UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE INFORMÁTICA

Characterization of Behavioural Patterns of Portuguese Blood Donors using Supervised and Unsupervised Learning

João Miguel Ferreira Castanheira

Mestrado em Ciência de Dados

Dissertação orientada por:

Prof. Sara Madeira

Prof. Pedro Monteiro

2021

"I repeat that we are explanation-seeking animals who tend to think that everything has an identifiable cause and grab the most apparent one as the explanation. Yet there may not be a visible cause; to the contrary, frequently there is nothing, not even a spectrum of possible explanations"

— Nassim Nicholas Taleb, *The Black Swan: The Impact of the Highly Improbable*

Acknowledgements

Firstly, I would like to thank the Fundação para a Ciência e a Tecnologia (FCT) for the national funds that supported this work within the project UIDB/50021/2020 and the project LAIfeBlood - Inteligência Artificial para a Gestão do Sangue, with reference DSAIPA/AI/0033/2019. I would also like to thank IPST for providing the data set without which this dissertation would have not been possible. The database complies with current legislation for the protection of data, in this case, from benevolent blood donors - their identity being unknown. Data were used for academic purposes only.

I would like to thank my supervisors, Professor Sara Madeira and Professor Pedro Monteiro, whose expertise and feedback played an important role in formulating the research questions and methodology. I am also grateful to Professor Sofia Teixeira, for her advice and for reviewing the text.

I would also like to thank my friends, who have always given me the necessary support over the years. I believe that it is only possible to achieve great things when we are surrounded by people who inspire, encourage and support us.

Finally, a special thanks to my family: Mum, Dad, Brother and Rute.

Thank you.

Abstract

Blood donations are essential to save innumerous lives on a global scale on a daily basis. Without blood donations, many medical procedures cannot take place. Thus, the study of what motivates blood donors to donate and how they behave is important to ensure a stable and safe blood supply.

Several studies tried to understand the most important factors for blood donor return, by using mainly logistic regression models. Those studies identified several donor demographic characteristics as important factors to describe donors' future behaviour. However, in this dissertation it is argued that if models have a poor performance in the task for which they are trained for, the conclusions taken from them may be erroneous. Thus, this dissertation presents a contribution for the study of understanding blood donor behaviour by using the most recent machine learning, evaluation and interpretability techniques.

In this dissertation, several machine learning experiments are implemented aiming to predict blood donors return one year following a given donation, gaining insights about blood donors future behaviour and which factors influence it the most. Primarily, the blood donations dataset is split according to several geographic characteristics. Each segment is further split into blood donations from new and experienced donors (i.e. those who donated more than once). For experienced donors several features regarding their past behaviour are computed. Finally, different machine learning models are trained on top of each segment.

Our results suggest that donor's demographics, as well as features regarding the donation, are not enough to predict donor return. As such, it is not possible to estimate the impact that donor's demographics have on donor's future behaviour. However, models trained over experienced donors performed significantly better than those trained over new donors data, due to the impact of past behaviour features. However, even with past behaviour features the machine learning models do not achieve outstanding scores in predicting donor future behaviour, and, as such, this work demonstrates that both demographics and past behaviour features are insufficient to accurately explain future behaviour.

Keywords: Blood Donor; blood donation; machine learning;

Resumo Alargado

Doações de sangue são uma componente essencial dos sistemas de saúde. Sem doações de sangue, muitos dos procedimentos médicos dados como garantidos não se podem realizar. Uma oferta constante de sangue é necessária de modo a assegurar que os hospitais têm acesso a sangue suficiente para cobrir todas as suas necessidades. Recentemente têm-se observado alterações na demografia de vários países desenvolvidos, com as taxas de natalidade a decrescer, a esperança média de vida a aumentar e, consequentemente, a um envelhecimento da população. Uma população mais envelhecida traduz-se num aumento de síndromes e doenças crónicas, levando posteriormente ao aumento da procura por doações de sangue. Posto isto, e de modo a assegurar uma oferta constante de sangue, é imperativo estudar o comportamento dos dadores, bem como o que os motiva a continuar a doar.

De modo a perceber o comportamento dos dadores de sangue diversos estudos têm utilizado técnicas de modelação e previsão para compreender o seu comportamento futuro. A maioria dos estudos utilizam regressões logísticas, de modo a prever se um determinado dador iria voltar um ano após uma dada doação. Posteriormente, os coeficientes de regressão têm sido analisados de modo a perceber como determinadas características impactam o comportamento futuro dos dadores. Diversos estudos identificam fatores demográficos como importantes para um dador voltar a doar, tais como a idade, o género, o peso corporal ou a profissão.

No entanto, nenhum destes estudos providencia métricas de desempenho dos seus modelos de previsão. Nesta dissertação argumenta-se que se os modelos preditivos têm um mau desempenho na tarefa para o qual são treinados, então as conclusões retiradas destes podem ser erróneas. Para além disso, os estudos acima mencionados utilizaram, na sua maioria, centenas ou poucos milhares de registos de doações de sangue. Neste sentido, esta dissertação apresenta uma contribuição para o estudo do comportamento dos dadores de sangue ao utilizar as mais recentes técnicas de modelação, previsão e avaliação de modelos preditivos, bem como ao utilizar um número consideravelmente maior de registos de doações de sangue, quando comparado aos estudos relacionados.

Diversos modelos de aprendizagem automática são implementados, de modo a prever se um determinado dador volta a doar um ano após uma dada doação. A literatura demonstra que, em diversos países, o comportamento dos dadores de sangue tende a mudar de acordo com a zona geográfica onde a doação é feita. Por exemplo, existe evidência de que dadores que doam em zonas rurais tendem a doar mais frequentemente do que aqueles que doam em zonas urbanas. Posto isto, nesta dissertação as doações de sangue são segmentadas de acordo com diversos critérios geográficos, de modo a analisar se existem diferentes comportamentos/padrões de doação em diferentes segmentos.

De seguida, cada segmento é dividido em doações de novos dadores e de dadores experientes (aqueles que doaram mais que uma vez). Diversas entidades nacionais e internacionais responsáveis por gestão de sangue aconselham a que o estudo do comportamento dos dadores seja segmentado de acordo com dadores novos e experientes. Deste modo, é possível perceber e modelar o comportamento destes grupos, bem como entender quais os fatores mais importantes para doações futuras entre estes dois tipos de dadores, e se estes são diferentes entre si. Por último, diversos modelos de aprendizagem automática são treinados sob cada segmento.

Os dados utilizados na presente dissertação foram disponibilizados pelo *Instituto Português do Sangue e da Transplantação* (IPST), e correspondem a milhões de doações de sangue realizadas em Portugal. As tarefas de aprendizagem automática realizadas podem ser descritas como tarefas de classificação binária, que têm como objetivo prever se um determinado dador volta a doar um ano após uma dada doação. Para tal, diversas características demográficas dos dadores – tais como idade, género, situação laboral, estado civil e tipo de sangue – são utilizadas como variáveis independentes. Para além destas variáveis, são também utlizadas variáveis relacionadas com a doação em si, referentes ao local e quando a doação foi realizada.

Adicionalmente, a literatura apresenta evidências de que para dadores experientes, variáveis relacionadas com o seu comportamento passado são as mais importantes para modelar o comportamento futuro. Consequentemente, diversas variáveis relacionadas com o comportamento passado são processadas para dadores experientes.

Os resultados obtidos revelam que, ao contrário do que foi anteriormente documentado na literatura, factores demográficos e informação específica da doação em si são insuficientes para prever com precisão o comportamento futuro dos dadores. Para além disso, não são registadas diferenças consideráveis entre modelos treinados sobre diferentes segmentos geográficos.

No entanto, é possivel diferenças significativas no desempenho de modelos treinados sobre os dadores novos e os dadores experientes – estes últimos apresentando um melhor desempenho. Utilizando as técnicas mais recentes de interpretabilidade de modelos de classificação, observa-se que as variáveis mais importantes para prever o comportamento futuro de dadores experientes são as variáveis referentes ao seu comportamento passado, o que se encontra em linha com a literatura existente. Porém, dadas as métricas de desempenho dos modelos treinados sobre dadores experientes e a análise feita às mesmas, não é possível afirmar que os modelos conseguem claramente distinguir entre aqueles que voltam passado um ano e aqueles que não voltam.

Não obstante, duas experiências são desenhadas e implementadas de modo a despistar o real impacto das variáveis referentes ao comportamento passado dos dadores experientes na tarefa de prever o seu comportamento futuro. Numa primeira experiência, estas variáveis são removidas dos dados de treino e novos modelos são treinados. Os resultados desta experiência demonstram que o desempenho dos classificadores decresce significativamente quando estas variáveis são removidas, o que realça a importância das mesmas. Uma segunda experiência é desenhada na qual apenas as variáveis do passado são utilizadas como dados de treino. Os resultados demonstram que não existem diferenças significativas entre os classificadores treinados com todas as variáveis (demografia do dador, informação da doação e variáveis do passado), e aqueles treinados apenas com variáveis do passado.

Posto isto, é provável que o comportamento dos dadores seja modelado por variáveis que não estão presentes nos dados. Outra hipótese é que o fenómeno de doar sangue tenha um elevado grau de aleatoriedade. Graus elevados de aleatoriedade num fenómeno geram dados com elevados níveis de heterogeneidade, o que dificulta a tarefa de aprendizagem por parte dos modelos. De modo a avaliar o grau de heterogeneidade dos dados e a encontrar grupos homogéneos de dadores de sangue, técnicas de aprendizagem não supervisionada são utilizadas.

Dois grupos de novos dadores são encontrados. O primeiro grupo contém dadores tendencialmente mais novos, que doam durante a semana, tendencialmente nas grandes áreas metropolitanas ou capitais de distrito. O segundo grupo contém dadores tendencialmente mais velhos, que fizeram a sua primeira doação ao fim-de-semana, maioritariamente fora das grandes áreas metropolitanas e capitais de distrito, e que apresentam uma probabilidade significativamente maior de voltarem para uma segunda doação, em comparação com os elementos do primeiro grupo.

Dois grupos de dadores experientes foram também encontrados, tendo o algoritmo feito o agrupamento maioritariamente com base na sua experiência passada. O primeiro grupo contém dadores com um número significativamente maior de doações passadas, e com um maior número médio de doações por ano. Estes dadores são tendencialmente do sexo masculino, mais velhos, casados e empregados. O segundo grupo contém dadores que fizeram um número significativamente menor de doações no passado, e que tendem a doar menos frequentemente em comparação com o primeiro grupo. Este grupo contém uma percentagem significativamente maior de dadores do sexo feminino, solteiros, mais novos e estudantes, em comparação com os elementos do primeiro grupo.

Palavras-chave: Dador de Sangue; doação de sangue; aprendizagem automática;

Contents

List of Figures

List of Tables

Acronyms

SNS *Sistema Nacional de Saúde*

IPST *Instituto Português do Sangue e da Transplantação*

CART Classification and Regression Tree

RBC Red Blood Cells

ARIMA Autoregressive Integrated Moving Average

RFM Recency, Frequency and Monetary

MLP Multilayer Perceptron

SVM Support Vector Machine

EDA Exploratory Data Analysis

RAFT *Reorganização Administrativa do Território das Freguesias*

OSM Open Street Maps

INE *Instituto Nacional de Estatística*

CST *Centro de Sangue e Transplantação*

TIPAU *Tipologia das Áreas Urbanas*

APU *Áreas Predominantemente Urbanas*

AMU *Áreas Maioritariamente Urbanas*

APR *Áreas Predominantemente Rurais*

TPB Theory of Planned Behaviour

ROC Receiver Operating Characteristic

AUC Area Under the ROC Curve

EU European Union

SHAP SHapley Additive exPlanations

WHO World Health Organization

BFT *Brigada Freguesia TIPAU*

DFT *Dador Freguesia TIPAU*

BD *Brigada Distrito*

MA *Metropolitan Area*

RF Random Forest

ADB AdaBoost

GB Gradient Boosting

AML *Área Metropolitana de Lisboa*

AMP *Área Metropolitana do Porto*

LIME Local Interpretable Model-agnostic Explanations

Chapter 1

Introduction

This chapter provides an overview about the importance of blood donation to the health sector, as well as the importance of understanding which factors contribute most to the donors return. We address how the state of the art literature approaches this problem, and how in the present work we contribute to overcome their shortcomings and limitations.

1.1 Overview

Blood is a scarce resource, and its role in healthcare is fundamental, with blood donations being essential to save innumerous lives on a global scale on a daily basis.

It is estimated that 85 million blood transfusions are carried out annually across the globe, translating to nearly three blood transfusions per second [106]. Without blood donations, many of the medical procedures we otherwise take for granted could not occur. However, blood centers all over the world are suffering from a high shortage of blood supply [4]. The modern lifestyle, ever-increasing mobility and accompanying higher accident rates, and incidences of natural and human-made disasters (such as wars, earthquakes, etc.) have led to an ever-rising demand for blood transfusions [4]. The margin between blood supply and blood transfusion is now narrower than it has ever been in the last two decades [76]. This has led some experts to predict that, in the absence of appropriate interventions, total demand for red blood cells will surpass total supply in the near future [76].

One major factor to take into account while considering the needs of blood management is the changes observed in the demographics of developed countries. Most European countries are currently seeing an increasing number of senior inhabitants, while at the same time the number of young inhabitants is decreasing. This development may lead to a serious imbalance between blood supply and consumption [67]. Currently, the maintenance of a safe blood supply level is provided by a small number of volunteers, and their retention is difficult. The study of the retention of donors and their motivation to give blood at a higher frequency remain important challenges that have not been given the necessary attention in the

scientific literature [42]. To ensure constant and adequate blood supplies it is needed to plan programs to recruit and retain first-time blood donors to overcome the difficulties imposed by those who stop donating due to age, illness or positive results in a screening test[53, 92].

From a blood collection agency perspective, experienced donors, i.e., those who have donated more than once, provide some key advantages over first-time donors and play a major role in providing adequate blood supplies. First, they provide a relatively stable and comparatively safe supply of blood. Second, a body of regular donors provides the long-term opportunity for blood collection agencies to save on costs associated with the continual recruitment of new donors [65]. Moreover, the cost-benefit of repeated donation is better since those donors require less effort from the blood centers to be convinced to return [61]. Finally, the recruitment of experienced donors is easier due to better cooperation [75] and fewer adverse reactions [91]. To improve blood supply stability, the rate of experienced donors should be increased [53]. As such, these experienced donors are the real assets of blood transfusion centers. Since every few donors become experienced donors, recognizing the factors that affect blood donation will lead to a better understanding of potential regular donors. Hereupon, understanding donor return behaviour, and the factors that most affect their return is crucial for determining more effective measures to recruit and retain blood donors [91].

1.2 Motivation

Multiple studies have analysed which factors affect donor return the most. Those studies could be separated into two different types: studies that searched for the importance of psychological variables for donor's return; and studies that searched for the importance of donor's demographics and other non-psychological variables. Regarding psychological factors, intention [6, 39, 40] and altruism [41, 54, 66, 96] were identified as the most important factors. On demographics side, donor's age [42, 60], gender [60], geographic location [60], past behaviour [10, 21, 32, 33, 51, 78, 89, 90], body weight and job class [56] were identified as important factors. Despite the importance of psychological variables, the present thesis is focused on finding the importance of demographics and non-psychological variables for donor's return.

The studies that searched for the importance of donor's demographics have used mainly logistic regression models to predict blood donor return. The methodology of those studies could be described as follows: authors have access to data regarding several hundred, or few thousand of blood donations; logistic regression models are used to predict blood donor return after a given period of time (typically, a year); afterwards, the learned coefficients are used to explain which factors most affect donor return. However, from the review of literature that was made for the present thesis, it was possible to observe that none of those studies used any kind of model validation technique. The data is simply used to feed simple linear models, and after that, conclusions are drawn from the model without evaluating whether they actually fit the data. However, as stated by Breiman [15], if models are a poor emulation of the

phenomenon that generated the data, the conclusions drawn from it cannot be trusted. Hence, there is a gap in the literature regarding the real impact that demographics and other non-psychological variables have on donor return. This thesis aims to explore that gap.

1.3 Objectives

Annually, the *Sistema Nacional de Saúde* (SNS), through the IPST, which is the Portuguese public entity responsible for collecting and managing blood, receives around two hundred thousand blood donations from fifty thousand donors. However, the number of blood donations has been decreasing in the last few years. To help overcome these two trends, the increasing demand for blood products and the decrease in blood donations, it is critical to perform research on donors data. Machine learning and data mining techniques can be applied to analyse blood donation data to discover new insights and predict donors behaviour.

For this study, data from IPST was used. It is composed of around five million donations dating from 1970 to 2019. The raw dataset includes 47 columns regarding donor information (date of birth and gender), donation information (donation date, blood type and blood facility center type), geographic information (district, council, postal code) and more than 20 features regarding adverse reactions. This dataset provides a unique opportunity by which demographic characteristics, geographic distributions and donation patterns of blood donors could be identified and analysed.

This thesis aims at finding which demographics and other non-psychological variables affects donor return, by using the most recent machine learning and interpretability techniques, over the IPST data. By using a machine learning approach and state-of-the art evaluation techniques, one can ensure that the conclusions are trustworthy, in contrast to the above-mentioned current research techniques. The objectives of this dissertation can be described as follows:

- To study whether machine learning models can learn the relationship between donor demographic's, as well as other features, and the target variable regarding donor return within a period of one year;
- To use the most recent machine learning interpretability techniques, to extract knowledge from the models.

Therefore, this dissertation presents a contribution to the research topic in three distinct ways:

- By providing an overview of the literature, regarding the importance of demographics and other non-psychological variables for donor return, and reviewing it;
- By using the most recent machine learning techniques: both machine learning models, validation and interpretability techniques;

- By using a considerably higher amount of data, when compared to the related literature. With such vast amount of data, and by using the right model validation methods, more reliable conclusions regarding the importance of donor demographics on donor return could be taken.

1.4 Thesis outline

The remainder of this thesis is organized as follows. Chapter 2 reviews previous literature that uses as input blood donations datasets. Chapter 3 introduces the IPST data; describes it was processed in order to increase data quality; finally it includes an Exploratory Data Analysis (EDA) section. Chapter 4 describes the data used for the machine learning experiments; the classifiers used; the evaluation methods adopted; finally, the results are presented and discussed. Chapter 5 describes the use of unsupervised learning techniques to search for homogeneous groups of blood donors. Finally, Chapter 6 summarises the main conclusions of this work, and presents possibilities for further developments.

Chapter 2

Related Work

This chapter introduces the related work on research topics that used blood donation data as input. The first step of the work for this dissertation was reviewing the literature to search for scientific questions to explore, and, as such, some sections in this chapter are outside the scope of this dissertation.

The remainder of this chapter is organized as follows: Section 2.1 gives a quick overview of the literature that studied blood donors demographics; Section 2.2 regards scheduling policies; Section 2.3 covers the topic of predicting blood donor supply; Section 2.4 covers the topic of using clustering algorithms to find subgroups of blood donors; finally, Section 2.5 gives a more in-depth overview of the literature regarding the usage of blood donation data for predicting and understanding blood donors future behaviour, which is the scope of the present dissertation.

2.1 Donor demographics

The demographic characteristics of the blood donor populations are dynamic, reflecting changes in the general population's demographics, changes to the donor selection guidelines and periodic marketing campaigns [60]. Understanding those changes provides essential information for monitoring donor recruitment and provides evidence for work programs to encourage donors to donate more frequently and recruit and retain first-time donors to compensate for those who stop donating. Nonetheless, it is essential to note that different studies use data from different populations, with different characteristics, so it is customary to observe different results regarding demographics among different studies.

Studies on the relationship between age and donor status (first-time or experienced donor) show different results. Charbonneau et al. [20] indicated that the likelihood of donating increases with age, while two other studies stated that younger people are more likely to donate [49, 81]. Wiersum-Osselton et al. [105], and Gemelli et al. [38] reported that younger people were more likely to be first-time donors when compared to the elderly. They both reported that experienced donors were older than first-time donors. In the United States, where the minimum eligible donor age is 16, the highest return rates were

found for donors between the age of 16 and 18 [72]. Two studies also indicated that the elderly were more likely to be experienced donors than younger people [20, 49].

Regarding sex and their relationship with donations, multiple studies found that women were more likely to be donors than men [9, 20, 49].

Concerning education, Charbonneau et al. [20], and Atsma et al. [9] found that higher education is related to a higher propensity to donate. Regarding donor status, Gemelli et al. [38] results indicate that students were more likely to be first-time donors than experienced donors, while the group of experienced donors had a higher number of retired people and professionals. Some studies also evaluate the impact of employment status and income on donors behaviours. Priller et al. [81] results show that donors did not differ from non-donors in terms of their employment status, as well as people with high family income were more likely to be donors than people with low family income. One study, in the Netherlands [9] also found a correlation between marital status and donations. Their analysis has shown that married people are more likely to be experienced donors than divorced and never-married people.

The papers cited above have studied the influence of several individual characteristics on the frequency of blood donation. However, few studies have analysed regional variations and how they influence the modelling of blood donation frequency.

Crawford et al. [23] performed a study using data provided by the Hema-Quebec, which is the organization responsible for blood management in Quebec, Canada. They computed two variables, the donor's region of residence and the distance from that region of residence to the collection sites, and modelled how they influence the frequency of blood donations. Besides those two variables, they also used demographic characteristics such as donor's age and sex, and variables regarding reasons for deferral. To model the impact of regional variables on blood donation frequency, they used two modelling techniques: a negative binomial regression model and an ordinal logistic regression model. The former was used to model the frequency of blood donation made by the donors over a period of five years, with the dependent variable being the total number of donations per donor minus one, and the latter aimed to model a discrete ordinal variable regarding the frequency of blood donations derived from the total number (sum) of donations per donor. Their results indicate that the inclusion of the measurement of the distance between donors place of residence and location of the collection site increases both the frequency of blood donations (negative binomial regression model) and the possibility of the donors moving into the high-frequency donor's category (ordinal regression model).

2.2 Scheduling policies

Some blood centres around the world have open blood donation sessions, while others use fixed appointments. Suboptimal scheduling policies by blood donation centres contribute significantly to some blood supply problems because it could lead to prolonged waiting time for donors and a stressful working situation for the blood bank staff. Making each donor's donation experience as enjoyable as possible is

therefore probably a key factor for the recruitment and retention of donors [14].

Bosnes et al. [14] performed a study where data regarding fixed appointments from the Blood Bank of Oslo was used. They proposed a logistic regression model that outputs the probability that a given donor arrives at his appointment. A set of explanatory variables was used to characterize the donor, such as age, sex, previous donation pattern, previous short-time temporary deferrals, number of cancellations or time since the previous donation. The dependent variable is the response to a scheduled donation appointment, i.e., a boolean indicating whether the donor arrived or not. Their results show the estimated relationship between each explanatory variable and the probability of arrival. They concluded that the probability of arrival was higher for donors whose appointment was made through personal contact than for donor who has received a written invitation. They also found a strong correlation between donor age and the dependent variable, with their study showing that as donor age increases, the probability of arrival also increases.

On the other hand, Testik et al. [98] study the case where non-fixed appointments were used in Hacettepe University Hospitals blood centre. Blood centre without fixed appointments for collecting blood often experiences nonconstant donor arrival rates, which vary due to month-of-year, day-of-month, day-of-week, time-of-day, as well as many other factors. When a constant workforce size is employed in such blood centres, there is either idle personnel or donor satisfaction could be compromised due to long waiting times [98]. Testik et al. [98] used data mining to identify patterns that indicated significantly different daily and weekly arrival rates for blood donors and considered these factors to plan an adaptive work schedule for the facility. They considered two types of donor arrival patterns: daily arrival patterns (pattern between days) and hourly arrival patterns (patterns within the daily arrival patterns). To uncover those patterns, they used Two-Step clustering and Classification and Regression Tree (CART), respectively clustering and classification tasks. The Two-Step clustering method compresses the records into a set of clusters and CART merges these subclusters into larger clusters by using hierarchical clustering. The results found that the arrival rates to the blood centre varied based on ten distinct hourly patterns found within three identified daily patterns (Monday-Thursday, Friday and Saturday-Sunday) [4, 98].

2.3 Predict donations

Planning the future blood collection efforts must be based on adequate forecasts of transfusions demand [76]. An accurate prediction of the future demand is necessary to plan better blood drives, staffing of blood collection facilities and promotion campaigns. Time series machine learning models could be applied for better forecasts of blood demand.

Pereira et al. [76] use time-series data from the Hospital Clinic of Barcelona, Spain. Their data consisted of the monthly demand for Red Blood Cells (RBC) from January 1988 to December 2002. They split the data into training and test sets and applied three time-series methods: Autoregressive Integrated Moving Average (ARIMA), the Holt-Winters family of exponential smoothing models, and

one neural network-based method. Their results have shown that ARIMA was the model which fits best. Over 1-year time horizon, forecasts generated by ARIMA laid within the 10% interval of the real RBC demand in 79% of the months.

Another study also applied ARIMA, but this time for monthly supply forecast [4]. They used two years of data, but after some analysis, the authors found that their series was a random walk and therefore could not be predicted.

Other machine learning techniques have been used to forecast donor supply and demand, in addition to time sequence models.

Concerns have been raised that the general population's ageing in developed countries will increase blood products demand. Borkent et al. [13] developed different mathematical models to assess trends in blood demand and supply and predict how these trends developed over time. Their first demand model assumes that the mean number of transfusions per inhabitant per age and sex is constant. The second demand model incorporated both demographic changes and trends in clinical RBC use over time. The first model predicts an increase of 23% in RBC demand over the following years while the second model predicts a decrease in RBC demand by 8% in the same period. They concluded that the second model provided a much better estimate in RBC demand. The supply model used annual donor retention rates, donor recruitment rates, and the number of donations per donor year as independent variables. They conclude that despite an ageing population, the number of donations is more affected by donor retention rates and the influx of new donors from recruitment activities than by the general population's demography.

Santhanam et al. [88] applied CART to a public blood donation dataset [108] available on UCI Machine Learning Repository [29]. The dataset contains data about 748 donors randomly sampled from data owned by the Blood Transfusion Service Center in Hsin-Chu City, Taiwan. It consists of an extension of the Recency, Frequency and Monetary (RFM) marketing model. Five variables can be found in the dataset: R (Recency - months since the last donation); F (Frequency - total number of donation); M (Monetary - whole blood donated in c.c.); T (Time - months since the first donation) and the dependent variable representing whether the donor donated blood in March 2007 (1 if yes, 0 otherwise). Santhanam et al. [88] CART model performed very well in the binary classification task, achieving high accuracy and 99% precision. Darwish et al. [24] also uses the same dataset and applied a MLP and a Support Vector Machine (SVM).

2.4 Clustering

Clustering is the process of grouping a set of data objects into multiple groups or clusters, so that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters. Dissimilarities and similarities are assessed based on the attribute values describing the objects and often involve distance measures [46].

Ashoori et al. [8] use the K-means clustering method to describe blood donor behaviours. They used demographic variables such as age, blood donation status, blood type, gender, educational background and marital status. The clustering results identified two different clusters, where the first cluster contains just male donors, tendency older, with high education levels, O positive blood type, and all of them married. The second cluster contains lower-age donors than the first cluster, multiple blood types and genders, and most of the donors never married.

Apparicio et al.[5] also use clustering techniques, but this time to detect spatial clusters of high or low blood donation rates in Canada, according to donors sex and age. The detection of spatial clusters of blood donation rate is an important issue, especially for targeting spatial units with significantly low donation rates, where it could be possible to increase the numbers of donors [5]. They used data for over five years, then these data were aggregated for each of the municipalities and counties in Canada, according to sex and age groups. To detect spatial clusters, they used Kulldorff's scan statistics method [59]. Their results indicate several geographic areas with high or low blood donation rates for each group.

2.5 Retaining donors

A constant blood supply is needed to help ensure that hospitals have access to enough blood to meet their current and future needs [4]. One of the significant goals of blood transfusion centres is to find and get to know people who can donate healthy blood and maintain them. These donors are the real assets of blood transfusion centres. The study of the retention of donors and their motivation to give blood at a higher frequency remain important challenges that have not been given the necessary attention in the scientific literature [42].

Multiple studies have identified altruism as the most important psychological motive for donating [41, 54, 66, 96]. External factors such as social pressure and incentives to donate, such as gifts or rewards, have also been described as essential motivators [54, 78, 85]. Despite the importance of those psychological factors, understanding other factors such as socio-demographic ones, and the importance of past behaviour, could give a better overview of what truly motivates donors behaviour.

2.5.1 TPB and the role of past behaviour

Although the need for a better understanding of blood donor behaviour has been noted to be of key importance for blood collection agencies internationally, much of the previous applied research on donor behaviour has failed to draw clearly or systematically on contemporary theories of behavioural decision making [65].

Ferguson et al. [31, 34, 35] stated that the Theory of Planned Behaviour (TPB) [3] has been one of the most abiding theories in predicting blood donation intention and behaviour. A central factor in the TPB is the individual intention to perform a given behaviour [3]. The TPB is based on the assumption

that intention is the most important factor in predicting behaviour. Intention, in turn, is proposed to be influenced by attitude (positive or negative evaluations of performing the behaviour), subjective norm (perceptions of social pressure for the performance of the behaviour), and perceived behavioural control (perceptions of control over performing the behaviour) [65]. This model proposes that the more one engages in a behaviour, the more likely one will continue [89].

As one of the most commonly used tool for modelling beliefs-behaviour relations, TPB and its components have been successfully used by many researchers to predict blood donation intentions and behaviours. Bagozzi [10] developed a questionnaire that allowed the measurement of the constructs contained within the TPB framework, which showed that blood donations, i.e., the behaviour in the context of the TPB, was mainly predicted by intention, although the importance of intention decreased as the past frequency of blood donation increased. Similarly, Charng et al. [21] developed a study in which 658 donors were followed over a period of 7 months, and their results showed that intention was the most important determinant of blood donations. Other studies [6, 39, 40] have reached similar conclusions regarding the importance of intention as a determinant of behaviour.

Despite the importance of intention to donate blood, several studies have highlighted the major role of past frequency of blood donation, which some authors define as *habit*, as a determinant of future behaviour. As well-known by the social sciences, the frequency of past behaviour is a standard indicator of habit strength, and therefore it is one of the best predictors of future behaviour [74]. Charng et al. [21] define habit as the semi-automatic performance of a well-learned behaviour. The more often an individual performs a behaviour, the more likely it is that the behaviour will become a habit. Therefore, they state that past repeated behaviours can be an indicator of habit, and their findings suggest that an index of habit may add quality to a predictive model, by expecting that a past repeated behaviour could directly affect future behaviour.

Ferguson et al. [33] studied the efficacy of 6 different psychological variables (intentions) on predicting the number of blood donations in a sample of 630 blood donors over a 16-17 month period. One of the factors used was the previous number of blood donations. They distinguish between two different donor groups: occasional and regular donors, with the former defined as having made four or fewer donations (including the current donation), and the latter as having made five or more donations (including the current donation).

Their results have shown that, for regular donors, past behaviour was the best predictor of future behaviours. They have also added a quadratic term of the past behaviour, which significantly improved the power of the past behaviour to predict future behaviour. However, for occasional donors, past behaviour was not a significant predictor of future behaviour. Fergunson et. al [33] concluded that it might be the case that regular donors settle into a consistent pattern of cyclical donation only after a specific frequency of donation has been achieved. Other studies supported that hypothesis [51, 78, 91], by saying that, as donors continue to donate, they develop a blood donor role identity, which appears to occur after the third or fourth donation.

In a posterior study, Ferguson [32] studied the relationship between donor personality traits and donor

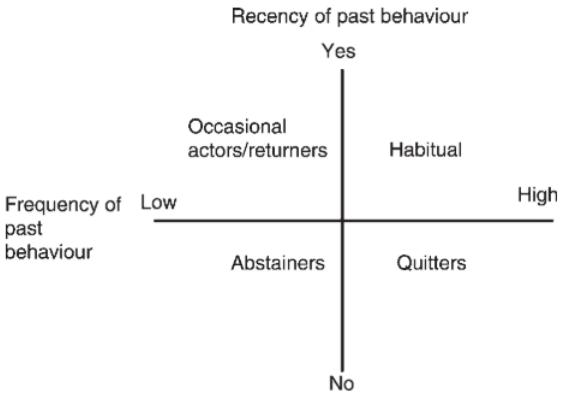

Figure 2.1: Behavioural patterns from the frequency and recency of past behaviour. Image from [32].

past behaviour. The frequency of previous donations as an index of donor past behaviour was previously used [21, 33], however, according to Ferguson [32] this primary quantity does not capture several other features related to the donor behaviour. Therefore, Ferguson [32] computed four different indices: frequency, years since the first donation, eligibility donation rate, and recency and studied how those different factors related with each other and with some personality traits. Frequency regarded the number of previous donations made by the donor. The eligibility donation rate was calculated by dividing the number of donations made by the number of years since the donors were eligible to donate, and, according to Fergunson [32] it provides a useful guide to donors' current life span rate. To examine the more subtle aspects of donor behaviours, they also included a factor that regards how many years each donor had been donating. As previously noted by Omoto et al. [73], the number of years since the first donation may capture something of the nature of donor general volunteer commitment. Lastly, recency considered the number of months from the previous donation to the current donation.

As noted previously by Perugini et al. [77], a distinction between frequency and recency of past behaviour could be drawn. Whereas for first-time donors, the recency and frequency of past behaviour should be equivalent, more experienced donors may vary concerning the number of previous donations, with some having donated recently and some not having donated for a while [32]. Therefore, Perugini et al. [77] suggest that the overall relationship between past and recent behaviour should be orthogonal, i.e., they measure different aspects of the past behaviour.

Similarly, Ferguson [32] results also suggest that recency and frequency were orthogonal. Given the independence of recency and frequency of past behaviour, they produced a scheme that describes blood donation behavioural patterns. Figure 2.1 shows that scheme.

According to Ferguson [32] scheme, those acting recently with a lower past frequency of behaviour may be seen as 'occasional' donors. Habitual behaviour may be defined in terms of a combination of acting recently with a high frequency of past behaviour [32]. Donors with a high frequency of past donations, but who have not donated recently are described as 'recent quitters'. Those who have neither

acted recently nor performed the behaviour frequently may be seen as 'abstainers'.

2.5.2 The role of demographics

Godin et al. [42] aim at finding factors associated with repeated donations among regular and new donors. They developed a questionnaire that allows for the construction of concepts within the TPB framework. Despite the three psychological variables that regard the TPB, i.e., attitude, subjective norm, and perceived behavioural control, they have also included other psychosocial variables known to contribute to the explanation of behaviour and intention in the context of blood donation: descriptive norm, personal identity, level of satisfaction and moral norm. Descriptive norm is expressed as the prevalence of the behaviour which prevails in donors social environment. Personal identity represents the degree to which a person perceives the appropriateness of adopting behaviour for someone in a position similar to his or her own in a social network. The level of satisfaction regards donors last donation. Finally, moral norm measures the sense of personal obligation toward adopting the behaviour [42].

Beyond those variables, they have also included past donor frequency of donation over the last 24 months and demographic variables such as age, sex, level of education and donor civil status. Logistic regression was therefore used to predict whether each donor completed a donation at the 6-month follow-up period, and different models for new and regular donors were created. A total of 2389 blood donors answered the questionnaire, and their results show that, for regular donors, past behaviour was the most crucial factor in predicting donor behaviour, followed by intention. For new donors, the most important predictors were intention and age. Concerning age, their results show that, for both groups, older donors were more likely to donate in the future when compared to new donors. None of the other demographic variables made a significant contribution to the model in either group [42]. In summary, their results support the idea that different promotion strategies should be adopted to increase repeated blood donation among experienced and new donors.

To understand the determinants of donor return, Schlumpf et al. [89] developed a survey in 2003, which was answered by 7905 donors. All the factors measured in the survey were ranked as possible predictors of actual return within 12 months by using logistic regression. Remind that, according to TPB, intention is the best predictor of behaviour. However, according to Schlumpf et al. [89] it is unclear whether additional factors are important behavioural influences. To model the intention to return, their survey included the question "How likely is that you will give blood again within the next 12 months", where responses ranged from "very unlikely" to "very likely" on a 5 step scale. Despite the intention to return, their survey included 30 questions regarding donor donation history, requests and/or appeals for blood donations, donation experience, motivating factors, the convenience of donation sites, altruistic behaviour, empathetic concern and donor identity. To determine which factors most contribute to donor return, blood donations from 2003 to 2004 were examined, and logistic regression was used to predict the actual return within 12 months.

For feature selection, they developed a two-stage process. The first stage consisted of using a stochas-

tic gradient boosting classifier to rank, according to the factor's importance, the best possible predictive factors of actual return behaviour, in order of importance. With that, they choose the 15 most substantial predictive factors of actual return in order of importance. After that, they developed a series of logistic regression models and used the most important factors identified by the gradient boosting machine as independent factors, separately, in three groups. For each group, if any of the three factors were significant, the next group of three factors with lesser predicting strength (per the gradient boosting analysis) than the previous three was added into the model. The process was stopped if all three factors in a group were considered not significant ($p > 0.05$) [89]. Features no longer significant at that stage were eliminated from the model. At the end of the process, they ended up with 8 features considered essential for the task. Their results show that the prior donation frequency, intention to return, donation experience, and convenient location to donate appear to predict donor return significantly.

Kheiri et al. [56] study investigated, among other things, which factors affect donor return. The study was conducted in Iran, and 846 donors who had donated blood for the first time from March 2008 to March 2009 were sampled. The data used included age, gender, body weight, marital status, education, place of living, job class, and blood-related factors such as blood type and RH. A logistic regression model was used, with the dependent variable being binary: 1 for those who returned to donate at least once, and 0 otherwise. According to their model, body weight and job class had a significant effect on return to donation.

Lattimore et al. [60] presented the demographic characteristics and geographic distribution of blood donors in England and North Wales during 2010 and 2011. National statistics were used to estimate the rates of blood donors by region per 1000 residents. Their dataset included donor demographic data, such as the donor postal code, date of birth, ethnicity, donation date and type. New donors were identified by looking at the first donation for each donor. The postcodes were used to map individuals to their geographic area of residence, and that area of residence was therefore used to compute a discrete variable regarding the type of area of residence: urban or rural. Their results show that the residents in rural areas gave a significantly higher number of donations when compared with those living in urban areas.

They have also used a logistic regression model to study which demographic factors mostly contribute to new donors coming back following up 6 months of their first donation. Their models have shown that men were more likely than women to return within 6 months after their first donation; regarding age, the odds of returning increased with each successive age group. Furthermore, new donors residing in rural areas were more likely to return than those living in urban areas. As well known in the literature, minorities populations (such as immigrants, refugees, and individuals with non-white ancestry) are underrepresented in the blood donors population in many western countries [57, 71, 99]. Lattimore et al. [60] argue that the lowest repeat donation rate among those living in urban areas could be due to the more ethnically diverse population in those same areas.

Other studies have analysed the correlation between the first donation and subsequent donations. These studies differ from the ones mentioned above in two ways: they used mainly survival analysis methods; and, instead of using a dataset of blood donations to predict donor return within a given time,

they used datasets of blood donors, to study the relationship between the first donation and subsequent donations.

James et al. [51] determined that the likelihood of attempting a subsequent donation during the first year correlated positively with being older, male, RH negative and having a history of prior donation attempts. Perhaps the most important finding of their study is the identification of changes in the probability of donor return depending on the time that has elapsed since the first donation. This finding strongly suggests that there may be a specific 'window of opportunity' associated with donor characteristics, resulting in an increased propensity to attempt to donate blood [51]. In another prior study, James et al. [50] also conclude that the length of the interval since the previous donation affected the likelihood of a subsequent donation. Other studies had reached similar conclusions, for instance, Ownby et al. [75] also found that older donors, as well as those who were more highly educated and those with an RH negative blood type, returned more frequently [91]. The same study also reported that for first-time donors, shorter intervals between the first and the second donation were associated with an increased donor return [75]. These relationships between inter-donation intervals and the number of future donations may prove useful in understanding return behaviour and developing donor recruitment, and retention strategies [75].

Alkahtani et al. [4] performed a similar study, where they analysed which factors contribute most to retaining donors. To do that, they used machine learning algorithms such as logistic regression, Random Forest (RF) and SVM to develop and evaluate models for classifying blood donors as return and non-return. They used data collected by a public hospital in Saudi Arabia, with the features for the classification task being donors demographics such as age, gender, nationality and city, as well as first donation date and period in a month (from the first to last donation dates, within the period under study). Their three classifiers reached AUC of around 94%.

2.5.3 Discussion

Classification models have two main goals [15]: explainability and predictability. The former aims to provide some kind of understanding between the input features and the output variables, while the latter aims to forecast the output for a new given input as accurately as possible [86]. For example, simple linear classification models, such as linear or logistic regressions, are easy to understand and explain but typically perform worse than non-linear models [18, 109]. On the other hand, non-linear models, such as SVM or MLP, are capable of capturing more complex relationships in the data, and therefore they often outperform simple linear models in predictive performance [18, 86]. However, these models are often considered black boxes because they do not provide any direct explanation for their predictions [18]. As a result, they do not provide enough insight into the classification, which is needed in sensitive domains, such as life and social sciences. Thus, simpler models, such as linear ones, are still preferred in many areas due to their simplicity and interpretability.

Regarding the understanding of the determinants of donor behaviour, one is more concerned with explainability than predictability. Thus, it is more beneficial for blood agencies to know the patterns and

determinants of blood donation than using the models to predict whether a donor would donate in the future. Furthermore, with the patterns and determinants identified, blood agencies can better recruit and maintain blood donors.

The only above-mentioned study that used complex models was the study made by Alkahtani et al. [4], in 2019. They used non-linear models such as SVM and RF. However, they did not use any explainability technique, and thus, no conclusions regarding the importance of demographics were taken. Furthermore, they computed their target variable from a column that was further used as a feature during the training. This is a bad practice, because by doing so, it would be easy for any classifier to learn the rule that was used to create the target variable. Thus, their models are probably ignoring the demographic variables and just looking at that feature.

Nevertheless, different studies [42, 56, 60, 89] analysed the determinants of donor's return by using regression models, which are explainable. However, none of them used any model validation technique.

According to Breiman [15] predictive accuracy is the primary criterion for how good the model is in fitting the data. Breiman [15] states that, if models are a poor emulation of the mechanism that generated the data, i.e., if the model does not fit well the data, then the conclusions drawn from it may be wrong.

2.5.3.1 The two modelling cultures

Breiman [15] contrasted the 'two cultures of modelling', which are summarized in Figure 2.2. Figure 2.2 a) shows a vector of input variables **x** and a box which represents a 'nature function', i.e., the underlying mechanism that generated **y** given **x**. According to Brieman [15], there are two different approaches towards the goals of explainability and prediction: the data modelling approach and the algorithmic modelling approach, i.e., the machine learning approach.

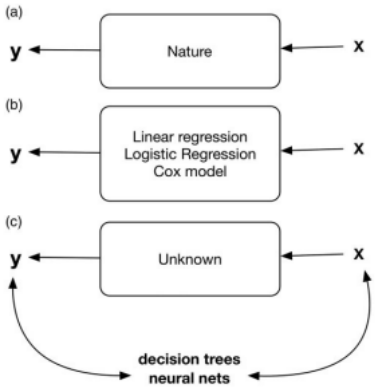

Figure 2.2: The two cultures of statistical modelling. a) the starting point; b) the data modelling approach; c) the machine learning approach. Source: Breiman [15] and Veltri [102].

The data modelling culture assumes a stochastic data model for the nature functions. In this approach, the values of the model parameters are estimated from the data and then it is used for explainability or

prediction. Figure 2.2 b) shows that approach. It has as input **x**, then simple linear models, such as linear as logistic regressions, are applied to understand or predict **y**. According to Breiman [15], that approach has at its heart the belief that, by looking at the data, one can reasonably invent a good parametric class of models for the complex mechanism devised by nature that generated the data. As stated by Breiman [15]:

> "With data gathered from uncontrolled observations on complex systems involving unknown physical, chemical, or biological mechanisms, the a priori assumption that nature would generate the data through a parametric model selected by the statistician can result in questionable conclusions that cannot be substantiated by appeal to goodness-of-fit tests and residual analysis."

In other words, the data modelling approach has little consideration on whether the model could have generated the data on hand or whether it fits the data. Instead, it uses goodness-of-fit tests and residual examinations as a model validation technique. However, problems have been identified regarding the usage of these techniques to quantify how well a given model fits the data. For instance, goodness-of-fit tests have demonstrated not to reject linearity when the data is non-linear, while residual analysis was shown limited to four or five predictors [15]. Willmott [107] also criticized the usage of some of those methods as a model validation technique. He showed that the usage of statistical tests, such as the R^2, to evaluate models predictive abilities, and as such their ability to fit the data, were insufficient. As a result, misleading conclusions may follow from data models that pass goodness-of-fit tests and residual checks [15], but that are not evaluated on a test set.

On the other hand, there is the algorithm modelling approach or the machine learning approach, as seen in Figure 2.2 c). The analysis in this culture considers that nature produces data in a black box whose insides are complex, mysterious, and at least, partly unknowable [15]. Unlike the data modelling approach, the machine learning approach assumes that the observable data is drawn independent and identically distributed from an unknown distribution. The goal of this approach is to find a function $f(\mathbf{x})$, that for future **x**, $f(\mathbf{x})$ will be a good predictor of **y**.

The machine learning approach assumes predictive accuracy on unseen data as the primary criterion for model validation. This is the most obvious way to see how well a given model fits the data: the model is trained on one dataset (training set), while another small dataset (test set) is used for model validation. This is known in the literature as the holdout method. Its starts by splitting the dataset into training and test sets. Then, the model is trained on the training set and therefore used to make predictions on the test set. After that the ground truth **y** is compared with the model predictions **y'**. The closeness of **y** and **y'** is, therefore, a measure of how good the model is in emulating the mechanism that generated the data. In resume, the model is trained using the training data, and then it is used to make predictions on a test set, and afterwards is evaluated regarding how good its predictions were.

Those two different approaches in dealing with data modelling problems pose a dilemma: models that best emulate nature in terms of predictive accuracy, such as neural networks, RF or SVM, which

are commonly used by those who follow the machine learning approach, are also the most complex and inscrutable. However, according to Breiman [15], this dilemma is not well framed because framing the question as the choice between interpretability and predictability is an incorrect interpretation of what the goal of statistical modelling is. Breiman [15] stated that the goal of statistical modelling is not just interpretability, but accurate information:

> "The point of a model is to get useful information about the relation between the response and predictor variables. Interpretability is a way of getting information. But a model does not have to be simple to provide reliable information about the relationship between predictor and response variables; neither does it have to be a data model."

In other words, one must first ensure that a model fits the data well, and after that, be concerned about interpretability. Otherwise, misleading conclusions could emerge when the model is interpreted without ensuring that it emulates the data.

Nonetheless, the machine learning approach and its methodology also has some problems. As stated by Flask [37], the perfect model for a given dataset often does not exist. In many cases, the data is 'noisy' – examples may be mislabelled, or features may contain errors – in which case it would be detrimental of trying too hard to find a model that correctly classifies the training data because it would lead to overfitting, and hence not generalising to new data [37]. Other times the used features are not informative concerning the task being solved. Based on that, machine learning model selection and validation are crucial tasks, with the former being defined as the process of searching for the optimal values for a given set of model hyperparameters.

Some problems were identified with the usage of the holdout method for model validation. It was shown that it might introduce some bias in the predictive accuracy, for instance, due to the random way the test set is chosen [37]. The cross-validation technique [95] proved to overcome this issue. It can reduce the bias imposed by the holdout method by training and testing in multiple folds of the dataset. It works as follows: the data is randomly split into k equal-sized folds, and k - 1 folds are used for training, whereas the remaining fold is used for model validation. This process is repeated k times by training k models and using each fold once for model validation. In the end, the average test set performance, and its standard deviation, are computed. By averaging overtraining sets, one gets a sense of the learning algorithm's variance and its generalization capabilities.

However, if a single procedure of k-fold cross-validation is used for both model selection and validation, the same test set is used for two different tasks, and therefore the model tends to overfit the test data, which could lead to an optimistic bias in model validation. Nested cross-validation was shown to reduce these biases [101]. The nested cross-validation has an inner loop cross-validation nested in an outer cross-validation loop. While the inner loop is responsible for model selection, the outer loop is responsible for model validation. Varma et. al. [101] have shown that this approach can significantly reduce the bias when compared to a single cross-validation procedure for both model selection and validation.

Chapter 3

Data

The raw data used in this thesis was provided by IPST and consisted of a CSV file with around 2.15GB size, 5 787 731 records and 55 columns. It contains blood donations collected in Portugal by blood centres from 1970 until August 2020.

Each record contains the date of collection (without time) and features regarding donors demographics, such as donor sex, race, nationality, marital status, profession, place (district, council and parish) and donors date of birth, blood type, the first four digits of the postal code as well as the postal location regarding donors place of living.

The dataset also includes columns about the collection site, donation type and adverse reactions. Regarding the collection site, it includes columns regarding its type (fixed, mobile or advanced site), the place where the collection was done (such as a school, fire department or a company) and the location of the collection site (district, county and parish). Regarding the donation type, the dataset included one column that indicates whether the donation was standard, apheresis or autologous. It also included other columns regarding the blood component collected, such as whole blood or platelets. Finally, regarding adverse reactions, the dataset included 18 columns.

The data processing and analysis were performed using Python[1] as programming language and PowerBI[2] for some data visualizations. Seaborn [3] and Matplotlib [4] Python packages were used for data visualization.

This chapter is organized into two different sections. Section 3.1 presents the preprocessing made to the dataset provided by the IPST, to increase data quality. It includes tasks such as cleaning the data, outliers removal and computing new variables. Section 3.2 includes the EDA made to the whole dataset to get an overview of the dataset and its main columns.

[1] https://www.python.org/
[2] https://powerbi.microsoft.com/
[3] https://seaborn.pydata.org/
[4] https://matplotlib.org/

3.1 Data preprocessing

Low-quality data will lead to low-quality mining results [46]. Garbage in, garbage out, is the fundamental principle of data analysis [94]. The road from raw data to a clean, analysable data set can be a long one [94].

Real-world data tend to be dirty, incomplete, and inconsistent. Data preprocessing techniques can improve data quality, thereby helping to improve the efficiency of the subsequent mining process [46]. Data preprocessing is an important step in the knowledge discovery process, because quality decisions must be based on quality data [46].

3.1.1 Noisy data

Just because something is written down doesn't make it correct [94]. Noisy data can be defined as unwanted variance or some random error that occurred in a measurable variable [12]. There are many possible reasons for noisy data. The data collection instruments used may be deficient. There may have been human or computer errors occurring at the data entry level. Also, errors in data transmission can also occur [45]. Therefore, noisy data in columns considered crucial for different data analysis and mining tasks were identified and removed.

Data regarding blood donations made before 1995 were removed since they represent just 1194 blood donations. The dataset also contained 138592 blood donations made in 2020, until the end of August. That data was removed since it was incomplete, i.e. one does not have access to the blood donations made during all the year; and because 2020 was an atypic year due to the COVID-19 pandemic, which can have an impact on future data analysis.

Regarding donor demographics, noise was identified in the donor's date of birth column. It contained donors born centuries ago, or in the future. Donors age was computed based on the blood donation date and donors birth date, for each donation. Figure 3.1 shows a boxplot regarding the donors age distribution. It is possible to find outliers, such as donors with negative age or above 100 years old. According to the IPST, only people over 18 can donate blood. The dataset contains 1327 blood donations in which donor age is below that threshold. Based on that, those records were removed. Also, and according to IPST, donors over 65 years of age can donate blood with a doctor's authorisation. The World Health Organization Europe stands that the standard age limits for blood donation are 18-65 years of age, but in some European countries, it is 17-70 [1]. Based on that, 2104 blood donations made by donors with more than 70 years of age were also removed.

3.1.2 Missing data and data discrepancies

It is often found that many of the tuples do not have any recorded values for some attributes [12]. Multiple features that we consider critical for data mining and analysis contained missing data. This subsection

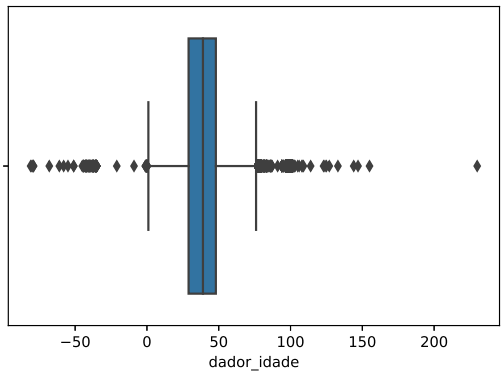

Figure 3.1: Donor's age distribution.

will handle missing values regarding columns considered crucial for a different kind of data analysis. A more in-depth analysis of the missing values is done in each chapter, considering the data used for each task.

Regarding demographics, donors sex had 56 and donors race had 1 911 433 (33,59%) missing values. The two columns regarding blood type and Rh factor type had 277 465 (4,87%) missing values together. The marital status had 50 387 missing values (0,89%). Finally, donors postal location and postal code had, combined, 1320 missing values.

The donation type column had 1 039 080 (18,26%) missing values, while the column regarding the blood component collected had 1 065 161 (18,72%) missing values. The collection site type had 470 495 (8,27%) missing values, and the place where the collection was made had 369 453 (6,49%) missing values.

According to [45], different approaches could be adopted for dealing with missing data, such as filling the missing values manually; ignore them; use a global constant; use the attribute's mean; or use the most probable value by using inference based tools.

3.1.3 New columns

Age and age groups were calculated based on donation date and donors date of birth. New donors were identified by assigning each record with a flag indicating if it corresponds to that donor's first donation. For new donors, the interval between the first and second donations was calculated.

According to IPST, a regular donor is one who donates at least once in a year. Based on that, donations from regular donors were identified by assigning a flag indicating if the current donor made a previous donation within the previous 12 months.

The raw data did not include a blood type feature, but two different features regarding blood group

('A', B', 'AB' or 'O') and Rh factor (positive or negative blood). Therefore, based on those two columns, the blood type was computed.

The dataset included a variable regarding donors profession. However, that variable contained 2 991 unique values, with almost 200 of them containing only one record. Therefore, a new variable regarding the employment status was computed based on that one.

3.1.4 Georeferencing donor's address

Georeferencing the donor's address is an important task that would add more quality to the data. Therefore, this task is needed to obtain valuable geographic information regarding the donor's place of residence, such as the district, council and parish. With that information, one could further analyze it to extract valuable insights, such as the importance of the donor region of residence for frequent blood donations.

To obtain an estimand of donor's place of residence's, two columns from the blood donation dataset were used: one column regarding the first four digits of the donor's postal code and a column regarding the donor's postal location. To better understand the preprocessing process, one needs to understand how the Portuguese territory is organized.

3.1.4.1 Portugal administrative territory & postal codes

The Portuguese territory has a complex and unique administrative structure that consists in districts (distritos) that are sub-divided in counties (municípios), which in turn are sub-divided in parishes (freguesias) [17]. In 2013, Portugal suffered a *Reorganização Administrativa do Território das Freguesias* (RAFT). Prior to that reorganization, there were 18 districts in the mainland, plus two autonomous regions (Madeira and Açores), sub-divided into 308 counties which, in turn, were subdivided into 4 260 parishes [17].

RAFT was a territory reform implemented in 2013, that consisted in the aggregation and/or the merger of parishes and/or the change of its boundaries. The main reasons to implement RAFT were the settlements with international entities like the European Commission, the European Central Bank and the International Monetary Fund – Troika – that intervened in the country to help overcome the 2008 crisis [17]. In addition, by implementing the territory reform, the country would reduce expenses by increasing local efficiency. Prior to RAFT implementation, Portugal had 4260 parishes. RAFT has reduced a total of 1 168 parishes to a total of 3 092 parishes.

Regarding the postal codes, in Portugal they are formed by a sequence of seven digits, in the format 0000-000, followed by a postal location up to 25 characters. The first digit represents one of the nine postal regions, followed by two digits that regards the postal distribution centres. The fourth digit is zero if the post code represents a capital of a municipality, otherwise it is zero. In the case of a designated address, it could take any other value. The last three digits sort designated addresses and buildings.

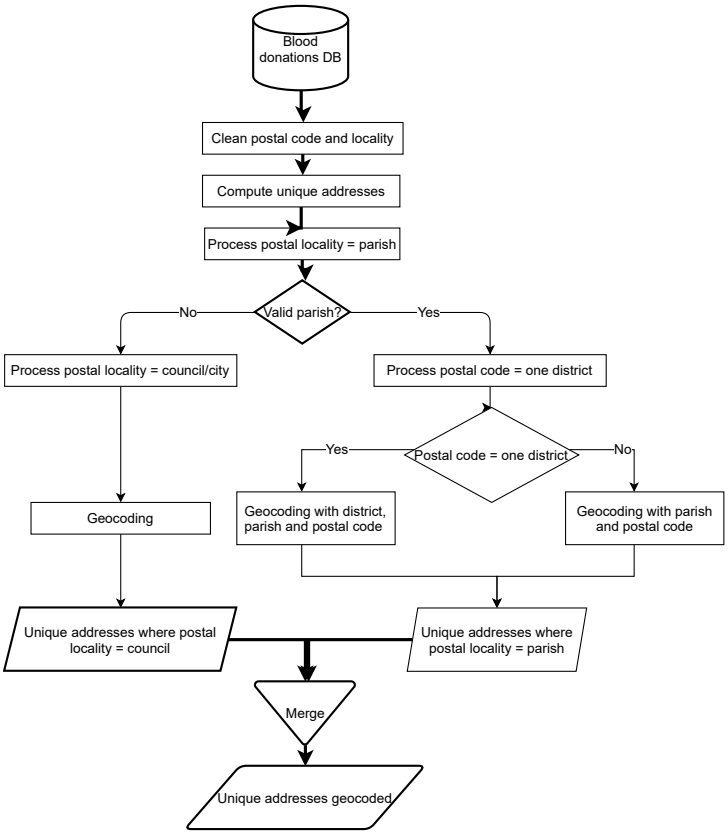

Figure 3.2: Processing donor's address.

It is important to note that the postal location could have different meanings regarding the person's living place. For instance, those living in urban centres usually fill that value with the city's name or the council where they live. On the other hand, those living in rural areas could fill that value with the parish or village names. This turns the process of computing the donor's place of living even more difficult.

Also, in many cases, the first four digits of the postal code could not be directly associated with one district or council. If they were directly associated, inferring the district or council from the first four digits of the postal code would be easier in the sense that one postal would have only one district or council. For instance, Aveiro and Coimbra districts both have postal codes beginning with 3020.

3.1.4.2 Processing donor's address

Figure 3.2 is a flow chart that describes the processing made.

Because both columns contained some noise and discrepancies, the first stage of this process consisted of cleaning them. The postal code contained some records with letters instead of numbers. Also, invalid postal codes, such as ones one more than four digits, were found. The postal location also had invalid

values such as numeric digits, double white spaces, or accents in the wrong positions. Therefore, those two columns were cleaned.

After that, all the unique addresses regarding the combination of those two columns were computed. The cleaned postal location column was then matched against known parish names to check if they corresponded to any known parish. The current parish names, i.e., the parish after the RAFT, were obtained by using external data from the Portuguese Open Data portal [5]. To get the old parishes names, i.e., those after the RAFT, data from the *Direção Geral do Território* [6] was used. The matching process has taken into account the parishes' administrative reorganization in Portugal RAFT, by matching the postal location with both the old and new parishes names.

The matching process could be described as follows:

1. Stop-words, such as 'de,' 'do,' or 'das' were removed by using a Python package [7] that contained lists of stop-words per language. This step was done to avoid cases in which the same postal location was written with different stop-words. For instance 'Aldeia dos Capuchos', and on the other side, 'Aldeia do Capuchos';

2. The cleaned postal location was matched with known parish names, by using regular expressions. If all the words in the cleaned postal location match with a known parish name (old or new parish), then that postal location was assumed to be a valid parish;

3. At the end of that process we get the donor's addresses with valid parish names.

There were 8 733 unique addresses regarding the combination of the postal location and the first four digits of the postal code. From those unique addresses, 4 693 of them (53,7%) matched a known parish name.

3.1.4.3 Valid parish

For those addresses in which the postal location match known parish name, geocoding was further applied to get the district (when needed) and council/city of the given parish. The geocoder used was the LocationIQ [8]. LocationIQ is an open-source system that builds on top of Open Street Maps (OSM) data to provide a geocoding service. It receives a query, i.e. a named location (full addresses or named areas) and resolves it into a structured format that includes district, council/city, parish, and latitude/longitude of their whereabouts.

For the addresses in which the first four digits of the postal code match a single district, that district was obtained by merging it with external data [9] and given to the geocoder in order to increase the accuracy

[5] https://dados.gov.pt/en/datasets/freguesias-de-portugal/
[6] https://www.dgterritorio.gov.pt/cartografia/cartografia-tematica/caop?language=en
[7] https://pypi.org/project/stop-words/
[8] https://locationiq.com/
[9] http://centraldedados.pt/codigos_postais/

of its predictions. In that cases, the query used for the geocoder contained the district, parish and postal code. For the addresses in which the postal code corresponded to more than one district, the query contained just the parish and the postal code.

3.1.4.4 Parish not valid

For those addresses where the postal location did not match a known parish, it was checked whether they matched any known city/council. As said above, that could happen mostly in cases in which the donor's residence is in an urban area. In order to match the postal location with known council/city names, open data was used, and the same cleaning and matching process described above was applied.

At the end of that process, we got the postal location's addresses matching a given city/council name. After that, those records were merged with open data to get the respective district. Then, the LocationIQ geocoder was used to get the latitude and longitude of that city/council.

In total, from the 8 733 unique addresses, in 208 of them, the postal location matched with a known city/council name.

3.1.4.5 Lost data

At the end of that process, there were 4 996 (56,8%) addresses for which the postal location matched a known parish, and 208 addresses for which the postal location matched a given city/council. So, in total, there were 5 204 (59,2%) addresses for which, at least, the city/council of the donors place of residence was known. Those addresses were then geocoded in order to get the latitude and longitude regarding the place of residence.

At the end of that process, there were 3 580 (40,8%) addresses for which the postal location did not match a parish or a city/council. This was due to different reasons:

- For some addresses, the postal location did not match any known parish or city/council, it contained just random text;
- In other cases, despite they match a parish or a city/council, the postal code of the address was in a different region of the given postal location, or vice versa. Those records were not take into account, because one could not be sure whether the postal location or the postal code is wrong;
- Some addresses, instead of having the locality name written, had an abbreviation;
- Some addresses had spelling errors;
- Some addresses, instead of having the parish or city name written, had the village or the road name.

3.1.4.6 Merge with donations database

The process's final stage consisted of merging the unique addresses dataset geocoded with the blood donations dataset. In total, 4 324 578 blood donations (around 75,9%) ended with the district, council and parish regarding the donor's place of residence. 4 860 844 blood donations (around 85,4%) ended just with the district and council/city regarding the donor's place of residence.

3.1.5 Georeferencing collection site

Information regarding the place where each donation was made could be very useful. One could then analyse and extract valuable insights from that information. For instance, Crawford et al. [23] used collection site geo-information to study the importance of the distance between the donor's region of residence and the collection site and donor's frequency of donation.

IPST provided some geo-information regarding the collection sites, such as the full address, latitude, and longitude. They provided that data for a total of 1 409 collection sites. That data was then merged with the blood donations database to obtain each donation's collection site latitude and longitude. That process is represented in the flowchart below.

The first stage of the process consisted of computing, for each donation, the collection site unique identifier. It was computed based on three existing columns presented on the blood donations dataset. Next, those columns were cleaned and merged to create the unique identifier column. From all blood donations accounted in the dataset, 377 245 of them (around 6,7%) did not have a valid collection side ID. After that, the unique ID's of the collection sites, presented in the blood donations database, were computed and merged with an external file provided by the IPST that contained the geo-information for some collection sites.

As said above, IPST provided geo-information regarding 1 409 collection sites. However, after computing the unique collection site IDs from the blood donations dataset, it was possible to note that there were a total of 5 853 different collection sites in it. It means that, in the blood donations dataset, there were blood donations from 4 444 collection sites for which geo-information regarding the collection site was not available.

After merging the unique collection site ID's with the blood donations dataset, it was possible to check that those 1 409 collection sites with geo-information corresponded to 2 254 127 blood donations (around 40%). On the other hand, those 4 444 collection sites for which geo-information was not provided corresponded to 3 390 183 blood donations (around 60%).

3.1.6 Rural or Urban Areas

Above was described the data processing made to obtain geo-information regarding the collection site and the donor's address. In total, 75,9% of the blood donations were geocoded regarding the donor's

address and around 40% regarding the blood collection site address. With that information, one could then merge it with external and open data to increase the data quality and to increase the quality of the analysis made.

Lattimore et al. [60] assigned their donor's postcodes data to rural or urban settlements to increase the quality of their analysis and to try to evaluate whether donation frequencies were different in those different regions. They conclude that donor's residents in rural areas returned more frequently in the first 6 months after their first donation when compared with those residents in urban centres. A similar approach was adopted to retrieve information regarding the donor's residence based on that study.

The concentration of both the population and the economic activity in urban areas led to the development of a classification mechanism to delimit representative units of the urban dimension. As a result, *Instituto Nacional de Estatística* (INE) developed a territorial nomenclature about the degree of urbanization of the Portuguese parishes called TIPAU, which if free available [10]. It classifies parishes tripartitely in *Áreas Predominantemente Urbanas* (APU), *Áreas Maioritariamente Urbanas* (AMU) and *Áreas Predominantemente Rurais* (APR).

The blood donation data was merged with TIPAU data to obtain the degree of urbanization regarding both the collection site place and donors place of living. A variable named BFT was created. This variable indicates the degree of urbanization of the parish where the donation was made. Regarding donor's place of living, a variable named DFT was created, which indicates the degree of urbanization of the parish regarding donor place of residence. By adding this information to the blood donation data, one could further test, for instance, if different donor's behavioural patterns exist in areas with different levels of urbanization.

3.1.7 Shapefiles and spatial data

In order to explore the geographic distribution of blood donations, shapefiles were used to get the shape of the Portuguese councils and districts, and afterwards, they were converted to TopoJSON to plot them on PowerBI. The districts [11] and councils [12] shapefiles were obtained via the Portuguese public administration open data portal [13] .

A shapefile is a standard geospatial data format for Geographic Information System developed and regulated by the Environmental Systems Research Institute. It can describe points, lines, and polygons. A point is zero dimension and has only the property of location. A line is one-dimensional and has the property of length and location. A line has two end points and may have additional points in between to mark the line's shape. A polygon is two-dimensional and has the properties of area, perimeter, and location [19].

[10] https://dados.gov.pt/en/datasets/tipologia-das-areas-urbanas-2014/
[11] https://dados.gov.pt/pt/datasets/distritos-de-portugal/
[12] https://dados.gov.pt/pt/datasets/concelhos-de-portugal/
[13] https://dados.gov.pt/en/

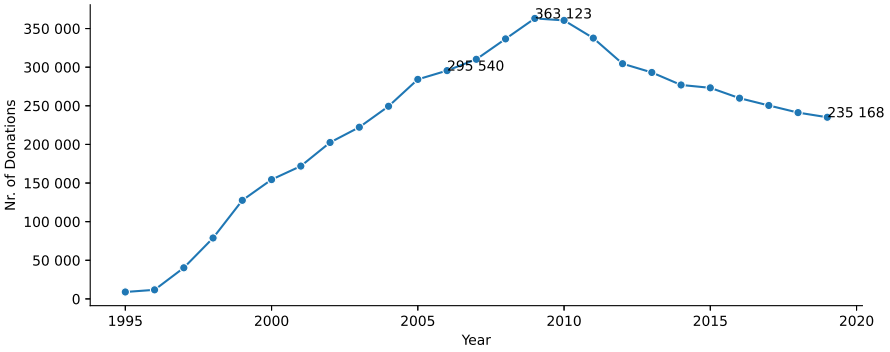

Figure 3.3: Total number of donations across time.

The shapefile is not a single file but a collection of files stored in the same directory with a common filename prefix. It contains three mandatory file: the main file (*.shp*), the index file (*.shx*) and the database file *.dbf*. The first two describe the geometry: the *.shp* file stores the feature geometry, with each record describing a shape with a list of its vertices. The *.shx* file maintains the feature geometry's spatial index, and the *.dbf* file contains feature attributes with one record per feature [30].

To plot a shape map in PowerBI, one needs to use a TopoJSON file format instead of a shapefile. So, the districts and councils shapefiles were converted to TopoJSON by using a public service [14].

3.2 Exploratory data analysis

EDA is the process of extracting meaningful information from the data, by investigating the dataset, elucidating subjects, and visualizing outcomes. The term EDA was first coined by John W. Tukey for describing the *act of looking at data to see what it seems to say* [69]. EDA is an approach to data analysis that applies a variety of techniques to maximize specific insights into a dataset, reveal an underlying structure and extract significant variables from it [70]. It allows data visualization in order to understand it as well as to create hypotheses for further analysis and focuses on creating a synopsis of data or insights for the next steps in a data mining project [70]. In this section, EDA is performed by analyzing the main columns of the dataset.

Figure 3.3 shows the total number of blood donations during time. By analyzing it, it is possible to note that total blood donations have decreased over the last ten years. It had a spike in 2009 with 363 123 donations, and since then, it has decreased continuously. In 2019, 235 168 donations were made, representing a decrease of 35% over 10 years.

[14] https://mapshaper.org/

3.2.1 Regular and new donors

Figure 3.4 (A) show the total number of donations from regular donors. Recall that, according to IPST, a regular donor donates at least once a year. Blood donations from regular donors were identified by assigning a key, for each blood donation, if the donor had donated in the previous 12 months, counting from the day of the given donation.

It is possible to see that 3 970 745 (69,7%) blood donations were assigned as donations from regular donors. IPST should make efforts to maintain those donors since they are their major asset. Also, according to the World Health Organization (WHO) [97], regular donors are safer than first-time donors because they are informed, committed and regularly screened for diseases.

On the other hand, Figure 3.4 (B) shows the number of donations that came from new donors. It is possible to see that 4 648 228 (81,6%) blood donations did not come from new donors.

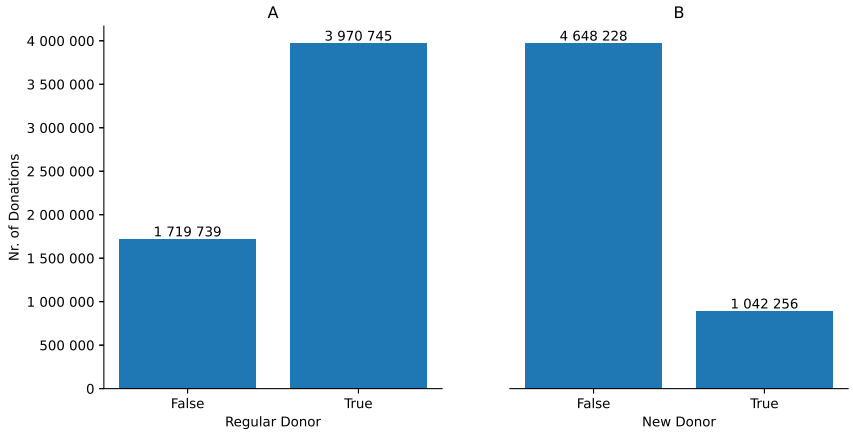

Figure 3.4: Donations from experienced donors (A) and new donors (B).

Figure 3.5 shows the number of new donors for each year. It is possible to see that the number of new donors per year is decreasing in the last 10 years, as happens with donations. It had a spike in 2005, accounting for 60 791 new donors that year. In 2009 there were 60 774 new donors, while in 2009, there were 26 901, which represents a decrease of around 44% in the total number of new donors in ten years.

3.2.2 Donation type

According to IPST, a donor is an individual with healthy habits and behaviours who voluntarily appears in the blood service to donate blood. There are two types of blood donations: homologous and autologous. Homologous donations are when a person donates blood to be used by another individual. They could be divided into whole blood (i.e. standard) and apheresis donations. Apheresis donations differentiate from

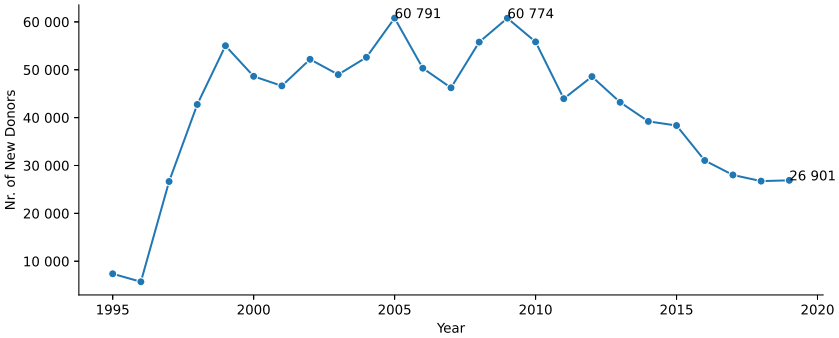

Figure 3.5: New donors over time.

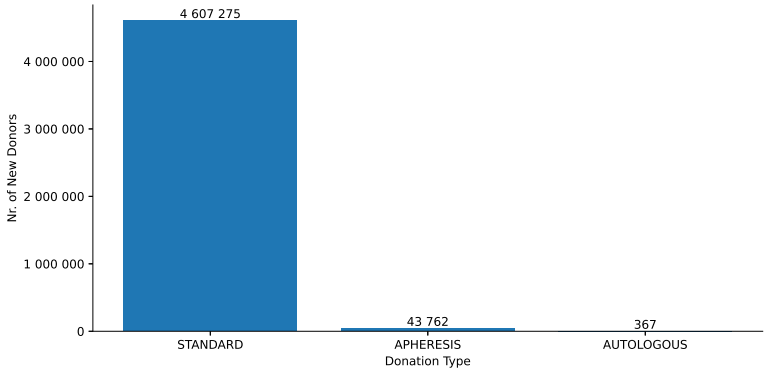

Figure 3.6: Donation types.

whole blood donations because just a specific blood component is donated, such as plasma, platelets, or red or white blood cells. On the contrary, autologous blood donations are when a person donates blood for their own use. Figure 3.6 shows the distribution of the donation type column.

It is possible to observe that 99,1% of blood donations are whole blood donations. The dataset also included 43 762 apheresis and 367 autologous donations.

3.2.3 Demographics

Table 3.1 contains data regarding three donor demographic characteristics and their respective distribution in four different populations: the population of all blood donations, the donations from new donors and regular donors, and the population of donations in which adverse reactions were recorded.

By analysing the total population, females accounted for 48,14% of total donations while males accounted for 51,85%. The age group with more donations is between 35-44, accounting for 27,58% of

Table 3.1: Donations demographics.

	All donations		**New donor donations**		**Regular donors donations**		**Adverse reactions**	
Male	2 950 728	(51,85%)	507 377	(48,68%)	2 100 394	(52,89%)	10 208	(41,06%)
Female	2 739 756	(48,14%)	534 879	(51,31%)	1 870 351	(47,11%)	14 655	(58,94%)
18 - 24	860 616	(15,12%)	357 167	(34,27%)	418 155	(10,53%)	7 576	(30,47%)
25 - 34	1 282 580	(22,54%)	269 809	(25,89%)	812 367	(20,46%)	7 145	(28,74%)
35 - 44	1 570 357	(27,59%)	231 084	(22,17%)	1 143 282	(28,79%)	5 732	(23,05%)
45 - 54	1 287 792	(22,63%)	136 337	(13,08%)	1 018 961	(25,66%)	3 227	(12,98%)
55 - 70	689 139	(12,11%)	47 859	(4,59%)	577 980	(14,57%)	1 183	(4,76%)
A+	2 083 593	(38,49%)	315 340	(38,57%)	1 516 114	(38,51%)	9 735	(39,15%)
A-	404 045	(7,53%)	59 939	(7,33%)	294 501	(7,48%)	1 986	(7,99%)
O+	1 920 077	(35,47%)	288 593	(35,3%)	1 398 848	(35,54%)	8 349	(33,58%)
O-	407 771	(7,53%)	58 724	(7,18%)	298 396	(7,59%)	1 922	(7,73%)
AB+	154 972	(2,86%)	24 332	(2,97%)	111 820	(2,84%)	702	(2,82%)
AB-	30 199	(0,56%)	4 569	(0,56%)	21 825	(0,55%)	124	(0,5%)
B+	341 876	(6,31%)	55 270	(6,75%)	244 219	(6,20%)	1 624	(6,53%)
B-	70 486	(1,3%)	10 891	(1,33%)	50 531	(1,28%)	3 28	(1,32%)

total donations. The blood types with more donations were A+ and O+ with 38,49% and 35,47% of total donations, respectively.

A total of 1 042 256 (18,32%) blood donations came from new donors. Of those, 51,31% came from female donors, and 48,68% from male donors. It is possible to observe a considerable difference between the total population's age distribution and the new donor's population distribution. While in the total population the age group between 18-24 accounted for just 15,12% of total donations, this ratio increases to 34,27% for new donors. This means that new donors tend to be younger. Regarding blood type, there is no significant difference between its distribution in the total population and new donor's population.

Among all donations, 3 970 745 (69,78%) came from regular donors. They are more likely to be females (52,3%) than males (47,1%). Also, regular donors are generally older than the new donor's population. That difference is highlighted in Figure 3.7. Based on this analysis, more efforts should be made by IPST in order to retain new donors and turn them into regular ones.

A total of 24 837 (0,43%) adverse reactions were recorded. Among this total, 41,06% were from male donors, while 58,95% were from female donors. The majority of adverse reactions were recorded for donors within the 18-24 age group (30,47%). There is no significant difference between the blood type distribution among the population with adverse reactions and the total population. Just 28,82% of

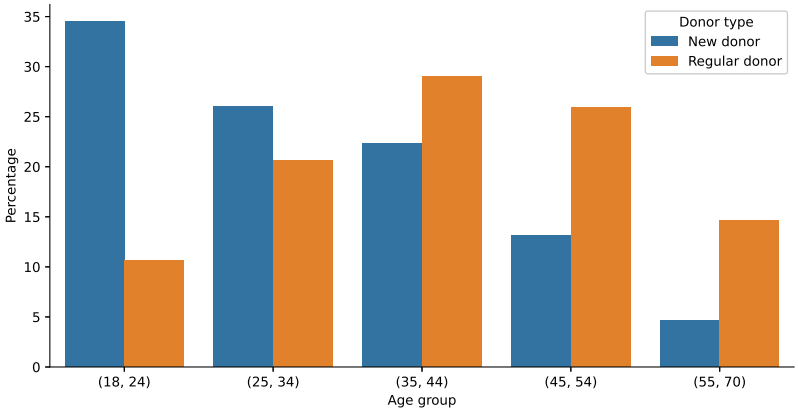

Figure 3.7: Age distribution by donations from new/regular donors. The values were normalized.

the reactions were recorded in new donors.

Figure 3.8 shows donors marital and employment status distribution, according to the number of total donations. It is possible to observe that around 61% of total blood donations came from married donors, while around 29% came from single donors. Figure 3.8 (B) regards the employment status, which is a variable that was computed based on the profession column available on the raw dataset. It is possible to observe that most of the donations came from employed blood donors, while just a few came from unemployed and students.

3.2.4 Collection site type

According to IPST internal documentation, donations sessions are organized in what is called 'harvest sessions'. Those harvest sessions could be described as a period in which a team of IPST professionals aims to collect blood or blood components.

Figure 3.9 shows the IPST harvest sessions types. They could be of two different main types: mobile and fixed. The mobile ones are located in a temporary (physical space) or mobile place (a van, for instance) situated outside a blood establishment, requiring the relocation of equipment, human and material resources, and these sessions last a specific time window. The fixed harvest sessions are regular, i.e. sessions that last the whole year, and they could be of two different types: fixed station, which means inside an IPST blood centre, or in an advanced station that is usually outside IPST blood centres.

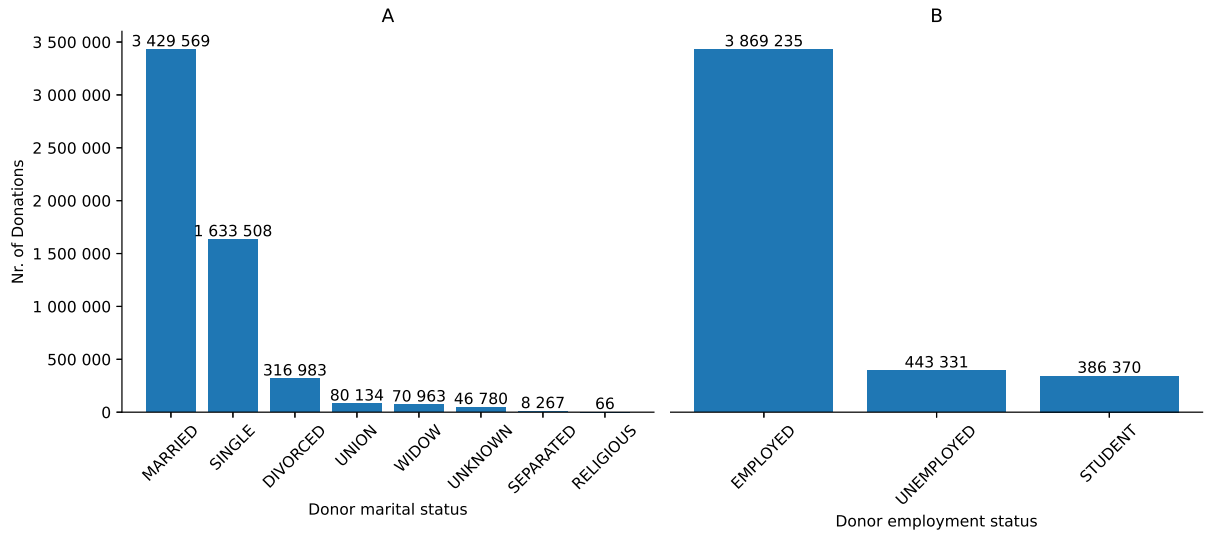

Figure 3.8: Donors marital (A) and employment (B) status.

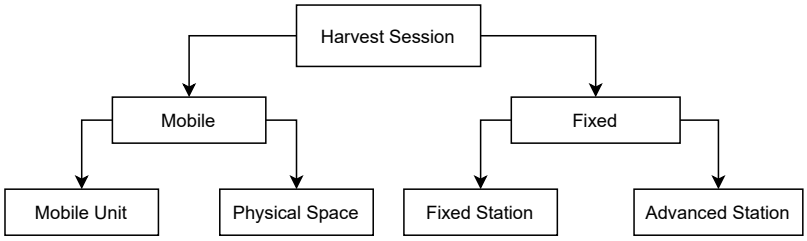

Figure 3.9: IPST harvest sessions types.

Figure 3.10 (A) shows the number of blood donations by collection site type. It is possible to see that 4 029 323 (70,8%) of the blood donations had a collection type of 'OTHER', while the rest 30% had values regarding the real types of collection sites that can be seen in Figure 3.9. The values assigned with 'OTHER' are probably a kind of default value taken from the transnational software from where the data was taken.

The variable regarding the collection site place initially had a cardinality of 19. However, 9 of those unique values were assigned less than 1 000 records. Because of that, a category called 'DIVERSOS' was created, which merges all those values. Figure 3.10 (B) shows the distribution of that variable. It is possible to observe that 2 583 367 donations had assigned 'OUTRA' value, which might be a default value. Following that, 827 954 donations were made in schools, while around 734 835 and 368 792 donations have taken place in localities, i.e. small villages, and companies.

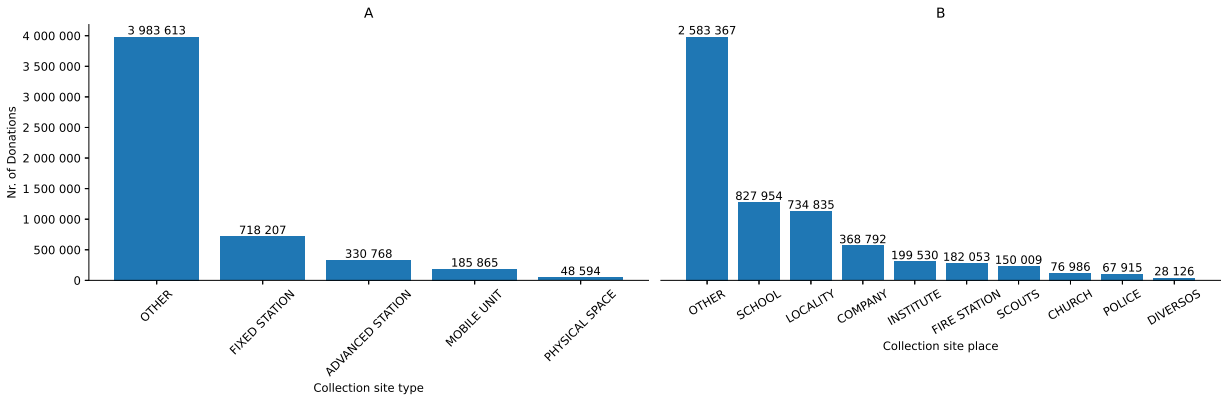

Figure 3.10: Collection site type (A) and place (B).

3.2.5 Geographic distribution

Figure 3.11 shows the geographic distribution of the blood donations regarding the collection site address. Lisbon, Porto, and Aveiro are the three districts with the most significant number of donations, accounting for a total of 1 095 769 (21,76%), 1 027 642 (20,32%), and 802 110 (15,86%) donations, respectively. On the other hand, Beja is the district with fewer donations, accounting for just 2 992 (0,07%) donations.

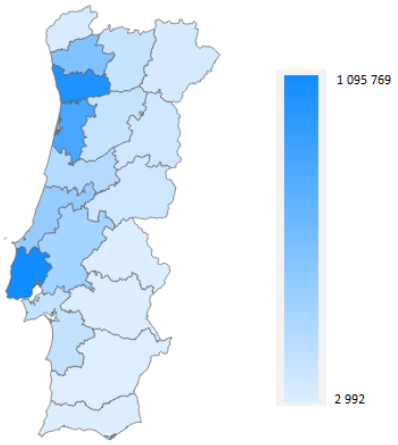

Figure 3.11: Blood donations geographic distribution.

Figure 3.12 shows the geographic distribution of the percentage of blood donations from experienced donors for each district. Aveiro is the district where the most considerable percentage was identified. Among the 824 410 total donations, 654 864 (79,43%) were from experienced donors. Following Aveiro comes Vila Real and Porto districts, where 77,48% and 71,66% of their total donations were from experienced donors. Lisbon, which is the district with more donations, accounts for just 61,04% of those from

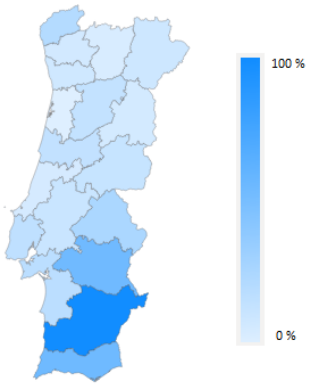

Figure 3.13: Geographic distribution of the percentage of blood donations from new donors.

experienced donors.

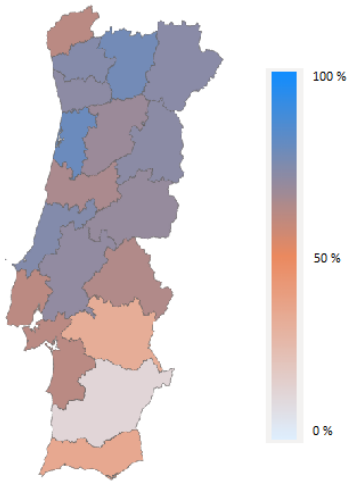

Figure 3.12: Geographic distribution of the percentage of blood donations from experienced donors.

Figure 3.13 shows the percentage of blood donations that came from new donors. As can be seen, Beja is the district with the highest percentage of new donors. From those 3 874 total donations, 3 090 (79,76%) were from new donors. Despite being the districts where more donations were made, Porto and Lisbon had a small percentage of new donors. In Lisbon, just 22,41% of the total donations were from new donors, while this ratio decreases to 17,18% in Porto.

Porto and Braga were the districts with more adverse reactions recorded. Among the total reactions identified, 37,62% and 20,11% were in Porto and Braga, respectively. Lisbon accounts for 13,33% of the adverse reactions recorded. Regarding blood type, there is no significant difference in its geographical

distribution.

3.2.6 Donations per 1000 residents

According to the WHO Europe, the average number of total blood donations in the 30 European countries that reported comparable data, rose from 34.7 per 1000 population in 2008 to 36.5 in 2010. In total, the average number of donations across the European Region ranges from 6 to 67.6 per 1000 inhabitants, with Denmark reporting the highest blood donation rate [1]. WHO Europe states that, for a country to be self-sufficient in national blood supplies, a country is estimated to need to maintain a minimum average of 20–25 experienced donors per 1000 inhabitants.

The two ratios mentioned above, the number of donations per 1000 residents and the number of experienced donors per 1000 residents, were computed and analysed. External data regarding the Portuguese population for each year was obtained from PORDATA [15]. That data was then processed to obtain just the population eligible to donate blood, i.e. those from 18 to 70 years old.

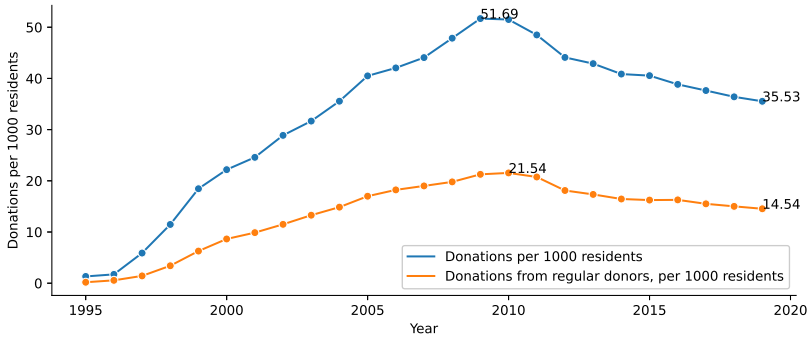

Figure 3.14: Donations per 1000 residents: in the whole population; and just donations assigned as regular donors.

By analysing the Figure 3.14 it is possible to see a similar trend to the one observed in Figure 3.3, with the number of donations per 1000 residents increasing until 2009 and decreasing since then. In 2009 it had a spike with around 51.69 blood donations per 1000 residents. In 2019 that ratio has dropped to 35.53. The number of experienced donors per 1000 inhabitants per year follows a similar pattern, with a spike in 2010 with an average of 21.54 experienced donors per 1000 residents. In 2019 it decreased to 14.54 experienced donors per 2000 residents, which is quite below the WHO Europe's recommendation for the country to be self-sufficient regarding its blood needs.

[15] https://www.pordata.pt/Portugal/Popula%c3%a7%c3%a3o+residente++estimativas+a+31+de+Dezembro+total+e+por+grupo+et%c3%a1rio-7

By taking into account the donors region of residence, the number of donations per 1 000 residents was computed. A ratio between the number of donations throughout 1995 to 2019 and the regional population for each council was established, which allowed determining the mean number of donations per 1 000 residents, per district and council. To compute the number of donations per 1 000 residents, the population estimates for each council were obtained via the PORDATA open data portal[16]. This data was then cleaned and merged with the blood donations dataset, and the ratios were computed.

Figure 3.15 shows the geographic distribution of the number of donations per 1 000 residents. Beja accounts for just 1.84 donations per 1 000 residents, being the district with the lowest ratio, while Aveiro is the district where the highest ratio was observed, accounting for 61,19 donations per 1 000 residents. Lisbon and Porto, the districts where more donations are made, account for around 28 and 29 per 1 000 residents.

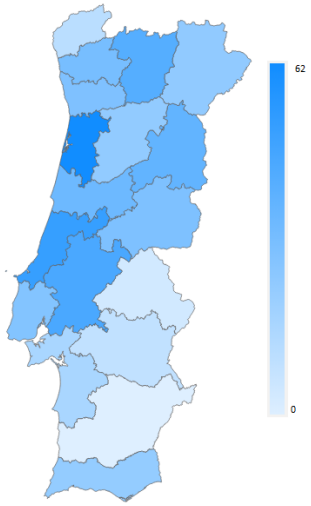

Figure 3.15: Donations per 1000 residents.

3.2.7 Rural of Urban areas

Figure 3.16 shows the distribution of the degree of urbanization for both donor's place of residence (Figure 3.16 (A)), and the collection site parish (Figure 3.16 (B)). Regarding DFT, it is possible to observe that most donations were made in APU and AMU areas. The same happens with BFT distribution, in which just 4% of the donations were made in collection sites located in rural areas.

[16] https://www.pordata.pt/Municipios/Popula%c3%a7%c3%a3o+residente++estimativas+a+31+de+Dezembro+total+e+por+grupo+et%c3%a1rio-137

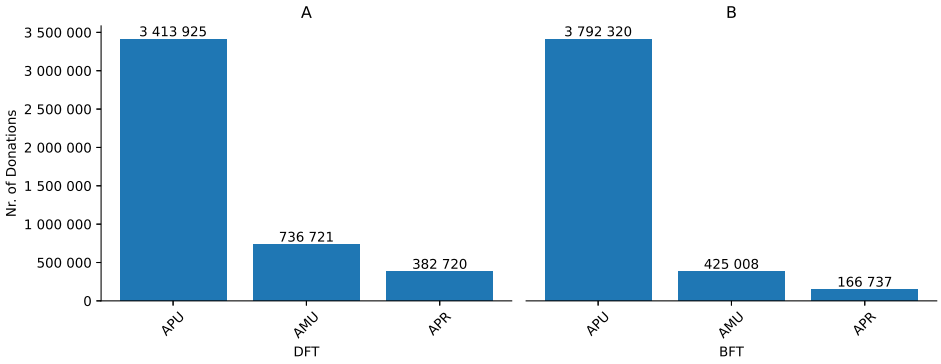

Figure 3.16: Donations per DFT and BFT.

Chapter 4

Predicting and understanding factors contributing to donor return

As we have seen in the Section 2.5.2, multiple studies analysed the determinants of blood donors return, by using mainly logistic regression models. In this task we propose to analyse the factors influencing donor return by using the most recent machine learning techniques. We hypothesise that a machine could learn the relationship between donor demographics, as well as other features, and the target variable indicating donor's return. This task could be described as a binary classification task, in which we wanted to predict whether a donor would donate blood in the following 52 weeks (one year) following their first donation.

Both the European Union (EU) and IPST make a distinction between new and experienced donors. The guideline on epidemiological data on blood transmissible infections (EMA/CHMP/B-WP/548524/2008) [1] from the European Medicines Agency recommends the EU members to distinguish between new and experienced donors. IPST followed the EU guidelines and adopted the same distinction [2]. The scientific literature also made a distinction between the two types of donors, to try to understand which factors mostly contribute to future donations among those different segments. For instance, there is a strong evidence [10, 21, 32, 33, 51, 78, 89, 90] that donor's future behaviour for experienced donors is mainly predicted by their past behaviour, while for new donors those metrics are not available.

Based on that, we believe that it makes sense to distinguish between new and experienced donors by creating different machine learning models for each group, to understand which factors motivate those two different groups of blood donors. Therefore, this task aims to separately identify the determinants of donor's future behaviour among those two groups.

The remainder of this Chapter is organized as follows. Section 4.1 describes the data used and Section 4.2 draw the methodology followed. Section 4.3 shows and discusses the results obtained. Finally, Section 4.4 analysis the feature importance for some of the models trained, by using the most recent model interpretability techniques, and Section 4.5 illustrates the impact that donors past behaviour features had

[1] https://www.ema.europa.eu/en/epidemiological-data-blood-transmissible-infections

[2] http://www.ipst.pt/files/IPST/INFORMACAO_DOCUMENTACAO/manual_gestao_dadores.pdf

on the classifiers.

4.1 Data

The data used for this classification task consists of 583 728 blood donations made between 2016 and 2019. It was decided to use just data for four years due to different reasons. First, 2020 data was removed because it was an atypical year due to the COVID-19 pandemic [3]. Also, it would not make sense to analyze older data when new data is available. It would be valuable for IPST to know the current donor behaviours/patterns and not the old ones. Furthermore, the data from the last years had less missing values when compared to older data, which could enrich our analysis. There is also a computational reason. In total IPST provided around 5.8 million records, dating from around 1970 until 2020. This is a massive amount of data, which would require a substantial computational cost to train machine learning models with all of that. Despise that, it would not make sense to search for donor's behaviour for such a long time because their behaviours could have changed during those 50 years.

The features used for this classification task are presented in Table 4.1. It contains the features used for both the new and experienced donors segments. In addition, Table 4.2 contains the additional features that were used just for the experienced donor's segment.

Each record consists of a blood donation, and there were available features regarding five different subgroups. The first subgroup is about donor's demographics, including features such as donor's gender, age, nationality, civil status, working situation and blood type. The second subgroup is about donor's geographic information, such as their place of residence. That subgroup includes just the metric TIPAU, which gives a sense of the urbanization level of the donor's place of residence.

Then, the other three subgroups include information about the donation itself. For example, there are features about when the donation was made (Donation - temporal), features containing donation geographic information (Donation - geographic), such as if it happened in a metropolitan area, in a capital district or at a given CST (Lisbon, Porto or Coimbra). Finally, the last subgroup contains features regarding the donation that are neither temporal nor geographic, such as if an adverse reaction was recorded, the blood facility centre type, and the place where the donation was made.

Table 4.2 contains the features that were computed just for the experienced donor's segment because they measure aspects of the past behaviour, which are not available for new donors. Those features are indicators of the donor's past behaviour at the moment of each donation and were first computed by Ferguson [32]. According to Ferguson [32], adding those metrics to the machine learning models could likely increase their performance.

[3] https://en.wikipedia.org/wiki/COVID-19_pandemic
[4] Donor's demographic features
[5] Donor's geographic features
[6] Donation temporal features
[7] Donation geographic features

Table 4.1: Dataset description.

Subgroup	Feature	Type	Description
Donor - Demo.[4]	Gender	Boolean	Male or Female
	Age	Ordinal	Bins were created to reduce feature variance
	Blood Type	Nominal	A+, A-, B+, B- ,AB+,AB-,O+,O-
	Nationality	Boolean	Portuguese or other
	Civil status	Nominal	Married, Single, Divorced or Widowed
	Working situation	Nominal	Student, Employed or Unemployed
Donor - Geo.[5]	Parish TIPAU	Nominal	Predominantly Urban, Medium urban or Rural areas
Donation - Temp.[6] .	Month	Ordinal	Month of the year when donation happened
	Week day	Ordinal	Weekday when donation happened
	Week of month	Ordinal	Week of month when donation happened
	Weekend	Boolean	True for if the donation happened in a weekend, False otherwise
	Semester	Boolean	True for if the donation happened in the first semester, False otherwise
Donation - Geo.[7]	CST	Nominal	Lisbon, Coimbra or Porto
	District capital	Boolean	True if the donation happened in a district capital, False otherwise
	Metropolitan area	Nominal	Lisbon or Porto metropolitans areas, or 'Other' otherwise
	Parish TIPAU	Nominal	Predominantly Urban, Medium urban or Rural areas
Donation - Other	Reaction	Boolean	True if an adverse reaction was recorded, False otherwise
	Facility Centre Type	Nominal	4 possible values regarding fixed or mobile station
	Place	Nominal	10 possible values such as: schools, companies, fire stations, etc
	Class	Boolean	True if donor returned within 52 weeks from the first donation, False otherwise

Table 4.2: Features for experienced donors.

Subgroup	Feature	Type	Description
Past variables	Frequency	Ordinal	Number of previous donations made by the donor
	Years donating	Ordinal	For how many years the donor donate
	Eligibility rate	Continuous	The division between frequency and number of years passed since the donor was 18
	Recency	Ordinal	Number of months since the previous donation

The features presented in Table 4.2 measured the past behaviour in different perspectives. Ferguson [32] observed that their data had shown that recency and frequency were orthogonal, i.e., they measure entirely different aspects of the past behaviour. Based on that, the Person correlation was computed to analyse the correlation between those four variables regarding the donor's past behaviour. Figure 4.1 shows the correlation plot.

It is possible to see that the variables years donating and recency have a correlation very close to zero. It makes sense because they measure different aspects of donors' past behaviour: whilst 'years donating' measures how long a donor donates, 'recency' regards how many months had passed since the last donation. As opposed to Ferguson [32], the data used for this task do not show orthogonality between recency and frequency (i.e. nr_past_donations): although the correlation is small (-0.19), it still exists. Nevertheless, there is a high correlation between (0.71) frequency and years donating, which is

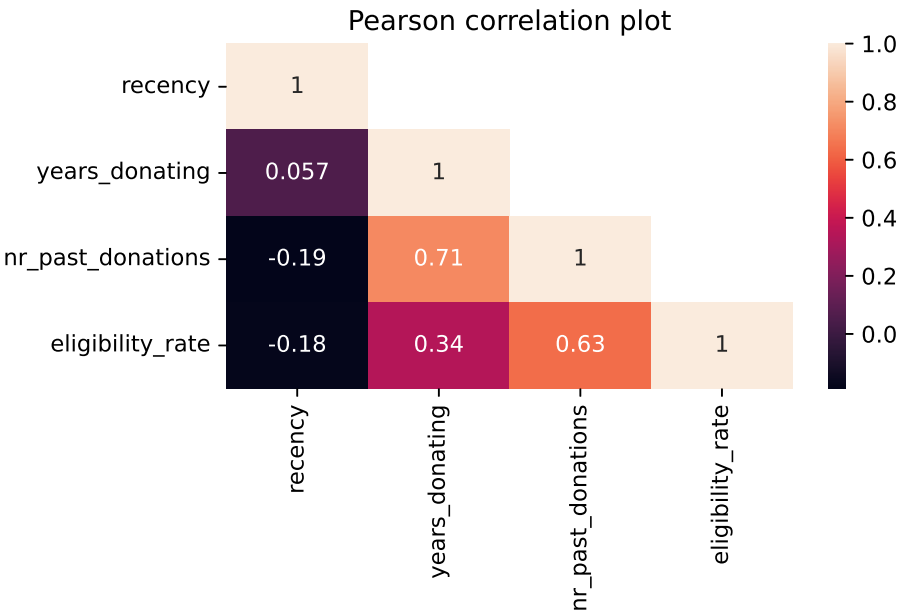

Figure 4.1: Past behaviour features correlation.

understandable: those who donated most in the past tend to be those who donate for a longer time.

4.2 Methodology

Figure 4.2 shows an overview of the methodology used in the present task. Using whole blood donations from 2016-19, the first stage of the methodology consisted of manual donor's stratification. First, the data was manually segmented according to multiple geographic criterion's to search for different donor's behaviours in different geographic regions. Then, two separate datasets were created for each geographic segment: one of new donors and the other of experienced donors.

For each of those datasets, different classifiers were therefore trained, and for each of those, different data processing pipelines were adopted. For a given model/pipeline combination, the models are build using a stratified 5 * 10-fold cross-validation scheme, evaluated, and the best parameters were chosen using random search [11].

4.2.1 Manual Donor's Stratification

The input for the Data Segmentation process is composed of whole blood donations made in Portugal between 2016 and 2019. The first stage of the data segmentation process consisted of creating different

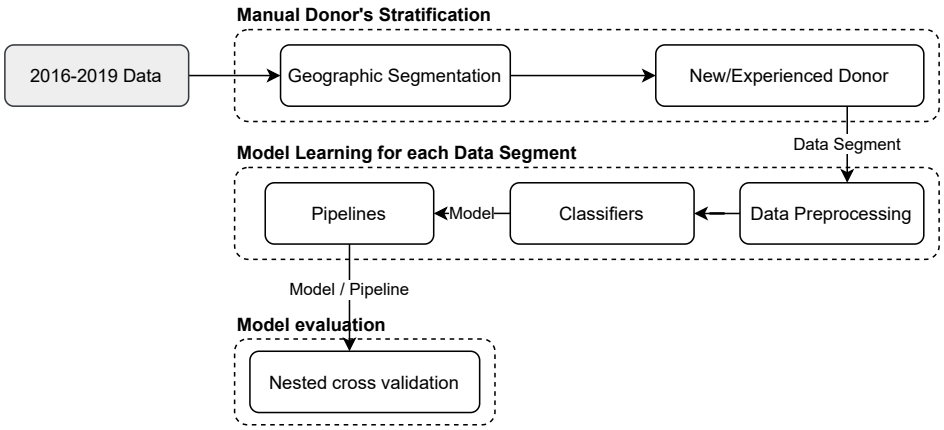

Figure 4.2: Methodology overview.

segments according to different geographic features available in the dataset, thus creating specialized models for each geographic segment. This segmentation process aimed to search for different donor's behaviours/patterns in different geographic regions. The process of segmenting the dataset and creating different models for different segments is commonly used in many fields. For example, in order to predict housing costs based on affinity to economic assets, predictive models could be built at the level of counties, districts, states or countries [52], depending on the requirements of the application and data availability.

Nevertheless, Lattimore et al. [60] found that the donation frequency varied across geographic region, with the lowest donation rate among those residents in London. They stated that that difference could reflect the more diverse, younger and mobile population hosted in London when compared to other regions. Other studies also observed differences in donor behaviour between people from rural and urban areas, with those donating in urban areas reporting lower donation rates [103, 104]. Moreover, in Section 3.2, geographical differences were also observed among the Portuguese blood donor population. For instance, despite Lisbon being the district with the most significant number of donations, Aveiro, Vila Real, and Porto have the highest proportion of donations from experienced donors. It was also seen that Lisbon had a more considerable percentage of new donors when compared to those other three districts. Also, it was seen that experienced donors tend to be older when compared to new donors.

The Portuguese donor population's geographic and demographic differences indicate different donor behaviours in different geographic regions. Because of that, it was decided to segment the data according to geographic characteristics. The geographic features used for the data segmentation are the following:

- *Brigada Distrito* (BD): The dataset used for this task contained blood donations made in 16 of the 18 Portuguese districts. The two missing districts are Beja and Portalegre because there is no data available for those two geographic regions in the period under study;
- CST: There are 3 CST's in total: Lisboa, Porto and Coimbra;

- BFT: There are three TIPAU segments: APU, AMU and APR. Data segments were created regarding the TIPAU categories of the collection site parish;
- DFT: Segmentation according to TIPAU was also made, but this time regarding the categorization of the donor's parish place of residence;
- MA: There are two MA's: AML and *Área Metropolitana do Porto* (AMP). A third segment was created that included all the donations made outside the two MA's.

In total, 28 geographic segments were created, and each of them was then split into donations from new and experienced donors, ending with a total of 56 segments for which machine learning experiences were trained on top.

4.2.2 Data Preprocessing

The data preprocessing stage included different tasks, such as dealing with data missing data or constant values, and feature encoding and normalization.

Regarding encoding, the nominal features (see Table 4.1) were encoded by using one-hot encoding, whilst for the ordinal features ordinal encoding was used, in order to keep the natural order between the values. The ordinal features were also normalized to a range of [0,1]. The binary features were transformed into 1 when the value is True, and 0 otherwise. Concerning missing data, the rows that included missing values were removed, for each data segment.

Also, features which are constant, i.e., features in which its values are the same for all the records, were removed. The existence of features with constant values happened due to the data segmentation process. For instance, as can be seen in Table 4.5, the dataset for Aveiro had 49 features (after one-hot encoding), whilst the dataset for Bragança had 41. An analysis of the dataset showed that, for Brangaça, three features were constants: 'Facility centre type', CST and MA, whilst for Aveiro there were none features with constant values. Also, for the dataset that regards Aveiro, MA feature could have two possible values: 'Porto' and 'Other', because Porto's MA include some parts of the Aveiro district. On the other hand, the same feature for the Bragança segment is a constant 'Other', because Bragança is not part of any MA. This process of removing constant features induced datasets with different number of features.

4.2.3 Model Learning & Evaluation for each Data Segment

Each data segment was imputed to 4 different classifiers: RF, Gradient Boosting (GB), AdaBoost (ADB) of RF's, and a MLP. Next, different pipelines regarding resampling techniques and feature selection with chi-squared tests were tested for each classifier.

Table 4.3: Pipelines.

Resampling	**Chi-squared**
No	Yes
No	No
SMOTETomek	Yes
SMOTETomek	No
TomekLinks	Yes
TomekLinks	No

Two different resampling techniques, one of oversampling and the other of undersampling, were tested: SMOTETomek and TomekLinks [100]. The former is an oversampling approach that follows the idea presented in [26], where SMOTE [22] oversampling algorithm is applied first, followed by the application of TomekLinks undersampling algorithm as a data cleaning method. The latter is an undersampling algorithm that aims to decrease the size of the majority class. Notwithstanding those two resampling techniques, the original dataset, i.e., the dataset without any resampling technique applied, was also used.

As a feature selection method, chi-squared tests were taken independently for each feature and the 7 most important features, i.e., the 7 features most dependent on the target variable, were chosen as dependent variables for model training.

In total, for each classifier 6 different pipelines were tested: the original dataset plus the 2 different resampling techniques, and for each of those resampling techniques, feature selection with the chi-square method was applied or not. Table 4.3 shows all the pipelines tested for each classifier.

A random search approach was performed to find the best parameters for each model/pipeline combination. The parameters and corresponding ranges are detailed in Table 4.4. The best parameters were chosen according to the best average AUC across the nested 5*10-fold cross-validation classification results. This way, model generalization error was estimated by averaging 5 different models test set scores trained over different dataset splits.

For model evaluation (the outer cross-validation loop), several metrics were retrieved such as the AUC, F-measure, predictive accuracy, sensitivity and specificity. AUC is an evaluation metric that provides information about how much the model can distinguish between classes, by using different classification thresholds [8]. The usage of different thresholds provide insights into the model's trade-off within sensitivity and specificity. Sensitivity and specificity refer to the model's accuracy regarding the positives and negatives, respectively. F-measure focus on the analysis of the positive class: it is the the weighted harmonic mean between positive predictive value and true positive rate, also known as precision and sensitivity, respectively, in the literature [36].

[8] The classification threshold used in this dissertation was 0.5

Table 4.4: Parameters and corresponding values testes for each model.

Classifier	Parameter	Range
RF	Nr of trees	range(50,500,50)[9]
RF	Criterion	{Gini Index, Entropy}
GB	Boosting stages	range(50,500,50)
GB	Learning rate	{0.01, 0.1, 0.2}
GB	Minimum samples to leaf	{20, 30, 40, 50, 60, 70, 80}
GB	Minimum samples to split	range(200,500,50)
ADB	Nr of trees	range(50,500,50)
MLP	Network sizes	{(10,), (20,), (40,), (60,), (100,), (10, 10), (20,20), (40, 40), (60,60), (100, 100)}[10]
MLP	Learning rate	{0.01, 0.05, 0.1}

In conclusion, there were a total of 56 data segments for which machine learning experiences were made. For each of those, 4 different classifiers were training, and for each classifier, 6 different pipelines regarding different resampling and feature selection techniques were applied, ending with around 1 300 machine learning experiences made. Those experiences were developed by using Python as programming language and scikit-learn [11] as the machine learning tool. To deal with class imbalance, the imblearn[12] python package was used. Section 4.3 shows the results of those experiences.

4.3 Results and Discussion

This section presents and discusses the results obtained by the proposed approach using manual donor segmentation and specific models for predicting and understanding donor's return behaviour, for each segment. Section 4.3.1 shows the results for the new donors, whilst Section 4.3.2 shows the results for experienced donors.

4.3.1 New donors

Table 4.5 shows the class distribution and the number of new donors per data segment, obtained by following the manual data segmentation approach proposed before. Recall that the manual data segmentation was made according to 5 different features. Bellow are the results are presented, according to each feature.

[9] A list of numbers starting in 50, ending in 500 and with a step of 50 between them

[10] (10,) stands for a network with 10 neurons and a single hidden layer; (10,10) stands for a network with 2 hidden layers and 10 neurons on each

[11] https://scikit-learn.org/stable/

[12] https://imbalanced-learn.org/stable

Table 4.5: Geographic data segments statistics, for new donors.

Segment	Value	Nr. Records	Nr. Features	Percentage of positives
BD	Aveiro	2 410	49	50,3%
	Bragança	470	41	48,7%
	Braga	4 295	44	47,5%
	Castelo Branco	934	42	51,2%
	Coimbra	1 290	43	40,5%
	Faro	232	35	18,1%
	Guarda	428	42	45,3%
	Leiria	1 267	47	45,9%
	Lisboa	11 001	50	41,8%
	Porto	7 302	49	48,5%
	Santarém	1 441	47	40,0%
	Setúbal	1 355	47	37,9%
	Viana do Castelo	674	42	40,2%
	Vila Real	904	42	58,8%
	Viseu	1 093	45	41,4%
CST	Porto	7 321	48	48,8%
	Lisboa	14 219	50	41,1%
	Coimbra	13 902	48	45,1%
BFT	AMU	2 222	49	46,4%
	APU	35 555	51	44,7%
	APR	665	47	51,1%
DFT	AMU	4 520	51	48,1%
	APU	28 840	51	44,0%
	APR	2 082	51	50,6%
MA	AML	11 405	48	41,1%
	AMP	6 656	48	45,3%
	OTHER	17 381	51	47,3%

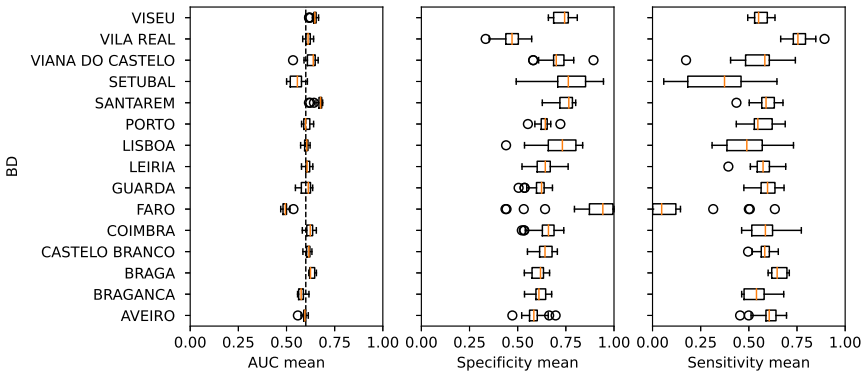

Figure 4.3: AUC, Sensitivity and Specificity mean results by collection site district, for new donor segment's. Recall that, for each model tested, 6 different pipelines regarding resampling and feature selection were tested (see Table 4.3). Furthermore, a 5*10 nested cross-validation procedure was used for each of those pipelines as a model selection and evaluation technique. This means that, for each model/pipeline combination, 5 different models were trained and tested inside the nested cross-validation procedure, each of them over different dataset splits. Afterwards, the model generalization error was estimated by computing the average test scores of the 5 different models. This plot shows the distribution of the averages of all the pipelines tested. A vertical line was placed at the AUC plot, at $x = 0.6$, to help in the interpretation.

By analysing the Table 4.5 it is possible to see that Lisbon and Porto have the biggest datasets. On the other hand, Faro and Guarda have the smallest datasets. Regarding class distribution, it is possible to observe that almost all segments have around 50% of positive cases, i.e., around 50% of new donors returned following one years after their first donation. However, Faro's segment contained just 18,1% positive cases, followed by Setúbal (37,9%), Santarém (40%), Viseu (41,4%) and Lisboa (41,8%).

4.3.1.1 *Brigada Distrito* (BD)

Figure 4.3 shows a boxplot that represents the distribution of the AUC, Sensitivity, and Specificity averages for each collection site district. That distribution includes the nested cross-validation average for all the models and pipelines tested.

By analysing the Figure 4.3, and by looking at the AUC means distribution, it is possible to see that they are all placed around 0.6. This means that, in general, in none of the data segments the trained models were truly capable of making a clear distinction between new donors who came after 12 months following their first donation and those who did not.

However, some differences can be noted between models trained for each district. For instance, models trained over Santarem data show an AUC median slightly better than models trained over blood donations made in other districts. On the other hand, models trained over Faro's data tend to be worse than

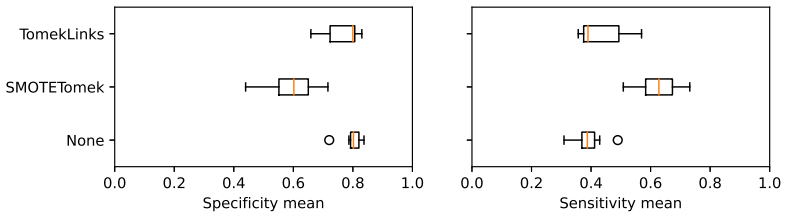

Figure 4.4: Sensitivity and specificity means scores distribution for Lisboa's data segment.

the others. This might be explained by the more accentuated class imbalance in Faro (18,1% of positive cases: see Table 4.5), and by the fact that the IPST did not provided access to all the data regarding Faro's district. Because of that Faro's results will be discarded in further analysis.

Regarding the Sensitivity and Specificity mean, things are a bit different. For instance, it is possible to see that the models for Vila Real outperform the other data segments in terms of true positive rate (sensitivity). There are a few models for Vila Real which are capable of correctly identify around 90% of new donor's who would return a year following their first donation, which is a quite good score in predicting the positives. Notwithstanding that, Vila Real is also the worst data segment regarding the true negative rate (specificity). This could be explained by the fact that Vila's Real segment has the most considerable percentage of positive cases (58,8%), which can cause models to overfit the positive data.

When looking at the Figure 4.3 it is also noteworthy that Setúbal and Lisboa's segments had a pretty big dispersion in their results when comparing it with the results for other segments. The distribution of the mean sensitivity scores for the Lisboa's segment had a minimum of 0.31, a median of 0.49 and a maximum of 0.73, whilst Setúbal's results had a minimum of 0.06, a median of 0.37 and a maximum of 0.65. The class imbalance might explain those dispersions in the two segments (41,8% and 37,9% of positive cases, respectively). Recall that the results presented in Figure 4.3 contain the results for all the models and pipelines combined, which includes models trained with and without resampling. Figure 4.4 shows the sensitivity and specificity means scores distribution by resampling strategy for the models trained on Lisboa's segment, while Figure 4.5 shows the same for the models trained on Setúbal's segment.

By analysing Figure 4.4 it gets clear that some of that dispersion could be explained by class imbalance (41.8% of positive cases). When no resampling strategy was used, the median of the specificity means was 0.80, while when the SMOTETomek oversampling algorithm was applied, it was 0.60. However, regarding sensitivity, when no resampling technique was used, the median was 0.39, whereas when oversampling was used, it was 0.63. It shows that by oversampling the dataset, models tend to be more balanced in their capacity to distinguish between both classes, while on the other hand, without resampling, models tend to overfit towards negative cases.

The same pattern could also be seen in Figure 4.5, which regards Setúbal's data segment. When no resampling strategy was used, the specificity median was 0.83, while the sensitivity median was 0.26.

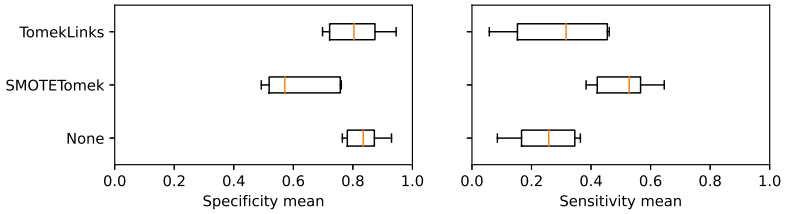

Figure 4.5: Sensitivity and specificity means scores distribution for Setúbal's data segment.

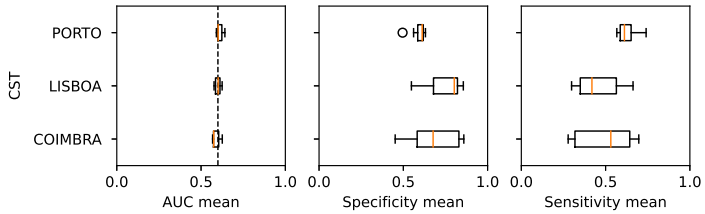

Figure 4.6: AUC, Sensitivity and Specificity mean results by CST.

However, when oversampling was applied, the specificity and sensitivities median were 0.57 and 0.53, respectively. The analysis of those two figures shows that the dispersion in the results in both Setúbal and Lisboa's segments could partly be explained by the slightly more accentuated class imbalance on those two regions. The remaining part could be explained by the different performances across the different models tested. However, none of the models had shown a noteworthy capacity to distinguish between the two classes clearly.

4.3.1.2 *Centro de Sangue e Transplantação* (CST)

Concerning the results for the CST segments, Figure 4.6 shows that model's trained for each segment reached very similar results regarding the AUC evaluation metric. It is possible to see that the AUC mean distribution for the three segments had a similar median, of around 0.6. Those are weak results for a binary classification task, which means that the trained models are not learning how to property separate between the two classes.

Regarding sensitivity, it is possible to see that the results for Porto's CST data segment have a median a little bigger than 0.6, which is a bit better when compared to the results for Lisboa and Coimbra CST's segments. Lisboa's CST sensitivity mean distribution has a median of around 0.42 and a maximum of 0.66, while Coimbra's CST has a sensitivity mean distribution with a median of 0.53 and a maximum of 0.69. Regarding the true negative rate, the opposite pattern than the one described above is noted: while Porto's CST is the segment in which the median of sensitivity means is higher, it is also the segment with the lowest median, concerning specificity (0.61), when compared to Lisboa and Coimbra medians (0.80 and 0.67 respectively).

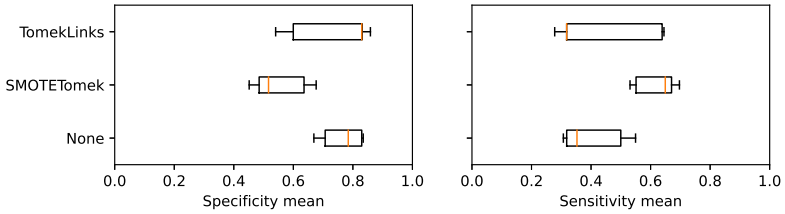

Figure 4.7: Sensitivity and Specificity mean results by resampling strategy, for Coimbra's CST.

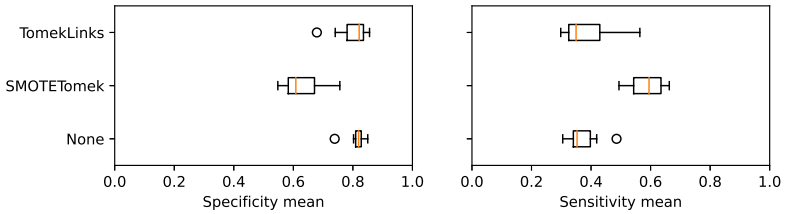

Figure 4.8: Sensitivity and Specificity mean results by resampling strategy, for Lisboa's CST.

When comparing the specificity and sensitivity scores distributions, it is noteworthy that Coimbra and Lisboa CST's had a greater dispersion, when compared to the results for Porto's CST.

Figure 4.7 shows the sensitivity and specificity mean scores distributions for Coimbra's CST, according to the three different resampling strategies adopted. It is possible to see that some of that dispersion could be explained by the different resampling strategies used. It is possible to see that by oversampling, models tend to score worse when predicting the negative cases, compared to when no resampling was used. However, regarding the positive cases, models tend to perform better when trained over the oversampled dataset, compared to then when no resampling was used. This is explained by the class imbalance (45,1% of positive cases). When the training data is balanced regarding the classes, models tend to be more balanced in distinguishing between the two classes. However, when trained over imbalanced datasets, models could, in turn, overfit towards the class with more records.

The same pattern is also visible for Lisboa's CST segment. Figure 4.8 shows the specificity and sensitivity means distribution, by resampling strategy. It is possible to see that some of the dispersion in the results for this segment could be exaplained by the different resampling strategies used. It is also possible to observe that, without resampling, models tend to overfit towards the negative cases, which might be explained by class imbalance (41,1% of positive cases). When oversampling was used, models tend to be more balanced in their capacity to distinguish between the two classes.

4.3.1.3 *Brigada Freguesia TIPAU* (BFT)

The results for the BFT are pretty similar to ones obtained in the CST segments. Regarding the distribution of the AUC means (Figure 4.9, the medians for the three segments are very close, around 0.6, which again

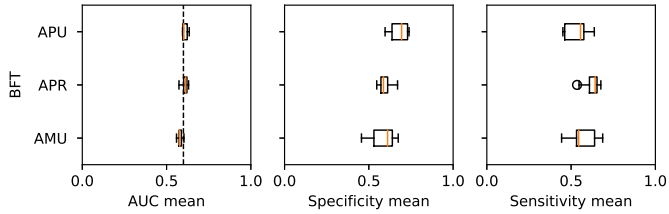

Figure 4.9: AUC, Sensitivity and Specificity mean results by BFT.

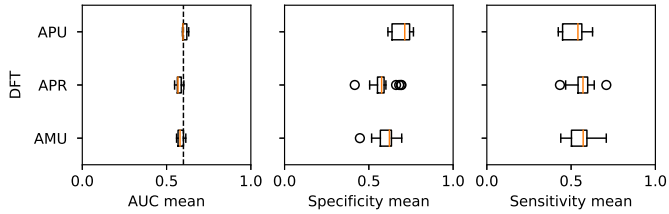

Figure 4.10: AUC, Sensitivity and Specificity mean results by DFT.

indicates a poor capacity of the trained models to learn to distinguish between the two classes. Regarding true and false positives rates scores, there is no noteworthy result for each BFT segment.

4.3.1.4 *Dador Freguesia TIPAU* (DFT)

Again, the results for the DFT segments are generally poor. Regarding the AUC (Figure 4.10), all the three segments reached a median of AUC means of around 0.6, with none of the segments outperforming any other in a noteworthy way. Regarding sensitivity and specificity, a similar pattern to the one identified in the BD and CST data segments is present here: segments with the the highest median of sensitivity means are the ones with the lowest median of specificity means, and vice versa. For instance, the APU segment had and median of sensitivity means of 0.54, being the lowest when compared to other segments. Despite that, it reaches the highest median regarding the specificity mean (0.71), when compared to the other segments.

4.3.1.5 Metropolitan Area (MA)

The results for the MA's segments follow the same pattern that the ones for the other segments (Figure 4.11). The median for the AUC means is around 0.6 for the three segments. The 'mirror behaviour' regarding the specificity and sensitivity is also seen here: the segments with the highest median regarding specificity means are the segments with the lowest median regarding sensitivity, and vice versa. The median of the specificity means for the AML segment were about 0.81, while the median of the sensitivity means was about 0.38.

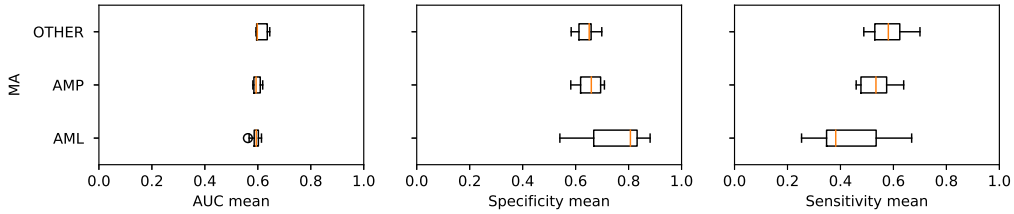

Figure 4.11: AUC, Sensitivity and Specificity mean results by MA.

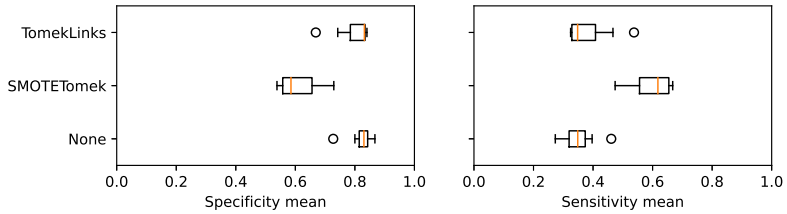

Figure 4.12: AML's sensitivity and specificity mean distribution by resampling.

It is also possible to observe a dispersion in the results for the models trained on the AML's segment, regarding the specificity and sensitivity mean distributions. As can be seen in Figure 4.12, that dispersion could be explained by the different resampling strategies used. Models trained on the original dataset, i.e., on the dataset without resampling, tend to overfit the negative cases, due to the class imbalance (41,1% of positive cases). When oversampling was used, models tend to be more balanced in their predictions: by oversampling, models tend to be better when predicting the positive cases, compared to when trained on the original dataset; however, they tend to perform worse in predicting the negative cases.

4.3.1.6 Discussion

In general, the results showed that the manual segmentation made for the new donors did not work well, given that no segment had a noteworthy performance compared to others. Furthermore, for all the five manual segmentation's made, the median AUC range of 0.55 - 0.6 shows a model's incapacity to clearly distinguish between donors who returned after a year following their first donation and those who did not.

The weak scores regarding the AUC show that the classifiers had a poor performance in the task for which they were trained, despite the different pipelines tested (regarding different classifiers, resampling strategies, and the usage of chi-squared test for feature selection). However, some segment's have shown a dispersion in the mean results regarding the true positive and negative rates, which turned to be explained by a more accentuated class imbalance in those segments and the effect of resampling on those cases

For instance, regarding the segmentation by district (BD), in which models were trained over blood donations made in each district, it was possible to observe a considerable dispersion in the specificity

and sensitivity mean results for Setúbal and Lisboa's segments, which turned to be explained by the class imbalance in those districts and the different oversampling strategies adopted. Overall, models registered specificity medians of around 0.6-0.7 and sensitivity medians of 0.5-0.6, which shows their incapacity to distinguish between the two classes. The results for the CST's segments also show a dispersion in the models' true positive and negative rate, which also turned to be explained by the class imbalance and the resampling strategies used.

The strategy to segment the new donor's dataset according to the typology where the donation occurred (BFT) and where donors lived (DFT) did not work as expected. By splitting the dataset according to those criteria, one wanted to test whether donors who lived or donated in urban areas had different determinants influencing donor return than those who lived or donated in less rural areas. However, none of the models trained over those datasets had shown a considerable performance.

Models trained for different metropolitan areas also did not show any noteworthy results. There was a dispersion in the specificity and sensitivity means distribution for models trained on the Lisbon metropolitan area, which was explained by a more accentuated class imbalance in that segment and the resampling strategies used.

4.3.2 Experienced donors

Table 4.6 shows the dataset size and class distribution for the experienced donor's segments.

By analysing it, it is possible to observe that the class distribution for the experienced donors' segments has considerable differences compared with the new donors' segments. While the former had a percentage of positives of around 40-50%, in which around 50% of the news donor returned following one year of their first donation, the latter shows a percentage of positives between 65 and 83% (excluding Faro's segment). Regarding the models for the BD segment, it is possible to observe that the Lisbon segment contained 67,1% of positive cases, while on the other hand, Vila Real contained 83,3% of positive cases. Regarding the dataset sizes of the BD segments, it is possible to see that Lisboa, Porto and Braga have the biggest datasets, while Faro, Viana do Castelo and Bragança have the smallest datasets.

4.3.2.1 *Brigada Distrito* (BD)

Figure 4.13 shows a boxplot representing the distribution of the mean AUC, Sensitivity and Specificity for all the models trained over experienced donors on the AML's segments.

Table 4.6: Geographic data segments statistics, for experienced donors.

Segment	Value	Nr. Records	Nr. Features	Percentage of positives
BD	Aveiro	32 621	54	79,6%
	Bragança	5400	45	79,2%
	Braga	49 655	48	77,6%
	Castelo Branco	8 202	46	76,5%
	Coimbra	13 120	47	71,1%
	Faro	474	40	40,3%
	Guarda	5 531	46	75,7%
	Leiria	20 177	51	75,6%
	Lisboa	86 084	54	67,1%
	Porto	69 835	53	76,6%
	Santarém	15 386	51	73,6%
	Setúbal	10 461	52	67,3%
	Viana do Castelo	3 985	46	79,9%
	Vila Real	14 658	46	83,3%
	Viseu	10 063	49	71,0%
CST	Porto	146 206	52	77,7%
	Lisboa	116 582	54	68,4%
	Coimbra	82 934	52	75,2%
BFT	AMU	28 870	53	72,9%
	APU	306 747	55	74,1%
	APR	10 105	51	73,9%
DFT	AMU	52 930	55	75,0%
	APU	267 440	55	73,6%
	APR	25 352	55	76,4%
MA	AML	86 361	52	66,5%
	AMP	67 368	52	76,6%
	OTHER	191 993	55	76,4%

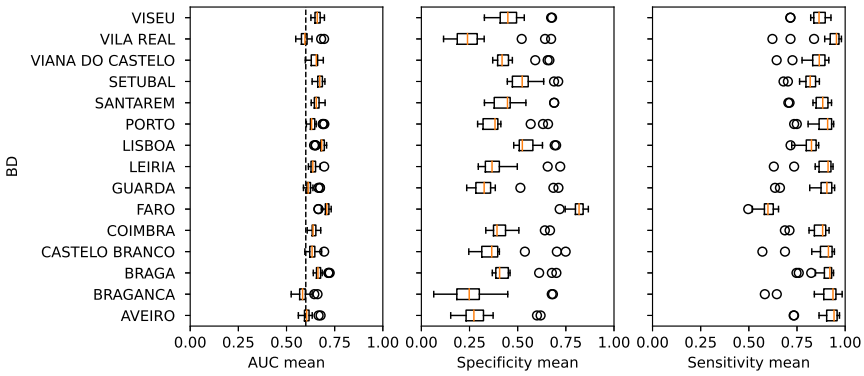

Figure 4.13: AUC, Sensitivity and Specificity distribution results by collection site district, for experienced donors.

An overview of Figure 4.13 shows that, for the experienced donors' segment, the median of the AUC's mean distribution is slightly better when compared to the medians for the new donors' segment (see Figure 4.3). Whilst the former had medians of around 0.6, the latter had, for some segments, medians of around 0.7. For instance, some classifiers trained over Lisboa and Setubal's data showed an increase in the AUC of 0.1 points, from 0.6 to 0.7. Nonetheless, the fact that AUC's of 0.7 indicate that the models are far from being perfect in their ability to distinguish between the two classes, an increase of 0.1 in the AUC median could still be considered a noteworthy improvement.

A deeper analysis of Figure 4.13 shows some differences within models trained over blood donations from different districts. For instance, models trained for Lisboa and Setúbal had a median of AUC means of 0.70 and 0.69, respectively, while models trained for Bragança and Vila Real had a median of AUC means of 0.58 and a.59, respectively.

When looking at the specificity and sensitivity mean's distribution it's possible to see that models tend to perform much better when predicting the positive cases, which might be due to the highest percentage of positive cases when dealing with experienced donors. For instance, the sensitivity means distribution for the Vila Real segment had a median of 95,4% and a maximum of 98,1%, which are pretty good scores in predicting the positive cases. However, regarding the prediction of the negative cases, it registered a median of 24% (recall that in Vila Real 83,3% of the experienced donor's returned a year following their last donation). On the other hand, models trained for Lisboa's segment (which contains 67,1% of positive cases) did not register such a huge difference between the medians of the specificity and sensitivity means.

The boxplots identified some outliers regarding the sensitivity and specificity mean distributions, which indicates a considerable dispersion in the results. Figure 4.14 a) shows those distributions for the models trained without resampling, while Figure 4.14 b) shows them for the models trained over the oversampled datasets. It is possible to observe that oversampling the data segments had a considerable impact on the results. Without oversampling models tended to be more imbalanced in their ability to distinguish between the two classes: pretty good at identifying the positive cases and pretty bad at identifying the

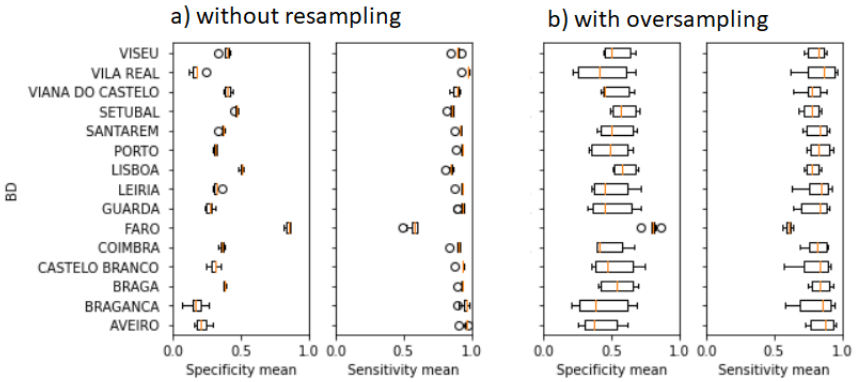

Figure 4.14: Sensitivity and Specificity distribution results by collection site district, for experienced donors. a) shows the results without resampling; b) shows the results with oversampling.

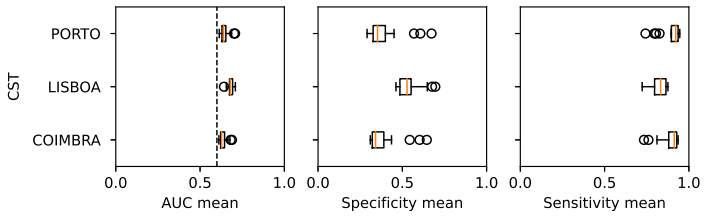

Figure 4.15: AUC, Sensitivity and Specificity distribution results by CST, for experienced donors.

negatives. On the other hand, by oversampling, the medians of the sensitivity means tend to decrease, while the medians of the specificity tend to increase. This makes sense given the class imbalance for the experienced donors' segments (see Table 4.6)

4.3.2.2 *Centro de Sangue e Transplantação* (CST)

Figure 4.15 shows the mean distribution scores for the CST's experienced donors segment. It is possible to observe that the results are slightly better, when compared to the new donors' segment (see 4.6), especially for the models trained over the Lisboa's CST: it had an AUC median of 0.58 in the new donors' segment, while for the experienced donors it registered an AUC median of 0.7.

Regarding the sensitivity mean distribution, it is possible to observe that Lisboa CST's had a median of around 0.8, while the models trained for Porto and Coimbra CST's registered a median of around 0.9. On the other hand, and regarding the specificity mean distribution, it is possible to observe that the models trained for Porto and Coimbra CST's tend to have small scores when compared to the Lisboa CST's segment. This might be explained by the differences in class imbalance within those three segments: Lisboa's CST segment had 68,4% of positive cases, while Porto and Coimbra CST's had 77,7% and 75,2%, respectively. This indicates, again, that when models are trained over a highly imbalanced dataset,

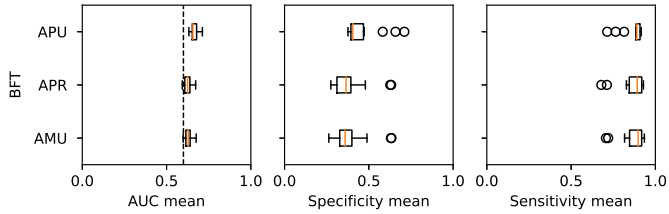

Figure 4.16: AUC, Sensitivity and Specificity distribution results by BFT, for experienced donors.

they tend to perform better when predicting the most representative class.

4.3.2.3 *Brigada Freguesia TIPAU* (BFT)

Concerning the results for the BFT's segment (Figure 4.16), it is possible to observe that training different models over different typologies, regarding the place where the donation was made, did not show any noteworthy effect. By comparing the AUC means distribution of the models trained over new (Figure 4.9) and experienced donors, one can possibly see that it just had an effect on the APU segment.

The mean AUC distribution has a median of around 0.62, with the APU segment performing a bit better (0.65), compared to other segments. Regarding the true and false positive rate scores, the models tend to perform better when predicting the positive scores, with the median of sensitivity means of around 0.9 in all segments. Model scores for the negative cases are bad, with a specificity median for the APR and AMU segments of around 0.35, while the APU segment had registered a median of 0.4 (which explains the better performance regarding its AUC).

4.3.2.4 *Dador Freguesia TIPAU* (DFT)

Figure 4.17 shows the results regarding models trained for DFT's segments. Recall that DFT segments concern the typology of the donor's address: rural, urban or predominantly urban areas. By analysing the figure 4.17 it is possible to observe no noteworthy difference between models trained for different typologies. As happened with other segments, the models tend to perform better when predicting the positive cases.

By comparing these results with the ones obtained in the new donors' segment (see Figure 4.10), it is possible to see that all the segments had an improvement in the AUC median of around 0.05.

4.3.2.5 Metropolitan Area (MA)

Figure 4.18 shows the results for the MA's segments. It is possible to observe that the models trained for the Lisboa MA perform worse when predicting the positive cases when compared to Porto and others' MA's. For instance, regarding the sensitivity mean distribution, models trained for Lisboa had a median of 81.7%, while Porto's models had 91.2% and the models trained on the rest of the country ('OTHER'

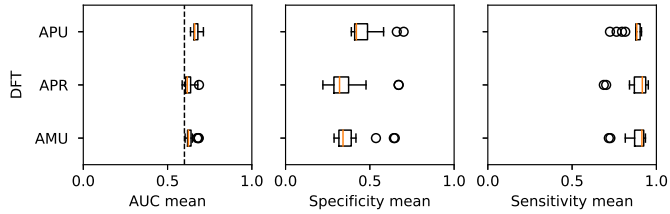

Figure 4.17: AUC, Sensitivity and Specificity distribution results by DFT, for experienced donors.

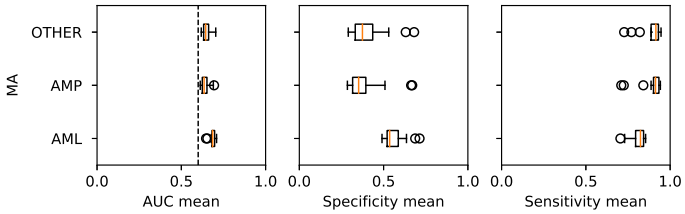

Figure 4.18: AUC, Sensitivity and Specificity distribution results by MA, for experienced donors.

in Figure 4.18) scored a median of 91%. This pattern was also observed in the results for the new donors (see Figure 4.11).

On the other hand, Lisboa's models tend to perform better in the prediction of the negative cases, when compared to the other segments. This is consistent with the other results: models that tend to have a good performance in predicting the positive cases are bad in predicting the negative cases.

By comparing those with the ones obtained for the new donors' segment (see Figure 4.18), is possible to observe, as it was in the other segments, that in general models performed better when dealing with experienced donors.

4.3.3 Discussion

As happened with the new donor's segment, the experienced donor's segment results shows that the geographic segmentation made did not work as expected. The idea of splitting the data according to multiple geographic characteristics was to test if different donor's behaviours or patterns existed in different geographic regions, but that hypothesis could not be verified given the results obtained. No segment had outstanding results, compared to other segments.

In general, the results for the experienced donor's segments improved compared with the results for the same geographic segment, but regarding new donors. The AUC's median for the new donors segmented ranged mainly between 0.55 and 0.6, while the AUC's for the experienced donors ranged between 0.6 and 0.7 approximately. The mean AUC distributions had also shown relatively small dispersions from the median, whereas sensitivity and specificity mean distributions tended to show more considerable dis-

persion's, with the boxplots identifying many outliers. This could be explained by class imbalance and the impact that the different classifiers and resampling techniques tested had on it.

The boxplot analysis also showed that the performance improvement was due to an increased model's capacity to predict the positives correctly: in general, sensitivity and specificity medians ranged within 0.8-0.9 and 0.3-0.4, respectively. That increased capacity to distinguish the positive cases, when compared to the new donors' segment, might be explained by:

1. A more accentuated percentage of positives: while just 40 to 50% of new donors return a year following their first donation, that percentage increases to around 70-80% when dealing with experienced donors;

2. The 4 additional features that regard donor's past behaviour (see Table 4.6). The literature (see Section 2.5) suggests that donor's past behaviour is the best predictor of future behaviour. It could be the case that those 4 additional features are helping model's to improve their ability to distinguish between the two classes.

Nevertheless, some slight differences could be noted in the model's performance, regarding different segments. For instance, models trained over Lisboa's data always had slightly better results when compared to models trained in other regions. This could be seen in the results obtained by district (BD), in which Lisboa's segment had the highest AUC median, and also by the results obtained by segmenting the dataset according to CST's and MA's: models trained over Lisboa's CST outperformed the others regarding its AUC median, and models trained over Lisboa's metropolitan area also did better when compared to models trained on other regions.

Considering the differences in the AUC medians of model's trained over new and experienced donors, and given the importance that the literature gives to donor's past behaviour as a predictor of future behaviour, the impact that the features that regard donor's past behaviour had on model's performance should be better analysed. Section 4.4 analysis model's feature importances.

4.4 Feature importance analysis

Analysing the determinants of donor return in the context of this dissertation and the methodology adopted means analysing the feature importance. Feature importance is a measure of individual contributions of the corresponding feature for a particular classifier, and it is the most commonly used model explanation technique [87]. It allows the extraction of relevant knowledge (knowledge that provides insights into the problem being solved) from a machine learning model, concerning the relationships learned by it.

Generally, feature importance can be divided into global and local importance's [44]. While global feature importance measures the importance of the feature for the entire model, local importance measures the contribution of a feature to the results of a trained model on a specific input [87]. An example of global

feature importance are the coefficients learned in logistic regressions models. Examples of local feature importance are Local Interpretable Model-agnostic Explanations (LIME) [82] and SHAP [62]. These last two are also considered model-agnostic methods because instead of interpreting what the models actually learned (which is not possible when using black-box models), they extract post-hoc explanations by treating the original model as a black box. The separation of interpretability from the model thus frees up the model to be as flexible as necessary for the task, enabling the use of any machine learning approach - including, for example, complex models such as neural networks [83].

It was decided to use SHAP as a model interpretation technique. SHAP was initially proposed by Shapley [93] in 1953, and it is based on game theory. Lundberg and Lee [62] introduced the usage of Shapley's approach for the interpretation of machine learning models. SHAP builds on the assessment of a combination of features: it calculates the predictive performance of every combination of features, with and without a given feature. Then, it measures the change in the model prediction when conditioning a given feature, thus allowing to quantify the contribution of each feature value for the model prediction.

4.4.1 Model selection

For feature importance analysis, it was decided to use models trained over the experienced donor segments because, as shown before, they tend to perform slightly better when compared to those trained with new donors data. Within all the models trained, it was decided to choose the best models trained over Lisboa and Aveiro and perform an analysis of their feature importance. Models trained over Lisboa and Aveiro were chosen due to different reasons. First, the idea was to select models trained on different data segments to check if the most important features were the same. Then, those two segments are different regarding dataset sizes and their percentage of positive cases. For example, Lisboa has the biggest dataset regarding the segmentation made by districts, while at the same time, it has one of the lower percentages of returning donors (67,1%), while Aveiro has one of the biggest (79,6&).

Given that multiple classifiers and pipelines were tested for each data segment, a criteria to choose one classifier within all the trained ones need to be established. Within all the classifiers and pipelines tested, it was decided to choose the classifier/pipeline combination that reached the highest AUC mean (see tables A.1 and A.2 in Appendix A). MLP's trained over data segments that used all the features, and that were oversampled, were chosen as the best classifier/pipeline combination.

Recall that a nested 5*10 cross-validation procedure was used as a model evaluation technique. This means that 5 different models were trained (on the outer loop) for each classifier/pipeline combination. Then, to select the 'best' of them, they were all fed with data from 2015 (unseen until then), their AUC's were evaluated, and the classifier with the highest AUC was chosen.

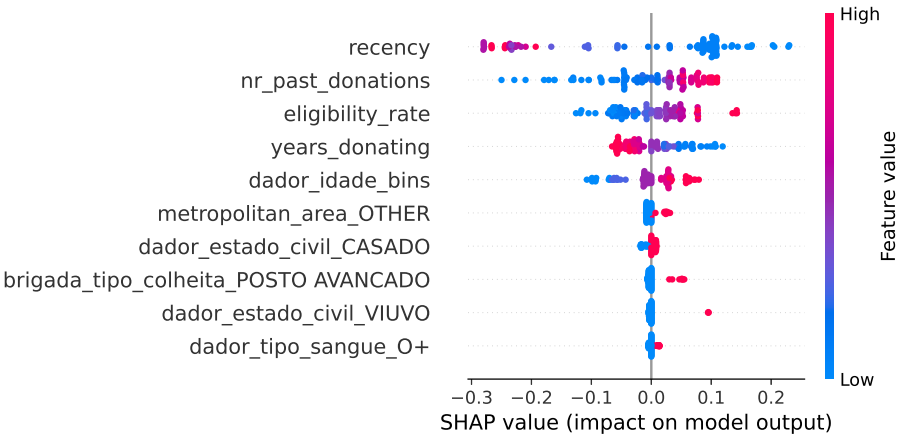

Figure 4.19: SHAP summary plot, for the MLP trained over Aveiro's data. This plot combines feature importance with feature effects [68]. Each point is a SHAP value (x-axis in units of log-odds), which measure the impact on model predictions, for a feature (Y-axis) and an instance. The colour represents the feature value from low to high. Features are ordered according to their importance, and just the 10 most important are shown.

4.4.2 Results

Figure 4.19 shows the SHAP summary plot for the MLP trained over Aveiro's data. It is possible to observe that the 4 most important features are the ones that regard donors' past behaviour. Specially recency, which is the feature that has the greatest impact on the model prediction. The summary plot shows that it has a negative impact on donor return when its values are pretty high.

The frequency of past donations (i.e., nr_past_donations) and the eligibility rate are the next two most important features, and higher values of these features lead to higher SHAP values, which correspond to a higher probability of donor return. In contrast, lower values of these features lead to a lower probability of donor return. The number of years that had passed since the donor's first donation (i.e., years_donating) also had a significant impact on return. Lower values of this feature correspond to a higher probability of donor return, while higher values had a negative impact on it. Regarding donor's demographics, donor's age is the most important feature, with being older increasing the probability of donor return.

Figure 4.20 shows the SHAP summary plot for the MLP trained on Lisboa's data. It is possible to observe that the most important features for Lisboa also regards donor past behaviour. The patterns observed are similar to the patterns observed in Aveiro: higher values of frequency of past donations lead to a higher probability of donor return; higher values of recency and years donating lead to a lower probability of donor return. However, the eligibility rate was considered the 9th most important feature in Lisboa, while for Aveiro it was 3rd.

In Section 4.3.2 it was seen that the results of models trained over experienced donors were, in general,

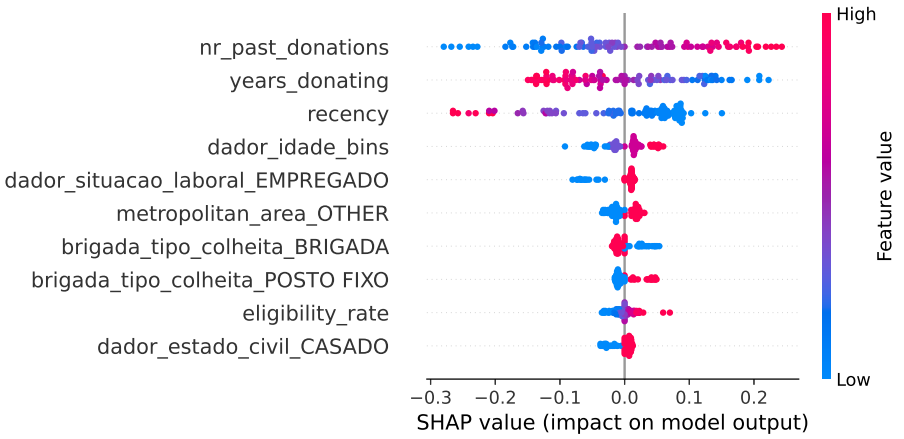

Figure 4.20: SHAP summary plot, for the MLP trained over Lisboa's data.

slightly better when compared to the results for the same segment but regarding new donors. In this section, it was seen that, for models trained on Aveiro and Lisboa's data, the most important features, according to the SHAP method, were the features regarding donor past behaviour.

However, recall that the models trained for Aveiro and Lisboa had AUC's means (see Tables A.1 and A.2 in Appendix A) of 0.68 and 0.71, approximately, which does indicate that they were far from being perfect in finding the function that maps the features to the target variable. In other words, the best models found were not capable of truly distinguishing between blood donors who return and those who do not. However, and as discussed in Section 2.5.3, if models are a poor emulation of the mechanism that generated the data, the conclusions drawn from it may be misleading. Therefore, one should not extract conclusions from a weak classifier and take them as ground truth. Hereupon, to analyse the real impact that past features had in experienced donors segments, in Section 4.5 models regarding experienced donors are trained, but without past behaviour features.

4.5 Training without past behaviour features

In Section 4.4 it was seen that the most important features, for models trained over Aveiro and Lisboa's experienced donor segments were features that regarded donor past behaviour. However, since the models were far from being perfect in predicting donor return, the results of the SHAP analysis could not be taken as ground truth. The classifiers learned to give more importance to the past behaviour columns; however, they are indeed weak classifiers, so one has to question its conclusions.

To analyse the impact that the past behaviour feature had on the results, new machine learning models were trained: *without* past features and with *just* past features. Afterwards, they were fed with blood donations from 2015 (unseen during the training pipeline), and their performances were compared. Those

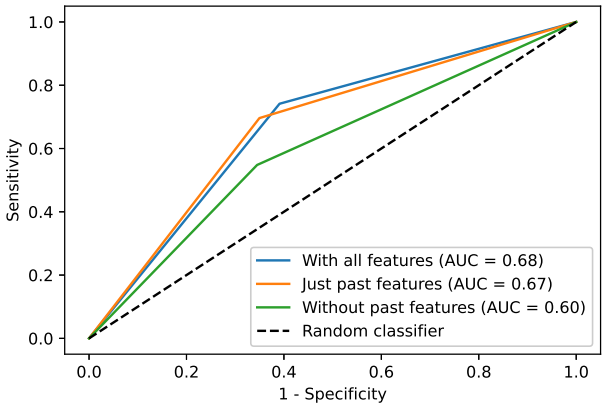

Figure 4.21: ROC curves for models trained over Aveiro's experienced donors.

validation sets included 11 872 and 86 084 blood donations for Aveiro and Lisboa, respectively. Finally, DeLong statistical test [27] was used to make comparisons between the AUC's. It tests whether two models performances are significantly different while accounting for uncertainty due to randomness.

In Section 4.4 it was seen that the models that reached the highest AUC mean for experienced donors in both Lisboa and Aveiro's segments were MLP trained over oversampled datasets. Hereupon, AUC's and oversampling were just, and the models were trained without and with just past behaviour features.

Figure 4.21 shows the ROC curves for models trained over Aveiro's data, with all the features, with just the past features, and without the past features. We can see that the AUC for the MLP trained without the past features was 0.60. However, when the past features were added it increased to 0.68. By training the MLP just with the past features, it reached an AUC of approximately 0.67. Table B.1 (Appendix B) shows the results for the DeLong statistical tests. It is possible to observe a significant increase in model performance ($p < 0.001$) when the past features were added. However, there is no significant difference ($p = 0.61$) in model performance when trained with all the features or with just the past features.

Figure 4.22 shows the ROC curves for models trained over Lisboa's data. It is possible to see that by training with the past features, or with all the features, the model had the same AUC: 0.71. By removing the past features, it decreased 0.06 points, to 0.65. Table B.2 (Appendix B) shows the p values obtained. It is possible to observe that same that was observed in Aveiro: there is a significant increase in models performance when the past features are added ($p < 0.001$), and no significant differences were recorded between models trained with all features and with just the past behaviour features.

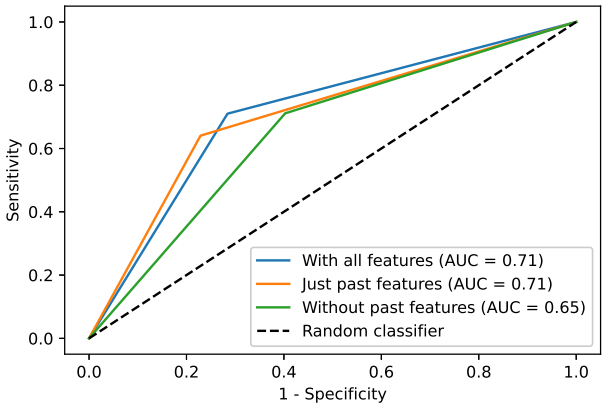

Figure 4.22: ROC curves for models trained over Lisboa's experienced donors.

4.6 Conclusions

In Section 2.5 multiple studies that tried to identify the determinants of donors return were presented. Much of the literature identified psychological variables, such as intention to donate, as essential factors for donor return. Moreover, past behaviour was also identified as one of the best predictors of future behaviour. Regardless the importance of psychological variables, in this chapter we were interested in measuring the importance of donor's demographic and geographic characteristics in its future behaviour. Multiple studies tried to identify demographic determinants of donor return by using logistic regression models and analysing their learned coefficients. However, in Section 2.5.3 it was argued that if models were not accurate at predicting donor return, then the conclusions drawn from them could not be trusted. Based on that, in this study, a machine learning approach was followed.

In Section 4.2 the proposed methodology was presented. The data was first segmented according to different geographic characteristics, and within each segment, it was split into blood donations from new and from experienced donors. For the latter group, four different variables that regard donor past behaviour were computed. The aim of this manual segmentation procedure was to identify if different donor behaviours exist in different geographic regions. By splitting each segment into blood donations from new and experienced donors, one wanted to test whether different determinants existed within those two groups of donors.

In general, the results were not promising. Models trained for new donors segments registered AUC's between 0.5 and 0.6, while it increased to 0.6-0.7 for experienced donors segments. No outstanding differences were noted within models trained over different geographic segments. Since the models were in general weak, using feature importance analysis to extrapolate conclusions regarding donors behaviour should not be done. Thus, is not possible to confirm the hypothesis that different donor behaviours exist in different geographic regions.

Nonetheless, the difference in the AUC's of models trained over new and experienced donors deserved a deeper analysis. Section 4.4 demonstrated that, for experienced donors, the most important variables regarded donor past behaviour, while Section 4.5 revealed that removing the past behaviour features from Aveiro and Lisboa datasets decreased models performance significantly. However, it was also observed that the performance of models trained with just the past behaviour features was not significantly different from the performance of models trained over all the features (including past behaviour features), which suggests that the other features are not relevant for predicting donor return.

Several studies cited in Section 2.5 identified several demographic factors as determinants of donor return. Godin et al. [42] identified age as an important factor, Kheiri et al. [56] identified donor's body weight and working status, and Lattimore et al. argued that genre, age, and typology of the place of living (rural or urban) are significant factors. However, as discussed previously, those studies did not provide any model validation technique, and therefore, their conclusions may be wrong. To the contrary, our results suggest that no demographic or geographic factor significantly impacts donor return.

Regarding the donor past behaviour, our results are consistent with the literature. Multiple studies have identified past behaviour as the best predictor of donor behaviour, and our results suggest that those features are the best predictors of future behaviour, when compared to the demographic and geographic features. However, given the performance of models trained with just past features, our results suggest that those features alone are not enough to explain donors future behaviour.

Several motives may have contributed to the general model's weakness in predicting donor return. Regarding the methodology adopted, different approaches could have been followed. For instance, different machine learning models could have been trained, such as probabilistic or linear models. Also, the hyperparameter space could be increased within the grid search. Regarding feature selection, chi-square statistical tests were used. However, this method is just suitable for categorical features (both nominal and ordinal), and the experienced donors' segment contained one continuous feature. Nonetheless, chi-square tests are applied to each feature independently, not considering relationships between features. Thus, other methods could have been applied.

However, despite the methodology adopted, if the features are not informative about donor future behaviour, the model's performance will never be outstanding. Nonetheless, a factor that must be considered is heterogeneity. It might be the case that different homogeneous subgroups of blood donations exists, that may vary widely from each other. In Chapter 5, clustering algorithms are applied to the dataset to find homogeneous groups of both new and experienced donors.

Chapter 5

Using unsupervised learning to find homogeneous groups of blood donors

As discussed in Section 4.6, one of the reasons for the difficulty of the classifiers to perform the task of predicting donor return following 12 months after a given donation might be the heterogeneity of the dataset. Heterogeneity is related to the concept of dispersion, meaning that the population under study has a high variability between its elements, which could, in turn, make the learning process more challenging. As stated by Karpatne [52], one of the major challenges in applying the predicting learning method in real-world applications is the heterogeneity in populations of data instances, i.e., different groups of data instances may show different nature of predictive relationships.

As stated Section 2.5.3, one of the underlying assumptions in the predictive learning framework is that the data instances in the training set are identical to each other, belonging to a common yet unknown distribution. Hence, the training instances are generally considered independent and identically distributed. In other words, the training and test sets are assumed to contain instances belonging to a single common population, thus sharing identical (or homogeneous) relationships between input and output variables [52]. However, those assumptions of homogeneity are violated in a number of real-word applications, because most real-word systems are composed of a plurality of data populations, with varying properties of predictive relationships in every population, which, in turn, limits the effectiveness of predictive models.

For instance, in the problem of predicting a patient's risk for a particular disease given their healthcare records, different populations of individuals from diverse socio-economic and ethnic backgrounds can show considerable variations in the relationship between their healthcare records and disease risk [52]. A solution for this problem could be the implementation of different models of disease risk for the different populations. Pires et al. [80, 79] also showed that in the problem of prognostic prediction of neurodegenerative disease, patient stratification through clustering algorithms proved to be a key tool to deal with the heterogeneous nature of the disease.

In order to learn predictive models in the presence of heterogeneity, one approach can be to first

divide the entire dataset into homogeneous partitions by grouping instances based on similarities between their explanatory variables, by means of clustering algorithms. This would result in groups of instances with similar values of explanatory variables, which are likely to share common predictive relationships between explanatory and target variables [52]. Clustering is the process of partitioning a set of data objects in subsets, that share similarities between objects within each cluster, but are dissimilar to objects in other clusters. In other words, clustering allows the identification of homogeneous groups, i.e., heterogeneous groups consisting of homogeneous elements.

This chapter describes the task of clustering blood donors, in order to search for homogeneous groups. In Chapter 4, the blood donations were segmented according to different geographic characteristics. In this chapter, however, we are interested in grouping blood donors, rather than blood donations. The identification of different homogeneous groups in blood donors could be helpful for IPST, for instance for developing different policy strategies for different blood donors' segments.

The remainder of this chapter is organized as follows. Section 5.1 briefly describes the data used for this task. Section 5.2 describes the methodology adopted for clustering both the new and experienced donors segment. Finally, Section 5.3 shows the results obtained.

5.1 Data

The dataset used for this task is the same that was used in Section 4. It consisted, originally, of 583 728 blood donations made between 2016 and 2019, 54 403 of whom were from new donors, while the remaining 529 325 donations were made by a total of 181 940 blood donors. Since we are interested in clustering blood donors, and not blood donations, the blood donations from experienced donors were grouped to create a blood donors dataset. The process of transforming the dataset is described in detail in Section 5.2.

5.2 Methodology

Figure 5.1 shows an overview of the methodology followed in the present task. Using whole blood donations from 2016-19, the first step consisted of splitting the dataset into blood donations from new and experienced donors.

For the new donors' segment, each record corresponds to a single donor. In other words, for that segment, the blood donations dataset is also a blood donor dataset because single donation for each donor is present. Since we are interested in finding homogeneous blood donors groups, there was no need to transform the donations dataset into a donors dataset. However, for the experienced donors' segment, there were multiple donations for each donor, and thus, it had to be transformed into a blood donors dataset. Hereupon, different methodologies were adopted concerning those two typologies.

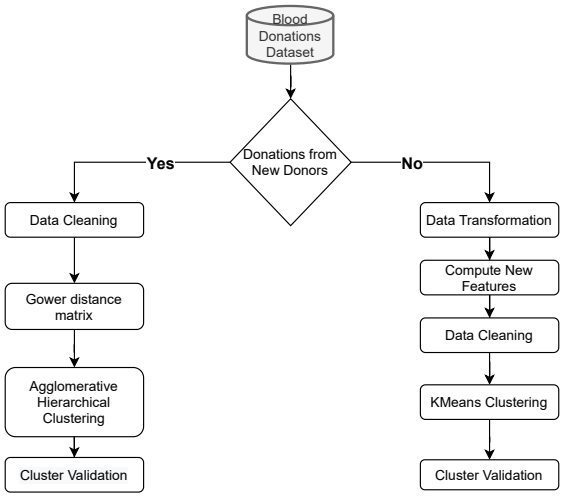

Figure 5.1: Clustering Methodology overview.

5.2.1 New donors

The methodology followed for the new donors data segment is described as follows:

1. A total of 18 979 records with missing values were removed. The final dataset contained 35 428 records;

2. The features used for the clustering task can be seen in Table 4.1.

3. In Table 4.1 it is possible to see that most features are nominal, except for four ordinal features. Based on that, it was decided to compute the Gower distance [43] matrix: it is a distance metric that measures the similarity between two objects containing both numeric and non-numeric attributes, which uses the Manhattan distance for numerical variables, and the Dice coefficient [28] for nominal variables;

4. The Gower's distance matrix was then used to feed Ward's Agglomerative Hierarchical Clustering algorithm. As suggested by Kaufman Rousseeuw [55], the ordinal variables were normalized into a [0,1] range and treated as numeric values;

5. For clustering evaluation, three different metrics were used: the Silhouette score [84]; Caliński and Harabasz index [16]; and the Davies-Bouldin score [25]. The Silhouette score is calculated using the mean intracluster distance and the mean nearest-cluster distance for each instance. It has a range within [-1,1], and the higher the value, the higher the clustering performance. The Caliński and Harabasz index measures the ratio between the within-cluster dispersion and the between cluster dispersion, and, as with the silhouette score, higher values of this index correspond to a

higher clustering performance. Finally, the Davies-Bouldin index measures the average similarity between clusters, and values closer to zero indicates a better partitioning. Hierarchical clustering requires the apriori specification of the number of clusters k. To choose the best number of clusters, different values of k were tested, ranging from 2 to 10. For each k, the three evaluation indexes were computed, and the best number of clusters was chosen using a majority vote, i.e., the k that performed best in at least two of the indices was chosen;

6. After choosing the number of clusters, it was necessary to characterize the blood donors subgroups. Thus, statistically significant differences across clusters were tested, using chi-square tests for the nominal and binary variables, and Mann–Whitney U test [64] for the ordinal variables. Also, some plots are presented to help visualise the differences across clusters.

5.2.2 Experienced donors

For the experienced donors, however, the methodology followed was different because those donors had more than one donation. Given that, it was decided to transform the blood donations dataset into a blood donors dataset. Therefore, the methodology followed for the experienced donors segment is described as follows:

1. The first stage consisted of transforming the blood donations dataset into a blood donors dataset. Table 4.1 contains the original dataset used as input for the transformation process. The first stage consisted of creating dummy variables for both boolean and nominal variables. Regarding donor demographics columns (e.g. gender; age; blood type; civil status and working situation), the last donation from each donor was used to get each donor demographic characteristics. For this experience, it was decided to remove the temporal information regarding each donation (e.g. weekday, weekend, week-of-month, month, semester);

2. Then, for variables regarding donation information (e.g. CST, the district capital, metropolian area, parish TIPAU, reaction, facility centre type and place), dummy variables were created, the dataset was further grouped by donor, and the dummies were summed. Figure 5.2 shows an example of that transformation process. At the end of the process, we got the number of times each donor donated in each blood centre type, for instance;

3. New variables regarding donor's past behaviour were computed: total number of donations; mean number of donations per year; and average interval time between donations, in days;

4. Originally, 181 940 experienced donors donated during the period under study. After transforming it, and removing records with missing data, it ended up with 179 326 records and 48 dimensions;

5. The data was normalized to a [0,1] range;

a)

Blood donations dataset	
Donor ID	**Blood Centre Type**
2	Advanced Station
2	Advanced Station
2	Fixed Station

b)

Blood donors dataset					
Donor ID	**BCT_Other**	**BCT_Advanced_Station**	**BCT_Fixed_Station**	**BCT_Mobile_Unit**	**BCT_Physical_Space**
2	0	2	1	0	0

Figure 5.2: Nominal features transformation example. a) represents the original dataset (e.g. the donations dataset); b) shows the final step, in which dummy variables were first created, then the data was grouped based on donor ID, and finally, the dummies values were summed.

6. Given the dataset size, it was impossible to compute the Gower distance matrix. It would be necessary to compute a 179 326 * 179 326 distance matrix, which was not feasible due to the required computational cost. Thus, the K-means [63] clustering algorithm was used. It is a very efficient algorithm for clustering large datasetes, which does not require a prior computation of a distance matrix;

7. The Silhouette, Caliński and Harabasz, and Davies-Bouldin indexes were used for clustering evaluation. As of hierarchical clustering, the K-means algorithm also requires the apriori specification of the number of clusters k. Thus, different values of k were tested, and the best one was chosen using a majority vote rule.

5.3 Results

This section describes the evaluation of the blood donors subgroups obtained by applying the proposed clustering methodology.

5.3.1 New donors

Table 5.1 shows the results obtained by clustering the new donors data with Gower's distance measure and Hierarchical clustering. As stated previously, the choice of the number of clusters was based on majority vote rule, i.e., the best value of k is the one that has the highest score in the majority of the metrics considered. Hereupon, the values were best for $k = 2$.

The clustering analysis resulted in 2 blood donors subgroups. Table 5.2 shows a characterisation of each cluster and of the whole dataset used for clustering. It is possible to observe that Cluster 1 joined 25

Table 5.1: Values obtained for clustering evaluation metrics Silhouette, Calinski-Harabasz, and Davies-Bouldin scores, for new donors.

k	Silhouette Score	Calinski-Harabasz	Davies-Bouldin
2	0.119	2 863.940	2.090
3	0.093	1 914.518	2.476
4	0.077	1 609.086	2.393
5	0.067	1 291.457	2.369
6	0.044	1 125.282	2.319
7	0.042	983.746	2.340
8	0.088	1 631.109	2.373
9	0.087	1 501.852	2.450

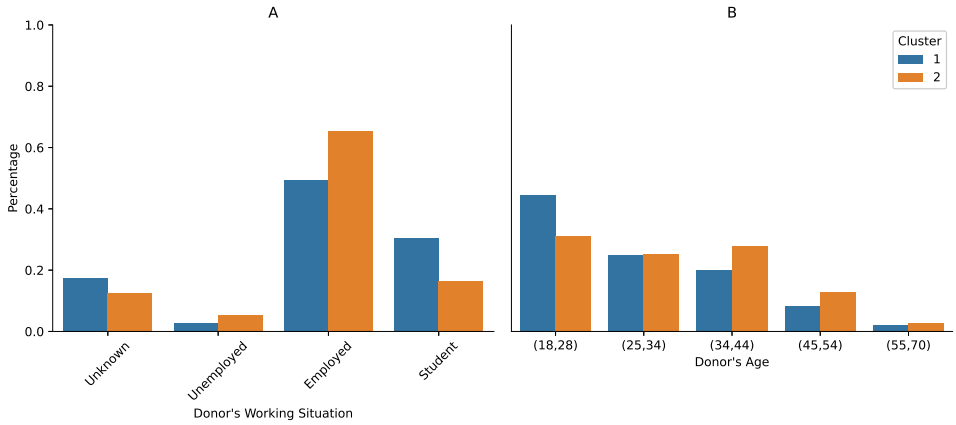

Figure 5.3: Donor's Age (A) and Working Situation (B) distributions, by cluster. The values were normalized, to get a better understating of the differences across distributions.

875 blood donors, while the Cluster 2 grouped 9 553 donors. Regarding statistical significance, all the variables with the exception of Gender, Blood Type, and Reaction had p-values < 0.01.

Figure 5.3 shows donor's age and working situation distributions by Cluster. It is possible to observe that Cluster 1 has more students in comparison with Cluster 2. On the other hand, Cluster 2 has a bigger percentage of employed donors. Regarding age, donors in Cluster 2 tend to be older than those in Cluster 1.

Figure 5.4 shows the distribution of three variables regarding the place where the donation was made. It is possible to observe that regarding the TIPAU topology (Figure 5.4 - A), almost 100% of the blood donors within Cluster 1 had made their donation in APU areas, while that value decreases to around 90% for the Cluster 2 members, with the remaining 20% being split across AMU and APR areas. Regarding

Table 5.2: Characterisation of the 2 clusters obtained from the new donors segment. The ordinal variables are described as medians, and the nominal as modes. p-values for the comparison of the characteristics across clusters.

Characteristic	**Cluster 1**	**Cluster 2**	**Dataset**	**p-value**
Number of Donors	25 875	9 553	35 428	-
Gender	Female	Female	Female	> 0.05
Age	(25,34)	(25,34)	(25,34)	< 0.01
Blood Type	A+	A+	A+	> 0.05
Nationality	Portuguese	Portuguese	Portuguese	< 0.01
Civil Status	Single	Single	Single	< 0.01
Working Situation	Employed	Employed	Employed	< 0.01
Blood Centre Parish TIPAU	APU	APU	APU	< 0.01
Month	June	July	June	< 0.01
Week day	Wednesday	Saturday	Thursday	< 0.01
Week of Month	2	2	2	< 0.01
Weekend	False	True	False	< 0.01
Semester	1	2	1	< 0.01
CST	Lisboa	Porto	Lisboa	< 0.01
District Capital	True	False	False	< 0.01
Metropolitan Area	Other	Other	Other	< 0.01
Donor Parish TIPAU	APU	APU	APU	< 0.01
Reaction	False	False	False	> 0.05
Facility Centre Type	Brigada	Brigada	Brigada	< 0.01
Place	Outra	Outra	Outra	< 0.01
Return after a year	False	True	False	< 0.01

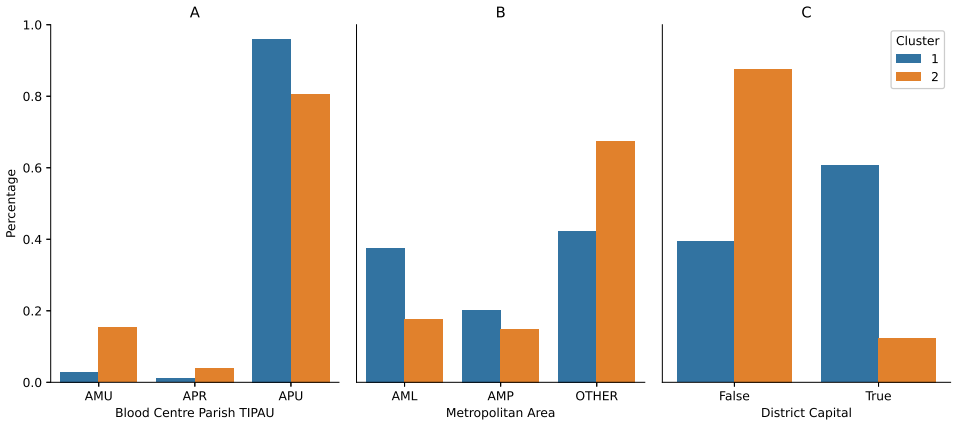

Figure 5.4: Blood Centre Parish TIPAU (A), Metropolitan Area (B), and District Capital (C) distributions, by cluster. The values were normalized, to get a better understating of the differences across distributions.

metropolitan areas, it is possible to observe that more than 60% of the donations in Cluster 2 were made outside AML and AMP areas, while this value stands around 40% in cluster 1 members. Thus, donors in Cluster 2 tend to give more outside the metropolitan areas when compared to donors in Cluster 1. Finally, a noteworthy difference in the distribution can be seen in Figure 5.4 (C): more than 80% of donations from Cluster 2 were made outside the district capitals, while this value decreases to around 40% for the Cluster 1 members. In summary, those three plots show a trend: Cluster 2 members tend to donate outside big metropolitan areas and district capitals compared to Cluster 1 members.

Figure 5.5 shows the Weekday (A) and Weekend (B) distributions by clusters. It is possible to observe a remarkable difference regarding the behaviour of those two types of donors: more than 90% of the Cluster 1 members donate during the week, while more than 80% of those in Cluster 2 donate during the weekend.

Finally, Figure 5.6 shows the distribution, by cluster, of the variable that regards donor return following a year after its first donations. It is possible to observe that donors within Cluster 2 came back for a second donation more often than those in Cluster 1. Just 42% of those within Cluster 1 come for a second donation, while this number increases to 53% for cluster 2 donors.

5.3.2 Experienced donors

Table 5.3 shows the results obtained by clustering the experienced donors' data with the K-means algorithm. Given the majority rule, the number of clusters chosen was 2.

Cluster 1 grouped 63 038 blood donors, while Cluster 2 grouped 116 288 donors. Statistical tests were used to compare the distribution of the features across clusters to understand the two blood donors subgroups formed by the K-means. The Kruskal–Wallis H test [58] was used for the continuous features,

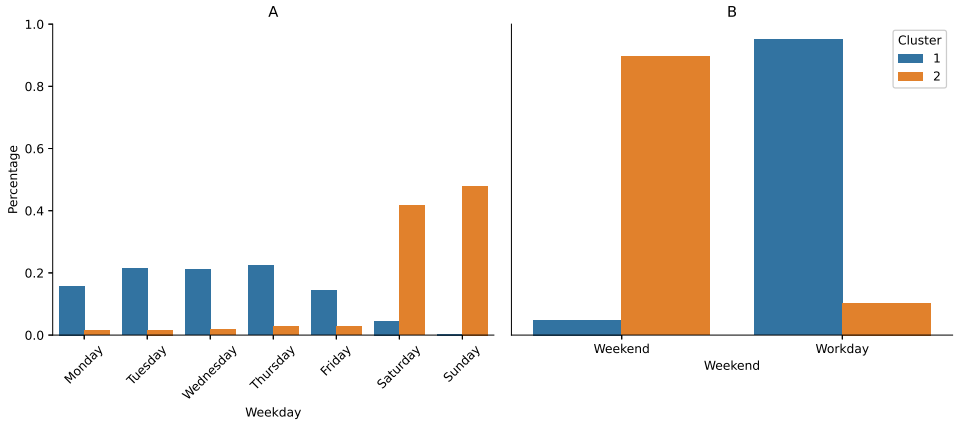

Figure 5.5: Weekday (A) and Weekend distributions by cluster.

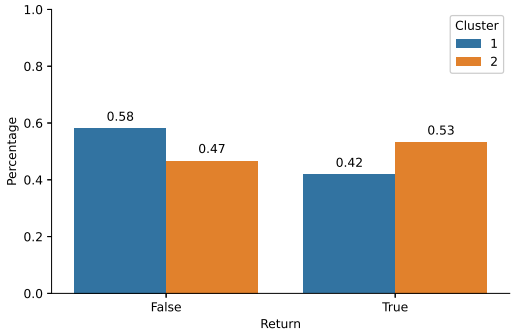

Figure 5.6: New donor's return for a second donation, following a year after its first donation, by cluster.

Table 5.3: Values obtained for clustering metrics Silhouette Score. Calinski-Harabasz Index and Davies-Bouldin Score, for experienced donors.

k	Silhouette Score	Calinski-Harabasz	Davies-Bouldin
2	0.136	12 297.778	3.534
3	0.039	10 194.992	3.466
4	0.052	9 374.231	3.003
5	0.048	8 927.629	2.909
6	0.074	8 440.446	2.663
7	0.068	7 752.553	2.885
8	0.064	7 724.620	2.490
9	0.068	7 381.700	2.447

Table 5.4: Characterization of the two resulting clusters obtained from the experienced donors' segment, according to donor demographics and past behaviour. Ordinal and continuous features are described as median (inter-quartile range), and nominal variables as %. p-values for the comparison of the characteristics across clusters.

Feature	**Cluster 1**	**Cluster 2**	**Dataset**	**p-value**
Male %	56.33	43.65	48.11	<0.01
Female %	43.67	56.35	51.89	<0.01
Age	45-54	35-44	35-44	<0.01
Single %	14.11	49.14	36.83	<0.01
Married %	78.03	41.82	54.55	<0.01
Employed %	88.51	63.27	72.15	<0.01
Unemployed %	8.3	4.86	6.07	<0.01
Student %	0.83	14.77	9.87	<0.01
Number of donations	23.0 (15.0-32.0)	6 (4.0-11.0)	14.14 (5.0-21.0)	<0.01
Mean donations per year	1.93 (1.59-2.33)	1.18 (0.75-1.73)	1.56 (1.0-2.0)	<0.01
Average interval time between donations	199.0 (167.0-243)	367 (249.0-609.0)	435.48 (196.0-461.0)	<0.01

while Chi-squared tests were used for nominal and ordinal features.

Table 5.4 shows the characterization of each cluster, according to donor demographics and past behaviour features [1]. By analyzing it, it gets clear that the K-means algorithm grouped the dataset according to the past behaviour. Donors in Cluster 1 have a median number of donations of 23, while those in Cluster 2 have a median of 9. The median number of donations per year of Cluster 1 members is 1.93, while this value decreases to 1.18 for Cluster 2 members. Regarding the average interval between donations, 50% of the donors within Cluster 1 donated every 199 days or less, while this indicator increases for one donation every 367 days (or less) for Cluster 2 members. This analysis shows that Cluster 1 members could be described as those who donate more regularly.

The difference across clusters is also significant regarding donors demographics. Despite the dataset having more female than male donors, those who donate more regularly tend to be males. Cluster 1 members also tend to be older and have a higher percentage of married and employed donors than Cluster 2 members. On the contrary, Cluster 2 members tend to be younger, almost half of them (49,14%) are single, and have significantly fewer employed donors than Cluster 1 members.

Regarding blood type, no significant differences among clusters were observed. Concerning the features that regard the donation information, for which dummies variables were created and their values were summed after grouping by donor ID, no significant difference among clusters was recorded.

[1] For a better representation, just the significant different characteristics across clusters are shown.

5.4 Conclusions

In several different domains, the literature has shown that an effective technique to increase models performance was the identification of subgroups in a given population and the posterior model training on top of each subgroup. In Chapter 4 the data was segmented according to different geographic characteristics, and machine learning models were trained on top of each segment. However, in this chapter, one was interested in applying unsupervised learning techniques to find homogeneous groups of both new and experienced blood donors.

Different approaches were followed regarding new and experienced donors. For new donors, the donations dataset could also be considered a donor's dataset because each donor just had a single donation. In this case, Gower's distance matrix was calculated and used to feed a Hierarchical Clustering algorithm. For experienced donor's, however, the procedure followed was different because multiple donations were present for each donor. Thus, the blood donations dataset was transformed into a blood donor's donors dataset: dummies were created for each nominal and boolean variable, the data was grouped by donor ID, and the dummies were summed up.

For new donors, two subgroups were found. Despite the Silhouette Scores being low for all the k values tested, it was decided, by majority vote, to select $k = 2$. The cluster analysis performed had shown that two subgroups found were majorly different regarding the day of the week in which donations were made. Cluster 1, which contained 25 875 donations (73% of the dataset), was majorly composed of donations made during the week, while Cluster 2 grouped donations made during the weekend. However, independent statistical analysis had shown that the elements within those two subgroups differed regarding donors demographics. Donors from Cluster 1 tend to be younger, to donate more in metropolitan areas and district capitals, and to donate during the week. Cluster 1 members tend to return less after one year following their first donation than those within Cluster 2. On the contrary, Cluster 2 members tend to be older, to donate during the weekend, and to donate more outside district capitals and the two metropolitan areas. Members of Cluster 2 also tend to return more often after one year following their first donation than members of Cluster 1.

For experienced blood donors, two subgroups were also found: Cluster 1 with 63 038 donors (35,1%) and Cluster 2 with the remaining 116 288 (64,9%) donors. The analysis had shown that those subgroups were significantly different regarding their past behaviour and demographics. Cluster 1 contained donors who donated more in the past, which tend to be male, older, married and employed, and tend to donate more regularly (one donation every 207 days, on average). On the other hand, Cluster 2 grouped those who donated less regularly (one donation every 503 days, on average). Statistical analysis had shown that those who donate less regularly had a significantly higher percentage of females, singles, youngsters, and students when compared to Cluster 1 members.

The results of the clustering tasks might be helpful for IPST. For instance, the clustering of new donors has shown that those who tend to return for a second donation are generally older. However, to maintain a constant and adequate blood supply, efforts should be made to retain young donors by turning

them into regular ones. The clustering of experienced donors had shown a similar pattern: those who donate more frequently are generally older. It might be related to the notion of habit, which is defined in the literature (see Section 5.4) as a semi-automatic performance of a well-learning behaviour. Those who donate more frequently might have formed a habit, while those who donated less have not. Thus, IPST should make efforts to increase the frequency of donation of those who donate less frequently.

The conclusion of this chapter deserves a discussion regarding the clustering algorithms used. Different approaches for clustering the blood donors datasets could have been followed. However, clustering mixed-type data (i.e. data consisting of numerical and categorical variables) is always a challenging task, for which few clustering methods are available [7]. Many clustering algorithms can only handle data that contain either numeric or categorical features [2]. Clustering algorithms group data points into clusters using some notion of similarity, which can be as simple as Euclidean distance for numeric columns. However, computing distance-based similarity measures for categorical data is a challenging task [2].

Nonetheless, the K-prototypes [48] algorithm could have been used. It integrates the K-means, and the K-modes [47] processes to cluster mixed-type data, and, as such, it uses Euclidean distance for numeric values and the Hamming distance for categorical values to measure the similarity between data points. However, evaluating the results of this method is not a straightforward task. As stated by Rabea et al. [7], many cluster validation indices are not suitable for mixed-type data. For instance, the Silhouette Score, one of the most used metrics for clustering evaluation, uses Euclidean distances to measure how far each data point is from its cluster and the nearest cluster. However, using a single distance metric for evaluating clusters that were grouped according to two different distance metrics is not a good approach. Therefore, Rabea et al. [7] implemented an extension for multiple evaluation metrics in order to cover mixed-data types. However, those extensions require the computation of distance matrices, which is not suitable for big datasets.

Chapter 6

Conclusions

This dissertation aimed to identify the determinants of blood donor return. Based on the experiences made, in which machine learning models were trained on top of different blood donations segments, it can be concluded that donors demographics, as well as temporal and geographic information regarding the donation (e.g. when and where the donation was made), are not sufficient to predict donor return. However, the experiences made for donors who have made more than one donation (i.e. experienced donors), and for which it was possible to compute features regarding their past behaviour, has shown that past behaviour features are the best available predictors of donors future behaviour. Nonetheless, it was possible to demonstrate that even the past behaviour features are not enough to predict donors future behaviour.

This dissertation was motivated by a gap identified in the literature. Several studies tried to identify the determinants of donor return by using mainly logistic regression models. Some of them identified demographic factors as important for donor return, such as donor age, sex, or education level. However, none of those studies provided model evaluation metrics. As such, it was argued that if models are weak at emulating the phenomenon that generated the data, i.e., if models are not good at predicting donor return, then the conclusion drawn may be misleading. Thus, this dissertation aimed at finding the determinants of donor return by using the most recent machine learning and model evaluation techniques.

In the literature, it was shown that, in different countries, donors behaviour changed according to the geographic region where the donation was made. For instance, some studies identified different donor behaviours according to the level of urbanization concerning where the donation was made. Those studies stated that those who donate in rural areas tend to return more frequently than those who donate in urban areas. Given that, it was decided to split the donations dataset according to different geographic characteristics and train machine learning models on top of each segment to search for different donors behaviours in different regions. Additionally, the literature also made a distinction between new and experienced donors to find the determinants of return for those two different types of blood donors. Thus, for each data segment, machine learning models were trained for both new and experienced donors.

The models' evaluation results were not promising. In general, models tended to be weak in predicting

donor return. Furthermore, no significant differences in models performances were recorded across the geographic segments. Given the general weakness in model performance, it was not possible to analyse feature importance to search for different determinants in different geographic regions.

However, regarding the segmentation according to new and experienced donors, it was possible to observe significant differences in models trained on top of those two types of donors. The model trained for experienced donors performed significantly better than those trained for new donors. Furthermore, an analysis of feature importance for experienced donors, using the state-of-the-art model agnostic techniques, has shown that the most important features in predicting donor return were the ones that regarded donor past behaviour.

To understand the impact of past behaviour features on models performance, models for experienced donors were trained with just the past behaviour features, and without the past behaviour features. The results have shown that, by training with just the past behaviour features, models reached the same performance as when trained with all the features (i.e. with donor demographics and features regarding the donation). On the contrary, models' performances decreased significantly when past behaviour features were removed.

The general weakness in predicting donor return might be due to different reasons. For instance, it might be the case that donors future behaviour is influenced by variables not available on the dataset. On the other hand, the act of blood donation might be somehow circumstantial or random. A high level of randomness in a given phenomenon would probably generate a dataset with a high level of heterogeneity, which in turn, turns the learning task more challenging for the machines. Thus, one was interested in using unsupervised learning techniques to search for homogeneous groups of blood donors, that might share similar characteristics. One of the major assumptions in the machine learning framework is that the training instances in the training set are identical to each other, belonging to a common distribution. Thus, training machine learning models on top of highly heterogeneous populations may lead to poor performances. Hence, unsupervised learning techniques were used to search for homogeneous groups of both new and experienced donors.

Regarding new donors, two subgroups were found. The first subgroup was composed mainly of donors who donate during the week, which tend to be younger, donate more in metropolitan areas and district capitals, and return less after one year following their first donation. The other subgroup contained donors who tend to be older, to donate during the weekend, and to donate more outside district capitals and metropolitan areas. Donors within that subgroup also tend to return more often one year following their first donated, than those in the first subgroup.

For experienced donors, two subgroups were also found. The clustering algorithm separated the experienced donors majorly according to their past behaviour. In one subgroup were donors who tend to donate more frequently, which tend to be male, older, married and employed. On the contrary, the other subgroup contained donors with significantly higher percentages of females, singles, students, and youngsters, compared with the first subgroup.

The two subgroups found, for both new and experienced donors, has shown that those who returned

more often for second donations, as well as those who donate more frequently, tend to be older. However, to maintain a safe blood supply, IPST should encourage new donors to donate more often, by turning them into regular donors.

6.1 Contributions

The conclusion of this dissertation deserves a reflection regarding its impact on the scientific field. This work presented a contribution to the research topic in several ways: by overviewing and reviewing the literature; by using considerably more data when compared with the related studies; and by using the most recent machine learning, model evaluation and interpretability techniques. The main findings of this dissertation are summarized as follows:

- Contrary to existing literature, our results demonstrate that donor demographics, as well as features regarding the donation, are insufficient to predict donor return;
- Regarding past behaviour features, our results are in line with the literature, which suggests that those are the best predictors of donors future behaviour;
- However, even with the past behaviour features, models achieved considerable low scores, meaning that they were unable to truly distinguish between donors who returned for a future donation and those who don't. As such, it is not possible to state that past behaviour features are accurate predictors of future behaviour;
- Two subgroups of new donors were found, by using clustering algorithms. An analysis of those subgroups found that those who return more often for a second donation tend to be older and to donate during the weekend, mainly outside district capitals and metropolitan areas;
- Two subgroups of experienced donors were also found. The clustering algorithm segmented the experienced donors according to how frequently they donated. An analysis of those subgroups found that those who donate more frequently tend to be male, older, married and employed.

6.2 Future Work

For future work, it would be interesting to apply machine learning models on top of the subgroups found by the clustering algorithms. Probably it would be easier for the machines to predict donor return in the subgroups that return more often and donate more frequently. As such, it would be interesting to find which factors most affect the return on donors that donate more often, as to how those determinants differ from those who do not return so frequently.

Given the general weak performance of the models trained in this work, it was not possible to estimate the impact that donors demographics, as well as the features regarding the donation, have on their future behaviour. Moreover, more research is needed to understand which factors influence donors return the most, and why some donors return more often than others.

References

[1] WHO/Europe | Blood safety - Data and statistics. URL https://www.euro.who.int/en/health-topics/Health-systems/blood-safety/data-and-statistics#. 20, 36

[2] A. Ahmad and S. S. Khan. Survey of state-of-the-art mixed data clustering algorithms. *Ieee Access*, 7:31883–31902, 2019. 78

[3] I. Ajzen. The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2):179–211, dec 1991. ISSN 07495978. doi: 10.1016/0749-5978(91)90020-T. 9

[4] A. S. Alkahtani and M. Jilani. Predicting return donor and analyzing blood donation time series using data mining techniques. *International Journal of Advanced Computer Science and Applications*, 10(8):113–118, 2019. ISSN 21565570. doi: 10.14569/ijacsa.2019.0100816. URL www.ijacsa.thesai.org. 1, 7, 8, 9, 14, 15

[5] P. Apparicio, M.-S. Cloutier, V. Chadillon-Farinacci, J. Charbonneau, and G. Delage. Blood donation clusters in Québec, Canada (2003-2008): spatial variations according to sex and age. *Vox Sanguinis*, 106(4):297–306, may 2014. ISSN 00429007. doi: 10.1111/vox.12082. URL http://doi.wiley.com/10.1111/vox.12082. 9

[6] C. J. Armitage and M. Conner. Social cognitive determinants of blood donation. *Journal of Applied Social Psychology*, 31(7):1431–1457, jul 2001. ISSN 00219029. doi: 10.1111/j.1559-1816.2001.tb02681.x. URL https://onlinelibrary.wiley.com/doi/full/10.1111/j.1559-1816.2001.tb02681.xhttps://onlinelibrary.wiley.com/doi/abs/10.1111/j.1559-1816.2001.tb02681.xhttps://onlinelibrary.wiley.com/doi/10.1111/j.1559-1816.2001.tb02681.x. 2, 10

[7] R. Aschenbruck and G. Szepannek. Cluster validation for mixed-type data. *Archives of Data Science, Series A*, 6(1):P02, 12 S. online, 2020. ISSN 2363-9881. doi: 10.5445/KSP/1000098011/02. 78

[8] M. Ashoori and Z. Taheri. Using clustering methods for identifying blood donors behavior. In *5th Iranian Conference on Electrical and Electronics Engineering (ICEEE2013)*, pages 4055–4057, 2013. 9

[9] F. Atsma, I. Veldhuizen, F. De Vegt, C. Doggen, and W. De Kort. Cardiovascular and demographic characteristics in whole blood and plasma donors: Results from the Donor InSight study. *Transfusion*, 51(2):412–420, feb 2011. ISSN 00411132. doi: 10.1111/j.1537-2995.2010.02867.x. URL https://pubmed.ncbi.nlm.nih.gov/20804526/. 6

[10] R. P. Bagozzi. Attitudes, intentions, and behavior: A test of some key hypotheses. *Journal of Personality and Social Psychology*, 41(4):607–627, 1981. ISSN 00223514. doi: 10.1037/0022-3514.41.4.607. URL /record/1982-09791-001. 2, 10, 39

[11] J. Bergstra, J. B. Ca, and Y. B. Ca. Random Search for Hyper-Parameter Optimization Yoshua Bengio. *Journal of Machine Learning Research*, 13:281–305, 2012. doi: 10.5555/2188385. URL http://scikit-learn.sourceforge.net. 42

[12] P. Bhatia. *Data mining and data warehousing : principles and practical techniques*. Cambridge University Press, Cambridge, United Kingdom New York, NY, 2019. ISBN 9781108727747. 20

[13] B. A. Borkent-Raven, M. P. Janssen, and C. L. Van Der Poel. Demographic changes and predicting blood supply and demand in the Netherlands. *Transfusion*, 50(11):2455–2460, nov 2010. ISSN 00411132. doi: 10.1111/j.1537-2995.2010.02716.x. URL https://pubmed.ncbi.nlm.nih.gov/20529000/. 8

[14] V. Bosnes, M. Aldrin, and H. E. Heier. Predicting blood donor arrival. *Transfusion*, 45(2): 162–170, feb 2005. ISSN 0041-1132. doi: 10.1111/j.1537-2995.2004.04167.x. URL http://doi.wiley.com/10.1111/j.1537-2995.2004.04167.x. 7

[15] L. Breiman. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *https://doi.org/10.1214/ss/1009213726*, 16(3):199–231, aug 2001. ISSN 0883-4237. doi: 10.1214/SS/1009213726. URL https://projecteuclid.org/journals/statistical-science/volume-16/issue-3/Statistical-Modeling--The-Two-Cultures-with-comments-and-a/10.1214/ss/1009213726.fullhttps://projecteuclid.org/journals/statistical-science/volume-16/issue-3/Statistical-Modeling--T. XIII, 2, 14, 15, 16, 17

[16] T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974. doi: 10.1080/03610927408827101. URL https://www.tandfonline.com/doi/abs/10.1080/03610927408827101. 69

[17] C. Carneiro. Administrative reorganization of the territory of the parishes in portugal : a political scheme? Master's thesis, Universidade Católica Portuguesa, http://hdl.handle.net/10400.14/31901, 7 2020. 22

[18] G. Casalicchio, C. Molnar, and B. Bischl. Visualizing the feature importance for black box models. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11051 LNAI:655–670, 2019. doi: 10.1007/978-3-030-10925-7_40. 14

[19] K.-T. Chang. *Introduction to geographic information systems*, volume 4. McGraw-Hill Boston, 2008. 27

[20] J. Charbonneau, M. S. Cloutier, and É. Carrier. Why Do Blood Donors Lapse or Reduce Their Donation's Frequency? *Transfusion Medicine Reviews*, 30(1):1–5, jan 2016. ISSN 15329496. doi: 10.1016/j.tmrv.2015.12.001. URL https://pubmed.ncbi.nlm.nih.gov/26764124/. 5, 6

[21] H.-W. Charng, J. A. Piliavin, and P. L. Callero. Role Identity and Reasoned Action in the Prediction of Repeated Behavior. *Social Psychology Quarterly*, 51(4):303, dec 1988. ISSN 01902725. doi: 10.2307/2786758. URL /record/1989-25838-001. 2, 10, 11, 39

[22] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357, June 2002. ISSN 1076-9757. 45

[23] M. S. Cloutier, P. Apparicio, J. Dubé, J. Charbonneau, and G. Delage. Regional variation in the modeling of donation frequency: The case of Héma-Québec, Canada. *Transfusion*, 52(11):2329–2338, nov 2012. ISSN 00411132. doi: 10.1111/j.1537-2995.2012.03604.x. 6, 26

[24] M. Darwiche, M. Feuilloy, G. Bousaleh, and D. Schang. Prediction of blood transfusion donation. In *2010 4th International Conference on Research Challenges in Information Science - Proceedings, RCIS 2010*, pages 51–56. IEEE Computer Society, 2010. ISBN 9781424448401. doi: 10.1109/rcis.2010.5507363. 8

[25] D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979. doi: 10.1109/TPAMI.1979.4766909. 69

[26] G. E. de Almeida Prado Alves Batista, A. L. C. Bazzan, and M. C. Monard. Balancing training data for automated annotation of keywords: a case study. *Revista Tecnologia da Informação*, 3(2):15–20, 2003. ISSN 15169197. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.10.2192&rep=rep1&type=pdf. 45

[27] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845, 1988. 64

[28] L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, July 1945. doi: 10.2307/1932409. URL https://doi.org/10.2307/1932409. 69

[29] D. Dua and C. Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml. 8

[30] E. ESRI. Shapefile technical description. *An ESRI white paper*, 4:1, 1998. 28

[31] E. Ferguson. Predictors of future behaviour: A review of the psychological literature on blood donation. *British Journal of Health Psychology*, 1(4):287–308, nov 1996. ISSN 1359107X. doi: 10.1111/j.2044-8287.1996.tb00510.x. URL https://bpspsychub.onlinelibrary.wiley.com/doi/full/10.1111/j.2044-8287.1996.tb00510.xhttps://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1111/j.2044-8287.1996.tb00510.xhttps://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/j.2044-8287.1996.tb00510.x. 9

[32] E. Ferguson. Conscientiousness, emotional stability, perceived control and the frequency, recency, rate and years of blood donor behaviour, sep 2004. ISSN 1359107X. URL https://pubmed.ncbi.nlm.nih.gov/15296679/. XIII, 2, 10, 11, 39, 40, 41

[33] E. Ferguson and P. A. Bibby. Predicting future blood donor returns: Past behavior, intentions, and observer effects. *Health Psychology*, 21(5):513–518, 2002. ISSN 0278-6133. doi: 10.1037//0278-6133.21.5.513. URL https://pubmed.ncbi.nlm.nih.gov/12211519/. 2, 10, 11, 39

[34] E. Ferguson and S. Chandler. A stage model of blood donor behaviour: Assessing volunteer behaviour. *Journal of Health Psychology*, 10(3):359–372, may 2005. ISSN 13591053. doi: 10.1177/1359105305051423. URL https://pubmed.ncbi.nlm.nih.gov/15857868/. 9

[35] E. Ferguson, C. R. France, C. Abraham, B. Ditto, and P. Sheeran. Improving blood donor recruitment and retention: Integrating theoretical advances from social and behavioral science research agendas, nov 2007. ISSN 00411132. URL https://pubmed.ncbi.nlm.nih.gov/17958528/. 9

[36] A. Fernández. *Learning from imbalanced data sets*. Springer, Cham, Switzerland, 2018. ISBN 978-3319980737. 45

[37] P. Flach. *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*. Cambridge University Press, USA, 2012. ISBN 1107422221. 17

[38] C. N. Gemelli, J. Hayman, and D. Waller. Frequent whole blood donors: understanding this population and predictors of lapse. *Transfusion*, 57(1):108–114, jan 2017. ISSN 15372995. doi: 10.1111/trf.13874. URL https://pubmed.ncbi.nlm.nih.gov/27774681/. 5, 6

[39] M. Giles and E. Cairns. Blood donation and Ajzen's theory of planned behaviour: An examination of perceived behavioural control. *British Journal of Social Psychology*, 34(2):173–188, 1995. ISSN 20448309. doi: 10.1111/j.2044-8309.1995.tb01056.x. URL https://pubmed.ncbi.nlm.nih.gov/7620844/. 2, 10

[40] M. Giles, C. McClenahan, E. Cairns, and J. Mallet. An application of the Theory of Planned Behaviour to blood donation: The importance of self-efficacy. *Health Education Research*, 19(4):380–391, aug 2004. ISSN 02681153. doi: 10.1093/her/cyg063. URL https://pubmed.ncbi.nlm.nih.gov/15155590/. 2, 10

[41] S. A. Glynn, S. H. Kleinman, G. B. Schreiber, T. Zuck, S. McCombs, J. Bethel, G. Garratty, and A. E. Williams. Motivations to donate blood: Demographic comparisons. *Transfusion*, 42 (2):216–225, 2002. ISSN 00411132. doi: 10.1046/j.1537-2995.2002.00008.x. URL https://pubmed.ncbi.nlm.nih.gov/11896338/. 2, 9

[42] G. Godin, M. Conner, P. Sheeran, A. Bélanger-Gravel, and M. Germain. Determinants of repeated blood donation among new and experienced blood donors. *Transfusion*, 47(9):1607–1615, sep 2007. ISSN 00411132. doi: 10.1111/j.1537-2995.2007.01331.x. URL https://pubmed.ncbi.nlm.nih.gov/17725724/. 2, 9, 12, 15, 66

[43] J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27(4): 857–871, 1971. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2528823. 69

[44] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5), aug 2018. ISSN 0360-0300. doi: 10.1145/3236009. URL https://doi.org/10.1145/3236009. 60

[45] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011. ISBN 0123814790. 20, 21

[46] J. Han, M. Kamber, and J. Pei. Data mining concepts and techniques, third edition, 2012. URL http://www.amazon.de/Data-Mining-Concepts-Techniques-Management/dp/0123814790/ref=tmm_hrd_title_0?ie=UTF8&qid=1366039033&sr=1-1. 8, 20

[47] Z. Huang. A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. 78

[48] Z. Huang and Z. Huang. Clustering large data sets with mixed numeric and categorical values. *IN THE FIRST PACIFIC-ASIA CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING*, pages 21—-34, 1997. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.94.9984. 78

[49] A. B. James, C. D. Josephson, B. H. Shaz, G. B. Schreiber, C. D. Hillyer, and J. D. Roback. The value of area-based analyses of donation patterns for recruitment strategies. *Transfusion*, 54(12):3051–3060, dec 2014. ISSN 15372995. doi: 10.1111/trf.12740. URL https://pubmed.ncbi.nlm.nih.gov/24912544/. 5, 6

[50] R. James and D. Matthews. The Donation Cycle: A Framework for the Measurement and Analysis of Blood Donor Return Behaviour. *Vox Sanguinis*, 64(1):37–42, jan 1993. ISSN 00429007. doi: 10.1111/j.1423-0410.1993.tb02512.x. URL http://doi.wiley.com/10.1111/j.1423-0410.1993.tb02512.x. 14

[51] R. C. James and D. E. Matthews. Analysis of blood donor return behaviour using survival regression methods. *Transfusion Medicine*, 6(1):21–30, 1996. ISSN 09587578. doi: 10.1046/j.1365-3148.1996.d01-46.x. URL https://pubmed.ncbi.nlm.nih.gov/8696444/. 2, 10, 14, 39

[52] A. Karpatne. *Predictive Learning with Heterogeneity in Populations*. PhD thesis, University of Minnesota, 2017. 43, 67, 68

[53] L. Kasraian. Causes of discontinuity of blood donation among donors in Shiraz, Iran: Cross-sectional study. *Sao Paulo Medical Journal*, 128(5):272–275, 2010. ISSN 15163180. doi: 10.1590/s1516-31802010000500006. URL https://pubmed.ncbi.nlm.nih.gov/21181067/. 2

[54] L. Kasraian and M. Maghsudlu. Blood donors' attitudes towards incentives: Influence on motivation to donate. *Blood Transfusion*, 10(2):186–190, 2012. ISSN 17232007. doi: 10.2450/2011.0039-11. URL https://pubmed.ncbi.nlm.nih.gov/22044949/. 2, 9

[55] L. Kaufman and P. J. Rousseeuw, editors. *Finding Groups in Data*. John Wiley & Sons, Inc., Mar. 1990. doi: 10.1002/9780470316801. URL https://doi.org/10.1002/9780470316801. 69

[56] S. Kheiri and Z. Alibeigi. An analysis of first-time blood donors return behaviour using regression models. *Transfusion Medicine*, 25(4):243–248, aug 2015. ISSN 09587578. doi: 10.1111/tme.12177. URL http://doi.wiley.com/10.1111/tme.12177. 2, 13, 15, 66

[57] E. F. Klinkenberg, E. M. Huis In't Veld, P. D. de Wit, A. van Dongen, J. G. Daams, W. L. de Kort, and M. P. Fransen. Blood donation barriers and facilitators of Sub-Saharan African migrants and minorities in Western high-income countries: a systematic review of the literature, apr 2019. ISSN 13653148. URL https://onlinelibrary.wiley.com/doi/full/10.1111/tme.12517https://onlinelibrary.wiley.com/doi/abs/10.1111/tme.12517https://onlinelibrary.wiley.com/doi/10.1111/tme.12517. 13

[58] W. H. Kruskal and W. A. Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952. doi: 10.1080/01621459.1952.10483441. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1952.10483441. 74

[59] M. Kulldorff. A spatial scan statistic. *Communications in Statistics - Theory and Methods*, 26:1481–1496, 06 1997. doi: 10.1080/03610929708831995. 9

[60] S. Lattimore, C. Wickenden, and S. R. Brailsford. Blood donors in England and North Wales: Demography and patterns of donation. *Transfusion*, 55(1):91–99, jan 2015. ISSN 15372995. doi: 10.1111/trf.12835. URL https://pubmed.ncbi.nlm.nih.gov/25178387/. 2, 5, 13, 15, 27, 43

[61] P. London and B. M. Hemphill. The motivations of blood donors. *Transfusion*, 5(6):559–568, 1965. ISSN 00411132. doi: 10.1111/j.1537-2995.1965.tb01206.x. URL https://pubmed.ncbi.nlm.nih.gov/5860077/. 2

[62] S. M. Lundberg and S.-I. Lee. A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf. 61

[63] J. MacQueen. Some methods for classification and analysis of multivariate observations. *https://doi.org/*, 5.1:281–298, jan 1967. URL https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992. 71

[64] H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50 – 60, 1947. doi: 10.1214/aoms/1177730491. URL https://doi.org/10.1214/aoms/1177730491. 70

[65] B. M. Masser, K. M. White, M. K. Hyde, and D. J. Terry. The Psychology of Blood Donation: Current Research and Future Directions. *Transfusion Medicine Reviews*, 22(3):215–233, jul 2008. ISSN 08877963. doi: 10.1016/j.tmrv.2008.02.005. URL https://pubmed.ncbi.nlm.nih.gov/18572097/. 2, 9, 10

[66] A. H. Misje, V. Bosnes, O. Gåsdal, and H. E. Heier. Motivation, recruitment and retention of voluntary non-remunerated blood donors: A survey-based questionnaire study, 2005. ISSN 00429007. URL https://pubmed.ncbi.nlm.nih.gov/16262757/. 2, 9

[67] A. H. Misje, V. Bosnes, and H. E. Heier. Recruiting and retaining young people as voluntary blood donors. *Vox Sanguinis*, 94(2):119–124, dec 2007. ISSN 00429007. doi: 10.1111/j.1423-0410.2007.01004.x. URL http://doi.wiley.com/10.1111/j.1423-0410.2007.01004.x. 1

[68] C. Molnar. *Interpretable Machine Learning*. 2019. XIV, 62

[69] S. Morgenthaler. Exploratory data analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):33–44, jul 2009. ISSN 1939-5108. doi: 10.1002/wics.2. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.2. 28

[70] S. Mukhiya and U. Ahmed. *Hands-On Exploratory Data Analysis with Python*. Packt Publishing, 2020. ISBN 9781789537253. URL https://books.google.pt/books?id=GSR2zQEACAAJ. 28

[71] E. L. Murphy, B. Shaz, C. D. Hillyer, P. Carey, B. S. Custer, N. Hirschler, J. Fang, and G. B. Schreiber. Minority and foreign-born representation among US blood donors: Demographics and donation frequency for 2006. *Transfusion*, 49(10):2221–2228, oct 2009. ISSN 00411132. doi: 10.1111/j.1537-2995.2009.02271.x. URL https://pubmed.ncbi.nlm.nih.gov/19555415/. 13

[72] E. P. Notari Iv, S. Zou, C. T. Fang, A. F. Eder, R. J. Benjamin, and R. Y. Dodd. Age-related donor return patterns among first-time blood donors in the United States. *Transfusion*, 49(10): 2229–2236, oct 2009. ISSN 00411132. doi: 10.1111/j.1537-2995.2009.02288.x. URL https://pubmed.ncbi.nlm.nih.gov/19903284/. 6

[73] A. M. Omoto and M. Snyder. Sustained helping without obligation: Motivation, longevity of service, and perceived attitude change among AIDS volunteers. *Journal of Personality and Social Psychology*, 68(4):671–686, 1995. doi: 10.1037/0022-3514.68.4.671. URL https://doi.org/10.1037/0022-3514.68.4.671. 11

[74] J. A. Ouellette and W. Wood. Habit and Intention in Everyday Life: The Multiple Processes by Which Past Behavior Predicts Future Behavior. *Psychological Bulletin*, 124(1):54–74, 1998. ISSN 00332909. doi: 10.1037/0033-2909.124.1.54. 10

[75] H. E. Ownby, F. Kong, K. Watanabe, Y. Tu, and C. C. Nass. Analysis of donor return behavior. *Transfusion*, 39(10):1128–1135, 1999. ISSN 00411132. doi: 10.1046/j.1537-2995.1999.39101128.x. URL https://pubmed.ncbi.nlm.nih.gov/10532608/. 2, 14

[76] A. Pereira. Performance of time-series methods in forecasting the demand for red blood cell transfusion. *Transfusion*, 44(5):739–746, may 2004. ISSN 00411132. doi: 10.1111/j.1537-2995.2004.03363.x. URL https://pubmed.ncbi.nlm.nih.gov/15104656/. 1, 7

[77] M. Perugini and R. P. Bagozzi. The role of desires and anticipated emotions in goal-directed behaviours: Broadening and deepening the theory of planned behaviour. *British Journal of Social Psychology*, 40(1):79–98, Mar. 2001. doi: 10.1348/014466601164704. URL https://doi.org/10.1348/014466601164704. 11

[78] J. A. Piliavin and P. L. Callero. *Giving blood: the development of an altruistic identity.* Johns Hopkins University Press, 1991. 2, 9, 10, 39

[79] S. Pires, M. Gromicho, S. Pinto, M. Carvalho, and S. C. Madeira. Predicting non-invasive ventilation in als patients using stratified disease progression groups. In *2018 IEEE Interna-*

tional Conference on Data Mining Workshops (ICDMW), pages 748–757, 2018. doi: 10.1109/ICDMW.2018.00113. 67

[80] S. Pires, M. Gromicho, S. Pinto, M. de Carvalho, and S. C. Madeira. Patient stratification using clinical and patient profiles: Targeting personalized prognostic prediction in ALS. pages 529–541. Springer International Publishing, 2020. doi: 10.1007/978-3-030-45385-5_47. URL https://doi.org/10.1007/978-3-030-45385-5_47. 67

[81] E. Priller and J. Schupp. Social and economic characteristics of financial and blood donors in germany. *DIW Economic Bulletin*, 1(6):23–30, 2011. 5, 6

[82] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778. 61

[83] M. T. Ribeiro, S. Singh, and C. Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016. 61

[84] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. ISSN 0377-0427. doi: https://doi.org/10.1016/0377-0427(87)90125-7. URL https://www.sciencedirect.com/science/article/pii/0377042787901257. 69

[85] D. Royse and K. E. Doochin. Multi□gallon blood donors: who are they? *Transfusion*, 35(10):826–831, 1995. ISSN 15372995. doi: 10.1046/j.1537-2995.1995.351096026363.x. URL https://pubmed.ncbi.nlm.nih.gov/7570912/. 9

[86] M. Saarela and S. Jauhiainen. Comparison of feature importance measures as explanations for classification models. *SN Applied Sciences*, 3, 123. doi: 10.1007/s42452-021-04148-9. URL https://doi.org/10.1007/s42452-021-04148-9. 14

[87] M. Saarela and S. Jauhiainen. Comparison of feature importance measures as explanations for classification models. *SN Applied Sciences*, 3(2):272, 2021. ISSN 2523-3971. doi: 10.1007/s42452-021-04148-9. URL https://doi.org/10.1007/s42452-021-04148-9. 60

[88] T. Santhanam and S. Sundaram. Application of CART algorithm in blood donors classification. *Journal of Computer Science*, 6(5):548–552, may 2010. ISSN 15493636. doi: 10.3844/jcssp.2010.548.552. URL https://thescipub.com/abstract/jcssp.2010.548.552. 8

[89] K. S. Schlumpf, S. A. Glynn, G. B. Schreiber, D. J. Wright, W. Randolph Steele, Y. Tu, S. Hermansen, M. J. Higgins, G. Garratty, and E. L. Murphy. Factors influencing donor return. *Transfusion*, 48(2):264–272, feb 2008. ISSN 00411132. doi: 10.1111/j.1537-2995.2007.01519.x. URL https://pubmed.ncbi.nlm.nih.gov/18005325/. 2, 10, 12, 13, 15, 39

[90] G. B. Schreiber, A. M. Sanchez, S. A. Glynn, and D. J. Wright. Increasing blood availability by changing donation patterns. *Transfusion*, 43(5):591–597, may 2003. ISSN 00411132. doi: 10.1046/j.1537-2995.2003.00388.x. URL https://pubmed.ncbi.nlm.nih.gov/12702179/. 2, 39

[91] G. B. Schreiber, U. K. Sharma, D. J. Wright, S. A. Glynn, H. E. Ownby, Y. Tu, G. Garratty, J. Piliavin, T. Zuck, and R. Gilcher. First year donation patterns predict long-term commitment for first-time donors. *Vox Sanguinis*, 88(2):114–121, feb 2005. ISSN 00429007. doi: 10.1111/j.1423-0410.2005.00593.x. URL https://pubmed.ncbi.nlm.nih.gov/15720609/. 2, 10, 14

[92] G. B. Schreiber, K. S. Schlumpf, S. A. Glynn, D. J. Wright, Y. Tu, M. R. King, M. J. Higgins, D. Kessler, R. Gilcher, C. C. Nass, and A. M. Guiltinan. Convenience, the bane of our existence, and other barriers to donating. *Transfusion*, 46(4):545–553, apr 2006. ISSN 00411132. doi: 10.1111/j.1537-2995.2006.00757.x. URL https://pubmed.ncbi.nlm.nih.gov/16584430/. 2

[93] L. S. Shapley. *17. A value for n-person games*. Princeton University Press, 2016. 61

[94] S. S. Skiena. *The Data Science Design Manual*. Springer Publishing Company, Incorporated, 1st edition, 2017. ISBN 3319554433. 20

[95] M. Stone. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, jan 1974. ISSN 2517-6161. doi: 10.1111/J.2517-6161.1974.TB00994.X. URL https://onlinelibrary.wiley.com/doi/full/10.1111/j.2517-6161.1974.tb00994.xhttps://onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1974.tb00994.xhttps://rss.onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1974.tb00994.x. 17

[96] I. M. B. Suárez, A. Fernández-Montoya, A. R. Fernández, A. López-Berrio, and M. Cillero-Peñuela. How regular blood donors explain their behavior. *Transfusion*, 44(10):1441–1446, oct 2004. ISSN 00411132. doi: 10.1111/j.1537-2995.2004.04102.x. URL https://pubmed.ncbi.nlm.nih.gov/15383016/. 2, 9

[97] W. H. O. H. L. Technology, B. S. Unit, G. B. S. Initiative, I. F. of Red Cross, and R. C. Societies. Consensus statement on how to achieve a safe and adequate blood supply by recruitment and retention on voluntary, non-remunerated blood donors, geneva, 8-11 april 1991, 1993. 29

[98] M. C. Testik, B. Y. Ozkaya, S. Aksu, and O. I. Ozcebe. Discovering blood donor arrival patterns using data mining: A method to investigate service quality at blood centers. *Journal of Medical Systems*, 36(2):579–594, apr 2012. ISSN 01485598. doi: 10.1007/s10916-010-9519-7. URL https://link.springer.com/article/10.1007/s10916-010-9519-7. 7

[99] R. A. Thomson, J. Bethel, A. Y. Lo, H. E. Ownby, C. C. Nass, and A. E. Williams. Retention of 'safe' blood donors. *Transfusion*, 38(4):359–367, 1998. ISSN 00411132. doi: 10.1046/j.1537-2995.1998.38498257374.x. URL https://pubmed.ncbi.nlm.nih.gov/9595018/. 13

[100] I. Tomek et al. An experiment with the edited nearest-nieghbor rule. 1976. 45

[101] S. Varma and R. Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics 2006 7:1*, 7(1):1–8, feb 2006. ISSN 1471-2105. doi: 10.1186/1471-2105-7-91. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-7-91. 17

[102] G. A. Veltri. Big data is not only about data: The two cultures of modelling. *Big Data & Society*, 4(1):2053951717703997, 2017. doi: 10.1177/2053951717703997. URL https://doi.org/10.1177/2053951717703997. XIII, 15

[103] T. Volken, A. Buser, A. Holbro, T. Bart, and L. Infanti. Blood donor to inactive donor transition in the Basel region between 1996 and 2011: a retrospective cohort study. *Vox sanguinis*, 109(2):155–162, aug 2015. ISSN 1423-0410. doi: 10.1111/VOX.12269. URL https://pubmed.ncbi.nlm.nih.gov/25900049/. 43

[104] C. Weidmann, S. Schneider, D. Litaker, E. Weck, and H. Klüter. A spatial regression analysis of German community characteristics associated with voluntary non-remunerated blood donor rates. *Vox Sanguinis*, 102(1):47–54, jan 2012. ISSN 00429007. doi: 10.1111/j.1423-0410.2011.01501.x. URL https://pubmed.ncbi.nlm.nih.gov/21534984/. 43

[105] J. C. Wiersum-Osselton, T. Marijt-Van Der Kreek, A. Brand, I. Veldhuizen, J. G. Van Der Bom, and W. De Kort. Risk factors for complications in donors at first and repeat whole blood donation: A cohort study with assessment of the impact on donor return. *Blood Transfusion*, 12(SUPPL.1), 2014. ISSN 17232007. doi: 10.2450/2013.0262-12. URL https://pubmed.ncbi.nlm.nih.gov/23867173/. 5

[106] E. P. Williams, P. R. Harper, and D. Gartner. Modeling of the collections process in the blood supply chain: A literature review. *IISE Transactions on Healthcare Systems Engineering*, pages 1–12, 2020. ISSN 24725587. doi: 10.1080/24725579.2020.1776426. URL https://www.tandfonline.com/doi/abs/10.1080/24725579.2020.1776426. 1

[107] C. J. Willmott. ON THE VALIDATION OF MODELS. *Physical Geography*, 2(2):184–194, 1981. doi: 10.1080/02723646.1981.10642213. URL https://doi.org/10.1080/02723646.1981.10642213. 16

[108] I.-C. Yeh. UCI Machine Learning Repository: Blood Transfusion Service Center Data Set, 2008. URL https://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center. 8

[109] A. Zien, N. Krämer, S. Sonnenburg, and G. Rätsch. The Feature Importance Ranking Measure. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5782 LNAI(PART 2):694–709, 2009. doi: 10.1007/978-3-642-04174-7_45. URL https://link.springer.com/chapter/10.1007/978-3-642-04174-7_45. 14

Appendix A

Results for experienced donors, Aveiro and Lisboa

Table A.1 shows the results for all the classifiers/pipelines regarding experienced donors who donated in Aveiro.

For some experiences, the ADB classifier thrown an error during the training. The error is related with the communication protocol between processes (models were trained using multi-processing[1]), when using pickle[2] objects of a considerable size, under non-Windows environments. This bug[3] was fixed on Python 3.8, however, this models were trained under Python 3.7.3, and it was not possible to retrain them.

[1] https://docs.python.org/3/library/multiprocessing
[2] https://docs.python.org/3/library/pickle
[3] https://bugs.python.org/issue35152
[4] All the features available, or the 7 most significant features according to the chi-squared tests

Table A.1: Results for all classifiers and pipelines trained on experienced donors in Aveiro.

Classifier	Sampling	Features used[4]	Mean AUC	STD AUC
Random Forest		All	0.604963	0.006343
Random Forest		7	0.599390	0.007160
Random Forest	SMOTETomek	All	0.610784	0.009516
Random Forest	SMOTETomek	7	0.617384	0.007790
Random Forest	TomekLinks	All	0.628555	0.009536
Random Forest	TomekLinks	7	0.612555	0.008458
MLP		All	0.569029	0.030757
MLP		7	0.561436	0.021200
MLP	SMOTETomek	All	**0.676893**	0.010182
MLP	SMOTETomek	7	0.666397	0.011323
MLP	TomekLinks	All	0.604214	0.015083
MLP	TomekLinks	7	0.588613	0.017479
Gradient Boosting		All	0.592055	0.007713
Gradient Boosting		7	0.576888	0.006481
Gradient Boosting	SMOTETomek	All	0.603308	0.006992
Gradient Boosting	SMOTETomek	7	0.633432	0.011989
Gradient Boosting	TomekLinks	All	0.615536	0.008599
Gradient Boosting	TomekLinks	7	0.597102	0.005924

Table A.2: Results for all classifiers and pipelines trained on experienced donors in Lisboa.

Classifier	Sampling	Features used	Mean AUC	STD AUC
Random Forest			0.678459	0.004292
Random Forest		7	0.643043	0.004291
Random Forest	SMOTETomek		0.679678	0.004714
Random Forest	SMOTETomek	7	0.650002	0.004179
Random Forest	TomekLinks		0.696704	0.003621
Random Forest	TomekLinks	7	0.651755	0.005047
MLP			0.678380	0.010670
MLP		7	0.682102	0.010132
MLP	SMOTETomek		**0.708058**	0.005488
MLP	SMOTETomek	7	0.707395	0.006365
MLP	TomekLinks		0.694556	0.007865
MLP	TomekLinks	7	0.683363	0.006582
Gradient Boosting			0.680328	0.005562
Gradient Boosting		7	0.679177	0.006393
Gradient Boosting	SMOTETomek		0.685398	0.006623
Gradient Boosting	SMOTETomek	7	0.703071	0.005785
Gradient Boosting	TomekLinks		0.698690	0.005676
Gradient Boosting	TomekLinks	7	0.685632	0.009922

Appendix B

DeLong statistical test p-values

Table B.1: DeLong statistical test p-values, for Aveiro's models

	Just past features	Without past features	With all features
Just past features			
Without past features	2.58E-28		
With all features	0.61	9.07E-32	

Table B.2: DeLong statistical test p-values, for Lisboa's models

	Just past features	Without past features	With all features
Just past features			
Without past features	1.62E-17		
With all features	0.48	6.23E-32	