	**4.24%**	**94.88%**	**46.87%**	**43.58%**	**27.66%**	**69.03%**	**39.14%**	**98.01%**	**7.22%**	**0.04%**	**38.85%**	**0.00%**	**0.00%**	

Table 3: Confusion Matrix for Deep Learning 10 epochs

	true religion	true home	true lodging	true transportation	true other	true education	true food	true work	true healthcare	true lifestyle_and_beauty	true leisure_and_culture	true shopping	true service	class precision
pred. religion	9921	0	3399	457	0	103	55	52	6	1	1	0	0	**70.89%**
pred. home	865	868230	2755	585	285919	53	158250	0	583	61	0	0	0	**65.91%**
pred. lodging	24321	0	289978	79247	0	16357	140586	223	149571	66688	3273	10855	2651	**37.00%**
pred. transportation	24046	0	82970	140326	0	13315	83291	31	37453	43486	5398	3920	2762	**32.11%**
pred. other	124	9381	245	76	58432	12	5255	0	60	17	0	1	0	**79.39%**
pred. education	1192	0	6183	4173	0	162155	35990	9730	3520	2487	334	349	112	**71.68%**
pred. food	19258	4130	130630	94937	11548	29942	232326	6288	77991	65332	6952	12320	3461	**33.42%**
pred. work	792	0	4718	3838	0	1267	27888	411915	2549	2670	139	462	127	**90.26%**
pred. healthcare	3	0	962	116	0	39	800	0	1479	306	0	27	5	**39.58%**
pred. lifestyle_and_beauty	13	0	1458	808	0	36	1546	0	833	5886	15	705	1	**52.08%**
pred. leisure_and_culture	0	0	0	233	0	0	5535	0	0	0	11423	0	0	**66.45%**
pred. shopping	0	0	0	0	0	0	0	0	0	0	0	0	0	**0.00%**
pred. service	6	0	10	68	0	7	97	0	12	41	78	0	1	**0.31%**
class recall	**12.32%**	**98.47%**	**55.41%**	**43.20%**	**16.42%**	**72.62%**	**33.59%**	**96.19%**	**0.54%**	**3.15%**	**41.37%**	**0.00%**	**0.01%**	

Table 4: Confusion Matrix for Deep Learning 200 epochs

	true religion	true home	true lodging	true transportation	true other	true education	true food	true work	true healthcare	true lifestyle_and_beauty	true leisure_and_culture	true shopping	true service	class precision
pred. religion	21	0	23	4	0	0	1	0	0	0	0	0	0	**42.86%**
pred. home	748	864884	2362	535	279526	55	156228	0	580	26	0	1	0	**66.28%**
pred. lodging	41325	0	269109	76131	0	15605	104361	282	124879	57496	3455	8285	2143	**38.28%**
pred. transportation	18808	0	79702	123369	0	9455	77719	152	36816	36475	3464	4614	2121	**31.42%**
pred. other	126	16857	353	91	76373	13	10148	0	93	72	0	0	0	**73.35%**
pred. education	734	0	2892	2112	0	151259	27426	5497	1235	1225	149	174	43	**78.48%**
pred. food	17956	0	159863	117511	0	40722	276060	4117	102388	85271	9003	14705	4682	**33.17%**
pred. work	795	0	4626	3896	0	6066	29651	418173	2543	2778	116	472	123	**89.12%**
pred. healthcare	1	0	3059	458	0	69	4087	1	4926	1986	0	3	1	**33.76%**
pred. lifestyle_and_beauty	27	0	1319	522	0	42	386	17	597	1646	13	385	7	**33.18%**
pred. leisure_and_culture	0	0	0	235	0	0	5552	0	0	0	11413	0	0	**66.35%**
pred. shopping	0	0	0	0	0	0	0	0	0	0	0	0	0	**0.00%**
pred. service	0	0	0	0	0	0	0	0	0	0	0	0	0	**0.00%**
class recall	**0.03%**	**98.09%**	**51.42%**	**37.98%**	**21.46%**	**67.74%**	**39.92%**	**97.65%**	**1.80%**	**0.88%**	**41.33%**	**0.00%**	**0.00%**	

Table 5: Confusion Matrix for Deep Learning 500 epochs

	true religion	true home	true lodging	true transportation	true other	true education	true food	true work	true healthcare	true lifestyle_and_beauty	true leisure_and_culture	true shopping	true service	class precision
pred. religion	42996	0	0	0	0	620	9822	10	0	1	0	1	0	**80.44%**
pred. home	0	836490	0	0	246046	0	142094	0	0	0	0	0	0	**68.31%**
pred. lodging	1472	0	168960	15282	0	6053	41813	312	15422	15926	1967	1730	629	**62.68%**
pred. transportation	348	0	1638	28362	0	679	6707	6	2113	1140	127	285	87	**68.36%**
pred. other	0	45251	44	0	107369	0	15774	0	0	0	0	0	0	**63.74%**
pred. education	151	0	744	1028	0	87912	10252	46	576	320	18	29	6	**86.97%**
pred. food	34802	0	347058	276003	2484	83734	412650	0	205651	166757	20696	26122	8267	**26.05%**
pred. work	772	0	4864	3952	0	42752	33214	427865	2524	2686	119	471	131	**82.38%**
pred. healthcare	0	0	0	224	0	1536	18027	0	47771	0	519	1	0	**70.17%**
pred. lifestyle_and_beauty	0	0	0	0	0	0	6	0	0	145	0	0	0	**96.03%**
pred. leisure_and_culture	0	0	0	13	0	0	1260	0	0	0	4167	0	0	**76.60%**
pred. shopping	0	0	0	0	0	0	0	0	0	0	0	0	0	**0.00%**
pred. service	0	0	0	0	0	0	0	0	0	0	0	0	0	**0.00%**
class recall	**53.38%**	**94.87%**	**32.29%**	**8.73%**	**30.17%**	**39.37%**	**59.66%**	**99.91%**	**17.43%**	**0.08%**	**15.09%**	**0.00%**	**0.00%**	

Table 6: Confusion Matrix for Decision Tree Model

	true religion	true home	true lodging	true transportation	true other	true education	true food	true work	true healthcare	true lifestyle_and_beauty	true leisure_and_culture	true shopping	true service	class precision
pred. religion	43031	0	0	0	0	552	7948	9	0	1	0	1	0	**83.49%**
pred. home	0	854914	0	0	205046	0	122108	0	0	0	0	0	0	**72.32%**
pred. lodging	1471	0	238325	12975	0	3714	54385	33	10586	9382	841	1723	627	**71.34%**
pred. transportation	223	0	881	32820	0	577	5565	51	802	758	60	185	50	**78.19%**
pred. other	0	26827	44	0	150843	0	9261	0	0	0	0	0	0	**80.68%**
pred. education	207	0	1863	1212	0	136209	15877	46	1490	557	33	64	17	**86.44%**
pred. food	34933	0	277904	274374	10	81219	416380	3	170406	146960	19738	26230	8307	**28.59%**
pred. work	673	0	4262	3449	0	636	29805	428089	2234	2361	101	431	117	**90.67%**
pred. healthcare	1	0	18	9	0	267	20110	2	88533	4	15	2	2	**81.25%**
pred. lifestyle_and_beauty	2	0	10	13	0	111	7728	6	6	26951	55	0	0	**77.26%**
pred. leisure_and_culture	0	0	0	12	0	1	2452	0	0	0	6770	0	0	**73.31%**
pred. shopping	0	0	1	0	0	0	0	0	0	1	0	3	0	**60.00%**
pred. service	0	0	0	0	0	0	0	0	0	0	0	0	0	**0.00%**
class recall	**53.43%**	**96.96%**	**45.54%**	**10.10%**	**42.38%**	**61.00%**	**60.20%**	**99.96%**	**32.30%**	**14.41%**	**24.52%**	**0.01%**	**0.00%**	

Table 7: Confusion Matrix for Decision Tree Model with Bagging

	true religion	true home	true lodging	true transportation	true other	true education	true food	true work	true healthcare	true lifestyle_and_beauty	true leisure_and_culture	true shopping	true service	class precision
pred. religion	2413	0	3875	897	0	92	957	0	984	421	34	7	46	**24.81%**
pred. home	0	805552	0	0	235395	0	147186	0	0	0	0	0	0	**67.80%**
pred. lodging	36492	0	266441	95453	0	15742	132564	356	115247	67880	7658	13183	2209	**35.37%**
pred. transportation	5983	0	43504	72728	0	8601	46959	67	24538	24148	3020	2661	1163	**31.16%**
pred. other	0	75234	0	0	120128	0	18157	0	0	0	0	0	0	**56.26%**
pred. education	887	955	6715	5008	376	160873	35460	61	2585	2217	613	345	113	**74.41%**
pred. food	32063	0	183585	130007	0	34051	256422	7471	111641	70314	6942	9088	3797	**30.33%**
pred. work	728	0	4902	3955	0	788	29050	420170	2784	2706	140	465	131	**90.20%**
pred. healthcare	990	0	5540	5681	0	1115	8201	9	9695	6168	518	372	150	**25.22%**
pred. lifestyle_and_beauty	973	0	7272	10122	0	1806	10656	74	6128	10904	985	1371	665	**21.40%**
pred. leisure_and_culture	1	0	371	260	0	2	3586	0	4	5	7626	4	2	**64.29%**
pred. shopping	3	0	1058	405	0	79	1734	6	424	1428	56	1141	3	**18.01%**
pred. service	8	0	45	348	0	137	687	25	27	784	21	2	841	**28.75%**
class recall	**3.00%**	**91.36%**	**50.91%**	**22.39%**	**33.75%**	**72.05%**	**37.08%**	**98.12%**	**3.54%**	**5.83%**	**27.62%**	**3.98%**	**9.22%**	

Table 8: Confusion Matrix for Naïve Bayes

	true religion	true home	true lodging	true transportation	true other	true education	true food	true work	true healthcare	true lifestyle_and_beauty	true leisure_and_culture	true shopping	true service	class precision
pred. religion	41929	0	0	0	0	125	9528	10	0	0	95	0	0	**81.12%**
pred. home	73	731623	530	118	167987	3	147053	0	164	34	0	0	0	**69.84%**
pred. lodging	13763	0	362740	136130	0	30915	227657	296	198672	106090	5377	15970	4405	**32.92%**
pred. transportation	7657	0	59709	97315	0	7425	53050	15	11262	19158	2214	1966	1232	**37.29%**
pred. other	41	150116	2801	49	187910	3	18934	0	60	13	2	0	0	**52.21%**
pred. education	580	0	4506	3059	0	158267	31867	8723	2552	1956	242	302	104	**74.60%**
pred. food	15526	0	83818	82740	0	24942	162485	1353	54889	55155	6086	9784	3186	**32.50%**
pred. work	964	2	6140	4991	2	1506	32515	417842	3692	3431	185	540	186	**88.53%**
pred. healthcare	8	0	3056	221	0	96	2012	0	2758	1128	7	69	7	**29.46%**
pred. lifestyle_and_beauty	0	0	5	3	0	3	1	0	8	8	0	8	0	**22.22%**
pred. leisure_and_culture	0	0	3	238	0	1	6517	0	0	2	13405	0	0	**66.47%**
pred. shopping	0	0	0	0	0	0	0	0	0	0	0	0	0	**0.00%**
pred. service	0	0	0	0	0	0	0	0	0	0	0	0	0	**0.00%**
class recall	**52.06%**	**82.97%**	**69.32%**	**29.96%**	**52.80%**	**70.88%**	**23.49%**	**97.57%**	**1.01%**	**0.00%**	**48.55%**	**0.00%**	**0.00%**	

Table 9: Confusion Matrix for Neural Networks

	true religion	true home	true lodging	true transportation	true other	true education	true food	true work	true healthcare	true lifestyle_and_beauty	true leisure_and_culture	true shopping	true service	class precision
pred. religion	43027	0	0	3	0	604	9807	0	0	1	0	0	0	**80.51%**
pred. home	0	871119	0	0	259876	0	149222	0	0	0	0	0	0	**68.04%**
pred. lodging	5924	0	242859	51927	0	8427	43286	6	44620	35754	2553	5566	1803	**54.86%**
pred. transportation	16	0	41	28272	0	106	5340	0	32	35	8	6	2	**83.50%**
pred. other	0	10622	26	0	96023	0	8689	0	0	0	0	0	0	**83.24%**
pred. education	290	0	2774	1921	0	172961	31789	12	1340	1120	138	171	44	**81.37%**
pred. food	30578	0	273092	238801	0	40498	395911	1	177595	141978	11255	22441	7144	**29.56%**
pred. work	705	0	4513	3702	0	686	29842	428220	2410	2569	107	452	127	**90.47%**
pred. healthcare	0	0	3	4	0	1	10240	0	48059	2	0	1	0	**82.42%**
pred. lifestyle_and_beauty	1	0	0	0	0	2	1146	0	0	5516	0	0	0	**82.76%**
pred. leisure_and_culture	0	0	0	234	0	0	6347	0	0	0	13552	0	0	**67.31%**
pred. shopping	0	0	0	0	0	1	0	0	1	0	0	2	0	**50.00%**
pred. service	0	0	0	0	0	0	0	0	0	0	0	0	0	**0.00%**
class recall	**53.42%**	**98.80%**	**46.41%**	**8.70%**	**26.98%**	**77.46%**	**57.24%**	**100.00%**	**17.54%**	**2.95%**	**49.08%**	**0.01%**	**0.00%**	

Table 10: Confusion Matrix for Random Forest 100 trees

	true religion	**true home**	**true lodging**	**true transportation**	**true other**	**true education**	**true food**	**true work**	**true healthcare**	**true lifestyle_and_beauty**	**true leisure_and_culture**	**true shopping**	**true service**	**class precision**
pred. religion	43025	0	0	3	0	601	9802	0	0	0	0	0	0	**80.52%**
pred. home	0	871115	0	0	255330	0	149221	0	0	0	0	0	0	**68.29%**
pred. lodging	5918	0	241347	51885	0	8408	42988	4	43100	35158	2554	5568	1805	**55.01%**
pred. transportation	20	0	51	28316	0	86	5353	0	43	53	10	13	2	**83.41%**
pred. other	0	10626	26	0	100569	0	8676	0	0	0	0	0	0	**83.88%**
pred. education	291	0	2774	1928	0	172976	31788	1	1342	1127	138	171	44	**81.37%**
pred. food	30581	0	274568	238781	0	40525	396431	1	177704	142046	11252	22432	7142	**29.55%**
pred. work	706	0	4540	3713	0	687	29840	428233	2425	2574	107	451	127	**90.46%**
pred. healthcare	0	0	2	4	0	1	10100	0	49442	2	0	1	0	**83.02%**
pred. lifestyle_and_beauty	0	0	0	0	0	1	1068	0	0	6015	0	0	0	**84.91%**
pred. leisure_and_culture	0	0	0	234	0	0	6352	0	0	0	13552	0	0	**67.30%**
pred. shopping	0	0	0	0	0	1	0	0	1	0	0	3	0	**60.00%**
pred. service	0	0	0	0	0	0	0	0	0	0	0	0	0	**0.00%**
class recall	**53.42%**	**98.79%**	**46.12%**	**8.72%**	**28.26%**	**77.47%**	**57.32%**	**100.00%**	**18.04%**	**3.22%**	**49.08%**	**0.01%**	**0.00%**	

Table 11: Confusion Matrix for Random Forest 200 trees