Intelligent Profiling of Blood Donors in Ireland

Joanna Kossakowska
Institute of Technology, Tralee, Co. Kerry, Ireland

Follow this and additional works at: https://sword.cit.ie/allthe

Part of the Artificial Intelligence and Robotics Commons, and the Cardiovascular System Commons

Recommended Citation
Kossakowska, Joanna, "Intelligent Profiling of Blood Donors in Ireland" (2017). Theses [online]. Available at: https://sword.cit.ie/allthe/329

This Master Thesis is brought to you for free and open access by the Dissertations and Theses at SWORD - South West Open Research Deposit. It has been accepted for inclusion in Theses by an authorized administrator of SWORD - South West Open Research Deposit. For more information, please contact sword@cit.ie.
Intelligent Profiling of Blood Donors in Ireland

Joanna Kossakowska

Master of Science
Institute of Technology, Tralee
2017
INTELLIGENT PROFILING OF BLOOD DONORS IN IRELAND

BY

JOANNA KOSSAKOWSKA

Master of Science

Institute of Technology, Tralee

Supervisors:

Dr. Pat Doody

Mr Andrew Shields

A Thesis Submitted to Quality and Qualifications Ireland in Fulfilment of the Requirements for the Master of Science Degree

May 2017
ABSTRACT

The demand for blood products in Ireland is constantly rising due to population growth and population ageing. It is believed that within the next decade these two factors will present challenges to blood donor recruitment and the availability of blood supplies. Improving the retention of blood donors will have a positive impact on the availability of blood products. Identification of suitable donors with the potential for long-term donating can potentially enhance the predictability of blood supply levels.

This research proposes that the patterns of blood donation behaviours of donors can be isolated from blood donor databases held by blood collecting institutions. These databases typically include only basic information such as age, gender, dates of donations and deferrals from donating. Additionally, the analysis might be enriched with factors related to motivational factors such as intention and attitudes towards blood donation, perceived risks, and personality traits.

Machine learning is the fundamental tool used to analyse these blood donation datasets in an attempt to isolate patterns of donor behaviour and characteristics. It is hypothesised that the patterns discovered can be used to predict an individual’s propensity for blood donation and as a consequence forecast future donations. This analysis attempts to create profiles of different types of donors such as short-term and regular donors, which may help predict the likelihood for individuals to become long-term blood donors.

The experiments completed in this thesis include the application of machine learning algorithms for the purpose of donor classification and the prediction of blood donation behaviours. The datasets used in this study has been kindly provided by the Irish Blood Transfusion Service. Identification of suitable algorithms for profiling of blood donors may aid the forecasting of long-term blood donations and the efficient management of blood stocks. Furthermore, the results of this research may assist policy makers in creating policies to retain future regular donors.
ACKNOWLEDGEMENT

The completion of this study would not have been possible without the support and guidance of Mr Andrew Shields and the expertise of Dr. Pat Doody. I would like to thank both my supervisors for their patience in reviewing my work. Their continued support led me in the right direction.

I would like to thank to Mr Stephen Cousins, the National Donor Services Manager at the Irish Blood Transfusion Service, for providing his knowledge and expertise in blood donation in Ireland.

I would like also to extend my thanks to my family and friends for all your love and support along the way. Special thanks to my Mom and Dawid. I could not have gotten through it without your encouragement and wise words.
# TABLE OF CONTENTS

Abstract ............................................................................................................................. 1  
Acknowledgement............................................................................................................. 2  
Table of Contents .............................................................................................................. 3  
List of Figures ................................................................................................................... 7  
List of Tables .................................................................................................................... 9  

**CHAPTER 1 - INTRODUCTION** .................................................................................. 11  
1.1. The power of blood .......................................................................................... 11  
1.2. Research potential ............................................................................................. 12  
1.3. Research objectives .......................................................................................... 13  
1.4. Thesis structure .................................................................................................. 14  

**CHAPTER 2 – BLOOD DONATION** ............................................................................ 15  
2.1. Introduction ...................................................................................................... 15  
2.2. Blood Donors ................................................................................................... 16  
2.2.1. Voluntary donations .................................................................................. 17  
2.2.2. Irish Blood Donor restrictions ................................................................... 17  
2.3. Factors influencing future donations ................................................................ 18  
2.3.1. Previous donation experience ................................................................... 18  
2.3.2. Convenience .............................................................................................. 20  
2.3.3. Altruism in blood donation ....................................................................... 20  
2.4. Sociodemographic characteristics of blood donors .......................................... 21  
2.4.1. Age and gender ......................................................................................... 21  
2.4.2. Other sociodemographic factors ............................................................... 23  
2.5. Blood donor motivation vs. behaviour ............................................................. 24  
2.5.1. Theory of planned behaviour (TPB) ......................................................... 24  
2.5.2. Motivational determinants of donation behaviour .................................... 25  
2.5.3. History of donation in determining of donors return behaviour ............... 26  
2.5.4. TPB applied to Irish donors ...................................................................... 28  
2.6. Personality traits ............................................................................................... 29  
2.6.1. Five-factor personality model .................................................................... 29  
2.6.2. Five-factor model in blood donation .......................................................... 29
6.2.5. Pre-analysis with multiple regression .....................................................121
6.2.6. Summary .................................................................................................122
6.3. Predicting three-month donations .................................................................123
  6.3.1. LDA ........................................................................................................123
  6.3.2. QDA ........................................................................................................124
  6.3.3. CART ......................................................................................................124
  6.3.4. Multi-Layered Perceptron (MLP) ...........................................................127
  6.3.5. Feed-Forward Neural Networks (FF) ......................................................127
  6.3.6. Probabilistic Neural Network models (PNN) .........................................128
  6.3.7. Summary .................................................................................................128
6.4. Predicting one-year donations .......................................................................130
  6.4.1. CART ......................................................................................................131
  6.4.2. LDA ........................................................................................................133
  6.4.3. MLP ........................................................................................................133
  6.4.4. Summary .................................................................................................134
6.5. Predicting one-year donations of active donors .............................................135
  6.5.1. LDA ........................................................................................................136
  6.5.2. QDA ........................................................................................................136
  6.5.3. CART ......................................................................................................137
  6.5.4. MLP ........................................................................................................138
  6.5.5. Summary .................................................................................................139
6.6. Evaluation of results - summary.....................................................................141
  6.6.1. Comparison of methods ...........................................................................141
  6.6.2. Analysis of the discovered patterns ........................................................144
CHAPTER 7 - CONCLUSIONS ..................................................................................147
  7.1. Main findings and contributions ..................................................................148
    7.1.1. Regular donor profiling ........................................................................148
    7.1.2. Future donor profiling ...........................................................................150
  7.2. Future work .................................................................................................151
Appendix A .............................................................................................................153
Appendix B ..............................................................................................................162
List of References ..................................................................................................173
LIST OF FIGURES

Figure 1 TPB model........................................................................................................25
Figure 2 Example of a directed acyclic graph (DAG).....................................................36
Figure 3 Example of a cyclic graph .............................................................................36
Figure 4 Example of a Bayesian network. ..................................................................37
Figure 5 Bayes classifier assigning variable X into one of two groups: green and blue. 38
Figure 6 LDA class separation example ....................................................................39
Figure 7 An example of a decision tree classifying inputs into class A and B based on age and gender.................................................................42
Figure 8 Group impurity/entropy of different levels....................................................44
Figure 9 Information gain presented on a decision tree..............................................44
Figure 10 Structure of a biological neural cell..........................................................48
Figure 11 A single neutron.......................................................................................49
Figure 12 ANN architecture.....................................................................................50
Figure 13 Examples of linearly and non-linearly separable classification problems.....51
Figure 14 PNN architecture....................................................................................54
Figure 15 PC 1 and PC2 perpendicular in a 2-dimentional space.................................57
Figure 16 Examples of directions of variance (eigenvectors) in a dataset....................58
Figure 17 PCA chart illustrating correlations between model inputs (Table 1) ..........80
Figure 18 PCA illustrating the correlations between inputs and the distribution of class objects (targets set 1)........................................................................81
Figure 19 PCA illustrating the correlations between inputs and the distribution of class objects (targets set 2)........................................................................81
Figure 20 CART classification results with targets set 1............................................86
Figure 21 C5.0 Classification results with targets set 1.............................................87
Figure 22 C5.0 classification results with targets set 1.............................................88
Figure 23 Classification results for models with increasing number of inputs and with targets set 1......................................................................................92
Figure 24 CART classification tree (targets set 2)......................................................98
Figure 25 C5.0 classification results with targets set 2.............................................100
Figure 26 Classification results for models with increasing number of inputs and with targets set 2....................................................................................104
Figure 27 Distribution of class objects from dataset used for predicting donations (PCA)............................................................................................................117
Figure 28 Generated correlations between inputs used for predicting donations (PCA)............................................................................................................118
Figure 29 Multiple RF regression: inputs to targets set 2.........................................121
Figure 30 Multiple RF regression: inputs to targets set 2 excluding deferrals number and the length of the last deferral.........................................................122
Figure 31 CART tree generated for predicting three-month donations ......................126
Figure 32 CART tree generated for predicting one-year donations............................132
Figure 33 CART tree generated for predicting one-year donations of active donors... 138
Figure 34 PCI results summary for classification inputs and targets set 1 ................. 153
Figure 35 Summary of LDA model generated for donor classification (targets set 1). 153
Figure 36 Summary of LDA model generated for donor classification (targets set 2). 154
Figure 37 CART classification tree (targets set 1)..................................................... 158
Figure 38 C5.0 classification tree (targets set 1)........................................................ 159
Figure 39 CART classification tree (targets set 2)........................................................ 160
Figure 40 C5.0 classification tree (targets set 2)........................................................ 161
Figure 41 Regression: donations number vs donor gender............................................. 162
Figure 42 Regression: donations number vs. length of intervals between first two donations .......................................................... 162
Figure 43 Regression: years of donating vs. number of donations................................. 163
Figure 44 Regression: first donation age vs. months between first two donations........ 163
Figure 45 Regression: next donation vs. time since the last donation............................ 164
Figure 46 Regression: average donation interval vs. donating period............................. 164
Figure 47 Regression: gender vs. average donation interval ............................................. 165
Figure 48 Regression: donor gender vs. donating period ................................................ 165
Figure 49 Regression: average deferral length vs. donating period............................... 166
Figure 50 Regression: average deferral length vs. number of donations.......................... 166
Figure 51 Regression: number of deferrals vs. number of donations............................ 167
Figure 52 Regression: average deferral length vs. donor age......................................... 167
Figure 53 Regression: average deferral length vs. donor gender..................................... 168
Figure 54 Regression: first donation age vs. donating period........................................ 168
Figure 55 CART tree generated for predicting three-month donation............................. 171
Figure 56 Tree generated for predicting one-year donations (active donors only)............ 172
### LIST OF TABLES

Table 1  List of inputs used for classification of donors ................................................................. 78
Table 2  Donor Classification results of DT algorithms for target set 1 ........................................... 94
Table 3  Donor Classification results of ANN models for targets set 1 ............................................. 95
Table 4  Donor Classification results of DT algorithms for target set 2 ........................................... 106
Table 5  Donor Classification results of ANN models for targets set 2 ........................................... 107
Table 6  List of inputs used for predicting donations ........................................................................... 116
Table 7 Three-month donation prediction with LDA, QDA, CART, MLP, and feed-forward models .......................................................................................................................... 130
Table 8 One-year donation prediction with LDA, CART and MLP models ........................................ 135
Table 9 One-year donation prediction with LDA, QDA, CART and MLP models (active donors) .......... 140
Table 10 Donor classification results with CART (targets set 1) .......................................................... 154
Table 11 Donor classification results with C5.0 (targets set 1) ............................................................ 154
Table 12 Donor classification results with RF (targets set 1) ............................................................. 154
Table 13 Donor classification results with Feed-Forward network (NNET) (targets set 1) ................. 155
Table 14 Donor classification results with MLP and back propagation (targets set 1) ...................... 155
Table 15 Donor classification results with MLP (RSNNS) and back propagation (targets set 1) ......... 155
Table 16 Donor classification with MLP (RSNNS) and back propagation with momentum (targets set 1) ............................................................................................................................ 155
Table 17 Donor classification with PNN (targets set 1) ....................................................................... 156
Table 18 Donor classification results with CART (targets set 2) .......................................................... 156
Table 19 Donor classification results with C5.0 (targets set 2) ........................................................... 156
Table 20 Donor classification results with RF (targets set 2) ............................................................. 156
Table 21 Donor classification with Feed-Forward network (NNET) (targets set 2) ......................... 157
Table 22 Donor classification with MLP (targets set 2) ....................................................................... 157
Table 23 Donor classification results with MLP (RSNNS) and back propagation (targets set 2) ...... 157
Table 24 Donor classification with PNN (targets set 2) ....................................................................... 157
Table 25 Donation prediction with LDA (three-month) ..................................................................... 168
Table 26 Donation prediction with QDA (three-month) ................................................................... 168
Table 27 Donation prediction with CART (three-month) .................................................................. 169
Table 28 Donation prediction with MLP (three-month) ................................................................... 169
Table 29 Three-month donation prediction with Feed-Forward network (NNET) ......................... 169
Table 30 Three-month donation prediction with PNN ...................................................................... 169
Table 31 One-year donation prediction with CART .......................................................................... 169
Table 32 One-year donation prediction with LDA ............................................................................ 169
Table 33 One-year donation prediction with MLP ............................................................................ 170
Table 34 One-year donation prediction with LDA (active donors only) ............................................ 170
Table 35 Three-month donation prediction with QDA (active donors only) .......... 170
Table 36 Donation prediction results with CART (active donors only) .............. 170
Table 37 Donation prediction results with MLP (active donors only) ............... 170
CHAPTER 1

Introduction

1.1. The power of blood

Every year blood transfusions save millions of human lives. Transfusions are used to replace lost blood and blood products and are a vital part of numerous medical conditions such as surgical practices, trauma, and cancer (O’Reilly 2004). Transfusions make it possible to perform increasingly complex medical and surgical procedures that are becoming more and more popular. They result in the improved life expectancy and quality of life of thousands of patients suffering from many acute and chronic conditions (WHO 2010).

National blood transfusion service institutions are responsible for ensuring access to safe blood transfusions for every patient within their particular jurisdiction. They are charged with ensuring sufficient blood supply as well as the quality of all blood collected. The blood from transfusions must be tested against diseases transferable through blood to
avoid disease transmission from a donor to the patient (IBTS 2010). The tested blood supplies must be properly managed to ensure consistent availability of sufficient supplies of every blood type. Managing blood supplies to meet demand is complex and costly as all blood components have a limited shelf-life and storage of is very expensive (WHO 2010) (Nagurney et al. 2012) (IBTS 2013). The proper management of blood supplies requires constant replenishment.

The challenge that national blood transfusion institutions are constantly facing is retaining the pool of regular and voluntary blood donors since these are the most efficient and reliable source of transfusable blood. Developing ways of identifying future and indeed regular donors would increase the predictability of donations, and thus, improve manageability of blood supplies (Godin et al. 2007).

1.2. Research potential

Research has proven the effectiveness of data mining techniques in profiling objects and people and indeed even predicting their behaviours (Palomino-Garibay et al. 2015) (Zhe et al. 2015) (Kandias et al. 2016). Data mining sub-disciplines such as machine learning and predictive analytics have been previously applied in numerous medical applications. One particularly successful application was in the prediction of illnesses based on patients’ symptoms and demographic data (Malpani & Lu 2011).

Data mining has already been used in the analysis of existing blood donors’ data in order to identify patterns of donors’ personal and sociodemographic features, and their associated donating behaviours (Boonyanusith & Jittamai 2012) (Ashoori & Taheri 2013) (Ramachandran et al. 2011). Further research on profiling of blood donors is needed to enable predictability of blood supplies (Boonyanusith & Jittamai 2012) (Sharma & Gupta 2012) (Santhanam & Sundaram 2010) (Sundaram & Santhanam 2011).

Previous research on blood donors’ data has been mainly based on historical data of donations and questionnaires for retrieving motivational or sociodemographic factors (Boonyanusith & Jittamai 2012) (Sundaram & Santhanam 2011) (Ferguson, C. France, et al. 2007). Motivational factors of regular blood donors have been subject to thorough
research in the field of psychology and also by the application of machine learning techniques (Mostafa 2009). Research suggests that combining knowledge of the sociodemographic conditions of individuals with the analysis of motivational factors could improve the analysis (Holdershaw et al. 2003). Some research was also performed to identify character traits of regular donors. All these three types of analysis have been recommended to isolate features distinctive to individuals donating regularly (Ferguson 2004) (Masser et al. 2008).

Current donor acquisition and retention strategies need to be revised (Sojka & Sojka 2008). According to McMahon & Byrne, it is important that separate strategies be used to attract different types of donors (McMahon & Byrne 2008). Learning donors' motives and features may contribute to both the refinement of donor acquisition and retention policies and possibly the predictability of donations. Intelligent classification methods may facilitate distinguishing between groups of donors. Getting to know each of these groups' distinctive characteristics will allow institutions to adjust blood collection strategies to best fit each class of donors (Harrington et al. 2007).

1.3. Research objectives

This research aims to identify methods and algorithms to mine blood donor data and discover how to identify future regular donors. A number of machine learning algorithms are investigated and examined. The effectiveness of the algorithms are compared so as to determine the most efficient data mining models for predicting blood donations. Detailed objectives of the research are presented in the following research questions:

1) Can regular donors be isolated from the available donor databases using intelligent classification methods?
2) Can donors' future participation in blood donation be predicted?
3) Does the IBTS dataset include data patterns that could be used for profiling blood donors?
4) What factors influence the donation behaviour of the Irish blood donors?
1.4. Thesis structure

Chapter 2 contains a review of the current literature related to blood donor profiling, presenting relevant opinions and conclusions. It presents some background information on the blood donation situation in Ireland and highlights the need for analysing blood donor data. Factors that influence blood donation behaviour are also outlined.

Chapter 3 contains a review of the relevant literature pertaining to the machine learning techniques implemented in the experiments undertaken as part of this work.

Chapter 4 describes the datasets collected for the study and the ethical issues related to the availability. The approach taken and the methodology used in performing the research is presented.

Chapter 5 and Chapter 6 present the experiments performed. Chapter 5 experiments focus on blood donor classification. Chapter 6 experiments focus on predicting donations. Each chapter starts with a description of the pre-analysis of the relevant dataset generated from the database samples acquired from the Irish Blood Transfusion Services.

Chapter 7 contains the conclusions from the research performed. This chapter also includes suggestions for future work.
CHAPTER 2

Blood Donation

2.1. Introduction

The World Health Organization (WHO) and the Institutional Federation of Red Cross & Red Crescent Societies (IFCR) are the two bodies that provide guidance and support for countries working to establish efficient blood donor programs (WHO 2010). The WHO suggests that each state develop a national blood policy. This policy should specify blood bank structure and operations as well as a blood donor program that provides plans for new donor recruitment and regular donor retention. In Ireland, this responsibility is placed on the Irish Blood Transfusion Service (Office of the Attorney General 2014) (IBTS 2014b).

According to the WHO, blood donation by 1% of the population is the minimum needed to meet the most basic demands of developed countries (WHO 2010). It has been calculated that for Ireland approximately 3,000 donations per week is sufficient to meet
the demand of the population with over 1,000 people receiving transfusions in Irish hospitals every week. It is estimated that one-quarter of the Irish population will need a blood transfusion at some point in their life (IBTS 2014a). Based on a report from 2009, 38% of the Irish population have donated blood at least once in their life. More recent data reveals that each donor on average will donate 1.6 times a year, while up to 4 donations a year are suggested by the Irish Blood Transfusion Service. (IBTS 2014a)

Even developed countries with advanced systems of blood collection in place struggle to keep blood supplies at sufficient levels. The number of donations is subject to sudden falls for many reasons, for example, when new restrictions on donor eligibility are put in place. Each time new restrictions are introduced new donor recruitment and retention strategies to counteract the decrease in blood supplies are required (IBTS 2012) (IBTS 2013). It is believed that the best way to keep donations at a stable level is to increase the pool of regular donors by encouraging existing donors to donate again (Eder et al. 2009).

2.2. Blood Donors

The quality and safety of blood collected for transfusions is very important. The WHO suggests that all donations should come from voluntary donors and be non-remunerated. Voluntary donations have contributed to the prevention of diseases transferable through blood like HIV or hepatitis viruses (WHO 2010). This can be attributed to the fact that voluntary donors do not gain any economic benefits from their donations and therefore have no reasons to lie about their medical conditions and lifestyle. Paid donors usually come from poorer social groups. Predominantly they are donating for economic reasons. Voluntary donors are also considered to be healthier and to lead a healthier lifestyle. This has been confirmed by the better quality of their blood and the lower number of deferrals among their donations (Pennings 2005) (Eder et al. 2009). A donor being deferred from donating is considered not eligible to donate blood. A person may be given a lifelong deferral or he/she can be deferred for a particular period only. Whether a donor is deferred temporarily or permanently will depend on the specific reason for disqualification.
2.2.1. Voluntary donations

It has been shown that voluntary donors become regular donors far more often. (WHO 2010) (Holdershaw et al. 2003)

Extensive research has taken place to discover factors that make people donate blood voluntarily for the first time and return for repeat donations. Masser et al. explain that different factors determine the behaviour of new and existing voluntary donors. According to their research, the two groups respond to different retention strategies. (Masser et al. 2008) It is important to reiterate that maintaining the pool of current donors is more important and brings better results than recruiting new donors. This approach also has economic benefits. In the Irish blood donor program, each first-time voluntary donor must register with the IBTS. They must provide their personal details, and fill in a health and lifestyle questionnaire. Their medical history is then analysed by the IBTS staff. Prior to a donation being allowed a haemoglobin test must be made on a drop of blood taken from the candidate’s finger. If this test is successful, a blood sample is then sent for further tests to determine blood group and to test for blood transferable diseases such as HIV, hepatitis B, hepatitis C, HTLV and syphilis. Other supplementary tests are made if required for specific patient needs. (WHO 2010)

This process of first-time donor screening is expensive in terms of time and costs. For these reasons, more effort is required to determine the factors affecting existing donors’ return donation behaviour. (Masser et al. 2008) (Cimaroli et al. 2012) (Schreiber et al. 2003) (Saberton et al. 2009)

2.2.2. Irish Blood Donor restrictions

In Ireland, an eligible blood donor must be over 18 years of age and in case of a first time donor not older than 65. He/she must weigh between 50 and 120 kg and cannot have had illnesses such as angina, heart attack, cancer, or chronic fatigue syndrome. A person will not be admitted for a donation if he or she had a blood transfusion after 1 January 1980, or spent at least one year in UK between 1980 and 1996. Many other restrictions exist
that can disqualify a person as a donor permanently or temporarily. Questionnaires, face-to-face interviews and blood tests are used by the medical staff to assess potential blood donors. (citizeninformation 2016).

2.3. Factors influencing future donations

2.3.1. Previous donation experience

Donor recruitment programmes need to consider the donation experience as it significantly affects donation behaviour. The process of taking the blood takes from 8 to 15 minutes and usually the donor can return to their daily routine straight away. (IBTS 2013) Occasionally a donor post donation may feel vasovagal or unwell. They may also experience side effects of blood donations such as bruising, bleeding from the needle site or irritation of a nerve in the arm. These are not serious side effects and are not reported by the staff. Nevertheless, every year there are occasional serious cases where a donor is admitted to the hospital as a result of complications after donating blood (IBTS 2010).

Most of these side effects from donations are trivial. However, it has been proven they may have a strong influence on donors return behaviour (Ferguson, C. R. France, et al. 2007). Indeed, among the most often cited factors that discourage people from donating are vasovagal and other adverse reactions (Ferguson & Bibby 2002), negative experience from donations, (Piliavin 1990) and fear of donating (Ferguson, C. R. France, et al. 2007).

Based on France’s (France et al. 2005) study on the significance of vasovagal reactions, donors experiencing moderate and severe vasovagal reactions are less likely to re-donate by 50% or more. However, the likelihood of such vasovagal reactions is very low. 97% of all vasovagal reactions are considered light. The impact of light vasovagal reactions among first-time donors results in reduced return behaviour by 33% whereas return behaviour of experience donors is affected by light vasovagal reactions by just 20%. According to Piliavin “first the donors must deal with the negative aspects of donations, and later the inner motivations begin to have an impact” (Piliavin 1990). This means that
the more experienced the donor the easier for them is to deal with the unpleasant side effects.

Ferguson and Bibby, as well as Germain et al., state that negative experience as a result of impolite and unprofessional service/staff may also discourage donors from future donations (Ferguson & Bibby 2002) (Germain et al. 2007). Schriber et al. found that poor staff skills was highly rated as a reason for a donor not returning. Slightly less important reasons were physical side effects and the overall length of the process (Schreiber et al. 2006). According to Germain et al., a bad experience from previous donations is more significant for lapsed donors rather than for current regular donors (Germain et al. 2007). In fact, according to Ferguson and France, studies show that the less experienced the donor the higher the chance that the donor will not return to the clinic for a donation after a bad experience (Ferguson, C. France, et al. 2007). These studies found that even the act of observing others experiencing negative effects of a donation is considered a strong impediment to retention of donors. Ferguson and Bibby state that observing others faint caused a reduction in the number of future donations for occasional donors (Ferguson & Bibby 2002).

Anticipated anxiety was also discovered to be highly correlated with blood donation behaviour (Ferguson, C. R. France, et al. 2007). The fear of donating for first-time donors is generally a fear of the unknown. Long-time donors who have experienced or observed the negative effects of blood donation can also suffer from anticipated anxiety. Sojka et al. reported that the second most commonly reported obstacle, after laziness, to becoming a regular blood donor was a fear of needles (Sojka & Sojka 2003) (Schreiber et al. 2006) (Piliavin 1990). One method to reduce the anxiety and fears accompanying first-time donors is through education. Murphy and Healy stated that proper education should be particularly targeted at young people in secondary and third-level education. They suggest having separate strategies that address specific groups of society in order to improve their blood donation awareness. Research shows that the fear of donating is highest among people that have never donated blood and tends to drop after first donation (Murphy & Healy 2010).
2.3.2. Convenience

Convenience is one of the most often cited facilitating factors among donors. Piliavin writes that the lack of convenience decreases altruistic motivation (Piliavin 1990). Convenience is a complex term. It may refer to the location of the blood clinic, ease of access, opening hours, waiting time, and procedure length in time among others. According to Schreiber et al. the lack of a convenient place to donate is the most significant deterrent for all donors and even more important for first-time donors (Schreiber et al. 2006). Godin et al. confirms that among the strongest predictors of return behaviour are the practical concerns such as the convenient location of a donation point, convenient opening hours, or short waiting time before the procedure (Godin et al. 2005) (Schlumpf et al. 2008).

Previous donation experience and factors related to convenience need to be considered in the analysis of donor return behaviour before considering inner motivational factors. Negative experience and lack of convenience builds a perception of barriers (Schreiber et al. 2006) (Schlumpf et al. 2008). Many researchers report that donor recruitment programs should be focused mainly on removing the barriers that stop people from donating (Steele et al. 2008) (Schreiber et al. 2006) (Piliavin et al. 1982) (Germain et al. 2007). However, most programs for encouraging and improving retention of blood donors usually concentrates on their altruistic values (Piliavin & Charng 1990).

2.3.3. Altruism in blood donation

Piliavin has shown that blood donors have an altruistic approach to life and a desire for self-sacrifice (Piliavin & Charng 1990). Bani et al. suggests that female donations are even more likely to be influenced by altruistic concerns due to women being more sensitive to altruistic aspects and more influenced by the perception of the need for blood through periodic calls for donors (Bani & Giussani 2010). In the case of male and younger donors it is less likely that their donations are guided by altruism and social responsibility (Steele et al. 2008). It is common to categorise blood donation as an altruistic act. This is the reason why the majority of donor recruitment campaigns tend to target donors
altruistic motives. This approach is strongly rejected by Schreiber and Germain (Germain et al. 2007) who wrote that:

"Lapsed and current donors were affected in the same way by calls to their altruistic motives. Studies confirm that, lapsed donors do not engage in less altruistic behaviour than currently active donors do. Though, the author proves that effectiveness of calls to donors’ altruistic motives is often overestimated."

Also Oswalt confirmed the predominance of practical reasons over altruistic motives in relation to blood donation (Oswalt 1977). Cimaroli and Misje, through independent studies, (Cimaroli et al. 2012) (Misje et al. 2005) came to conclusions that altruism may prompt contemplation but does not cause behaviour.

Healy reported that most non-donors explain that they do not donate because, in fact, they were never asked to give blood or they had no convenient opportunity. Healy does not believe it is possible that non-donors are devoid of altruistic values but rather they have had no occasion to reveal it (Healy 2006). France and Ferguson claim that no thorough study has been ever conducted on how altruism motivates individuals to become blood donors and that the sociodemographic situation of individuals is the subject of the core analysis performed on blood donors (Ferguson, C. R. France, et al. 2007).

2.4. Sociodemographic characteristics of blood donors

Another set of factors that influence donating behaviour are sociodemographic factors. Age and gender are the sociodemographic characteristics that are most often considered in the analysis of blood donation behaviour. Other factors often considered are donor status, religion, and tendency toward volunteering.

2.4.1. Age and gender

Studies on determinants of blood donation behaviour agree that men are more likely to donate blood but according to Germain et al., women are more loyal donors (Germain et al. 2007). According to Godin, men aged between 50 and 70 years are most likely to donate blood regularly. In particular, men who are married or in a relationship, have
several children and it is known that a blood donation drive took place in their residential
area in last three months (G Godin et al. 2005). Alternate research completed by Morris
suggests that men between 35 and 55 years old, are well educated, married and work in a
management position are most likely to re-donate (Morris 2011). Morris infers that
programs created with the intention of increasing the pool of regular blood donors should
be specifically target young men as they are most likely to become regular donors and
keep donating through their life (Drackley 2010).

Misje et al. noted that among the youngest donor group (aged 17-25) female donations is
higher than male donations. In fact, women are often overrepresented among first-time
donors but become underrepresented among regular donors. According to Gillespie and
Hillyer, there is a significant decline in the number of women donating between four and
eight donations (Gillespie & Hillyer 2002) (Misje et al. 2010) (Misje et al. 2010). Based
on Healy's research, a significant fall in the number of female donors is noticeable in the
age group 25 and 45 years old. This can be attributed mainly to women becoming
pregnant and giving birth (Healy 2006). It has been estimated that only 42- 50% of these
women will return to giving blood after childbirth. The lower number of female donors
is also caused by a higher number of deferrals due to inefficient iron levels in their blood
or their body weight being under 50kg (Healy 2006). It is believed that these women cease
to donate due to practical obstacles and discomfort related to donation itself, but not due
to a loss of motivation (Misje et al. 2010).

Davey suggests that female donors should receive free iron supplements to improve their
haemoglobin levels. He believes it would increase the number of women eligible to
donate blood and consequently increase the pool of female regular donors (Davey 2004).
Female donors are also considered more prone to vasovagal reactions. However, in
general they are seen as much safer donors due to healthier lifestyles and are less likely
to engage in risky behaviours that might reduce the quality of their blood (Piliavin 1990)
(Schreiber et al. 2005) (Ringwald et al. 2010).

Ringwald et al. claim that specific recruitment strategies should be directed to young
females, so they become more familiarised with the process and less anxious of donating
blood (Ringwald et al. 2010). Morris in his study explains that the low percentage of
women regular donors is a result of the donor recruitment and retention strategies that are
addressed mainly to men (Morris 2011). Indeed, Healy writes that men and women have
different responsibility philosophy, whereas men prefer the individual responsibility,
women tend to prefer community responsibility (Healy 2006). Hence, men and women
respond differently to the same blood donor recruitment call.

2.4.2. Other sociodemographic factors

According to Saberton and Drackley, the level of education attained has a positive effect
on blood donation in developed countries such as US, Canada, Australia, and Ireland
(Saberton et al. 2009) (Drackley 2010). In most of the developed countries, the higher
economic status also has a positive impact on donation. According to Healy, higher
income also positively influences donations in Ireland (Healy 2000).

Studies carried out in countries with a relatively high immigration level and with very
diverse society (US and Canada) show, that immigrant status has a negative impact on
one’s future donations. Research in Canada performed by Saberton shows that this effect
is particularly strong in the case of immigrants whose ability to speak the local language
is weak, and who originate from countries outside of the EU. On the other hand, other
studies show that immigrants who do not speak the local language also often fail to
assimilate in society, but insist on their children to taking part in pro-social activities. This
has the effect that children of immigrants are quite likely to become blood donors
(Saberton et al. 2009).

Links between a person’s volunteering tendencies and blood donation tendency seem to
be obvious at first sight. However, some studies have shown that there is no connection
between them. (Godin et al. 2005) Godin found these two activities do not correlate with
each other. On the contrary, Gillespie revealed that “multi-gallon” (multiple) donors have
a greater tendency to participate in humanitarian volunteering efforts compared with
others (Gillespie & Hillyer 2002). According to Gillespie, religion can also has a positive
influence on donation behaviour but interestingly only in countries where the Red Cross
runs the blood supply (Gillespie & Hillyer 2002). Healy explains that this tendency results
from the way in which Red Cross selects their donors (Healy 2006). Although it a secular
organization, the Red Cross are focused mainly on religious organisations and often
approach people in their churches. In countries like Ireland, where the blood supply is the responsibility of the national organisations, religion has no clear effect on donor behaviour (Gillespie & Hillyer 2002).

2.5. Blood donor motivation vs. behaviour

Numerous studies have analysed blood donor behaviour with a primary focus on motivational factors rather than social status and demographics (Gillespie & Hillyer 2002) (G Godin et al. 2005). Theory of Planned Behaviour (TPB) is a method, which has become popular in attempts to profile blood donors as it helps to explain the relationship between human motives and behaviours.

2.5.1. Theory of planned behaviour (TPB)

The Theory of Planned Behaviour (TPB) was developed as a result of research on the relationships between human’s intentions and behaviour by Ajzen and Fishbein (Ajzen 1991).

According to the TPB, behavioural intention is influenced by three factors. These factors are attitude, subjective norms, and Perceived Behaviour Control (PBC). (Figure 1) A person’s attitude is based on their beliefs about the outcomes of the behaviour. The subjective norm of a person is determined by normative beliefs such as a significant others preference and personal motivation. PBC is defined by other factors that may facilitate the behaviour. For example, access to resources and opportunities to perform a particular behaviour successfully (Conner & Armitage 1998) (Holdershaw et al. 2003) (Sutton 1998).
Perceived Behavioural Control (PBC)

Although in TPB intention is considered the main predictor of behaviour, it does not always have enough power to trigger donation behaviour without strong perceived behaviour control. Intention may however be reinforced with PBC (Giles, M; Cairns 1995).

The authors of TPB consider their model open to extensions with other factors related to the particular type of behaviour (Ferguson, C. R. France, et al. 2007) (France et al. 2007) (Godin et al. 2007) (Conner & Armitage 1998). Masser et al. augmented the model by adding self-identity as a blood donor and anticipated regret from not donating to the determinants of intention to donate blood. Their model was successful in predicting 70% of blood donation behaviour (Masser et al. 2009).

2.5.2. Motivational determinants of donation behaviour

In 1990, Piliavin wrote that the intention to donate blood and the habit of donating blood are the only valid predictive factors of blood donation behaviour (Piliavin 1990). Since then numerous studies have been performed that have led to the development of models reporting improved efficiency for predicting donor behaviour. According to Godin (Godin et al. 2005) (Godin et al. 2007) the most important determinants of blood donation behaviour are a perception of control over the donation process and anticipated regret of failing to donate at the expected time. According to Masser’s work, moral norm, self-efficacy and anticipated regret affect one’s blood donation behaviour (Masser et al. 2009). Masser suggests that further improvements to the predictability of donation behaviour

Figure 1 TPB model.
may be accomplished with the addition of variables relevant to the blood donation process (donation anxiety), measures of the blood donation experience (self-identity as a blood donor), and the structural or organisational elements (perceived inconvenience, being too busy, blood donation process taking too long). Masser agreed with the conclusions that France presented in 2007 stating the need to incorporate variables like vasovagal reactions, donor satisfaction, and previous donation experience as a blood donor into the TPB model used to predict donation behaviour was important (France et al. 2007).

According to Holdershaw et al., to improve the TPB models’ efficiency behavioural variables should also be included next to the cognitive variables in the TPB model. They believe adding particular sociodemographic characteristics (age, gender, profession, education, environmental conditions) and past donation behaviour in the model resulted in a TPB model that is more efficient at predicting donations. The comparison of two models (cognitive and behavioural) that the researchers implemented suggest that a purely cognitive approach is better in predicting intention to donate, whereas the behavioural approach works better in predicting donation behaviour (Holdershaw et al. 2003).

### 2.5.3. History of donation in determining of donors return behaviour

As mentioned previously, a donor’s early impression of donating shapes their long term donation behaviour. A study led by Schreiber and Sharma (Schreiber et al. 2005) shows clear correlations between the first donations pattern and the return behaviour. They examined the frequency of donations in the first 12 months and proved it is highly correlated with long-term regular donations:

"Among those giving 1, 2, 3, 4 and > or = 5 donations in the first year, 4%, 11%, 21%, 32% and 42%, respectively, became regular."

Schreiber et al. confirm that donation patterns in the first year in conjunction with demographic variables can predict if an individual will become a long-term regular donor (Schreiber et al. 2005).
In a later study by Masser et al., it is suggested that blood collection agencies should focus on encouraging first-time donors to re-donate during the first eight months. They claim that the first three weeks after donation determines a donor’s intentions and behaviour over the next eight months. Within this post-first donation period, donor’s attitude, perceived control, and identity as a donor are the factors that correlate with return donation behaviour most strongly (Masser et al. 2012).

According to Masser, self-identity as a blood donor may appear after the first donation. However, a previous study by Schreiber et al. shows that self-identity as a blood donor starts much later, in fact after 3 or four donations (Schreiber et al. 2005). Piliavin reports that first-time donors are motivated by social obligation and a feeling of satisfaction that rewards the altruistic donation. Over the next number of donations, donors develop a sense of personal obligation, self-identity and self-satisfaction. Eventually, it may become a custom and routine (Piliavin et al. 1982).

The length of time between the first two donations is considered to have a substantial impact on future donation behaviour for first-time donors. According to Ownby et al., the shorter the donation interval between them the more likely it is that the donor will come back for subsequent donations (Ownby et al. 1999).

Previous studies have shown that donors who are temporarily deferred are more likely not to return for subsequent donations than those who donated successfully. According to Piliavin, this observation is especially true for first-time and new donors. Based on Piliavin’s study of US donors, among new donors who are temporarily deferred, only 2.8% returned for donation within 6 months compared with 27.3% of those who were not deferred. The more experienced a donor the less influence deferrals have (Ngoma et al. 2013). This may be explained by the growing self-identity as a donor (Piliavin 1987).

Development of self-identity as a blood donor is influenced mainly by repeated donation behaviour and the expectation from significant others to keep donating. Regular donors with well-developed self-identity typically influence their friends and family members and often encourage them to become blood donors. According to Piliavin, observing the habits of significant others results in taking their habits and repeat their behaviours. The
tendency to copy others' practices especially applies to children and young people (Piliavin 1990)

2.5.4. TPB applied to Irish donors

In a study of Irish donors by McMahon et al., an augmented TPB model was applied to the population of University College Galway staff and students (McMahon & Byrne 2008). The variables added to the basic TPB were self-identity, anticipated regret, moral norm, and past behaviour. Self-identity and anticipated regret were proven to be significant in predicting donations whereas moral norms and past behaviour were identified as not having a significant effect on donors' behaviour. The results of the study suggest that donors' behaviour was strongly influenced by a personal responsibility to donate. The results also suggested that appeals to moral norms and self-identity effect only existing donors.

McMahon suggests that low moral norm combined with a low response rate to questionnaires (35%) may be a result of the low number people eligible to donate blood among the population targeted in his study. This suggestion may indicate a lack of conviction in a personal obligation to donate blood among non-donors and a tendency to feelings of apathy toward blood donations among the general population. This study suggests that actions towards retention of donors in Ireland should appeal to their sense of moral norm and self-identity as a donor. McMahon also implies that the low number of people eligible to donate in the approached group could be caused by the diverse nationalities among the university staff and students (McMahon & Byrne 2008).

According to another research study of third-level students in Cork, the low number of donations in Ireland might be affected by insufficient knowledge about blood donation. The authors suggest that lack of essential knowledge about the donation process is the main reason why non-donors are afraid of it. According to Murphy and Healy, Irish society needs more education about the safety of the procedure and the demand for the transfusions. Increasing the awareness within the Irish society may in fact help to increase the number of donors in Ireland (Murphy & Healy 2010).
2.6. Personality traits

As well as sociodemographic and motivational factors examined previously, donors’ personality has been explored to examine possible correlations with blood donation behaviour. The main approach applied to such studies has relied on the application of personality models such as Five-Factor personality model.

2.6.1. Five-factor personality model

The Five-Factor personality model (Big Five personality model) suggests there are five broad domains that describe human personalities. The domains are extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience (Goldberg 1990). These Big Five are considered to represent the top level of the personality trait hierarchy. All narrower personality traits and behaviours are included on lower levels of the hierarchy (Paunonen & Ashton 2001). It has been proved that most of the personality-based consistencies in behaviour can be explained by the Five-Factor personality model (Paunonen & Ashton 2001).

In 1982, Goldberg invented a self-report inventory of trait-descriptive adjectives. He created the questionnaire items to provide more contextual information than just single adjectives (Saucier & Goldberg 2002). Multiple studies have confirmed that the Big Five personality model is the most reliable taxonomy in personality research and that Goldberg’s questionnaire allows for accurate and trusted personality description (Gow et al. 2005).

2.6.2. Five-factor model in blood donation

Not much research has been conducted to analyse blood donor behaviour using the five-factor personality model. The most interesting research has been performed by Ferguson in 2004. The study confirmed correlation between blood donation and conscientiousness, emotional stability, tendency to engage in health beneficial behaviours like exercise and
diet as well as the avoidance of risky behaviour such as driving fast. In fact, there is evidence that blood donation has some health benefits (Ferguson 2004).

Ferguson’s study confirmed the positive role of conscientiousness in prosocial behaviours like blood donation (Ferguson 2004). The study also found that the personality domains of conscientiousness and emotional stability were significant predictors of blood donation acts. Conscientiousness in particular is linked with increased donations, especially for male, long-time, regular donors. Emotional stability was shown to be far more significant than conscientiousness in the context of blood donation behaviour for women.

Higher levels of emotional and procedural control also showed as being related to higher frequency of donations. Surgency (extraversion) has a direct effect on males’ past blood donation behaviour. Results of Fergusons’ studies confirmed previously described differences between the factors that shape the blood donation behaviour of males and females (Ferguson & Bibby 2002) (Ferguson 2004).

In the citation below, Ferguson (Ferguson 2004) explains how to understand and use his findings:

"The relationship between personality and behaviour was different for male and female donors. For males, conscientiousness was related to behaviour and it may be that for males, there is a need to concentrate on issues pertaining to the procedures, order, and routines associated with donation. For female donors, emotional stability was related to donor behaviour, in this case suggesting focusing on emotional aspects of donations (e.g. saving life) may be more effective. It is also important for transfusion services to be aware of these gender differences related to personality and use this information in relation to developing programmes for donor recruitment."

The above findings reaffirm the need to consider differences between male and female attitudes towards blood donation when developing blood donor programmes.

2.7. Conclusions

Many researchers from numerous countries have conducted studies to analyse factors that determine blood donation behaviour. However, there have been some researchers that question the reliability of the findings.
Piliavin believes demographic characteristic of blood donors might give an incorrect view of the population that is willing to donate blood. In her opinion, if one group of a population is overrepresented in the analysed group it does not necessarily follow that people within a different demographic group are less willing to donate. She states that the results of studies on the demographic characteristic of blood donors are mostly influenced by “differential targeting of recruitment efforts and scheduling of mobile visits, rather than to real differences in the motivation of different groups” (Piliavin 1990).

Healy believes that the patterns of blood donation differ from country to country (Healy 2006). He has shown that substantial differences in donation rates are recorded even between neighbouring countries of similar culture. Healy suggests that they can result from differences in strategies, social organisations of blood donation supplies present in each of these countries and different types of people addressed. He claims that non-donating people are not necessarily less motivated, but they may not had the opportunity to donate (Healy 2006):

"The difference does not lie in their individual character but in the structure of the collection system that provides them with opportunities to donate"

The observations of Piliavin and Healy highlight how careful researchers need to be in formulating and interpreting the results of blood donor’s data analysis.
CHAPTER 3

Data Mining Concepts

3.1. Introduction

Data mining methods have been extensively used by researchers to analyse blood donor data in an attempt to identify patterns of donor features and behaviours. This chapter presents a review of the literature discussing the most promising techniques including machine learning and data mining for classification and predictive analytics. An explanation of the theoretical fundamentals of the most relevant data mining techniques is presented.
3.2. Data mining

The term "data mining" refers to a very broad area of knowledge discovery, further defined by Larose (Larose 2005) as:

"the process of discovering meaningful new correlations, patterns, and trends in large amounts of historical data, using pattern recognition technologies as well as statistical and mathematical techniques."

Another definition formulated by Giudici (Giudici 2005) states:

"Data mining is the process of selection, exploration, and modelling of large quantities of data to discover regularities of relations that are at first unknown with the aim of obtaining clear and useful results for the owner of the database."

According to Sholom and Weiss (Weiss & Indurkhya 1998), data mining algorithms can be divided into two categories, namely prediction and knowledge discovery. Prediction applications include classification, regression, and time series analysis, whereas knowledge discovery refers to as clustering, visualisation, deviation detection, database segmentation, association rules, summarization, and text mining (Li et al. 2009). Predictive data mining algorithms aim to identify patterns of associations between input variables and outcomes present in historical data. This process of association of inputs and outputs takes place during a learning process. Once the final associations are reached, the learned patterns are applied to new inputs in order to predict their outcomes.

Machine learning can be considered a subset of data mining. Written and Frank (Witten et al. 2016), defined the concept of machine learning as:

"...the technical basis of data mining. It is used to extract information from the raw data in databases—information that is expressed in a comprehensible form and can be employed for a variety of purposes. The process is one of abstraction: taking the data, warts and all, and inferring whatever structure underlies it."

According to Frank and Witten, models learn when they change their behaviour in a way that makes them perform better in the future (Witten et al. 2016). This is key to the whole concept of machine learning. Machine learning algorithms are designed to learn when presented with patterns. Each pattern instance is processed by the algorithm that then tries to predict its class. The predicted class is compared with the expected one and the difference between the actual and expected outcome is expressed as an error. This
measurement of error is used to change the behaviour of the algorithm when processing the next instance. When the learning process is complete, the algorithm should be capable of classifying the class of all new, previously unseen instances correctly.

The use of machine learning has become a traditional approach in predictive data mining, sometimes referred to as “predictive analytics”.

Data mining models are expected to use separate datasets for training and testing. For this purpose, the complete dataset is randomly partitioned into two parts – the training set and the testing set. If the dataset is too small cross-validation techniques may be used to compensate. (Shmueli & Koppius 2011).

The concept behind cross-validation is that the original dataset is randomly divided into $k$ equally sized chunks. The model is trained and tested $k$ times. In each out of $k$ runs, one of the chunks from the training set is omitted and it is used as a test set. Finally, the generalization errors from all $k$ runs are averaged to obtain the overall expected generalization error (Janert 2010).

It is especially important that the training dataset is class-balanced. Namely, the dataset should contain a similar number of occurrences of all classes in the data set. If one class occurs much less frequently than any of the other classes the accuracy metric may not present the true evaluation of the model. It is especially problematic if the less frequent class is the positive class. Then, the classifier may need additional metrics to evaluate the efficiency of the model in predicting the positive class. In addition to accuracy and error rate, precision and recall should be used.

Precision can be defined as the number of correct classifications among all instances labeled positive. It is also called a positive predictive value (PPV):

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

Recall is the fraction of correct classifications among all instances labeled as negative. In binary classification, recall is also referred as sensitivity.
3.3. Supervised learning

With supervised learning, the algorithm learns with the help of a "teacher". The presence of the "teacher" in the training process is the primary discriminant of supervised learning. As the training dataset contains the inputs and the corresponding class, the role of the teacher is to compare the generated answers against the expected prediction present in the training dataset. A prediction error is calculated based on the difference between the generated and expected results. This error is used to update the model, so that the prediction error made in the next run is lower than the previous. The aim of the training process is to achieve the lowest error possible (Abbott 2014). The mapping between the inputs and outputs is the key to supervised learning (Gollapudi 2016).

3.3.1. Bayes’ networks

A Bayesian network is the simplest form of a supervised classifier. It is a statistical model that represents a set of variables and their conditional dependencies via a directed acyclic graph (DAG) (Abbott 2014)

Bayesian networks are drawn as a network of nodes, one for each attribute of the behaviour being classified. The nodes are connected through directed edges in such a way that there are no cycles, see Figure 2. A cycle is a type of graph where every node has a parent and a child as shown in Figure 3.
Edges represent conditional dependencies between the nodes. Those nodes that are not connected to any other nodes in the network represent variables that are conditionally independent of each other.

Provided with a set of behaviour features, the network/graph is used to compute the probability that a behaviour is in a particular class. The behaviour is then assigned to the class with the highest probability.

Figure 4 shows a graphical representation of a sample Bayes' network. The network presents dependencies between four variables ("cloudy", "sprinkler", "rain", and "wet grass") and the conditional probabilities calculated for each variable based on the dependencies between them. Each node is represented by the circle with its conditional probabilities table placed next to it. The network predicts if the grass will be wet or not given the values of the other three variables ("cloudy", "sprinkler", "rain").
The conditional probability is the probability of a variable given that another variable is true, denoted \( P(X|Y) \) (Abbott 2014). In Bayesian networks, the conditional probability of interest is the probability of a node variable given the parent node variable is true. The sum of the conditional probabilities given the parent variable is true and false must be 1 (Bohm, 2010).

The conditional probability is calculated through application of Bayes’ theorem. According to the theorem, the probability of class \( Y \), given that \( X \) condition is true, is stated as:

\[
p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}
\]

*Equation 1*

The overall class probability calculation involves a mean calculation for all the instances representing a single class and computation of the covariance matrix that is later used to normalise the distance between the new inputs and the group mean. The covariance
matrix generalises the notion of variance to multiple dimensions by measuring the square of the standard deviation (the variance) on two inputs rather than a single one (Bishop 2006).

To compute covariance, the standard deviation of input 1 is multiplied by the standard deviation of input 2, and by the correlation between inputs 1 and 2. The covariance matrix is used to normalise the inputs (Abbott, 2014).

In Figure 5, the normalised units of covariance are represented by the rings around the means. The Bayes classifier classifies a new value (X) based on the calculated distances to the mean of each class. This distance is normalised/rounded to the nearest normalised unit of covariance before the shortest one is chosen. The mean for which the shortest distance from the variable is calculated becomes the solution with the highest probability. The Bayes’ network stores these probabilities in the node’s table (Gollapudi 2016).

The classifier uses all the tables to predict the probability of each class for a particular instance. Thus, the probability of a single instance is the product of all the individual probabilities from the various conditional probability tables calculated for a set of edges. Whereas, the probability of the dataset is the product of these products for all instances (Witten et al. 2016).

The learning process of the Bayes’ networks involves searching through the space of possible sets of edges, estimating the conditional probability tables for each set, as
described above, and measuring the resulting network’s quality by the probability of the data given the network, thus, using the “teacher”.

### 3.3.2. Linear Discriminant Analysis

Linear discriminant analysis is a supervised classification method that was developed by Fisher (Bohm & Günter Zech 2010) and is an example representation of the Bayesian approach towards decision-making.

The LDA algorithm reduces the dimensionality of data by maximising the outer variance between the groups and minimaxing the inner variance within the groups, so the groups are easier to distinguish (PennState 2015). The process of dimension reduction converts a set of data of many dimensions into data with lesser dimensions by removing random variables and ensuring that it conveys similar information. As a result, the groups should be easily separated as shown in Figure 6. In situations where the class distributions strongly overlap, simple discriminant analysis does not work.

![Figure 6 LDA class separation example](image)

When two or more groups are known, LDA classifies a new variable X into a group based on the observed values of several continuous variables. Group membership is predicted
based on parameters of the training dataset such as prior class probabilities (Equation 2), the population means (Equation 3) and variance-covariance matrix (Equation 4).

For group \( i \), prior class \( (\pi_i) \) probabilities are calculated according to the following formula, where \( i = 1, 2, \ldots, g \):

\[
p_i = \Pr(\pi_i)
\]

Equation 2

Then, the population mean \( (\mu_i) \) is estimated by the sample mean vectors:

\[
\mu_i = E(X|\pi_i)
\]

Equation 3

The variance-covariance matrix \( \Sigma \) is estimated by using the pooled variance-covariance matrix:

\[
\Sigma = \text{var}(X|\pi_i)
\]

Equation 4

Based on these parameters, the LDA algorithm identifies a combination of features that separate classes, and the attributes of objects that classify them as members of one class. The linear score function for each group \( i \) is computed as:

\[
s_i^l(X) = -\frac{1}{2} \mu_i^t \Sigma^{-1} \mu_i + \mu_i^t \Sigma^{-1} x + \log p_i = d_{i0} + \sum_{j=1}^{p} d_{ij} x_j + \log p_i
\]

where:

\[
d_{i0} = -\frac{1}{2} \mu_i^t \Sigma^{-1} \mu_i
\]

\[
d_{ij} = j^{th} \text{ element of } \mu_i^t \Sigma^{-1}
\]

Equation 5
The class membership of a new variable $X$ is decided based on the linear scores. The variable $X$ is assigned to the population with its largest linear score (Hastie et al. 2009).

### 3.3.3. Quadratic Discriminant Analysis

Quadratic discriminant analysis (QDA) is a non-linear version of LDA. In QDA, the classes are separated by a square surface rather than by a line as in LDA. LDA assumes classes have identical covariance matrices. (Giudici 2005) In QDA, a separate covariance matrix needs to be calculated for each class. In discriminant analysis, an input variable is classified into the group that has the smallest squared distance to the particular point. LDA simplifies the squared distance into a linear function. In QDA, the squared distance calculated for each group is not the same, and therefore not simplified into a linear function. (SAS Institute 2015) QDA should be used instead of the LDA when there is an assumption the variance matrices are not the same for the groups (Minitab 2015).

DA is considered a simple and efficient algorithm for solving prediction problems. It is related to PCA, as they both employ dimensionality reduction methods. Each method looks for linear combinations of variables that best explain the data, but LDA is more focused on finding the difference between classes than the similarities. Both algorithms are often used for feature selection. Unfortunately, models based on Bayes’ theorem are considered less efficient for when there is a large amount of inputs. In such cases, other models should be considered such as decision trees or artificial neural networks (ANN).

### 3.3.4. Decision trees

Decision trees are a widely used supervised modelling technique which recursively divides the training set of inputs until each decision consists only of inputs of one class (Delen 2015). They are considered a powerful classification method (Guazzelli 2012) and have gained popularity due to their simplicity, the clarity of their architecture and for producing processing rules for predicting target output variables.

Decision tree algorithms include two stages: learning and predicting. In the learning stage, a tree is built using training data. In the second stage, the resulting tree is used to generate
a prediction based on new inputs. Decision trees can use nominal and continuous inputs and can also handle missing values. They are recursive models and nonlinear predictors. They can be used for both classification and regression problems (Layton 2015).

They use mathematical techniques, which categorise a set of data records. Equation 6 illustrates how data is presented to a decision tree, where $Y$ represents the classification target variable, and $\mathbf{x}$ represents a vector composed of input variables, $x_1, x_2, \ldots x_k$:

$$ (x, Y) = (x_1, x_2, x_3, x_4 \ldots x_k) $$

*Equation 6*

The decisions taken by the DT algorithm are illustrated on a tree-like graph with nodes and edges, see Figure 7. Each node displays the value of each variable at the decision split. Edges join nodes together. Nodes are often referred to as leaves (internal and external), and edges are referred to as branches, hence the tree analogy.

*Figure 7. An example of a decision tree classifying inputs into class A and B based on age and gender.*

The outputs of a decision tree can be read as a series of “if-then-else” rules used to define the final prediction.
Decision tree algorithms differ in the way the splits are executed (Abbott, 2014). C4.5, C5.0 (successor of C4.5) and CART (Classification and Regression Trees) are the most popular variants of algorithms used to generate decision trees (Abbott 2014).

C5.0 uses Gain Ratio (normalised Information Gain). The Gain Ratio is used by the algorithm to select appropriate attributes as part of the splitting criteria. Information gain used by the C5.0 is based on the concept of entropy. Information gain is explained as:

Information Gain = Entropy(parent) - Weighted Sum of Entropy(Children)

or

\[ IG(T, a) = H(T) - H(T|a) \]

where T is the set of training examples, \( a \) is the variable the information gain is defined for while H() is entropy.

Equation 7
(Strickland 2016)

Entropy measures the level of impurity in a group of examples. It is calculated with by the following formula:

\[ I_E(f) = - \sum_{i=1}^{j} f_i \log_2 f_i \]

Equation 8
(Strickland 2016)
Examples of different levels of group impurity where two groups (green and pink) are considered are presented below:

Figure 8 Group impurity/entropy of different levels

Figure 9 illustrates information gain in a decision tree that classifies inputs into two groups, green and pink. It shows how the group impurity decreases in the child nodes after the split in the parent node:

Figure 9 Information gain presented on a decision tree.

The CART algorithm uses Gini Index in place of the Information Gain used by C5.0 (Gollapudi 2016). CART creates only binary trees, as each node may split into only two branches based on a given criteria. The algorithm always looks for rules that result in the best possible partition. The C5.0 algorithm is not limited to binary trees (Gollapudi 2016).
Gini impurity is based on Gini index and plays a similar role to entropy in information gain. Strickland (Strickland 2016) defined Gini impurity as:

"... a measure of how often a randomly chosen element from the set would be incorrectly labelled if it was randomly labelled according to the distribution of labels in the subset."

Gini impurity for a set of items with J classes, where \( i \in \{1, 2 \ldots J\} \) and \( f_i \) is the fraction labelled with class \( i \) is calculated as:

\[
I_G(f) = \sum_{i=1}^{J} f_i (1 - f_i) = \sum_{i=1}^{J} (f_i - f_i^2) = \sum_{i=1}^{J} f_i - \sum_{i=1}^{J} f
\]

\textit{Equation 9}

(Strickland 2016)

Because C5.0 trees use entropy, they are less efficient for data with an unbalanced class division. According to Abbot, C5.0 will not create any data splits for a dataset with a large imbalance of the targets. When all the splits are generated, the variables need to be selected (Abbott, 2014).

Trees built using C5.0 and CART are often very large and often result in complex trees that over-fit the data and do not generalise the training data (Abbott, 2014). Methods are available which reduce the number of nodes by pruning nodes that are less significant or important to the decision. Pruning may occur during tree creation, "pre-pruning" or after the tree has been created "post- pruning" (Frank 2000).

While there are many advantages of decision trees, they also have some important disadvantages. DTs are unstable in that they are sensitive to disturbed data. Different tree structures are produced because of differences in the distribution of similar data. Even small changes to the training dataset can have a substantial impact on the resulting tree structure. This effect is caused by the hierarchical nature of decision trees - an error on one level is easily propagated all the levels below (Theodoridis 2015). To avoid this, Seiffert & Hammer, recommend combining decision trees with neural networks. (Seiffert & Hammer, 2006). Decision tree algorithms are not considered suitable for predictive
learning on their own, as they seldom provide high enough levels of accuracy without utilising a boosting mechanism (Bohm & Günter Zech 2010).

3.3.5. Boosted trees – Random Forest (RF)

Boosted decision trees can be built using a technique referred to as a random forest (RF) algorithm. A RF classifier uses a number of decision trees to improve the classification rate (Theodoridis 2015). The main idea behind the concept of random forests is that multiple decision trees are built using a random ensemble of predictors (Witten et al. 2016). The trees are trained with a subset of the dataset while the remainder of the dataset is put into a bag to be used for testing, a practice termed “bagging”.

For training, each node in a tree is split according to Gini impurity, as in the CART algorithm (Breiman 2001). The trees generated for training should be as uncorrelated as possible. The algorithm must find the optimal number of variables as while too many improves the tree’s strength it increases the inter-tree correlation. During classification, each simple tree uses a random ensemble of predictor votes for the most popular class. A margin function is defined for each tree that measures the extent to which the average number of votes at a random vector X, Y for the correct class exceeds the average vote for all other categories. The larger the margin, the more reliable the classification. The margin is used to measure the accuracy of the RF class predictions (Breiman 2001).

Given an ensemble of classifiers $h_1(x), h_2(x) \ldots h_K (x)$, with the training set drawn at random from the distribution of the random vector $Y, X$, and where $I (\cdot)$ is the indicator function, the margin function is denoted as:

$$m_g(X, Y) = \text{avg}_k I(h_k(X) = Y) - \max_{j \neq Y} \text{avg}_k I(h_k(X) = j).$$

Equation 10

(Breiman 2001)

The classification of new input vectors is obtained using the majority vote of all trees. In regression, the final prediction is the average of all the predictions performed by all the
simple trees (Bohm, 2010). The final prediction is expressed as the following equation, where the index $K$ runs over the individual trees in the forest (statsoft 2015).

$$ Random Forest Prediction s = \frac{1}{K} \sum_{K=1}^{K} K^{th} \text{ tree response} $$

*Equation 11*

RF calculates the predictive strength of all variables during the training stage. Measuring variable importance starts with fitting the random forest model to the data. During the fitting process, the error is calculated on data not selected for training. For each data point, the error is recorded in a vector and averaged over all the simple trees. Features that produce large values for this score are ranked as stronger predictors than features, which provide small values (Breiman, 2001).

The results of random forest classification are not as easy to interpret as in the case of CART and C5.0 because the outcome cannot be plotted in a single tree. On the other hand, RF is not so prone to overfitting as simple decision trees. The RF algorithm is particularly suitable for datasets where one class is much larger than the other ones, that is, in cases where the prediction error between classes is very unbalanced (Breiman & Cutler 2004). Calculating feature importance is another benefit as it is useful for deciding the features that should be used for building the next classification models.

### 3.3.6. Artificial Neural Networks (ANN)

Artificial neural networks are an important machine learning technique. ANNs solve classification problems via the discovery of patterns in data. They are widely researched and have been proven to have huge potential for pattern discovery in many problem domains. (Bui et al. 2009). An important characteristic of ANN's is their ability to generalize what they have learned. They can however be prone to overfitting. Model overfitting is often a result of an overly complex training model, a relatively small training set or by the performance of too many epochs during the learning process. (Witten et al. 2011) When the model has been overfitted the network tends to performs well on the training dataset, but poorly on unseen data. (Jain, Jianchang Mao, et al. 1996)
ANNs have been inspired by biological neural networks and the way the human brain uses them to learn. A human brain consists of a large number of neural cells called neurons that process the information they receive. The cells massively interact with each other performing complex and parallel processing (Fausett 2010). The structure of a biological neural cell is illustrated in Figure 10. The arrows indicate the direction in which information moves along each cell while being processed and passed to the next cell. In the biological neural networks, the process of cell activation transmission results in the diffusion of chemicals called neuro-transmitters.

![Figure 10 Structure of a biological neural cell.](image)

Dendrites are the branching fibres that extend from the body of the neural cell. They terminate at synapses that perform the electrochemical communication between cells. Synapses receive the activation/information from other cells and present it to dendrites. The cell body is called the soma. It supports chemical processing of the information and the production of neurotransmitters. Inside the soma, the incoming activations are converted into the output activations. The axon transmits the activation from the soma to the synapsis of other neurons (Fausett 2010) (Jain, J Mao, et al. 1996).
In 1943 McCulloch and Pitts introduced a simplified form of a biological neuron and proposed its application in Artificial Intelligence (Kumar & Thakur 2012). The McCulloch and Pitts neuron is a mathematical function that simulates the activity of a biological neural cell (Fausett 2010).

Figure 11 illustrates a single neuron or processing element based on the design of McCulloch and Pitts.

![Figure 11 A single neuron](image)

In the McCulloch and Pitts neuron an incoming signal is presented to the input of the neuron. The summation function totals the input values multiplied by their corresponding weights. A transfer or activation function compares the output of the summation in the previous step to a Threshold Value. If the summation is higher than the neuron’s threshold value, the neuron is said to fire. (Bohm, 2010).

Graphically, an artificial neural network can be represented as a weighted, directed graph, built of multiple nodes that represent neurons as illustrated in Figure 11. Often the neurons are organised in a layered structure. Figure 12 presents an example of such a network. This network is a multi-layered neural network that consists of an input and an output layer as well as a number of hidden layers that are used for the processing of data between the input and output layers (Fig. 3.3.5-1) (Witten et al. 2016).
Figure 12 ANN architecture

The architecture of an ANN defines how the neurons or nodes of a neural network are organised and defines the type and direction of the connections between the nodes.

There are two main types of ANN architecture considered in this thesis:

- Feed-forward networks (data moves only in one direction, they response in the same way to each new input)
- Multi-layered feedback/recurrent networks (contain feedback loops so data can be moved in two directions and inputs to each neuron are updated by feedback paths) (Jain, Jianchang Mao, et al. 1996)

An early single-layer feed-forward network was called the Perceptron. It was one of the earliest types of neural networks that was capable of solving linearly separable problems only. Multi-layer networks overcome this weakness and solve problems that are not linearly separable (Fausett 2010). Multi-layered feed-forward neural networks are often referred to as Multi Layered Perceptrons (MLP) after the Perceptron. Examples of linearly separable and non-separable problems are presented in Figure 13.
During the learning process, networks are trained using examples of inputs fed into their input layers, and the corresponding outputs. Data in the ANN model moves along the layers and is processed at each node (Jain, Jianchang Mao, et al. 1996).

The learning algorithms applied to Multi-layered networks are generally variants of the back propagation algorithm (Rumelhart et al. 1986). Learning with back propagation is based on an error-correction rule. The algorithm consists of two phases, namely a forward pass and a backwards pass.

In the forward pass, the weights on the connections between nodes are initialized with random values. The first record in the dataset is applied as input to the network. This information is passed through the network, from the input layer to the hidden layer or layers and then on to the output layer. At each step, the inputs to each node are processed.

At each node in the hidden layer and the output layer, the weighted sum of the inputs is calculated according to Equation 12. This equation assumes the current node is k, the weight value is the weight on the connection between nodes i and k (first set of weights
\( W^{(1)} \) – between the input layer and first hidden layer) and the input is \( x_k \) (Bohm & Günter Zech 2010).

\[
u_i = \sum_k W_{ik}^{(1)} x_k
\]

_Equation 12_

An activation function transforms the weighted sum of inputs to an output signal. The Sigmoid function is usually used as the activation or transfer function for nodes in the back propagation algorithm. It is denoted by the following equation:

\[
s(u) = \frac{1}{e^{-u} + 1}
\]

where \( u \) is the weighted sum of inputs.

_Equation 13_

The forward pass propagates the input values from the input layer to the output layer and transforms the data at each node in the hidden and output layer.

At the first hidden layer the output is:

\[
x_j = s(\sum_k W_{jk}^{(1)} x_k)
\]

_Equation 14_

At the final output layer, the outcome is calculated as follows:

\[
y_i = s(\sum_j W_{ij}^{(2)} s(\sum_k W_{jk}^{(1)} x_k))
\]

_Equation 15_

The value calculated at the output layer is compared with the expected output given by the training dataset. The difference between the generated output (\( y \)) and desired output (\( y_t \)) is calculated by Equation 16 is referred to as the mean square error (MSE):
This calculated error is propagated backwards to update the weights of the network. This phase is considered the backward pass and its goal is to calculate a weight set that will decrease the difference between the output generated by the network and the target output. The output weights are updated first. Then, the hidden layers weights are also updated in the backward direction.

A single step in training includes one forward pass and one backward pass and is called an epoch. During one epoch, all records are processed once. Training a network may take thousands of epochs and is often considered a time-consuming process (Bohm, 2010).

The performance of the trained network can be tested with data not used during the learning process. This phase is called generalization testing. It measures how well the network can generalize what it has learned to recognise or classify unseen data. If the network provides correct responses to new inputs with a high degree of accuracy it is said that it is a good generaliser (Fausett 2010).

Probabilistic neural networks (PNN) are a variant of ANNs that are expected to learn much faster.

3.3.7. Probabilistic Neural Networks (PNN)

PNN is a multi-layered network that utilises feed-forward connections only. For pattern classification, PNNs use a probabilistic approach based on Bayes decision theory. Specifically the PNN uses a statistical algorithm called kernel discriminate analysis to form an estimate of the probability density functions (pdf's) of categories in a classification problem (Gorunescu 2006).

The PNN architecture has four layers, namely, an input layer, a pattern layer, a summation layer, and an output layer as shown in Figure 14.
According to Mishra and Das (Mishra et al. 2013), the PNN input layer is only used to distribute the input features to the neurons of the pattern layer. The pattern layer contains one neuron for each training instance. Each neuron of the pattern layer takes its input pattern vector and estimates the probability density function (pdf) for that input pattern. It calculates a product of the pattern vector with a weight vector, $Z_i = X_i W_i$, and then passes this sum through an activation function. In contrast to MLP, the activation function is not a sigmoid function, but is based on one of the two following equations, where $\sigma$ is a smoothing factor:

$$\exp[(Z_i - 1)/\sigma^2]$$

*Equation 17*

or:

$$\exp[-(W_i - x)^2(W_i - x)/2\sigma^2]$$

*Equation 18*

The pattern layer produces a vector of distances between the input vector and the training input vectors.
The summation layer neuron applies the result from the previous layer to compute the output for the single pattern. The number of nodes in the summation layer is equal to the number of classes in the training instances. Each summation node receives and sums the outputs from the pattern layer nodes related to the particular class according to the equation below, where \( i \) is the pattern number.

\[
\sum_i \exp[-\frac{(Wi-x)^2}{2\sigma^2}]
\]

*Equation 19*

The output layer node takes two inputs. It produces binary outputs related to two different categories by using the classification criterion, which only compares the two inputs and outputs the class with the bigger probability:

\[
\sum_i \exp[-\frac{(Wi-x)^2}{2\sigma^2}] > \sum_j \exp[-\frac{(Wj-x)^2}{2\sigma^2}]
\]

*Equation 20*

(Gorunescu 2006)

In contrast to MLP, PNN models learning is not an iterative process as it uses a one pass learning algorithm. As a result, learning is much faster than Multi-Layered networks trained using Back Propagation based algorithms (Mishra et al. 2013). A PNN model is guaranteed to converge to the optimal classifier as the size of the training set increases. PNNs do not need a validation dataset so all available data can be used for building the model (Mostafa 2009) Another advantage of PNNs is that training samples can be added and removed easily without the need for extensive retraining. PNN is also not subject to local minima which can be a problem of both Decision Trees and Back Propagation based ANN’s (Gorunescu, 2006).

One difficulty of training a PNN model is that the technique relies heavily on finding the proper value of the smoothing parameter, sigma. To find the optimal value for sigma many models may need to be built and tested. PNN models are also not considered as
general as MLP in regards to their application domains and require a highly representative training set (Mishra et al. 2013).

3.4. Unsupervised learning

Unsupervised learning refers to a number of techniques that look for relationships between inputs in order to assign each input to a particular cluster or class. Unsupervised learning methods are generally applied to datasets that contain inputs that do not have a labelled response. Clustering is a popular unsupervised learning technique that can be used for different purposes such as the pre-processing of data before further analysis, identification of important features, isolating distinct groups, identifying similar regions of data in a data set, visualisation of data, or data mining (Seiffert & Hammer 2006). Different tools are used to help understand clustering result, such as plots, graphs, calculated means, ranges, and standard deviation. While clustering techniques can be particularly useful, the result can often be difficult to interpret.

Clustering algorithms measure the similarity or dissimilarity between a pair of objects. The algorithms partition objects into homogeneous groups so that the similarities between objects of one group are large compared to their similarities to other groups. A dataset is successfully clustered if in a given space objects of the same group are placed close to each other. Clustering algorithms are not generally used to pre-process data. PCA is the primary tool in exploratory data analysis for the selection of variables for predictive models.

3.4.1. Principal Component Analysis (PCA)

Principal component analysis (PCA) represents unsupervised learning. It is often considered for the analysis of input variables and in particular for identifying the most important variables in a data set. The PCA algorithm is also used in descriptive analytics as a dimensionality reduction method (Siekmann 2003). It works best with the data that has linear relationships between inputs (Abbott, 2014). It attempts to reduce the number
of inputs needed to achieve the most optimal classification results. This is achieved by combining variables that are correlated with each other into groups called principal components (PCs). In this way, it identifies the core variables in the dataset.

Principal components are linear projections of the original inputs. They are new variables calculated from the original dataset and include the variance of the analysed inputs. The PCs are identified based on orthogonal transformations of variables. An orthogonal transformation is a linear transformation that preserves the lengths and angles between the vectors (Li et al. 2009). In two-dimensional space if PC1 is orthogonal (perpendicular) to PC2, like on Figure 15, there can be no correlations between them (Abbott 2014).

![Figure 15 PC 1 and PC2 perpendicular in a 2-dimentional space](image)

The PCA computation uses a singular value decomposition (SVD) method. Given a matrix $X$ ($N \times P$), generated from the analysed dataset, the method decomposes this matrix into a unitary matrix $U$ ($N \times P$), a diagonal matrix $S$ ($P \times P$) and a square matrix $V$ ($P \times P$). Matrices $U$ and $S$ have the same size as the original matrix $X$, and matrix $V$ has the size of the number of columns of the original matrix $X$ (Li et al. 2009). The output of the decomposition is the PC score matrix and can be written as:

$$ X = U \cdot S \cdot V^T $$

*Equation 21*

$X$ can be also written as a sum of $p$ covariance matrices, where $i$ refers to the number of columns in the matrices $U$ and $S$, and the number of the variables in the dataset:
These covariance matrices are used to find sets of new axes in the original space, the eigenvectors of the covariance matrix. The new axes are constructed from linear combinations of the features in the space in which the dataset is plotted as shown on Figure 16. They systematically break down the variance in the data points. The directions of the vectors show the extent of the distributions of the data points and are ordered from the most variance to the least (Shlens 2014).

![Direction of least variance](image)

*Figure 16 Examples of directions of variance (eigenvectors) in a dataset*

The result of this process is an ordered list of directions. The axes of directions along which there is the greatest variance are referred to as the "principal components", or PCs.
3.5. Machine Learning Application in Blood Donors Analysis

A number of data mining and machine learning experiments have been conducted on blood donor data samples with the aim to find donation patterns, classify donors and predict donations.

A research paper by Harper (2005) described results of experiments that compare the performance of popular data mining methods such as CART, DA, and MLP. According to the author, the accuracy of the CART and MLP models was similar. Harper claimed that the performance of CART was at least as good as the performance of the MLP and better than the DA. He observed the consistency of results obtained by CART models. Specifically, the accuracy of decision tree models was not prone to the changing datasets characteristics such as the number of records and variables. The run-time scored by decision trees was low in all cases. The ANN models achieved lower accuracy when datasets with high variance or deviance were used (Harper 2005).

Mostafa (2009) described an Egyptian study that compared MLP, PNN, and LDA models. Mostafa's paper is the only published study that describes attempts of blood donor classification with PNN. Although both PNN and MLP models turned out to be 100% efficient, PNN models proved much faster in training and simpler in application. The LDA model scored significantly lower (83.3%) than both PNN, and MLP models in terms of classification. The study conducted by Mostafa used only data extracted from questionnaires. The dataset included basic sociodemographic factors such as age, gender, and education level. It also contained information on donors' altruistic values and motivational factors (Mostafa 2009).

Santhanam et al. (2010) applied DT model using the CART algorithm to analyse a blood donor database that contained information on the donations of 748 randomly selected donors. The dataset included four features including recency – time from the last donation, frequency – number of donations, monetary – total blood donated, and months since the last donation. Based on these four inputs the experiment aimed to predict which donors would give blood within the next month. The decision trees correctly predicted
99.9% of donations and confirmed that future donations of particular donors may be predicted based on their history of donations (Santhanam & Sundaram 2010).

Boonyanusith and Jittamai (2012) used two methods to analyse blood donor data. The first method used a J-48 algorithm (a version of C5.0) to build classification decision trees. The next method generated MLP models using the backpropagation algorithm. Both attempts resulted in almost identical classification accuracy, 76.25% (J-48) and 75.75% (MLP) respectively (Boonyanusith & Jittamai 2012).

A paper by Khemphila presented a general comparison of several data mining models including decision trees generated with the CART algorithm, logistic regression classifiers (LR) and ANNs but not in the context of blood donation. The experiment aimed to develop profiles of individuals with heart disease based on a database from a medical centre, but using without any information on their heart conditions. In this case, the accuracy of the ANN model appeared the highest but again, only slightly higher than the accuracy of CART model (80.2 vs. 79.3) (Khemphila & Boonjing 2010).

The choice between decision trees and ANNs is based on what is most important for the researcher, either maximum accuracy, most likely to be achieved with ANN, or high model interpretability, most likely achievable with DT (Boonyanusith & Jittamai 2012). Researchers for whom training time matters are encouraged to consider the use of the PNN models (Mishra et al., 2013).

3.6. Conclusions

This chapter presented the theoretical background of data mining techniques and discussed the attempts of their application into blood donor analysis. Based on the results of the experiments described, PNN, MLP, and CART appear to be the most promising models for use in blood donors profiling. Hence, these models were the models chosen for the experiments conducted as part of this research. The results of the experiments are presented later in this thesis.
CHAPTER 4

Research Methodology

4.1. Introduction

This chapter presents the methodology of experiments used throughout the rest of the thesis to address the research objectives formulated in chapter 1. A number of different research methods will be adopted, using software implementations of the algorithms discussed in the literature review on the IBTS dataset. Each algorithm will be evaluated by measuring their classification performance on a labelled dataset.

This chapter starts with section 4.2, where the research objectives are restated, and is followed by section 4.3, which justifies the methods used in model analysis. Section 4.4 is where the software that has been used is described followed by a description of the research approach. In section 4.5 the ethics of dealing with potentially sensitive data are discussed and the approach taken to handle these ethical issues. Section 4.6, describes the datasets that has been used for all analysis and evaluations. In section 4.7 the chapter then
describes the evaluation methods used and ends with sections 4.8 summarising the chapter.

4.2. Statement of the research objectives

The primary goal of this thesis is to analyse the IBTS blood donor database using machine learning methods to isolate patterns in donation behaviour of regular donors. The hypothesis is that discoverable patterns are embedded in the database and adequate machine learning models will be able to predict other individuals’ donations and even identify future regular donors from samples of the database. The ability to predict donation numbers at particular time of the year would enable accurate blood collection planning and improve the management of the valuable supplies of blood. Donation behaviour will be identified and used for building profiles of blood donors.

This study attempts to answer the following research questions:

1) Can regular donors be isolated from the available donor databases using intelligent classification methods?
2) Can donors’ future participation in blood donation be predicted?
3) Does the IBTS dataset include data patterns that could be used for profiling blood donors?
4) What factors influence the donation behaviour of the Irish blood donors?

The above questions will be tackled through experiments focused on classifying donor into predefined groups, and predicting their classes based on their previous behaviour. This involves building data mining models that will use different machine learning algorithms for data classification, training them on blood donor datasets, testing their classification accuracy, and comparing the results. The next section briefly justifies the choice of the algorithms.
4.3. Justification of the choice of methods

According to the literature review in Chapter 3, ANN, PNN and decision trees are efficient in the classification of blood donors. Classification methods MLP and PNN deserve particular attention as they both demonstrate classification efficiency close to 100% in experiments run by Mostafa (Mostafa, 2009). Even though PNN is a well-known and recommended classification method due to its processing speed and ease of use, no other study was found that applied this algorithm to analysis of blood donors. Decision trees algorithms, like CART and C5.0, produce comparable but usually slightly weaker results in the reviewed papers dealing with classification of blood donors (Mostafa, 2009), (Harper, 2005), (Boonyanusith & Jittamai 2012) and (Khemphila & Boonjing, 2010). However, according to Bohm, boosted decision trees like RF are recommended over simple decision trees algorithms, as they generally provide better performance (Bohm, 2010). However, CART and C5.0 have been valued for their clarity, which RF lacks.

Therefore, both simple and boosted trees algorithms will be used to build classification models in attempts to find the answer to the first research question and prove that regular donors can be classified based on the available dataset.

To answer the second research question experiments with additional classification algorithms, namely LDA and QDA will be undertaken. These two techniques are traditional statistical methods recommended from the literature for pattern recognition and classification tasks (Mostafa 2009). They perform dimensionality reduction (LDA, QDA) and maximise the class separability. QDA is an extension of LDA and uses quadratic rather than linear functions (Pardoe & Cook 2007). The literature review discussed the study performed by Mostafa, which compared efficiency of LDA against ANN for blood donor classification (Mostafa, 2009). LDA and QDA will be used in the tasks aiming to predict donations.

Consolidated results of different classification methods are hoped to create broad conclusions on the dataset and the classification models. The outputs of both sets of experiments will be used to analyse the data patterns they discover. The existence of the
data patterns is the aim of the third research question. The patterns will be then applied to define profiles of regular and future donors.

Based on the literature review LDA is also a suitable tool for feature selection (Pardoe & Cook 2007), (Mostafa, 2009). PCA is another linear algorithm, often used for feature selection (Chen & Blackwell 2007), (Janert 2010). Both algorithms perform dimensionality reduction. The difference between them is that LDA does this while preserving all the class discriminatory information, whereas PCA always looks for the highest variance among the components. Considering how different the algorithms behave in order to achieve the same goal, they both should be implemented in the feature selection tasks, so the results can be compared and consolidated to formulate efficient results. The results will be used to answer the fourth research question.

4.4. Software

All classification models have been developed using algorithms provided by the publicly available R-project. The project provides a free software environment for statistical computing and graphical representation. It extends easily via packages available in the CRAN (Comprehensive R Archive Network) repositories. These downloadable libraries provide a variety of statistical functions and well-designed graphical techniques. One of the primary advantages to using R is the ability to manipulate the source code and add additional functionality when required. (The R Foundation 2017).

In particular this study uses R implementations of decision trees algorithms (CART, C5.0, and random forest) and several implementations of ANN. All the DT models, FF and some MLP models have been tuned with use of the “train” function provided by the “caret” library. (Kuhn 2015)

The “train” function requires setting of several parameters used for training and tuning. One of the tuning parameters is a summary function that specifies which summary metric will be used to select the optimal model. Two summary functions will be used in the
experiments. The "defaultSummary" function for predicting more than two classes, and
the "twoClassSummary" function will be used in case of the binary classification. The
functions take the observed and predicted values and estimate some measure of
performance (Kuhn 2015). The "tuneLength" parameter denoting the amount of
granularity in the tuning parameter grid will be set to 10 (Kuhn 2017).

Parameters used for training will be defined inside the "trControl" parameter, which holds
a list of values that define how the train function acts.

Decision tree models will be trained using the re-sampling method called "repeatedcv"
which performs repeated 10-fold cross-validation. Three different arguments will be
used with the "method" parameter, which specifies the training algorithm: "rpart" for
CART, "c5.0" for C5.0, and "rf" for random forest. A number of trees will need to be set
for the RF models. The parameter that specified the action to be taken if NAs are found
will be set to "na.omit" in order to omit the cases (Kuhn 2017).

The ANN models will be generated with "nnet", "RSNNS" and "pnn" libraries. Package
"nnet" provides a simple feed-forward neural network (FF) model, where information can
move only in one direction. The FF model will be tuned with the "train" function of the
"caret" package. The network has only one hidden layer and uses traditional back
propagation algorithm. The tuning parameters determine the size of models to be built
and the weight decay. For the "size" parameter, a vector with the numbers of nodes in the
hidden layer will be set to hold values 4, 5, 6, 7, 8. The "decay" parameter will be set to
use a vector holding values 0.1 and 0.05. The "na.action" will be set to "na.omit" and the
"maxit" parameter standing for the maximum number of iterations will be initially set to
400 (Venables 2016). The same setup will be used for tuning of the MLP models.

Other MLP models will be built using the "RSNNS" library thus they will need to be
tuned manually. Package "RSNNS" provides R implementation of the Stuttgart Neural
Network Simulator containing many standard neural networks models. Multi-layered
perceptron is one of them. The "mlp" function for building the MLP models has a long
set of available parameters. MLP models can be built for regression as well as
classification tasks. There is a wide choice of learning algorithms available for training
the MLP models with "RSNNS" (Bergmeir & Benítez 2014). For use with the MLP
models the inputs will be need to be scaled and normalized with function “normTrainingAndTestSet()” provided by the “RSNNS” package.

Another R package “pnn” provides a compact algorithm for building and training PNN models. Only one parameter needs to be set – the smoothing parameter called "sigma". The “pnn” library provides a way to tune PNN models by setting the “limits” parameter inside the “smooth()” function. In this way, a PNN model is trained with sigma values between the given limits. The best smoothing parameter is selected and the model smoothed with it can be tested with the “perf()” method. An advantage of the PNN algorithm over other ANN and DT algorithms is that it may use the full dataset for training (Mostafa 2009). The performance of the trained model is verified based on the same data. The model can be tested with new cases using the “guess()” function (Chasset 2016).

4.5. Ethics in data mining

Ethics is a broad area of study in social psychology, which concerns all aspects of human life. In data mining, ethics plays an especially important role in creation and disposal of the dataset. Ethical issues are subject to personal considerations, and this is why they are difficult to define and put in frames (Rahman & Ramos 2013). On the other hand, violating ethical codes and principles may be subject to legal actions. For this reason, organisations regulate their code of ethics to stay compliant with the legal regulations.

There are several fundamental rules that must be followed when data mining. Firstly, the individuals whose personal data is stored and used for analysis must confirm their consent for it. Secondly, the data may be used only in accordance with their knowledge. Thirdly, the organisation holding the data may not publish or give away personal data that allow for identification of the individuals. Fourthly, data mining cannot be used for racial, sexual, or religious discrimination. It must be taken into account, that not only the obvious features such as name, surname or address enable identification. Very often, features that individually are not harmful, when grouped with others, can identify the person. These variables cannot be published, so they have to be removed from the dataset and analysis. According to Witten, this requirement often negatively influences the results of the
analysis, as those removed variables are often the most significant once. In this way, the
dataset becomes useless (Witten et al. 2016).

In regards to the described aspects of ethics, the dataset received from the IBTS was
limited to dates of donations, deferrals, age, and gender. Additional sociodemographic
variables might increase the possibility of donor identification. For example, adding
donor location, and race or nationality might help to identify a donor who leaves in a
small residential area or town and is the only representative of a particular nationality.
Adding the distance to the blood collection point, or the profession may also increase the
probability of donor identification. In accordance with the discussed rules, the dataset had
to be limited to the current content. On this basis, new variables were generated, that were
subject to further analysis according to the methodology described in the following
sections.

4.6. Blood donors database

4.6.1. The original dataset

The dataset used for the experiments is a sample of the blood donor database provided by
IBTS. The original dataset is built of two separate sections that contain dates of donations
and deferrals assigned to donors’ identification numbers. The dataset includes also a few
additional attributes: donors’ age and gender. It contains information on 5000 donors
and 2015.

4.6.2. Data pre-processing and labelling

First, the two separate datasets were consolidated into one based on the donor ID columns.
Next, new variables were generated based on those existing ones. In result, the
information from the original datasets was formatted in a way it could be processed
efficiently by the classification models.
The new variables generation was subject to other researchers' findings described in the literature review chapter. They are believed to aid donor profiling, as they will act as donor characteristics. Any new information was added to the dataset in the result of the pre-processing what means the classifications and predictions that will be performed on the dataset will not be influenced by these changes.

Based on the dates of donations for each donor, number of donation was calculated and a list of intervals between the consequent donations was generated. Based on two deliverables the following variables were generated and added to the dataset:

- donation interval average – summarises lengths of all intervals between subsequent donations
- time since the first donation – number of days passed since the first donation
- time since the last donation – number of days passed since the first donation
- first two-donation interval length – number of days between the first two donations
- last two-donation interval length – number of days between the last two donations
- donating period – number of days between the first and last donations
- two-year donations number – number of donations within the last two years
- two-year donations interval average – summarises lengths of intervals between subsequent donations that happened within the last two years

Based on the dates denoting start and end of deferrals of each donor, number of deferrals was calculated and a list of deferrals' lengths was generated. Based on two deliverables the following variables were generated and added to the dataset:

- deferral length average – summarises lengths of all deferrals
- last deferral length – length of the last recorded deferral in days

Additionally, based on the birth dates of donors stored in the original datasets, the following new variables were generated:
• age group – donor’s current age
• first-donation age group – donor age at the time of their first donation

Donor age was denoted with five age groups:

1. < 25 years
2. 25 – 35 years
3. 35 – 45 years
4. 45 – 55 years
5. > 55 years

The following is the list of all the generated fields:

• gender: female (1) and male (2)
• age group
• donations number
• deferrals number
• time since the first donation
• time since the last donation
• deferral length average
• last deferral length
• first two donations interval length
• last two donations interval length
• donating period
• two-year donations number
• two-year donations interval average
• first donation age group

For classification tasks (Chapter 5), the following labels will be used to denote donor classes:

• L – lapsed donors, who have not donated blood for the last two years
• O – registered donors with 0 donations
• N – new donors with number of donations between 1 and 5
• R – regular donors with number of donations >= 5
• F – first-time donor with 1 donation

The predictive tasks (Chapter 6) will use two classes denoted as:

• Y – future donation
• N – lack of future donation

The strength and importance of the particular variables will be analysed and evaluated with PCA and LDA models as well as regression models. This was already explained in section 4.4.1. Variables considered important in blood donors classification and prediction will be added to the particular datasets. The experiments performed for answering questions 1 and 2 will use a slightly different set of input variables because they will try to qualify donors into different sets of classes.

All R algorithms used for classification of datasets will centre and scale the inputs within pre-processing. In order to centre the inputs, the mean value of each feature is removed. Next, the dataset will be scaled by dividing non-constant features by their standard deviation. Further data pre-processing of the datasets will depend on the type of classification in which the dataset will be used. In experiments for classifying regular donors, the whole dataset of 5000 records will be used (Chapter 5). However, experiments for predicting donation will require some records to be removed (Chapter 6). These will be rows that do not include donations recorded within the period used for predictions. Predictions of the three-month donations will be based on 4500, one-year donation predictions will use 4495 records, and finally the one-year predictions of active donors will use 669 records. The classification targets use in each set of experiments will be generated based on the data available in the dataset being analysed.

4.6.3. Training and testing datasets

For classification and predictive tasks, the datasets will be prepared with the function provided by “caret” library createDataPartition(). This function performs a stratified
random split of the data. The “p” parameter specifies the percentage of the dataset that will be used for training. The rest of the dataset will be used for testing. All experiments will use 0.75 of the dataset for training the supervised models, and cross-validation as the data resampling method. The training/testing functions will specify columns used as inputs and outputs.

For tasks using MLP models built with use of the “RSNNS” library the train and test datasets will be divided with use of “splitForTrainingAndTest()” function provided by the library. Also 75% of the dataset will be used for training, and 25% will be left for testing.

4.7. Research approach

4.7.1. Selecting predictive variables of donors database

Selection of the variables with the strongest predictive power will be performed with the application of PCA, LDA, and regression techniques. The strength and importance of the particular variables will be analysed and evaluated with PCA and LDA models. Regression is a popular tool for an analysis of the predictive strength of inputs and has previously been used in studies on blood donation (Karacan et al. 2013).

RF regression was recently performed in a study on authorship profiling by Palomino-Garibay et al. (Palomino-Garibay et al. 2015). Whereas, an example of LDA application into an analysis of donors’ features was described by Mostafa. (Mostafa, 2009) PCA application into feature selection for ANN models can be found in other research not focused on blood donation (Khan et al. 2001) (Glynn et al. 2006) (Kasraian & Tavassoli 2012).

PCA analysis will calculate the variance that each input variable introduces to the model. Those that explain the highest variance should be kept in the model for building classification and predictive models (Janert 2010).
Regression will be also used for the analysis of datasets. Two types of regression will be performed. The first one will examine correlations between pairs of variables. The second one will be a multiple regression that will explore dependencies between all inputs and the outputs. All the models will be tuned and evaluated with the root mean square error (RMSE).

All models involved in the same classification will need to use the same dataset in order to be compared efficiently. Furthermore, the techniques used for the analysis of the datasets (DLA, PCA, regression) will possibly produce different results and not always select the same features for the strongest predictors. In order to achieve the most efficient analysis of the features, the achievements of several techniques will need to be consolidated. Such approach may help to achieve one of the major aims of the research, which is to identify the features of donors and their donation habits that aid predictions of their further blood donation behaviour.

4.7.2. Finding regular donors in the dataset

In the attempt to answer the first research question, classification tasks will be divided into two parts. For each task, a different set of donor categories will be used. The classes were defined before they were used as the targets in the classifying models. The first set will include the following four categories:

- lapsed donors (L) (who did not donate in the last two years),
- first-time donors (F) (who are not lapsed and donated once),
- new donors (N) (who are not lapsed and donated 2 to 4 times),
- regular donors (R) (who are not lapsed and donated at least 5 times).

These categories were decided on the analysis of blood donors' history of donation by Schreiber (Schreiber et al. 2005). The most attention was paid to select the number for distinguishing between the new and regular donor. Based on the literature review in section 2.6.3 it was decided that five is the minimal number of donations that denotes the regular donor class. Regular donors' history of donations was discussed with the focus
on the study made by Schreiber who proved that the probability of becoming a regular donor significantly increases after the fifth donation.

According to the discussions with the IBTS staff, donors who did not donate within two years most probably will not donate again. At this point, they are qualified as lapsed donors (L). Based on this partition only donors with donation history within the last two years will be classified into three groups (F, N, and R). Donors that are not lapsed are considered active. Firstly, this classification will segregate lapsed donors from active donors. Secondly, it will identify the active donors with the highest chance to repeat their blood donation.

The second set of donor classes includes the following groups:

- first-time donors (F) (donated once in their whole life),
- new donors (N) (with 2 to 4 donations),
- donors (O) (with 0 donations)
- regular donors (R) (with more than 4 donations in their whole life)

This classification will allow for selecting groups of donors that were first-time, new or regular donors at some stage of their lives. Currently, they can be either active or non-active donors. The new “O” class defines donors who even though are present in the database did not complete any donation.

Two series of experiments will be performed to predict the predefined donor classes. In each set of experiments, different set of classes will be used with classification models. Additionally, the classification experiments will be divided into separate parts based on the type of data mining model used.

### 4.7.3. Predicting blood donations

To answer the second research question an almost identical dataset will be used for predicting donation behaviour. The difference is that the models will use a different set of target classes:
Here the primary goal of the experiments is in predicting donations within the next particular period. The prediction will be made for three months periods based on seasons, as they influence donations according to (Cimaroli et al. 2012). Experiments will then be attempted to predict donations within the whole year (December 2015, January 2016, February 2016). The final set of experiments will be executed in attempt to predict one-year donations of active donors only.

The dataset used for answering the first research question will be additionally pre-processed in order to be suitable for this classification. Information about donations within the period that will be subject to prediction will be removed. Information from the removed dataset will be used as the “teacher” for the supervised learning. In addition, donors that started donating after the analysed period will be removed from the dataset.

For predictions of the three-month donations, the models will analyse data gathered before December 1st 2014. The prediction targets will be generated based on donation record between December 1st 2014 and March 1st 2015.

The following predictive models will be built: LDA, QDA, CART, MLP, FF, and PNN. For predicting one-year donations, the targets will be generated based on donations between January 1st and December 31st 2014. The following predictive models will be built and compared: LDA, QDA, MLP, and CART. All models will be tuned as well.

The prediction of donations will be based mainly on the donation history, as other information about the donors could have not been provided by the IBTS in order to be used within this study due to the ethical issues described in section 4.5.
4.8. Summary

The methodology described above has been based on the literature review and previous models for classification of blood donors and similar classification problems. The algorithms and tools were examined and applied in preliminary experimentations in order to eliminate methods that appear ineffective in the context of the dataset. The presented techniques will be implemented in the experiments as they were shown to offer a high probability for success.

The presented methodology has some drawbacks due to ethical considerations. Namely, the dataset misses sociodemographic factors of donors, thus, all the classifications and predictions will be based on the donation history. In order to compensate additional variables were generated based on the original database and they were added to the dataset.

The way the original dataset provided by the IBTS was formatted was not suitable for data mining. Additional new fields were created with the view to retrieving more information about donors' donation habits. For example, to analyse the correlation between donor class and the length of intervals between donations or their last two-year donation habits.
CHAPTER 5

Classification Experiments

5.1. Introduction

This chapter describes the building of data mining models for the classification of blood donors, using the dataset obtained from the IBTS. The classification models were built using the variables from the dataset as inputs and the defined donor classes as outputs. The output variables were generated based on the number of donations recorded for each donor in the dataset. Separate models were generated for two different sets of classes. The two sets of experiments are presented in sections 5.3.3 and 5.3.4.

These experiments were performed with the aim of identifying data mining models, which are best suited for selecting regular donors. Donors with five and more donations are considered regular donors. The generated results were analysed to identify the input variables, which are the strongest classifiers within the dataset, especially in regards to regular donor class. Donor classification problems were tackled with several machine
learning algorithms. Efficiency of each model was estimated based on its accuracy p-
value and kappa. Each section provides a summary of the classification results of the most
efficient models verified with the testing dataset. The efficiency of the built models was
examined and compared to select the most efficient algorithms and data mining models
for classifying donors.

The primary goal of running the classification tasks was to provide the answer to the first
and third research question that together ask if the data contains patterns that can be used
to identify regular donors with application of machine learning techniques. The following
experiments attempted to find the best approach to categorising donors into four groups
and to prove that donors can be profiled based on the database available to the IBTS
(section 4.2). The donor classes were generated based on the number of previous
donations. Two different sets of donor classes were created. For each set of classes
separate experiments were performed. Section 5.3 includes experiments on the targets set
1 and section 5.4 describes experiments on the targets set 2. The sets of classes are further
presented in the following subsections.

Before running the classification experiment, the dataset was analysed to investigate the
correlations among the variables and their importance in terms of affecting the variance
of the dataset. Detailed pre-analysis results and outputs produced with the experiments
are included in Appendix A and described in the next section.

5.2. Pre-analysis of dataset for donor class prediction

The pre-analysis of the generated datasets was performed with use of PCA and LDA
methods as discusses in section 4.4. It aims to analyse the feature importance of the
dataset before the application to classification models. The information about each input
feature was expected to help in building the models as well as to generate a high-level
description of donor groups.
5.2.1. Dataset description

A donor class will be predicted based on the dataset consisting of the following 13 columns (Table 1) and 5000 records. The dataset includes one binary input (X11) and categorical variables that relate to donor's age (X2, X12). In both cases, age may be assigned one of the five categories (1 to 5). Within data pre-processing, three averages were added to the dataset as described in section 4.6.2 (X5, X6, and X13). In addition, the dataset includes deferrals count and six variables that refer to duration in time (time since the first donation, time since the last donation, last deferral length, first two donations interval, last two donations interval, and donating period). These features, based on the literature review, are considered important for discriminating between donor classes. The following table describes the input dataset in more detail.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1 deferrals number</td>
<td>Numerical</td>
</tr>
<tr>
<td>X2 first donation age group:</td>
<td>category: 1 (&lt;25), 2 (25-35), 3 (35-45), 4 (45-55), 5 (&gt;55)</td>
</tr>
<tr>
<td>X3 time since first donation</td>
<td>numerical (days)</td>
</tr>
<tr>
<td>X4 time since last donation</td>
<td>numerical (days)</td>
</tr>
<tr>
<td>X5 donation interval average</td>
<td>numerical (days)</td>
</tr>
<tr>
<td>X6 deferral length average</td>
<td>numerical (days)</td>
</tr>
<tr>
<td>X7 last deferral length</td>
<td>numerical (days)</td>
</tr>
<tr>
<td>X8 first two donations interval</td>
<td>numerical (days)</td>
</tr>
<tr>
<td>X9 last two donations interval</td>
<td>numerical (days)</td>
</tr>
<tr>
<td>X10 donating period</td>
<td>numerical (days)</td>
</tr>
<tr>
<td>X11 gender</td>
<td>binary: 1 (female), 2 (male)</td>
</tr>
<tr>
<td>X12 current age group:</td>
<td>category: 1 (&lt;25), 2 (25-35), 3 (35-45), 4 (45-55), 5 (&gt;55)</td>
</tr>
<tr>
<td>X13 two-year donation interval average</td>
<td>numerical (days)</td>
</tr>
</tbody>
</table>

Table 1 List of inputs used for classification of donors
As discussed in Chapter 4, donors were classified into two sets of donor classes (set 1 and set 2). Section 5.3 describes their classification into first-time (F), lapsed (L,) new (N), and regular (R) donor classes (target set 1). Section 5.4 describes donor classification into first-time (F), new (N) regular (R) and non-donors (O) (target set 2). All the groups were presented in Chapter 4. Before building donor classification models, the dataset was subject to pre-analysis with PCA and LDA, the discriminating algorithms for examining the importance of the features used as models’ inputs.

5.2.2. Pre-analysis with PCA

A PCA model generated with the above inputs produced 13 principal components (Figure 34 - Appendix A). PC1 and PC2 together explained 0.42 of the proportion of variance. Nine principal components together explained over 95% of the variance within the dataset. Whereas, eleven principal components explained 100% of the variance. According to the PC1 column, the most important features within the set of inputs are time since the first donation and donation interval average, then, the interval between first two donations, the interval between last two donations, and total donating period. According to PC2, the average deferral length is the most significant variable in the dataset.

The PCA generated correlations between the input variables and showed how different variables are collated with each class of donors. Figure 17 shows the produced vectors and their names.

Separate group distributions were drawn for classes of set 1 and set 2, see Figure 18 and Figure 19.
Figure 17 shows some correlations between different variables and the classes, whereas the illustrations below show relevant class distributions (Figure 18, Figure 19).

Based on the produced outputs, the correlations generated for classes of set 1 show that the group of lapsed donors (L) is positively correlated with the time passed since the last donation. The group of regular donors (R) is rather positively correlated with two-year donation average. Their deferral numbers are quite significant too (Figure 18).
The PCA output generated for classes of set 2 show that there are no clear differences between new (N) and regular donor groups (R). They are strongly overlapping (Figure 19). According to Figure 17 and Figure 19, new donors tend to be positively correlated with long donating periods and long intervals between donations. However, regular donors are most correlated with the longest donating period. New donors also tend to be positively correlated with the long periods since the last donation. New donors, as well as regular donors, tend to be positively correlated with the high number of deferrals that tend to be short for them. Gender and age only slightly matter for distinguishing regular donors from other groups. Long intervals between first two donations and long average interval are most correlated with new donors.
5.2.3. Pre-analysis with LDA

Pre-analysis with LDA was performed with the aim to analyse feature importance for discriminating between the target classes as planned in section 4.4.1. The pre-analysis produced a slightly different picture of correlations and feature importance than the one generated by the PCA. The first LDA model provided the descriptions of four classes of objects from the target set 1 by calculating the probabilities of the groups. The output showed quite high unbalance between the group probabilities. The probability of the lapsed donor class was 0.77, and the probability of new donor class was 0.3. The probability of the regular donor class was only 0.1 and probability of the first-time donor class was the lowest, only 0.3. This unbalance in the dataset may negatively influence future classifications.

According to the means calculated during the pre-analysis, the following four variables have the highest feature importance for predicting donor classes of set 1: donation interval average, deferral length average, last deferral period, and two-year donation average. Regular donor class has the highest means of the donating period and time since the first donation. The mean of their two-year donation interval average is 90 days, whereas for new donors it is 41 days. Regular donors have a slightly higher mean of their first donation age than first-time and new donors (40). Furthermore, the mean of the time since the first donation is the highest for regular donors. Whereas, the mean of the average donation interval is the lowest for new donors, who also have the highest mean of the interval between first two donations. Finally, the means of the last deferral length and average deferral lengths are much higher for lapsed donors than for other groups. Gender and current age do not seem to play important roles in predicting any of these four classes of donors as their means are very similar for each class (Figure 35 - Appendix A).

The next pre-analysis was generated for the dataset with the second set of outputs (set 2). This time the group probabilities and group means were calculated for the same set of inputs and classes of set 2. The pre-analysis showed that probabilities of the classes in the current set are balanced quite well in contrast to classes of set 1, where the probability of the lapsed donor class was much higher when compared to the other three groups. This means that the classification into the second set of classes may be easier for the algorithms. Thus, it will generate more accurate class predictions.
Based on the coefficients of linear discriminants, the number of deferrals is the strongest class discriminant in the dataset. According to the calculated means, the group of donors with no donations have the lowest mean of the number of deferrals but the highest mean of the average deferral length. They have the highest mean of their current age and the lowest mean of their first donation age. On the other hand, the regular donors have the highest mean of their number of deferrals and a higher mean of the last deferral length (104) than new (79 days) and first-time (69 days) donors, but not the non-donors. The non-donors have the highest mean of the last deferral length (542 days) - over five times higher than the regular donors. According to the calculated means, regular donors have slightly shorter means intervals between the first two donations and last two donations when compared to new donors (Figure 36 - Appendix A).
5.2.4. Summary

Based on the above pre-analysis, the length of intervals between donations has the strongest class discriminating power in the dataset for both sets of outputs. Similarly, the total donating period. The characteristics related to deferrals are also very important. Nevertheless, the number of deferrals needs to be accompanied by the average deferral length. Only when applied together they can help to discriminate between the donor groups and allow the classification models to identify the regular donors. Both LDA and PCA did not find important correlations between the donor group and age or gender. Nevertheless, according to PCA at least nine variables should be used for classification models as together they explained 95% of the variance in the dataset. Whereas, using the first 11 out of the 13 variables should allow for explaining 100% of the variance.

5.3. Donor classification for target set 1

The following sections describe experiments performed for classification of blood donors based on the dataset provided by IBTS and pre-analysed with LDA and PCA. The process included generation of several types of classification models based on decision trees and artificial neural networks.

The first set of experiments attempted to classify donors into four groups, previously described in section 4.6.1: lapsed (L), first-time (F), new (N), and regular (R) donors. This classification was designed for distinguishing between the active (R, N, F classes) and non-active donors (L class), and further categorising all the active donors into three groups: F, N, and R.

Initially, the simple DT algorithms were applied: CART and C5.0. A boosted tree algorithm, RF, was used to build the next classification models, as according to the literature review, boosted trees are expected to achieve better classification results than the simple trees. Next, the same data was classified with ANNs: simple, feed-forward perceptron networks (FF) and MLPs.
Below subsections describe result of tests performed on the models.

5.3.1. CART

A CART model was applied first. It was tuned with use of the “caret” library. For training and tuning of the CART model, the “method” parameter inside the train function was set to “rpart”. The rest of parameters was set as described in Chapter 4 for use with the decision trees.

The outcomes of the tuned model testing showed that 100% of first-time donors and lapsed donors were classified correctly (Table 10 - Appendix A). Only a few of the regular donors were misclassified as new donors. It misclassified 7 out of 48 new donors as regular donors and 2 out of 134 regular donors as new donors.

The winning CART tree (Figure 37 - Appendix A) based its conditioning mainly on six variables. The first condition analysed the average two-year donations interval. Subsequent conditions were based on the following variables: donating period, time since the first donation, age group, time since the last donation, and all donations interval average.

Most of the regular active donors identified by the CART model donated blood multiple times within the last 742 days (almost two years). (see Figure 20). The average donation interval was considered for differentiating between new, first and regular donors. According to the CART tree, the regular donors are those with a donating period of over 393 days (over a year) and an average donation interval of less than 938 days. Those donating within the last two years, but with donating period of less than 392 days were classified as new donors.
Summing up, the tuned CART algorithm performed very well in the classification of lapsed, first-time, new and regular donors. It had only a few misclassifications between new and regular donors. Nevertheless, the overall accuracy achieved with the algorithm was 0.98, which is very high.

5.3.2. C5.0

According to the classification results achieved with the C5.0 algorithm (Table 11 - Appendix A), the model wrongly classified a few new and lapsed donors as regular donors. Only 3 of the 136 regular donors were incorrectly classified as new donors, and 2 of 48 new donors were misclassified as regular donors. Only the first-time donors were
classified 100% correctly. The conditioning tree built by the tuned model used five variables for the decision splits (Figure 38 - Appendix A). The initial condition was based on the last two-year donations. Then other factors were considered: time since the last donation, time since the first donation and donation interval average.

Most of the donors classified as regular donors have donating period over 392 days. This time the average donation interval for regular donors is below 531 days. A small number of regular donors were found among those with a donating period of over 3518 days and an average donation interval of over 531 days. Both of these groups of regular donors donated within the last two years. (Figure 21)

Figure 21 C5.0 Classification results with targets set 1
Some regular donors were also identified within those with one donation in the last two years (average two-year interval equal to 0). Those donors have donating period between 984 and 1985 days, thus, are rather long-term donors. Their average donation interval is less than 452 days (Figure 22).

Overall, the classification with C5.0 was very successful. The tuned model’s accuracy was equal to 0.997.

5.3.3. RF

Next, classification into F, L, N, and R groups was performed with the RF algorithm. For training the method inside the train function was set to “fr”, all other parameters were set as described in Chapter 4.

The tuned RF model, similarly to the previous classifications, misclassified a few regular and new donors (Table 12-Appendix A). There were 3 out of 48 new donors misclassified.
as regular donors, and 5 regular donors out of 136 were misclassified as lapsed donors. The overall accuracy of the model was (0.99).

According to the classification model, the highest feature importance in regards to detection of regular donors was calculated for donating period, time since the last donation, time since the first donation and two-year donations intervals average.

Summing up, the boosted trees did not cause any major improvement in the classification accuracy this time, what would be a difficult task considering that the accuracy of the tuned CART and C5.0 models was close to 100%. RF classification accuracy was minimally lower than C5.0 that performed only slightly better than CART. Nevertheless, the ANN models were tested next to examine their classification abilities in the case of the blood donor dataset.

5.3.4. FF

Next, as the first ANN model, a feed-forward network was used for the classification. Networks of size from three to eight were trained and tested during the tuning process. The rest of the parameters was set according to section 4.3 where the settings for ANN models is described. The train function found the best model, which was then used for prediction.

The tuning process took 400 iterations. The best FF model selected within the tuning process had the decay of 0.1 and six neurones in the hidden layer. Predictions achieved with the model were highly accurate. The highest ratio of misclassifications was recorded for the regular donor class. Only 4 out of 136 regular donors were misclassified as lapsed or new donors, 2 out of 48 new donors were misclassified as regular or lapsed donors. Also, 6 out of 970 lapsed donors were misclassified as regular donors, and 1 out of 96 first-time donors was classified as a lapsed donor (Table 13 - Appendix A).

According to the results of the model testing, the tuned FF model was a very accurate classifier. It also performed a classification that was almost 100% accurate (0.99). Next,
the MLP models were applied to check if they could perform equally well in the classification of blood donors.

5.3.5. MLP

For the MLP model tuning, models of size four to eight were trained. The rest of the parameters was set according to Chapter 4.

The first model used traditional the backpropagation function for learning. The tuning process was much longer than in the case of the FF model. As many as 8000 iterations were required to select the best model. The best results were achieved by the model built with six nodes in the hidden layer. Only a minimal number of donor misclassifications were recorded for this model. Only 1 of 968 lapsed donors and 1 of 47 new donors and 6 out of 136 regular donors were misclassified (Table 14). The overall accuracy of the tuned model achieved within 400 iterations 0.96.

A multi-layered perceptron models with backpropagation learning algorithm was also trained without tuning but using the package “RSNNS” that allows for using multiple hidden layers. The train and test datasets were divided with the use of “splitForTrainingAnTest()” so 75% of the dataset was used for training, and 25% was left for testing. The inputs were scaled and normalised.

A selection of learning algorithm available for the “mlp” function was used in the classification trials. However, only the performance results of two of them were close to the performance of the model tuned with the “caret” package. These algorithms are the standard back-propagation algorithm and the backpropagation with momentum algorithm. For both the algorithms, the best models was built of two layers, with three nodes. The models’ top performance was the same. They were 0.97 accurate. They used the same learning rate of 0.1 and the number of iterations of 8000, which is very high. Unfortunately, the train and test dataset were selected in a different way than in previous experiments. The RSNNS algorithm uses its own functions to select them. That is why the “set.seed()” function could not select the same records for testing and training.
Both the algorithms achieved the same classification results. The outputs of the most accurate models selected manually revealed few misclassifications of first-time donors as one new donor (1 out of 377) and lapsed donors (9 out 377) and misclassification few lapsed donors as regular donors (4 out of 559). Also, some new donors were misclassified as first-time (4 out of 161) and lapsed (13 out of 161) donors. Several regular donors were classified as first-time donors (6 out of 151) and as lapsed donors (2 out of 151) (Table 15, Table 16 - Appendix A).

Nevertheless, the accuracy of all the MLP models was very high and almost identical, close to 100%. Running the tuned MLP classification was easier than manually building and comparing the RSNNS models. The PNN models described in the next section also required manual tuning, as the “caret” library does not support tuning of the PNN models. However, according to the literature review, they are expected to be much faster the MLPs.

5.3.6. PNN

The function used for training the PNN models requires only one parameter. It is the smoothing parameter called sigma. The model used the same train and test datasets as the previous model. The inputs were normalised with function “normalizeData()” provided by “RSNNS” library. The model was tuned with the smoothing parameter between 2.0 and 0.1.

The selected PNN model used sigma equal to 0.03. The model was tested with use of the “guessO” function and the test dataset. (Chasset 2016) According to the results of the testing, the model had a high accuracy of 0.92. The model misclassified 15 out of 968 lapsed donors as regular donors and 9 out of 48 new donors. The model misclassified 21 out of 136 regular donors as lapsed donors and 5 as new donors (Table 17 - Appendix A).

Summing up, the PNN algorithm performed very well in this classification. The experiment proved that PNN is a suitable algorithm for this particular classification and can be successfully in the classification of donors based on the data provided by the IBTS.
5.3.7. Efficiency based on inputs

Based on all the above descriptions of the resulted models, it can be concluded that MLP was among those achieving the most accurate classification results in testing. That is why MLP was chosen to test the accuracy of models in regards to the number of inputs used for training. The following plot shows how the classification results rely on the different inputs. The full set of inputs is included in Table 1 and described in section 5.2.1.

![Classification - Type 1](image)

*Figure 23 Classification results for models with increasing number of inputs and with targets set 1*

According to the plots presented in Figure 23, the first five variables (gender, age, donations number, deferrals number, and time since the first donation) were enough to achieve the best classification accuracy, where both accuracy and kappa were close to 1. The initial accuracy based on gender was quite high (0.78). Nevertheless, the model’s balanced accuracy and kappa slightly dropped when age was added to the set of inputs. Surprisingly, classification based only on donor gender resulted in a very high accuracy. Nevertheless, the minimal value of kappa proves the classification reliability was equal to the random choice of targets. Including the number of donations and deferrals into the inputs significantly improved all three measured variables. The number of donations
increased the overall accuracy and the balanced accuracy to 0.86 and kappa to 0.58. Adding the number of deferrals increased the accuracy to 0.97 and kappa to 0.94. The time since the last donation improved the accuracy to 0.99 and kappa to 0.97.

Concluding, to generate an efficient classification model for categorising blood donors according to the class set 1, only five variables from the dataset are needed.

5.3.8. Summary

The above set of classification tasks (section 5.3) attempted to classify donors into targets of set 1, thus, to categorise active donors into three groups of active donors (regular, first-time or new donors) and all non-active donors into the group of lapsed donors. The classification results showed that the DT models are highly efficient in categorising blood donors into targets of set 1 based on the variables provided by the IBTS. The differences between the models' accuracy are only minimal. Based on the testing results all the models achieved very high kappa and the smallest possible p-value. This confirms that the DT models were highly reliable and the dataset contains enough information to be used in such classification tasks (PsychologyWikia 2017) (Senn et al. 2016). There was no benefit from using the boosted trees over the simple trees as they all generated nearly 100% accurate classification results with kappa close to its maximum value (1) and very low p-value. (Table 2), what confirms that the results are highly reliable.

According to the trees displayed by the models, classes from set 1 were predicted based on the frequency and the recency of donations as well as the time between the first and last donations (average interval length, time since the last donation, donating period). Variables such as age and gender were not used for decision splits, what confirms they are not valuable classifiers.

Based on the overall results, the simple trees are recommended to the IBTS for isolating the long-time donors from the IBTS database as all of the DT algorithms provided accurate classifications, but CART and C5.0 additionally display the decision trees. The simple trees illustrate the conditions and decisions taken, show the identified patterns of
different donor groups, which can be used to generate profiles of the groups. Nevertheless, RF can be used to calculate feature importance for the additional analysis of the data and the groups.

<table>
<thead>
<tr>
<th></th>
<th>CART</th>
<th>C5.0</th>
<th>RF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.98</td>
<td>0.99</td>
<td>0.992</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.97</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>R class Sensitivity</td>
<td>0.96</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>R class Specificity</td>
<td>0.98</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>R class Balanced Accuracy</td>
<td>0.97</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>P-Value</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
</tr>
</tbody>
</table>

Table 2 Donor Classification results of DT algorithms for target set 1

The same sets of inputs and targets were applied to ANN models. The table below (Table 3) presents results obtained from testing of the following models: tuned FF networks, tuned MLP with standard backpropagation algorithm, "RSNNS" MLP with standard backpropagation algorithm and PNN. According to the results of models' testing, the tuning algorithm implemented with the R library ("caret") indeed helped to select the most accurate ANN model. All the selected FF, MLP, and PNN models generated almost 100% accurate classifications. Nevertheless, the FF and MLP models are more reliable according to the calculated kappa values equal to 1.
<table>
<thead>
<tr>
<th></th>
<th>Tuned FF (nnet)</th>
<th>Tuned MLP standard backprop</th>
<th>MLP (RSNNS) standard backprop</th>
<th>PNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.994</td>
<td>0.992</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.98</td>
<td>0.98</td>
<td>0.95</td>
<td>0.89</td>
</tr>
<tr>
<td>R class</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.97</td>
<td>0.96</td>
<td>1.00</td>
<td>0.80</td>
</tr>
<tr>
<td>R class</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specificity</td>
<td>0.999</td>
<td>0.999</td>
<td>0.998</td>
<td>0.94</td>
</tr>
<tr>
<td>R class</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Balanced</td>
<td>0.98</td>
<td>0.98</td>
<td>0.999</td>
<td>0.89</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3 Donor Classification results of ANN models for targets set 1

Summing up, all the models obtained through tuning with "caret" library are recommended for future donor classifications performed by the IBTS as the application of each of the models is fast, and the models are likely to produce very accurate classifications of donors based on their donation history. Nevertheless, the training time of CART models was the shortest. The tuning algorithm used 20 different trees to select the best one. Tuned FF and MLP model' training times were much longer than the time needed for training the CART and PNN models. In case of the FF and MLP networks the time strongly relied on the number of network sizes and the selection of weight decay values specified in the "train()" function.

The further analysis of the dataset used another set of outputs. Results of the analysis are described in the next section.
5.4. Donor classification for target set 2

This section presents experiments that applied DT algorithms and ANN for classifying donors into four groups of set 2 (F, N, O, R). The goal of the following set of experiments was to identify the individuals that were regular (R), new (N) or first-time (F) donors at some stage of their lives. Those could be current or lapsed donors. All other individuals should be classified as non-donor (O) as they have no successful donation. All the groups were described in section 4.6.2.

The experiments described below were executed with the view to identify the non-active regular donors, in contrast to the experiments presented in the previous chapter where only active regular donors were identified. Within these experiments, the dataset was searched for the existing patterns that refer to all the regular donors included in the dataset.

The experiments started from tuning simple DT models (CART, C5.0). Then, boosted trees (RF) were used as they are expected to produce more accurate classifications. Next, ANN models were applied. Simple feed-forward networks were tuned first. Within the next experiments, different types of multi-layered perceptron models were generated. The first MLP model was tuned with “caret” function. The other MPL models were built with the use of the “RSNNS” package. The “RSNNS” MLPs were generated using two different learning algorithms: the standard backpropagation algorithm and the back propagation algorithm with momentum. Below, the most significant test results achieved during the experiments are described.

5.4.1. CART

For training and tuning of the CART model the “method” parameter inside the train function was set to “rpart”. The rest of parameters was set as described in Chapter 4 as for the use of decision trees.
Twenty trees were used for tuning of the model. The total tuning time was very short. The best CART model was found within a minute. The tuned CART model misclassified 16 out of 364 new donors and 12 out of 286 regular donors as lapsed donors (Table 18 - Appendix A).

The CART model generated a conditioning tree (Figure 39 - Appendix A) where most of the regular donors were found in two scenarios. Both scenarios assume their donating period is over 1400 days. In the first scenario, it is assumed that donation interval average of regular donors is below 528 days. The second scenario assumes their donation interval average is over 528 days and donating period is over 2230 days (Figure 24).

Summing up, the algorithm performed very well in the above classification of blood donors achieving the top accuracy of 0.99. Based on the displayed tree, some general patterns related to regular donors were found by the algorithm. Another simple DT algorithm C5.0 was applied next for comparison, with the aim to identify patterns that would be more precise.
5.4.2. C5.0

For tuning of the C5.0 model the "train()" function was set to use the "rpart" method. The rest of the parameters was set as described in Chapter 4. Additionally, C5.0 related parameters were set as follows: trials = 20 (an integer specifying the number of boosting iterations), model = tree (the type of model output), and winnow = true (use of predictor winnowing). The tuning time was longer than in the case of the CART models. It took
several minutes. The tuned model misclassified 7 out of 286 regular donors and 3 out of 374 new donors.

The tuned C5.0 misclassified a few new donors with regular donors, and vice versa (Table 19- Appendix A). The tree generated by the C5.0 algorithm had many more levels when compared to the trees generated with CART (Figure 40 - Appendix A). However, almost the same variables were taken into account for creating the conditions as in the previous CART tree. The C5.0 tree in the first condition isolated donors with more than one donation. Next, it examined the time passed since the last donation. If it was less than 4194 days (11 years), the donors were further analysed. The total donating period and average donation interval were the variables used for most of the conditions. The majority of the regular donors were identified in a scenario where the time since the last donation was less than 4590, donation interval average was below 404 days, and donating period was over 1126 days (Figure 25).

The overall accuracy of the model was nearly 100% (0.99). The tree was more detailed than the CART tree, what makes the discovered patterns more precise. The algorithm proved to be a good choice for the classification. More information on the dataset was expected to be provided with the RF algorithm applied next.
Figure 25  C5.0 classification results with targets set 2
5.4.3. RF

For tuning the RF model, the “method” parameter inside the “train()” function was set to “rf”. Then the rest of the parameter was set according to Chapter 4.

The tuning time of the model was longer than in the case of the C5.0 model. Classifications of donors with the RF algorithm showed a few misclassifications of regular and new donors (Table 20 - Appendix A). 5 out of 368 new donors were misclassified as regular donors, and 5 out of 286 regular donors were misclassified as new donors.

If it comes to the feature importance estimated by the RF algorithm, donating period is the most significant variable for discriminating between the classes. The second most significant variable is donation interval average. Next in order are time since the last donation, time since the first donation, and the interval between the first two donations.

The RF tuned model performed very well in the classification of blood donors. The overall top classification accuracy of the model is 0.99. The boosted DT algorithm is unable to display the decision taken during classification. However, it provided some useful information about the strongest class predictors within the dataset.

5.4.4. FF

In the further attempt to select the best classifier for the dataset, a simple feed-forward model provided by “nnet” package was tuned inside the “train()” function of the “caret” library. The FF model was built with the parameters presented in Chapter 5.

The model selected as the most accurate one was built of four nodes in the hidden layer and 0.1 decay. This resulting model turned out to be also highly efficient in classifying donors. It misclassified only five objects. 1 of 370 new donors was misclassified as a regular donor, and 3 were classified as first-time donors. Only 1 out of 286 regular donors was classified as a new donor (Table 21 - Appendix A).
So far, the model achieved the best results. In the next experiment, a tuned MLP was built for comparison.

5.4.5. MLP

The first MLP model was tuned within the “train” function provided by the “cart” library. The “method” parameter was set to “mlp”. The rest of the parameters was set according to Chapter 4. The model had longest training time so far. The resulted MLP model had one hidden layer with four neurones in the hidden layer. Few misclassifications were recorded, 2 out of 342 lapsed donors and 2 out of 286 regular donors (Table 22 - Appendix A).

Next, based on the results of model tuning a similar MLP model was built with the algorithm provided by RSNNS package. Two algorithms achieved the best and the same results standard backpropagation and backpropagation with momentum algorithms. Both the models used 300 iterations, had one hidden layer of three nodes and the learning rate of 0.1. The rest of the parameters was set in the same way as described in section 5.3.4. The models misclassified 4 out 233 of regular donors and 2 out 339 of lapsed donors (Table 23 - Appendix A).

Summing up, all MLP models achieved an accuracy of nearly 100%. The tuning mechanism provided by the “caret” library made the model selection very efficient. On the other hand, training with the “mlp()” function provided allows for more precise configuration. (i.e. learning function).

5.4.6. PNN

The inputs for use with the PNN models were generated in the same way as described in section 5.3.6. Unfortunately, for this experiment the same test and train datasets as the ones used with the previous models could not have been reproduced. Although the same seed was set on the data, the generated test and train datasets had different selections of records.
The PNN model was trained and smoothed with sigma values between 2 and 0.1.

According to the results of the testing, the model had a high accuracy of 0.89. The model misclassified 23 out of 308 first-time donors, 49 out of 335 new donors as first-time donors and 34 as regular donors. Out of 300 regular donors 34 were misclassified as new donors. All non-donors were classified correctly (Table 24 - Appendix A).

Summing up, the PNNs model performance very high. However, it was lowered than in case of the models generated in the previous experiments.

5.4.7. Model efficiency vs. number of inputs

Next, similar to section 5.3.6, the MLP networks that provided very efficient classification were used to measure and compare the accuracy, kappa and R balanced accuracy of the classification models with a different number of inputs. The following plots show how the subsequent inputs influence the classification results (Figure 26 Classification results for models with increasing number of inputs and with targets set 2). The full set of the inputs is included in Table 1. These classifications were performed with the MLP models, as they proved suitable for solving the above donor classification problems.
The above plots were generated based on MLP classifications into the targets of set 2 (Figure 26). They show how the classification accuracy, kappa and R class accuracy were improving while the subsequent input variables were being added to the training model. The first trained model had only one variable from the dataset. The last one had all 13 variables.

The consolidated outputs, presented in Figure 26, clearly indicate that the first five parameters are enough to achieve the best classification accuracy for the targets of set 2. Each of them continuously improves the accuracy (number of deferrals - 0.42, time since the first donation - 0.57, time since the last donation - 0.7, and average donations interval - 0.92). Then, the fifth input (average deferrals length) minimally improved the accuracy to 0.99 and increased the kappa value from 0.95 to 0.99. Classification models with the first two inputs (deferrals number and first donation age) had a very low value of kappa (0.43) what means these models were unreliable. Inputs 2 (time since the first donation), 3 (time since the last donation) and 4 (average donation interval) increased kappa and the model accuracy most significantly.
This experiment proved that five of the used variables provide an efficient classification of blood donors to outputs of set 2.

5.4.8. Summary

The purpose of donor classification into the target set 2 was to identify the individuals that were regular, new, or first-time donors at some stage of their lives. Those could be current or lapsed donors.

Summing up the above classifications into targets of set 2 turned out to be equally easy for all the described models except the PNNs. All the DT models performed almost identically well, no matter if these were simple or boosted trees (Table 4). Based on the calculated kappa, accuracy, and p-value all these models were highly reliable. The results achieved with the simple trees were so good that it was needless to improve them. That is why the application of the RF models did not bring any benefits over the simple trees except the ability to calculate the feature importance.

Based on the displayed trees, similar sets of variables were used for the major decisions of all the DT algorithms. According to the DTs generated in this experiment, the total donating period and the average interval between donations were the most important predictors of all the classes. The next significant variables were the time passed since the first and the last donations. The importance of these variables was confirmed by the RF feature importance calculation.
### Table 4 Donor Classification results of DT algorithms for target set 2

<table>
<thead>
<tr>
<th></th>
<th>CART</th>
<th>C5.0</th>
<th>RF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>R class Sensitivity</td>
<td>0.97</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>R class Specificity</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>R class Balanced Accuracy</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>P-Value</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
</tr>
</tbody>
</table>

According to Table 5, all ANN models achieved very similar and equally well classification results. Again, the FF and MLP models provided very accurate predictions, nearly 100% accurate. Based on the very high kappa calculated for the FF and MLP models, their classifications were highly reliable. The PNN model's performance was slightly worse. Both the accuracy of 0.88 and kappa of 0.85 was lower than in case of the other models. Nevertheless, the kappa was still above 0.7 (0.85) what means the model predictions are reliable. Even though the predictions of first-time donors and new donors were less accurate, the balanced accuracy of the regular donor class was 0.93 that is still high.
<table>
<thead>
<tr>
<th></th>
<th>Tuned Feed-Forward (nnet)</th>
<th>Tuned MLP standard backprop</th>
<th>MLP (RSNNS) standard backprop</th>
<th>PNN</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>0.996</td>
<td>0.996</td>
<td>0.995</td>
<td>0.88</td>
</tr>
<tr>
<td><strong>Kappa</strong></td>
<td>0.995</td>
<td>0.99</td>
<td>0.992</td>
<td>0.85</td>
</tr>
<tr>
<td><strong>R class Sensitivity</strong></td>
<td>0.996</td>
<td>0.993</td>
<td>0.98</td>
<td>0.88</td>
</tr>
<tr>
<td><strong>R class Specificity</strong></td>
<td>0.999</td>
<td>1</td>
<td>1.00</td>
<td>0.96</td>
</tr>
<tr>
<td><strong>R class Balanced Accuracy</strong></td>
<td>0.998</td>
<td>0.996</td>
<td>0.99</td>
<td>0.92</td>
</tr>
</tbody>
</table>

*Table 5 Donor Classification results of ANN models for targets set 2*

**5.5. Evaluation of results – the summary**

The experiments described in this chapter aimed to classify donors into predefined classes of donors with the focus on the regular donor class. The dataset used with the data mining models was described in the methodology in Chapter 4. The dataset was pre-analysed with dimensionality reduction methods (LDA and PCA). The pre-analysis identified the features with the power to discriminate between the groups of donors present in the datasets and aid the profiling of the donor groups. Further, in Chapter 5, the performed experiments were presented that aimed to isolate patterns existing within the dataset that could be applied to differentiate between groups of donors and help to identify regular donors. The experiments were divided into two sets, each using a different set of inputs. Different classification methods were applied to classify blood donors. It was expected that based on the experiments’ result the most efficient classifiers would be selected. The following section includes more in-depth analysis of the models’ achievements.
5.5.1. Comparison of classification methods

Sections 5.3 and 5.4 described experiments that aimed to show how the blood donor database might be used for donor profiling through classification tasks. ANN and DT models were trained and tested for classification of donors into two, slightly different set of classes. The classification tasks were expected to produce outputs revealing if machine learning algorithms can select patterns of data related to regular donors from within the database owned by the IBTS and used for donor classification, thus to answer the research question 1 and 3.

In result, all applied DT algorithms performed extremely well in both types of classifications. Overall, the accuracy and reliability of the models were close to 100%. The main difference between the C5.0 and CART algorithms is the construction of the conditioning trees. CART tends to build simpler and smaller trees than C5.0. Generated CART trees had fewer levels, although they used the same variables for conditioning as the C5.0 trees. The more detailed C5.0 trees may prove to be useful in finding the profiles of donors. Even though the C5.0 models were expected to achieve worse performance with the unbalanced dataset used with the classes of set 1, all the applied models performed equally well. (Abbott, 2014) RF does not display any decision tree in contrast to CART and C5.0. Nevertheless, the boosted trees are expected to improve the classification accuracy obtained with simple DTs. However, in this set of experiments they appeared needless as the CART and C5.0 models turned out highly efficient classifiers.

If it comes to the ANNs, all the models, FF, PNN and MLP, were very accurate in the classification of donors. They managed to predict the classes from both sets of targets nearly 100% accurately.

Most of the MLP models used one hidden layer, even though MLP networks allow multiple hidden layers. Only for classification of inputs into classes of set 1 the "RSNNS" MLP needed two hidden networks and 8000 cycles to converge, while for classification of donors into targets of set 2, one hidden layer and 300 cycles were sufficient. All other tested networks used 300 or 400 cycles. As stated in section 3.3.5, the main difference between the FF and MLP models is in the way the information is propagated across the
network. The FF networks are simpler than MLPs in their nature, but MLPs should be more suitable for complex problems. Nevertheless, both the problems tackled in the above sections turned out to be simple enough to be solved equally well by the FF and MLP models.

The PNN model achieved slightly lower accuracy than the MLP models only in classification into classes of set 2 (0.88 vs. 0.99). Results of the first classification, for PNN and MLP were equally accurate (0.96 vs. 0.99). This agrees with Mostafa’s observation that the networks can be equally efficient but the PNNs are much faster and simpler to tune (Mostafa 2009). Which is why also Mishra recommended the PNNs after he used them to obtain 100% accuracy in classification of vehicles (Mishra et al., 2013).

Even though both Mostafa’s research and the current study used PNNs for classification of blood donors, the two approaches are quite different. First, Mostafa used a very different set of inputs for the classification models. This study was expected to analyse the datasets available to the IBTS. In result, the dataset was very different from the dataset used by Mostafa who based his experiments on data coming from questionnaires he designed. The questionnaires included some socio-demographic features and variables related to TPB (section 2.6.1). Second, Mostafa also tried to classify donors into a different set of classes. Similar to Santhanam and Sundaram, he tried to separate donors from non-donors, what will be the subject of the experiments presented in Chapter 6 of this thesis. As the PNN models achieved equal or almost equal accuracy as other ANNs and DT, models the research confirms the usability of all the algorithm for finding regular donors based on the data patterns hidden in the database of the IBTS.

Furthermore, the fact that presented examples of the ANN achieved similar classification results to the DT models agrees with the observation derived by Santhanam and Sundaram (Santhanam & Sundaram 2010). All the experiments performed by them (section 3.4) used blood donation history for inputs and achieved almost identical donor classification results with ANN and CART models. The dataset used in the current study is very similar to the dataset used by Santhanam et al (Santhanam & Sundaram 2010). However, they used this dataset to predict donors and non-donors. The experiments described in section 5.4 proved the data can also be used for donor classification into four rather than two classes and contribute to profiling of regular donor.

109
Also, examples of studies not related to blood donation confirm almost identical accuracy of DTs and ANNs. One that describes the comparison of ANNs and DTs for medical decision-making also confirms that the run-time of the ANN models can be significantly longer than the run-time of the DTs what was observed in the current research. The observed training time of the ANNs was even more than ten times longer than the training time of the DTs. Another research on donor classification that was performed on the population of India applied the approach proposed by Mostafa but used only ANNs and DT. The results showed higher classification accuracy of ANN models, which does not agree with the current study (Boonyanusith & Jittamai 2012).

The classification experiment performed in Chapter 5 confirmed that donation history is sufficient to select regular donors, which agrees with the statement of Ferguson and Bibby (Ferguson & Bibby 2002). This means the data patterns corresponding to regular donors are hidden in this data. This finding fully answers the first research question and partially the third research question. The patterns related to the regular donor class indeed exist in the donation history and are described in section 5.5.2. The experiments described in the next chapter will look for patterns related to future donors thus will complement the answer to the third research question.

For each of the two sets of experiments described in sections 5.3. and 5.4 an additional analysis was performed that aimed to compare the MLP classification models, each using a different number of inputs. The results revealed that only four input variables are needed to achieve the most accurate classification of donors. For classifying donors into classes of set 1, these are the time since the first donation, time since the last donation, donation interval average, and deferrals length average. Variables referring to deferrals are much more important for predicting targets of set 2. For the IBTS, it is recommended to use the information on deferrals for selecting the regular donors. Both the number of deferrals and the length of deferrals should be applied for the best results.

Most importantly, the applied methodology turned out successful for isolating the donor groups and isolating profiles of regular donors, even though the dataset and donor classes were different from what has been proposed by the previous research. Thus, Chapter 5 answers the questions corresponding to the first and third research aims positively.
Based on the current research, it is recommended to the IBTS to use tuned DT models with CART or C5.0 algorithms. They provide the same classification accuracy and efficiency as FF and MLP models but are easier to interpret thanks to the ability to display their decisions and show the discovered patterns of data relevant to each class. Possibly, they could use the boosted trees if the classification results need to be improved. The RF algorithm provides the feature importance scores that can help to understand the dataset.

5.5.2. Review of the discovered patterns

Furthermore, the DT models identified some general patterns in the database that correspond to the regular donor class. Most importantly, the long-time, active regular donors have multiple donations within last two years, their total donating period (between first and last donations) is longer than a year, and the average interval between donations is shorter than 500 days (Figure 21). This confirms the finding described by Ferguson and Bibby, saying that a regular donor may not donate blood every time he/she is expected to donate. In Ireland, this means every three months. Nevertheless, their donations may be cyclical and of various frequency but recurrent (Ferguson & Bibby 2002). Donors with 10 donations in two years (over 733 days) can be considered lapsed, as most probably they will not re-donate (Figure 38 - Appendix A).

During the second set of experiments, both active and non-active regular donors were identified. Most of the classified regular donors had average donation interval below 528 according to the CART tree and below 400 days according to the C5.0 tree. All the displayed trees show an important correlation. Namely, the longer the donating period, the longer the donation intervals of regular donors. This observation confirms another finding described by Ferguson and Bibby, saying that the relationship between subsequent donations may not be linear but quadratic (Ferguson & Bibby 2002). Furthermore, the discovered pattern proves that the number of donations is not enough to indicate if a person is a regular donor or not. This opposes the argument of multiple researchers, like Piliavin, that blood donation becomes a habit and after several successful
donations, the donor continues to donate very regularly, only if he/she is eligible, so the intervals tend to be equal (Piliavin et al. 1982).

The decision trees did not find any patterns showing the particular importance of intervals between the first two and last two donations. This observation contradicts the argument presented by Schreiber and Ownby that the interval between first two donations strongly affects the appearance of the following donations (Schreiber et al. 2005) (Ownby et al. 1999). Nevertheless, the average donation interval was most often used for decision splits of the classification trees, what confirms the importance of all the intervals lengths suggested by Schreiber (Schreiber et al. 2005).

Additionally, the long-time donors were found to have multiple deferrals but not as long as new and first-time donors who tend to have fewer but very long deferrals. Also, the time passed since the last donation was crucial for selecting the active donors. Namely, donors who have not been donating for the last two years were classified as non-active, lapsed donors.

The classification experiments described in this chapter achieved very good results regarding the models' accuracy and reliability, proving that the patterns identified within the data are very efficient. However, the patterns displayed by the decision trees are very broad and not specific enough to generate meaningful profiles of regular donors. Nevertheless, for the IBTS, it is recommended to isolate regular donors with the application of classes of set 1. This will limit the classification to active regular donors, with donations in the last two years, who are more likely to donate in the future.

Summing up, this part of the thesis answers the first research question proving that the regular donors can be selected from the IBTS database with machine learning algorithms. The most efficient classification methods were presented. The identified patterns could be used as a start to the generation of a regular donor profile.
CHAPTER 6

Predictive Experiments

6.1. Introduction

Chapter 5 demonstrated the use of the blood donor dataset for the classification of donors. Most of the data mining models achieved high classification accuracy. This chapter continues the analysis by the building of predictive models with machine learning algorithms. The models use a slightly amended set of inputs where the number of donations is added to the set of inputs.

Here the models tried to predict two classes of donors: future donor (Y) and non-donor (N). Future donor group will predict donors, which are:

- expected to donate blood within the particular period.
- non-donors, registered donors not expected to donate within the particular period.
The results of the predictions were analysed to find out which input variables were the strongest predictors of the donor class. This chapter focuses on answering the second and third research questions, which ask if future donations can be predicted with machine learning techniques and if the IBTS donor database includes strong data patterns that can make these predictions reliable.

The rest of this chapter is laid out as follows: Section 6.2 discuss the pre-analysis performed before running the main experiments. The dataset was preliminarily analysed with the aim to investigate the correlations among the variables and their influence on the variance in the dataset. Detailed pre-analysis results and the outputs are included in Appendix B.

The next three sections 6.3, 6.4 and 6.5 describe the building of machine learning models for predicting future donations of the registered donors. The experiments were divided into three parts, each performing slightly different prediction of donations. The generated models looked for the hidden patterns within the dataset to categorise all individuals into two groups: future donors (Y) and non-donors (N). The outputs produced by the main predictive experiments are included in Appendix B.

Finally, section 6.6 summarises and evaluates the results of all the experiments introduced in the previous three sections to demonstrate the potential of the machine learning algorithms for performing reliable predictions of blood donations based on the donor data available to the IBTS. The section concludes by identifying patterns in the dataset, which can help to sketch a profile of the future donor, which is the subject of the third research question.

6.2. Pre-analysis of dataset for predicting donations

The pre-analysis of the generated datasets was performed with use of the PCA, LDA, and RF regression as discusses in section 4. The aim is to identify the important features inside the dataset before applying the predictive models described in sections 6.3, 6.4 and 6.5.
The information about each input feature helps in building efficient models as well as to generate a high-level description of future donor and non-donor groups.

6.2.1. Dataset description

The input variables applied to the models predicting donations were slightly different from the input set used in the previous analysis described in Chapter 5. The dataset used for predicting donations contains 14 columns and 4500 records (Table 6). The dataset includes one binary input (X1) and two categorical variables relating to donor's age (X2, X4). In the results of data pre-processing, two averages were added to the dataset (X7, X8). The data includes three counts that refer to donations and deferrals recorded in the original IBTS database (X3, X4, X13) and six variables that describe duration in time (X5, X6, X9, X10, X11, X12). Section 4.6.2 described the pre-processing and labelling of the dataset in more detail.

The predictive models choose between the two possible outputs: future donors and non-donors. The dataset contains records of 244 donors and 4256 non-donor. The following table presents the input set used in the experiments presented in this chapter.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1 Gender</td>
<td>category: 1 (female), 2 (male)</td>
</tr>
<tr>
<td>X2 age group:</td>
<td>category: 1 (&lt;25), 2 (25-35), 3 (35-45), 4 (45-55), 5 (&gt;55)</td>
</tr>
<tr>
<td>X3 donations numbers</td>
<td>numerical (days)</td>
</tr>
<tr>
<td>X4 deferrals number</td>
<td>numerical (days)</td>
</tr>
<tr>
<td>X5 time since the first donation</td>
<td>numerical (days)</td>
</tr>
<tr>
<td>X6 time since the last donation</td>
<td>numerical (days)</td>
</tr>
<tr>
<td>X7 donation interval average</td>
<td>numerical (days)</td>
</tr>
<tr>
<td>X8 deferral length average</td>
<td>numerical (days)</td>
</tr>
<tr>
<td>X9 last deferral length</td>
<td>numerical (days)</td>
</tr>
<tr>
<td>X10 first two donations interval length</td>
<td>numerical (days)</td>
</tr>
<tr>
<td>X11 last two donations interval length</td>
<td>numerical (days)</td>
</tr>
<tr>
<td>X12 donating period</td>
<td>numerical (days)</td>
</tr>
<tr>
<td>X13 two-year donations number</td>
<td>numerical (days)</td>
</tr>
<tr>
<td>X14 first donation age</td>
<td>category: 1 (&lt;25), 2 (25-35), 3 (35-45), 4 (45-55), 5 (&gt;55)</td>
</tr>
</tbody>
</table>

Table 6 List of inputs used for predicting donations

6.2.2. Pre-analysis with PCA

PCA was performed on the dataset. Figure 27 displays the distribution of the two groups of donors according to PC1 and PC2. It clearly shows that the PCA did not derive any distinct clusters from the input set. Objects of the two classes strongly overlap. The fact that objects of class 1 (donors) constitute only a small portion of all the objects suggest that their classification may be difficult. The blue arrow in Figure 27 points to the place in the picture where the biggest group of 1s is located.
Figure 27  Distribution of class objects from dataset used for predicting donations (PCA)

Figure 28 illustrates the same distribution of all the objects. Additionally, it includes the variable vectors generated by the PCA.
Several important correlations have been found based on the PCA output. To start with, gender (X1), age group (X2) and the last deferral length (X9) strongly correlated with each other and slightly correlated with the lack of donation. Also, two-year donations number (X13), donations number (X3) and deferrals number (X4) are closely related. The high number of two-year donations (X13) and of all donations (X3), together with a slightly lower value of deferrals number (X4) indicate the high probability for future donation. The first interval between donations (X10), first donation age (X14), time since the first donation (X5), donation interval average (X7), and last interval between donations (X11) are also correlated with each other. Their high values indicate no future donation. According to the PCA output, the time since the last donation is not closely correlated with any other variable. This means that its high score is positively related to the lack of donation. The length of the donating period (X12) is also not clearly related to any of the inputs and targets. The results show that the first two principal components
(PC1 and PC2) explain 47 percent of the variance in the dataset. The first eleven components explain 96% of the variance and 13 components explain 100%.

6.2.3. Pre-analysis with LDA

LDA analysis was then performed on the dataset. This analysis showed some distinctive features of the two groups: future donors (Y) vs. non-donors (N). Firstly, the future donors have a high mean of total donations comparing to those that do not donate. (19.5 vs. 3.5) Secondly, future donors have a much lower mean of the time passed since the last donation (349.9 vs. 2385 days) and a lower mean of the average deferral lengths. For donors who experienced deferrals, their average deferral lengths were almost a half of those experienced by non-donors (120.3 vs. 222.9). The total donating period was longer for donors than for non-donors, what confirms the positive correlation between the length of the donating period and the probability of donation (3761.5 vs. 979). Nevertheless, particularly important is the number of donations within the last two years (3.8 vs. 0.2).

This suggests that several donations in the last two years significantly increase the chances that he/she might donate in the future. When considering deferrals, the number is twice as high for donors than for non-donors. According to the coefficients of the linear discriminants, the most important variables correlated with the considered classification results are: two-year donations, donations interval average, average donations interval of last two years, and first donation interval. Male gender was only a slightly stronger indicator of the donor class (1.7 vs. 1.5).

6.2.4. Pre-analysis with regression

As discussed in section 4.4.1, regression gives another view on the feature importance. In this study, regression was conducted on selected pairs of inputs variables that are likely to show meaningful relations according to the previous pre-analysis techniques.
The correlations were displayed on the generated plots included in Appendix B (Figure 41 - Figure 54). The following principal conclusions have been made based on the regression outputs.

Firstly, the regression results confirmed that the shorter the interval between the first two donations, the higher number of donations through life. Only those whose second donation happens soon after the first one become long-term donors (Figure 42 - Appendix B). Indeed, future donors have the shortest time since the last donation, up to two years at most (Figure 45 - Appendix B). Whereas, the analysis showed that long donating period does not always refer to numerous donations.

The group of donors with up to 20 donations within five years donate most often. According to the RF regression, long-term donors usually stop after 40 donations. Those who kept donating after 40 donations continue for many more donations (Figure 43 - Appendix B). Most of the donors with a long donating period had short intervals between donations (Figure 46 - Appendix B).

When it comes to donation intervals, the regression showed that the younger the donor, the more probable it is that they would donate second time within a year. The intervals are much longer for people who begin donating after the age of 55 (Figure 44 - Appendix B). The length of intervals between donations is only minimally different for both genders. There is a small tendency for women to have a higher average of interval lengths (Figure 47 - Appendix B). However, there is no clear difference in the total donating periods between males and females (Figure 48 - Appendix B). However, a long-term donor with over 30 donations is likely to be a male (Figure 41 - Appendix B).

Finally, the correlation between the time passed since the last donation and next donation occurrence is very strong. The shorter the time since the last donation, the more likely it is a person will re-donate. According to the regression output, for people who have not donated for the last two years, it is quite unusual to donate again (Figure 45 - Appendix B). This finding agrees with the finding derived from the analysis of the dataset described in Chapter 5 and confirms that the class of lapsed (L) donors generated for those experiments was created based on a valid assumption.
6.2.5. Pre-analysis with multiple regression

Another method for analysing the dataset is multiple regression, which allowed for displaying and interpreting the relations between many input variables and the outputs. The RMSE was used to describe the relationships and was calculated for each of the 14 models.

Figure 29 shows a very sudden increase of RMSE for two models with added fourth (deferrals number) and ninth (last deferral length) input variables.

![Regression RMSE](image)

*Figure 29 Multiple RF regression: inputs to targets set 2*

The second plot on Figure 30 shows the same regression results without the two variables. In this case, the plot shows the error continuously dropping. The lowest RMSE was achieved by the 11th model – with 11 inputs.
The input variables that positively influence the RMSE are time since the first donation, time since the last donation, two-year donations number, and donating period. The total donations number surprisingly was shown to have an adverse influence on the RMSE in contrast to the number of donations within the last two years.

6.2.6. Summary

The findings presented in this section revealed a lot of useful information about the dataset and the groups of donors. Different techniques resulted in obtaining slightly different pre-analysis outcomes. In the context of donation predictions, all the methods confirmed the high predictive power of the following inputs:

- time since the last donation.
- two-year donations number.
- donating period.
The performed PCA revealed that the full set of inputs, including the first 13 variables in the dataset, needs to be applied to predictive models as they explain 100% of the variance in the dataset.

6.3. Predicting three-month donations

Four different data mining models were used for predicting donations in the 3-month period of a winter season (December 2015, January 2016, and February 2016). Available donation data before these three months was used to predict occurrences of donations during period.

Two classes of targets were used: donor (Y) and non-donor (N). The predictive models used within this section were based on the LDA and QDA algorithms as well as some more sophisticated architectures like MLP using standard backpropagation and DTs using the CART algorithm.

The dataset used in the experiments was reduced according to the pre-analysis of the dataset performed in section 6.2. As discussed the dataset includes highly unbalanced numbers of donors vs. non-donors. Therefore, the prediction results achieved within this section are expected to demonstrate how well the different algorithms perform with unbalanced datasets.

6.3.1. LDA

The LDA model was implemented with the algorithm provided by the “caret” library in the R programming language. The arguments of the “train()” function were set according to section 4.3. Specifically the “method” parameter was set to “lda” and the iteration number to 300. The tuned model was tested with 205 previously unseen inputs. The model obtained an accuracy of 0.97 and the balanced accuracy 0.82. Its reliability was confirmed by kappa score close to 0.7 (0.68). The LDA model misclassified 19 out of 51 donors and 17 out of 1064 non-donors (Table 25 - Appendix B).
Based on the overall accuracy score, the model's performance was very high. Nevertheless, the algorithm did not cope sufficiently well with the unbalance within the dataset that was caused by the unequal group sizes. The algorithm was much more successful in the classification of the predominant class of non-donors.

6.3.2. QDA

A QDA model was trained and tuned in the same way as the LDA model in the previous experiment with one small exception, namely the “method” parameter inside the “train()” function was changed to “qda”.

The tuned QDA model was tested on the same test set as the LDA model. The model misclassified 8 out of 61 future donors but misclassified 95 out of 1064 non-donors. (Table 26 - Appendix B). The donor class specificity scored 0.89 while its sensitivity was 0.87 and the balanced accuracy was 0.89. Unfortunately, the p-value calculated for the QDA model using the test dataset reached the unwanted value (1). QDA obtained only slightly lower accuracy than the LDA (0.97 vs. 0.90) and higher balanced accuracy (0.87 vs. 0.82) and much lower kappa (0.46 vs. 0.68). This indicates that the QDA predictions were less reliable. Nevertheless, it is worth noticing that in contrast to LDA, the QDA model has a tendency to misclassify the predominant group of non-donors rather than the smaller group of donors.

6.3.3. CART

For training and tuning of the CART model, the “method” parameter inside the train function was set to “rpart”. The rest of parameters was set as described in Chapter 4.3. Twenty trees were used for tuning of the model the model was tested with the test dataset.

During testing, the CART model (Figure 55 - Appendix B) misclassified 28 out of 51 of the donors, but only misclassified 8 out of 1064 non-donors. This is similar to LDA where the model also tended to classify inputs into the non-donor class rather than the donor
Classification decisions were based on six variables in the following order:

- two-year donations,
- donation interval average,
- time since the last donation,
- total donation number,
- donation interval average,
- deferral length average.

The root node of the generated decision tree identifies the most significant attribute, which is the number of overall donations. This indicates that donors are most likely to have at least five donations in the last two years. The algorithm found donors in the following scenarios. All the scenarios are displayed in Figure 31.

1) The first scenario assumes their time passed since the last donation is over 46 days and their average two-year donation interval is less than 104 days.
2) The second scenario, donors are those with the average of two-year donation interval lengths above 104 days, the last deferral length below 29 and the total donation number of 13 at least.
3) The third scenario identifies donors with the average two-year donation interval longer than 104 days, the last deferral length above 29 and belonging to the fourth age group at most.
4) The fourth scenario, donors have less than 5 donations in the last two years, but not more than 22 overall donations, and their average deferral length is longer than 129 days.
5) The last scenario assumes donors have less than five donations in the last two years, more than 22 overall donations and their average deferral length is over than 129 days, but their average donation interval is over 180 days and their maximum number of deferrals is two.

The accuracy of the CART model was similar to the accuracy obtained by both LDA and QDA. The model scored very low on the sensitivity (0.52) and very high on the specificity.
(0.99). The low sensitivity shows the low predictability of the donor class. In the result, the model obtained the balanced accuracy that was the lowest of the three models (0.75). Nevertheless, the model observed quite detailed patterns related to future donors.

Figure 31 CART tree generated for predicting three-month donations
6.3.4. Multi-Layered Perceptron (MLP)

The MLP model with traditional backpropagation algorithm was tuned to examine the suitability of ANNs for predicting blood donations based on the available dataset. For testing and tuning the parameters were set according to section 4.3.

The models used for training were varied to have three to eight nodes in the hidden layer, the model used fourteen inputs. The best MLP model selected for testing used three hidden layer nodes the tested model incorrectly classified 23 out of 61 donors resulting in the model sensitivity of 0.69 and balanced accuracy of 0.8 (Table 28 - Appendix B).

Applying multiple hidden layers, different learning functions, and learning rates, the accuracy of the MLP model was not improved.

6.3.5. Feed-Forward Neural Networks (FF)

The feed-forward networks (FF) proved sufficient classifiers in the experiments described in Chapter 5. Therefore, they were also applied in the predictive experiments.

For training, the FF networks used a range of network sizes between three and eight. The best model selected through tuning was of size 4 and had the weight decay of 0.1. Based on the results of this testing, the model accuracy was equal to the accuracy of the MLP and LDA models but its kappa was only about 0.54. The model misclassified 36 out of 51 donors and only 4 out of 1064 non-donors. It shows that similar to MLP, LDA, and CART, the model tended to classify objects to the predominant class of non-donors (Table 29 - Appendix B).

The FF model achieved better results than the MLP model in predicting the non-donor class (1060 vs. 1180). However, the number of recognised donors was worse (25 vs. 38) resulting in the low sensitivity (0.4) and lower balanced accuracy of the model (0.7). Concluding, that the perceptron model is less efficient than the MLP model using the backpropagation algorithm.
6.3.6. Probabilistic Neural Network models (PNN)

The PNN model was trained with the same dataset as the previous models. The inputs were normalised with function “normalizeData()” provided by “RSNNS” library. The model was tuned with the smoothing parameter between 2.0 and 0.1. The most successful model had the smooth parameter equal to 0.03.

According to the results of the testing, the model had a high accuracy of 0.94. The model misclassified 34 out of 1064 non-donors and 25 out of 61 donors (Table 30 Three-month donation prediction with PNN. The accuracy of the PNN model was comparable with the other models. It did not improve the prediction results achieved with the previously described models.

6.3.7. Summary

Section 6.3 has described the application of five classification models for the prediction of the next three-month donations based on the dataset presented in section 6.2. The fourteen variables of the dataset were used as inputs into the binary classification models for predicting an individual’s donation or its lack in the forthcoming three-month period. The table below (Table 7) summarises the outputs of each applied algorithm. It presents the efficiency of the best predictive models that were selected based on the testing results: accuracy, kappa, p-value, and the balanced accuracy of the donor class. Additionally, the table includes the sensitivity and specificity of each model that further outputs its predictive ability showing the ability to recognise future donors and non-donors.

The summary shows that the best prediction accuracy was achieved with the LDA and MLP models. The LDA model achieved the highest kappa value and lowest p-value, what makes the model more reliable than the other ones. There is a strong unbalance between the sensitivity and specificity of most of the models. All models have specificity close to 1, whereas the sensitivity is much lower. The unbalance is the strongest for the FF model (0.4 vs. 0.99), which indicates that the model predicts future non-donors very accurately, but has lower ability to recognise future donors – which were represented by a significantly lower number of inputs.
On the other hand, the QDA model has the lowest unbalance between the sensitivity and specificity (0.87 vs. 0.91) showing the high accuracy of 0.908 (although the lowest of all the models). Nevertheless, kappa calculated for the QDA model was below 0.5. It suggests that the testing results are still much better than a random selection of classes, but it is lower than the required kappa of at least 0.7 (PsychologyWikia 2017). The p-value for the model reached the worst possible value of 1, what may disqualify the model or mean that no p-value could be selected for the models (Senn et al. 2016).

Kappa calculated for LDA model is the nearest to 0.7 of all the models. The LDA model sensitivity vs. specificity balance is second best in the table. Its sensitivity was close to 0.7, which suggests it has significant potential for improvement. The LDA model more accurately recognised future non-donors than donors and showed the tendency to classify donors as non-donors. Nevertheless, the model achieved the best balance between recall (sensitivity) and precision (0.71) – 71% of the predicted donors were true donors, and 69% of all donors in the dataset were recognised by the model. The best precision was observed for the FF model (0.86) which also had the lowest recall (0.4) meaning that only 40% of all donors in the dataset were classified correctly, but among the predicted donors 86% of them were true donors. The lowest precision (0.36) and highest recall (0.87) was observed for the QDA model. This confirms that the algorithm had a tendency to classify inputs to the positive class.

From the IBTS point of view, it might be beneficial to run two models with the opposite tendencies (LDA and QDA) and compare their prediction outputs. If it comes to the PNN model, even though its overall accuracy was very high, the balanced accuracy was much lower (0.77). The model predictions were unreliable as indicated by a kappa much below 0.7 (0.52) and high p-value of 0.42.
<table>
<thead>
<tr>
<th></th>
<th>LDA</th>
<th>QDA</th>
<th>CART</th>
<th>MLP</th>
<th>FF</th>
<th>PNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.968</td>
<td>0.908</td>
<td>0.97</td>
<td>0.96</td>
<td>0.96</td>
<td>0.94</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.68</td>
<td>0.46</td>
<td>0.61</td>
<td>0.64</td>
<td>0.54</td>
<td>0.52</td>
</tr>
<tr>
<td>P-value</td>
<td>0.00028</td>
<td>1</td>
<td>0.00048</td>
<td>0.002</td>
<td>0.002</td>
<td>0.42</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.69</td>
<td>0.87</td>
<td>0.52</td>
<td>0.62</td>
<td>0.4</td>
<td>0.59</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.98</td>
<td>0.91</td>
<td>0.99</td>
<td>0.98</td>
<td>0.99</td>
<td>0.96</td>
</tr>
<tr>
<td>Balanced Accuracy</td>
<td>0.82</td>
<td>0.89</td>
<td>0.755</td>
<td>0.8</td>
<td>0.7</td>
<td>0.77</td>
</tr>
<tr>
<td>Recall</td>
<td>0.69</td>
<td>0.86</td>
<td>0.52</td>
<td>0.62</td>
<td>0.4</td>
<td>0.87</td>
</tr>
<tr>
<td>Precision</td>
<td>0.71</td>
<td>0.36</td>
<td>0.8</td>
<td>0.69</td>
<td>0.86</td>
<td>0.51</td>
</tr>
</tbody>
</table>

Table 7 Three-month donation prediction with LDA, QDA, CART, MLP, and feed-forward models

6.4. Predicting one-year donations

The dataset used in experiments in this section used the same set of variables as the previous experiment. This time historical data up to January 2014 was used to predict donations for the period starting from 1st January to 31st December 2014. All the records containing information on donors who donated only after this date were removed from the dataset. In result, 4495 records were left in the dataset. Only 477 records had a positive donation outcome (Y) which again demonstrates that this dataset is unbalanced. According to LDA pre-analysis, the probability for predicting donor class Y was 0.106 and probability for the non-donor class was 0.894. Three different data mining models were selected to evaluate their predictive capabilities in regards to the available dataset: CART, LDA, and MLP with backpropagation.
This section presents prediction results achieved only with the three algorithms that performed best in the experiment predicting donations within the three-month period. Each of them represent different approach towards classification offered by DTs, DA, and ANNs.

6.4.1. CART

The CART model applied for predicting one-year donations was trained and tuned. The “method” parameter inside the “train” function was set to “rpart”. The rest of parameters were set as described in Chapter 4.3.

Twenty trees were used for tuning of the model. The tuned model used 300 cycles for training. According to the test executed on the model, it reached the accuracy of 0.93 and kappa close to 0.7. It misclassified 27 out of 119 donors and 46 out of 1014 non-donors achieving the sensitivity of 0.77, of specificity of 0.95 resulting in the balanced accuracy of 0.86.

The most accurate decision tree (CART) for predicting one-year donations is displayed below (Figure 32). Its decisions were based on four input variables: two-year donations, donation interval average, age at the time of the first donation and the last two donations interval. The individuals identified as future donors have at least one donation in the last two years and average donation interval less than 267 days. If the average is higher, the first donation age is in the first age groups, and the last two donations interval is less than 712 days.
According to the prediction results, the CART model had difficulties with distinguishing between the two possible targets (donor, non-donor). However, it managed to classify 92 of 119 donors correctly and failed in the classification of 46 out of 1004 non-donors (Table 31 - Appendix B).

This model proved to be an efficient method for predicting future donors with the balanced donor class accuracy of 0.86 and kappa close to 0.7. This is a promising result considering the unbalance within the groups’ sizes. Similar to the previous set of experiment (section 6.3) the algorithm identified some specific patterns related to future donors.
6.4.2. LDA

For tuning and training of the LDA model, the “train()” function arguments were set as described in section 4.3. The “method” parameter was set to “lda”. The model used 300 iterations for training.

The model obtained a high accuracy of 0.94 and balanced accuracy of 0.81. Its reliability was confirmed by kappa score close to 0.7 (0.68). The LDA model misclassified 40 out of 119 donors and 46 out of 1014 non-donors (Table 32 - Appendix B).

Based on the overall accuracy score, the model’s performance was almost identical to the performance of the CART model. The only difference between the models was observed in the sensitivity calculated for the models. It was slightly lower in the case of the LDA model (0.64 vs. 0.77). The difference between the sensitivity and specificity scores (0.67 vs. 0.99) show that the algorithm did not cope with the unbalance in the dataset caused by the unequal group sizes. Similar to CART, the LDA algorithm tended to assign objects to the predominant class of non-donors.

6.4.3. MLP

In the next attempt to predict next year donors, an MLP model with the traditional backpropagation algorithm was trained and tuned. For testing and tuning of the model, the parameters were set according to section 4.3.

The trained models used three to eight nodes in the hidden layer, as the model used fourteen inputs. The best MLP model selected for testing was using only three hidden layer nodes. Nevertheless, the tested model wrongly classified 40 out of 119 donors and 21 out of 1014 non-donors resulting in the model sensitivity of 0.66, specificity of 0.97 and balanced accuracy of 0.82 (Table 33 - Appendix B).

Summing up, the above MLP scores are identical with the scores calculated for the LDA model. The two models coped slightly worse with the unbalanced group sizes than the CART model.
6.4.4. Summary

Section 6.3.2 presented experiments that aimed to predict occurrences of donations within one year. Donations in the year 2014 were predicted based on donations in years 1965 to 2013. Table 8 describes the best predictive ANN models based on their testing results. The highest accuracy and kappa were calculated for the MLP model. The model also achieved best score for precision (0.79) and recall (0.66). Based on their values 79% of predicted donor were true donor, while 66% of all donors in the dataset were recognised by the MLP model.

The p-value is very low for all the models what is highly desirable because it suggests the testing results are far from what the null hypothesis tests predicted for the models. (Gelman 2013) All the models have very high and almost identical accuracy. Kappa close to 0.7 for all models suggests their predictions are trustful. Nevertheless, the CART model testing showed the best balance between its sensitivity and specificity (0.77 vs. 0.95). According to the specificity of 0.77 and the balanced accuracy of the donor class equal to 0.86, donation prediction with the CART algorithm is quite successful considering the highly unbalanced nature of the dataset related to the class sizes.
<table>
<thead>
<tr>
<th></th>
<th>LDA</th>
<th>CART</th>
<th>MLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.943</td>
<td>0.935</td>
<td>0.945</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.68</td>
<td>0.68</td>
<td>0.69</td>
</tr>
<tr>
<td>P-value</td>
<td>0.000000005281</td>
<td>0.000001315</td>
<td>0.0000000006328</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.66</td>
<td>0.77</td>
<td>0.66</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.97</td>
<td>0.95</td>
<td>0.97</td>
</tr>
<tr>
<td>Balanced</td>
<td>0.81</td>
<td>0.86</td>
<td>0.82</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>0.76</td>
<td>0.66</td>
<td>0.79</td>
</tr>
<tr>
<td>Precision</td>
<td>0.76</td>
<td>0.66</td>
<td>0.79</td>
</tr>
</tbody>
</table>

Table 8 One-year donation prediction with LDA, CART and MLP models

6.5. Predicting one-year donations of active donors

For predicting one-year donations, the dataset was amended. Namely, all records with no donations recorded within the previous two years were removed. This change was performed with the view to improving the predictability of donations. This pre-processing reduced the size of the dataset from 4495 to 669 records. This could be expected to have an influence on its ability to predict. Four algorithms were applied in this experiment: LDA, QDA, CART, and MLP.
6.5.1. LDA

The LDA model was tuned in the same manner as in the previous experiments. According to the confusion matrix summarising the classification results, 81% of the model’s predictions were accurate. The model misclassified 21 out of 55 donors and 10 out of 111 non-donors achieving the sensitivity of 0.62, specificity of 0.9 and balanced accuracy of 0.7 (Table 34 - Appendix B).

Concluding, the accuracy of the model was satisfactory, but kappa well below 0.7 (0.56) reveals low reliability of the model. The model tended to classify donors as non-donors, which reveals that it could not cope with the unbalance between the sizes of the donor and non-donor groups. This resulted in the significant unbalance between the sensitivity and specificity scores (0.62 vs. 0.0).

6.5.2. QDA

QDA was applied next. The QDA model was prepared in the same way as in section 6.2 and tested with the test dataset.

The tuned QDA model’s accuracy reached only 0.68 but its balanced accuracy was higher than in case of the LDA model (0.76). The model misclassified 12 out of 55 donors and 41 out of 111 non-donors (Table 35 - Appendix B). The model’s sensitivity of 0.78 and specificity of 0.64 show that the QDA algorithm was more successful in predicting donors than non-donors. This fact confirms the previously observed tendency of the algorithm to accurately classify objects of the less populated group. Nevertheless, a very low kappa of 0.37 was calculated for the model, what means the model predictions were not reliable.

Summing up, the model presented a different approach than LDA towards a classification into groups of unequal sizes. Even though it obtained higher model sensitivity, it turned out to be inefficient for the current classification problem, due to its very low reliability suggested by kappa calculated for this model.
6.5.3. CART

For training and tuning of the CART model, the "method" parameter inside the train function was set to "rpart". The rest of the parameters was set as described in Chapter 4.3. Twenty trees were used for tuning of the CART model. The total time needed for training and tuning was very short. The tested CART model (Figure 56 - Appendix B) misclassified 25 out of 55 donors, and only 15 out of 116 non-donors, resulting in the sensitivity of 0.64 and specificity of 0.86 and balanced class accuracy of 0.75, while the overall model accuracy was equal to 0.79. Again, the CART algorithm tended to classify objects into the predominant class of non-donors (Table 36 - Appendix B).

The displayed CART decision tree showed that the classification decisions of the model were based on six variables in the following order: two-year donations, donation interval average, time since the last donation, total donation number, donation interval average and deferral length average. (Figure 56- Appendix B) The first scenario finds donors that have more than five donations in the last two years. All other scenarios assume donors have less than six donations in their life. Most of them would have the time passed since the last donation between 109 and 243 days. The other scenario assumes donors have the time passed since the last donation shorter than 110 days, their average donation interval is less than 193 days, and the last interval length is shorter than 106 days.
Summing up, the model revealed quite specific patterns of donors existing in the dataset. Even though the model accuracy scored very high, the predictions could not be considered reliable based on the kappa of 0.51.

### 6.5.4. MLP

Because of the low accuracy of the three previously used algorithms (LDA, CART, QDA), ANNs were applied into this classification. As an example of ANNs, an MLP model was trained and tuned. For tuning, networks of different sizes were built as described in section 4.3.

The tuned models were built using four nodes in the hidden layer. Based on the confusion matrix, the prediction accuracy of the tuned model was very similar to the other models, 0.79. The MLP model misclassified 23 out of 55 donors and 21 out of 111 non-donors (Table 37 - Appendix B).
Summing up, the model's characteristics calculated based on the predictions were almost identical to the characteristics of the CART model, but the time the MLP model spent on training and tuning was much longer.

6.5.5. Summary

Section 6.5 presented the experiments applied to predict donations of currently active donors. Results of the models' testing are displayed in Error! Reference source not found.. Based on the table's content it turns out that predicting donations of the active donors was much harder for the algorithms than predictions performed in the previous two sections. This change results from the fact that the dataset used in the latest experiment was significantly reduced by removing all non-active donors.

Error! Reference source not found. describes each model's ability to predict donations of active donors. According to the features of each winning model displayed in the table, the best accuracy and p-value was calculated for the LDA and CART model as well as the MLP model with three nodes in the hidden layer. Unfortunately, the value of kappa calculated for all four models is lower than required to qualify their predictions as reliable. In all cases, kappa value is much below 0.7. The highest kappa was calculated for the LDA model (0.56). If it comes to the precision of predicting donor class, the LDA turned out to be the most accurate achieving precision of 0.77 with recall (sensitivity) of 0.62. This means that 77% of predicted donors were classified correctly. Recall of 0.62 means that 62% of all true donors were classified as donors.

Summing up, all four models were shown to be less efficient classifiers in combination with the dataset than those presented in Chapter 5. The QDA model had the highest sensitivity, which indicates that it recognised the highest percentage of future donors. Among all the tested models, QDA had the highest balanced accuracy, although only 1% higher than the CART model. The smallest difference between the accuracy and the balanced accuracy was observed for the CART model. The model also had the highest kappa and second highest sensitivity (after QDA).
This experiment confirmed that when dealing with an unbalanced dataset, the QDA algorithm tends to assign the inputs to the smaller group and misclassify the instances of the predominant group. Indeed, the QDA model achieved the highest recall score, meaning that it successfully classified the highest number of the true donors. However, only 51% of all the predicted donors were true donors. This reveals that the QDA classification model tended to assign inputs to the positive class, which is not always desirable. LDA, CART, and MLP represent the opposite behaviour and are more likely to label objects as negative class members. Nevertheless, from the IBTS point of view, it may be more beneficial to employ models assuring specificity rather than the sensitivity, so they will not expect more donations than will occur, overloaded blood stock is more desirable than the insufficient stock.

<table>
<thead>
<tr>
<th></th>
<th>LDA</th>
<th>QDA</th>
<th>CART</th>
<th>MLP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>0.81</td>
<td>0.68</td>
<td>0.79</td>
<td>0.79</td>
</tr>
<tr>
<td><strong>Kappa</strong></td>
<td>0.56</td>
<td>0.364</td>
<td>0.51</td>
<td>0.51</td>
</tr>
<tr>
<td><strong>P-value</strong></td>
<td>0.00002501</td>
<td>0.4</td>
<td>0.0004</td>
<td>0.0000000001736</td>
</tr>
<tr>
<td><strong>Sensitivity (Recall)</strong></td>
<td>0.62</td>
<td>0.78</td>
<td>0.64</td>
<td>0.59</td>
</tr>
<tr>
<td><strong>Specificity</strong></td>
<td>0.91</td>
<td>0.63</td>
<td>0.86</td>
<td>0.89</td>
</tr>
<tr>
<td><strong>Donor Class</strong></td>
<td>0.7</td>
<td>0.76</td>
<td>0.75</td>
<td>0.74</td>
</tr>
<tr>
<td><strong>Balanced Accuracy</strong></td>
<td>0.7</td>
<td>0.76</td>
<td>0.75</td>
<td>0.74</td>
</tr>
<tr>
<td><strong>Recall</strong></td>
<td>0.61</td>
<td>0.78</td>
<td>0.63</td>
<td>0.58</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>0.77</td>
<td>0.51</td>
<td>0.7</td>
<td>0.6</td>
</tr>
</tbody>
</table>

*Table 9 One-year donation prediction with LDA, QDA, CART and MLP models (active donors)*
6.6. Evaluation of results - summary

6.6.1. Comparison of methods

This section summarises the results of all the experiments described in Chapter 6 that dealt with the three prediction problems. The first task (section 6.3) approached the prediction of donations within the next three months related to the winter season (December, January, and February). The second task tackled prediction of blood donations within one year (section 6.3.2), and the final prediction task attempted to predict one-year donations of active donors only (section 6.3.3).

All the experiments introduced in the previous three sections were undertaken to evaluate the potential of machine learning algorithms for performing reliable predictions of blood donations based on the donor data available to the IBTS. The experiments predicting one-year and next three-month donations turned out to be difficult tasks for the chosen algorithms when compared to the classification problem described in Chapter 5, although their overall accuracy was still very high.

All the presented models obtained almost identical accuracy, except for PNNs, which scored slightly lower overall accuracy of 88%. Nevertheless, the balanced accuracy of all the models, based on the class sensitivity and specificity were between 70 and 80% for most of the models. Considering the unequal group sizes, the balanced accuracy is more reliable indicator of the models’ efficiency than their overall accuracy.

Similar experiments performed by Santhanam (section 3.4) showed identical accuracy of both ANN and CART models in predicting donors vs. non-donors. However, their predictive models were 100% accurate (Santhanam & Sundaram 2010). Even though MLP and CART offer identical prediction accuracy, for the IBTS the simple DT algorithms such as CART are recommended. They are simple to implement and tune. They learn fast and generate very accurate and reliable predictions based on the donation history. Finally, they display their decisions and the discovered data patterns.
Even though the PNNs were recommended by Mostafa for predicting of donations, the PNN models used in these predictions were much less efficient than presented in that research, which is due to the use of different datasets, described below. Nevertheless, the PNN and MLP models appeared to be equally efficient predictors of donations, which agrees with observations described by Mostafa and the current research findings presented in Chapter 5. His predictive models, both MLP and PNN, obtained 100% accuracy (Mostafa 2009).

However, Mostafa did not present the sensitivity and specificity obtained by the model for comparison. What is more, as it was already mentioned in section 5.5.1, the dataset Mostafa used did not include donation history. Thus, the datasets used in the two studies are very different. Mostafa classified donors based on the information retrieved from questionnaires (sex, age, educational level, altruistic values, perceived risks of blood donation, blood donation knowledge, attitudes toward blood donation, and intention to donate blood.) rather than from the blood collecting institution. The data was referring to socio-demographic information and the variables related to the TPB factors (section 2.6). Even though the current set of experiments used the same set of donor classes as in the case of Mostafa' research, the predictions based on the donation history of the IBTS donors were less accurate.

The balanced accuracy of the predictive models presented in Chapter 6 reached 80%. Considering the highly unbalanced group sizes this accuracy is still quite satisfactory. The completed experiments showed how donations can be predicted with use of the data available to the IBTS. Moreover, they demonstrate that donation history of donors could be used to predict their future donations. This agrees with the statements presented by Ferguson and Bibby (Ferguson & Bibby 2002) as well as Schreiber et al. who also believed that donation history influence the appearance of future donation (Schreiber et al. 2005).

The lack of the balance between the numbers of records representing both classes could have significant, negative influence on the classification accuracy. Among the 4500 objects used for predicting donations within three months, there were only 240 of the actual donors. The dataset used for predicting one-year donations had 477 donors among the 4495 records, which slightly improved the balance within the dataset. For the
experiment presented in section 6.5 that dealt with predicting donations of active donors only, all non-active donors were removed from the dataset. This significantly decreased the size of the dataset to 669 records and negatively affected the accuracy of the prediction, even though the unbalanced of the dataset was significantly reduced (221 donors vs. 446 non-donors). The dataset turned out to be too small for the algorithms to learn the patterns needed for the classification of the testing inputs. In result, the accuracy of the LDA model dropped from 95% (for 4500 records) to 75% and the balanced accuracy dropped from 81% to 70%. The smaller dataset proved insufficient for finding patterns of donation behaviour. In these circumstances, LDA, CART and MLP models performed the most accurate predictions.

The LDA models had higher kappa than the MLP models, which indicates their higher reliability. Even though the overall accuracy of the predictions was very high, the models did not deal efficiently with recognition of the future donor. All the models except QDA tended to classify objects as the member of the dominating class – the non-donor class. In result, the balanced accuracy for the donor class was much lower than the overall accuracy of the model. Nevertheless, the outcomes still show high chances for improving the prediction with the use of more balanced dataset.

Even though the dataset was overpopulated with non-donors, the achieved balanced accuracy of the prediction was quite good – over 80% with the whole dataset and over 70% in the case of the reduced dataset. It can be concluded that the presented methodology has a high potential for improvement if the dataset contains similar numbers of objects of the two classes.

Summing up, this chapter positively answered the second research question and confirmed that future donations could be predicted with the use of data mining techniques. For the IBTS it is recommended to implement predictions referring to shorter rather than longer periods, as in the above experiments the three-month predictions were more accurate than the one-year predictions. Moreover, CART trees turned out the most efficient technique for dealing with the presented prediction tasks. Not only did they prove to be among the most accurate algorithms, but also they aided the profiling of donors by displaying the corresponding data patterns. (TIME) Again, the most
meaningful patterns were illustrated by the CART tree generated for three-month predictions of donations.

6.6.2. Analysis of the discovered patterns

The experiments performed in Chapter 6 also aimed to identify patterns in the dataset that can help to sketch a profile of the future donor that is the subject of the third research question.

The CART trees used for donation prediction identified several patterns within future donor characteristics.

The time since the last donation was discovered to be an important variable for predicting donations. According to the displayed decision splits, the three-month donors have their last donations at least 46 days before the start of the considered period (Figure 31). And their average donation interval within the last two years is shorter than 104 days. Donors with a longer donation interval generally had a last deferral time of less than a month. However, deferrals only apply to donors who have had at least 13 donations.

The other pattern identifies donors that had less than five donations in the last two years and more than 22 overall donations with an average deferral length over 129 days. For these donors the average donation interval is more than 180 days, and their maximum number of deferrals is two, donors with long deferral length are typically over the age of 55. The long-term donors described by the last patterns are the most trustworthy donors that generally continue donations after the deferral period.

Another identified pattern shows future donors with less than five donations within two years, but with multiple donations during their life. These donors have at most three deferrals of the average length below three months. Whereas, their donation interval is longer that in the previous pattern – over 180 days.

The patterns identified during one-year predictions show that a future donor has at least one donation in the last two years and average interval below 267 days. The second
pattern says that the future donors have the average interval over 267 days and were below 25 years of age at the time of the first donation (Figure 32).

The described patterns confirm several of the findings announced in the previous studies. The short intervals between donations turned out to be a strong determinant of the donation within the next three months which agrees with observations described by Schreiber claiming that the shorter the interval, the more likely the next donation is (Schreiber et al. 2005). Indeed, models predicting three-month donations identified mostly the true regular donors, as they tend to donate blood every three months. This finding confirms the fact that regular donors may be led by the strong anticipated regret of not donating and self-identification as a blood donor (G. Godin et al. 2005) (Masser et al. 2009). Furthermore, only the long-time donors with over 22 donations are allowed to have up to three deferrals, but those would be no longer than three months, so the intervals between pairs of donations will be only twice as long as the usual intervals (about 180 days). This pattern confirms assumption made by Piliavin that experienced donors are more likely to continue donations after a deferral (Piliavin 1987). Even the shortest one can discourage a first-time or new donor from re-donating. Also according to Shaz et al., deferrals are correlated with donor returns (Shaz et al. 2010). That is why it is recommended to the IBTS to use deferrals information for predicting future donations.

Based on the decision trees output it has been observed that donors with more than five donations are strongly expected to keep donating in future, which confirms Piliavin’s assertion that after relatively few donations the individuals start to identify themselves as blood donors, which leads them to the following donation acts (Piliavin 1987). This also agrees with the finding described by Schreiber et al., saying that donors are highly probable to become regular donors after five donations. (Schreiber et al. 2005). This behaviour may also be affected by the anticipated regret of not donating discussed by Godin (G. Godin et al. 2005) and Masser (Masser et al. 2009).

Summing up, this chapter positively answered the second research question and confirmed that dataset obtained from the IBTS include data patterns that enable predictions of future donations. The DT models proved that the dataset received from the IBTS includes strong data patterns that can build solid and trustworthy profiles of future
donors. The third research question was answered with the description of the identified patterns of future donors.
CHAPTER 7

Conclusions

This research originated from concerns about insufficient control over the blood supplies in Ireland. The short shelf life of blood products and the unpredictability of donors' attendance at blood collection clinics has led to the wastage of unused blood products and insufficient stock levels. This thesis has examined the issues behind blood donation using data generated from the IBTS, which has been analysed using machine learning algorithms. Three research aims were defined to focus this current study.

- Firstly, confirm that data mining models and specific machine learning algorithms can be used to build classification models that are successful in identifying regular donors based on data obtained from the IBTS.

- Secondly, prove that future blood donations can be predicted with data mining techniques, based on the Irish donor database.

- Thirdly, identify data patterns that should be used to build a profile of blood donors in Ireland.
Fourthly, identify factors that influence the donation behaviour of the Irish blood donors

All four questions have been answered based on experiments performed and presented within Chapter 5 and Chapter 6 of this thesis.

The rest of this chapter presents a summary of the main research findings and contributions, finally concluding with a discussion of the possible future direction of this research area.

7.1. Main findings and contributions

Based on the literature reviewed it is clear that the profiles of regular blood donors differ from country to country (Piliavin 1990). Some researchers believe that the profile of blood donors is affected by the national organisations involved in the management of blood collection campaigns, as they address particular groups (Healy 2006). Nevertheless, studies agree that different campaigns are needed to reach different groups even in regards to gender. Moreover, blood collection strategies should be better adjusted to fit with the various lifestyles of potential donors to remove any perceived or real barriers (Gillespie & Hillyer 2002).

7.1.1. Regular donor profiling

Using data pre-analysis and the outputs produced by decision trees, a brief profile of regular donors was defined with reference to the following features: deferrals' length and numbers, intervals' between donations, the number of donations and donor age.

Firstly, regular donors were seen to have more deferrals that other donor groups however these deferrals tended to be quite short in duration in comparison to the other groups.
In particular, the pre-analysis with LDA demonstrates that the intervals between the first two donations of regular donors are shorter, when compared to new and lapsed donors. However, the patterns displayed by decision trees did not confirm these observations. The DTs identified that donors with five and more donations are more probable to become long-term donors and donate regularly but not all of them donate as frequently as every three months. This infers the conclusion that regular donors should be classed based on the number of donations as well as the frequency of donations.

When considering the age profile of donors, regular donors tended to be older than first-time donors and new donors at the time of their first donation. The majority of active regular donors are over 55 years of age. However, analysis of the dataset showed that age, as well as gender to be important factors in this dataset. For example, there are more male regular donors than regular female donors in the dataset.

When considering previous donation experience (section 2.4.1), the average donation interval proved more influential than the length of intervals between the first and last two donations for differentiating between new and regular donors. DT algorithms were used where the length of the intervals between the first two and last two donations were used to consider if regular donors have shorter donation intervals between these periods.

In summary, the classification experiments performed in section 5.3.1 proved that a regular blood donor can be successfully identified with data mining models when applied to their donation history. The answers the first research question posed. These findings match the research findings in other literature, which claims that regular donor status refers mainly to the number and frequency of previous donations (section 2.4.1). The DT classification models identified the frequency of donations as important and has almost 100% accuracy in identifying regular donors. The number of deferrals were determined to be less important for the classification of regular donors whereas the average donation intervals and the total donating period were the strongest predictors of the regular donor group. Importantly, the experiments confirmed that two-year history of an individual’s donations is sufficient for distinguishing between different classes of active donors.
7.1.2. Future donor profiling

The pre-analysis and decision trees described in Chapter 6 determined that to distinguish between future donors and non-donor's, analysis should be based on the following factors: the number of donations in the last two years, time passed since the last donation and the interval between the last two donations. This agrees with the finding described by Schreiber regarding the correlation between short intervals and likelihood to become a regular donor, and confirms that blood-collection institutions should focus on shortening these intervals (Schreiber et al. 2005). According to the decision trees, a two-year history of donations should be considered when looking for future donors rather than the full history of all donations.

The predictive models described in Chapter 6 were undertaken in an attempt to provide an answer to the second research question. Predicting future donations based on the existing data patterns was a more challenging task than the classification of donors described in Chapter 5. However, LDA, MLP, and CART algorithms managed to find patterns associated with the appearance of the next donations, and despite the high unbalance in the dataset, they identified the future donor with a balanced accuracy above 80%. Because of the high imbalance of class occurrences in the test dataset, it was difficult for the models to learn the patterns to recognise donors. The algorithms tended to incorrectly classify true donors. Nevertheless, they were still achieving a high prediction accuracy thanks to the properly classified large number of true non-donors. In result, the overall accuracy metric is not a reliable measure of the efficiency of the predictive models. For this reason, the classification results are additionally verified with the calculation of their precision and recall. The two variables were combined and provided a reliable description of the models' real abilities to recognise the positive class (donor) which was underrepresented in the dataset.

While the predictive models obtained lower accuracy and reliability than the classification models from Chapter 5 due to the highly unbalanced nature of the testing dataset, the results of the predictive tasks may be more useful for the IBTS. It should be mentioned that the prediction accuracy could be significantly improved by using a testing dataset with similar number of occurrences of both classes.
The models generated predictions for future donations irrespective of whether the donors were regular or not. Hence, it can be concluded that based on the predictive tasks a definition of a regular donor was found that turned out more accurate than the definition of a regular donor used as the classification target in Chapter 5, which was based on the number of donations only. Future donors tended to have intervals between donations close to three months and at least five donations in the last two years.

The major recommendation to the IBTS is to pay more attention to the donation predicting results described in Chapter 6 rather than trying to identify regular donors according to techniques employed in Chapter 5. The decision trees generated for the predictive tasks discovered more meaningful and specific patterns that help to create useful profiles of future donors who may turn out to be regular donors. These DT models not only prove the existence of the patterns in the IBTS dataset but also provide a useful visualisation of them.

### 7.2. Future work

This study demonstrated that machine learning has tremendous potential for building profiles of regular donors and enabling predictability of blood supplies. The classification models tested produced very successful results. However, the experiments could be still improved.

Firstly, additional variables in the dataset, especially those reflecting the sociodemographic situation of individuals would hugely increase the usability of the prediction and increase its scope. Socio-demographic data contains significant information related to the donation habits of blood donors in Ireland. Such a dataset could enable the prediction of potential future donors among individuals with no previous donation history and create potentially interesting in-depth descriptions of regular donors. Unfortunately, it was not possible to include sociodemographic information in this study from the IBTS database due to data protection issues discussed in section 4.5.
The prediction accuracy of models described in Chapter 6 could potentially be improved by fixing the balance in the dataset. Specifically having similar numbers of donors and non-donors in the dataset would make it easier for the models to learn to distinguish between the two groups of donors. The dataset received from the IBTS and used for the predictive tasks includes an overwhelming majority of non-donors in the periods used for predicting. In order to achieve better prediction, the groups of donor and non-donor existing in the dataset should be of similar size.
## APPENDIX A

### Importance of components:

<table>
<thead>
<tr>
<th>Component</th>
<th>Standard Deviation</th>
<th>Proportion of Variance</th>
<th>Cumulative Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCI</td>
<td>1.8473</td>
<td>0.2625</td>
<td>0.2625</td>
</tr>
<tr>
<td>PC2</td>
<td>1.4310</td>
<td>0.1575</td>
<td>0.4200</td>
</tr>
<tr>
<td>PC3</td>
<td>1.3765</td>
<td>0.1457</td>
<td>0.5657</td>
</tr>
<tr>
<td>PC4</td>
<td>1.2806</td>
<td>0.1262</td>
<td>0.6919</td>
</tr>
<tr>
<td>PC5</td>
<td>1.0489</td>
<td>0.0846</td>
<td>0.7764</td>
</tr>
<tr>
<td>PC6</td>
<td>0.9490</td>
<td>0.0694</td>
<td>0.8438</td>
</tr>
<tr>
<td>PC7</td>
<td>0.8033</td>
<td>0.0496</td>
<td>0.9132</td>
</tr>
<tr>
<td>PC8</td>
<td>0.6972</td>
<td>0.0373</td>
<td>0.9505</td>
</tr>
<tr>
<td>PC9</td>
<td>0.6558</td>
<td>0.0331</td>
<td>0.9836</td>
</tr>
<tr>
<td>PC10</td>
<td>0.5613</td>
<td>0.0242</td>
<td>0.9978</td>
</tr>
<tr>
<td>PC11</td>
<td>0.3469</td>
<td>0.0092</td>
<td>0.9990</td>
</tr>
</tbody>
</table>

### Group means:

<table>
<thead>
<tr>
<th></th>
<th>LD1</th>
<th>LD2</th>
<th>LD3</th>
</tr>
</thead>
<tbody>
<tr>
<td>deffno</td>
<td>1.903507e-02</td>
<td>-3.602768e-02</td>
<td>0.0559901360</td>
</tr>
<tr>
<td>agewhenstarted</td>
<td>1.453659e-01</td>
<td>8.303941e-01</td>
<td>0.1176924401</td>
</tr>
<tr>
<td>sinceFirstDon</td>
<td>-1.088042e-01</td>
<td>-4.534551e-01</td>
<td>1.3874015386</td>
</tr>
<tr>
<td>sinceLastDon</td>
<td>-1.089521e-01</td>
<td>4.530674e-01</td>
<td>1.3875473635</td>
</tr>
<tr>
<td>donIntervalAverage</td>
<td>-3.262061e-04</td>
<td>3.900973e-04</td>
<td>-0.000175712</td>
</tr>
<tr>
<td>defLenAverage</td>
<td>-3.205820e-04</td>
<td>-6.753909e-04</td>
<td>0.0001985713</td>
</tr>
<tr>
<td>lastDeferralLen</td>
<td>-3.269233e-04</td>
<td>7.681727e-05</td>
<td>-0.000202723</td>
</tr>
<tr>
<td>firstTwoDonationsInterval</td>
<td>2.012945e-05</td>
<td>1.751222e-04</td>
<td>0.0003843539</td>
</tr>
<tr>
<td>secondTwoDonationsInterval</td>
<td>-2.947551e-05</td>
<td>2.000168e-05</td>
<td>-0.0007342817</td>
</tr>
<tr>
<td>donatingPeriod</td>
<td>-1.082885e-01</td>
<td>4.531118e-01</td>
<td>1.3880715555</td>
</tr>
<tr>
<td>twoYearDonAverAge</td>
<td>3.095173e-02</td>
<td>-4.371428e-03</td>
<td>0.0082705013</td>
</tr>
<tr>
<td>gender</td>
<td>5.344517e-02</td>
<td>1.120068e-01</td>
<td>0.1969767104</td>
</tr>
<tr>
<td>agegrp</td>
<td>-2.349322e-01</td>
<td>-1.013395e+00</td>
<td>-0.1913296422</td>
</tr>
</tbody>
</table>

### Figure 34 PCI results summary for classification inputs and targets set 1

### Figure 35 Summary of LDA model generated for donor classification (targets set 1)
Group means:

<table>
<thead>
<tr>
<th></th>
<th>defno</th>
<th>ageWhenStarted</th>
<th>sinceFirstDon</th>
<th>sinceLastDon</th>
<th>donIntervalAverage</th>
<th>defLenAverage</th>
<th>lastDeferralLen</th>
<th>firstTwoDonationsInterval</th>
<th>lastTwoDonationsInterval</th>
<th>donatingPeriod</th>
<th>twoYearDonAverage</th>
<th>gender</th>
<th>ageGr</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>0.4728745</td>
<td>2.334413</td>
<td>3089.219</td>
<td>3089.219</td>
<td>0.0000</td>
<td>79.34737</td>
<td>68.89879</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>0.9264811</td>
<td>2.668088</td>
<td>4892.258</td>
<td>3907.648</td>
<td>0.682.7231</td>
<td>87.00000</td>
<td>78.72591</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>0.3507897</td>
<td>1.000000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>545.13300</td>
<td>542.26517</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>2.5116279</td>
<td>2.768303</td>
<td>4707.748</td>
<td>1348.380</td>
<td>0.0000</td>
<td>1348.380</td>
<td>103.65030</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Coefficients of linear discriminants:

<table>
<thead>
<tr>
<th></th>
<th>LD1</th>
<th>LD2</th>
<th>LD3</th>
</tr>
</thead>
<tbody>
<tr>
<td>defno</td>
<td>5.140176e-02</td>
<td>-0.063790182</td>
<td>0.034741527</td>
</tr>
<tr>
<td>ageWhenStarted</td>
<td>1.229888e+00</td>
<td>-0.8590656296</td>
<td>0.289670072</td>
</tr>
<tr>
<td>sinceFirstDon</td>
<td>1.600006e+00</td>
<td>1.8407257330</td>
<td>-1.1974302835</td>
</tr>
<tr>
<td>sinceLastDon</td>
<td>-1.689463e+00</td>
<td>-1.8410540164</td>
<td>1.1973684726</td>
</tr>
<tr>
<td>donIntervalAverage</td>
<td>2.711042e-04</td>
<td>0.0003131087</td>
<td>-0.001568677</td>
</tr>
<tr>
<td>defLenAverage</td>
<td>2.847069e-05</td>
<td>-0.0002829331</td>
<td>0.000190281</td>
</tr>
<tr>
<td>lastDeferralLen</td>
<td>-2.619340e-05</td>
<td>0.0002727075</td>
<td>-0.0001507049</td>
</tr>
<tr>
<td>firstTwoDonationsInterval</td>
<td>-1.266590e-04</td>
<td>-0.0002435483</td>
<td>-0.0002000498</td>
</tr>
<tr>
<td>lastTwoDonationsInterval</td>
<td>-1.006086e-05</td>
<td>-0.0001988372</td>
<td>-0.000722310</td>
</tr>
<tr>
<td>donatingPeriod</td>
<td>-1.689251e+00</td>
<td>-1.8404503061</td>
<td>1.1984333415</td>
</tr>
<tr>
<td>twoYearDonAverage</td>
<td>2.914765e+00</td>
<td>-0.0010475140</td>
<td>-0.002233351</td>
</tr>
<tr>
<td>gender</td>
<td>1.103487e+01</td>
<td>0.1081377218</td>
<td>-0.0492174067</td>
</tr>
<tr>
<td>ageGr</td>
<td>-1.422080e+00</td>
<td>1.1662926310</td>
<td>-0.3966688357</td>
</tr>
</tbody>
</table>

Figure 36: Summary of LDA model generated for donor classification (targets set 2)

---

Table 10: Donor classification results with CART (targets set 2)

<table>
<thead>
<tr>
<th>Prediction</th>
<th>F</th>
<th>L</th>
<th>N</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>96</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>L</td>
<td>0</td>
<td>968</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>N</td>
<td>0</td>
<td>0</td>
<td>41</td>
<td>2</td>
</tr>
<tr>
<td>R</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>134</td>
</tr>
</tbody>
</table>

---

Table 11: Donor classification results with C5.0 (targets set 2)

<table>
<thead>
<tr>
<th>Prediction</th>
<th>F</th>
<th>L</th>
<th>N</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>96</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>L</td>
<td>0</td>
<td>968</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>N</td>
<td>0</td>
<td>0</td>
<td>46</td>
<td>3</td>
</tr>
<tr>
<td>R</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>133</td>
</tr>
</tbody>
</table>

---

Table 12: Donor classification results with RF (targets set 2)

<table>
<thead>
<tr>
<th>Prediction</th>
<th>F</th>
<th>L</th>
<th>N</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>96</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>L</td>
<td>0</td>
<td>968</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>N</td>
<td>0</td>
<td>0</td>
<td>45</td>
<td>0</td>
</tr>
<tr>
<td>R</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>131</td>
</tr>
</tbody>
</table>
Table 13 Donor classification results with Feed-Forward network (NNET) (targets set 1)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Prediction</th>
<th>F</th>
<th>L</th>
<th>N</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>95</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>1</td>
<td>964</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>0</td>
<td>0</td>
<td>46</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>0</td>
<td>6</td>
<td>1</td>
<td>133</td>
<td></td>
</tr>
</tbody>
</table>

Table 14 Donor classification results with MLP and back propagation (targets set 1)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Prediction</th>
<th>F</th>
<th>L</th>
<th>N</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>96</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>0</td>
<td>967</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>0</td>
<td>1</td>
<td>46</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>130</td>
<td></td>
</tr>
</tbody>
</table>

Table 15 Donor classification results with MLP (RSNNS) and back propagation (targets set 1)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Prediction</th>
<th>F</th>
<th>L</th>
<th>N</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>367</td>
<td>0</td>
<td>4</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>9</td>
<td>555</td>
<td>13</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>1</td>
<td>0</td>
<td>144</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>143</td>
<td></td>
</tr>
</tbody>
</table>

Table 16 Donor classification with MLP (RSNNS) and back propagation with momentum (targets set 1)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Prediction</th>
<th>F</th>
<th>L</th>
<th>N</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>367</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>9</td>
<td>134</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>1</td>
<td>0</td>
<td>40</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>190</td>
<td></td>
</tr>
</tbody>
</table>

155
Table 17 Donor classification with PNN (targets set 1)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Prediction</th>
<th>F</th>
<th>L</th>
<th>N</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>96</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>0</td>
<td>952</td>
<td>4</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>0</td>
<td>0</td>
<td>39</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>0</td>
<td>15</td>
<td>2</td>
<td>110</td>
<td></td>
</tr>
</tbody>
</table>

Table 18 Donor classification results with CART (targets set 2)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Prediction</th>
<th>F</th>
<th>N</th>
<th>O</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>295</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>N</td>
<td>0</td>
<td>357</td>
<td>0</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>0</td>
<td>0</td>
<td>294</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>0</td>
<td>16</td>
<td></td>
<td></td>
<td>274</td>
</tr>
</tbody>
</table>

Table 19 Donor classification results with C5.0 (targets set 2)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Prediction</th>
<th>F</th>
<th>N</th>
<th>O</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>295</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>N</td>
<td>0</td>
<td>371</td>
<td>0</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>0</td>
<td>0</td>
<td>294</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>0</td>
<td>2</td>
<td></td>
<td>0</td>
<td>279</td>
</tr>
</tbody>
</table>

Table 20 Donor classification results with RF (targets set 2)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Prediction</th>
<th>F</th>
<th>N</th>
<th>O</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>295</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>N</td>
<td>0</td>
<td>368</td>
<td>0</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>0</td>
<td>0</td>
<td>294</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>0</td>
<td>5</td>
<td></td>
<td>0</td>
<td>281</td>
</tr>
<tr>
<td>Reference</td>
<td>F</td>
<td>N</td>
<td>O</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>295</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>0</td>
<td>369</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>0</td>
<td>0</td>
<td>294</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>285</td>
<td></td>
</tr>
</tbody>
</table>

*Table 21 Donor classification with Feed-Forward network (NNET) (targets set 2)*

<table>
<thead>
<tr>
<th>Reference</th>
<th>F</th>
<th>N</th>
<th>O</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>678</td>
<td>2</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>N</td>
<td>0</td>
<td>337</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>O</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>R</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>229</td>
</tr>
</tbody>
</table>

*Table 22 Donor classification with MLP (targets set 2)*

<table>
<thead>
<tr>
<th>Reference</th>
<th>F</th>
<th>N</th>
<th>O</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>284</td>
<td>49</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>N</td>
<td>23</td>
<td>252</td>
<td>0</td>
<td>34</td>
</tr>
<tr>
<td>O</td>
<td>0</td>
<td>0</td>
<td>305</td>
<td>0</td>
</tr>
<tr>
<td>R</td>
<td>1</td>
<td>34</td>
<td>0</td>
<td>265</td>
</tr>
</tbody>
</table>

*Table 23 Donor classification results with MLP (RSNNS) and back propagation (targets set 2)*

*Table 24 Donor classification with PNN (targets set 2)*
Figure 37 CART classification tree (targets set 1)
Figure 38 C5.0 classification tree (targets set 1)
Figure 13: CART classification tree (target set 2)
Figure 40 C5.0 classification tree (targets set 2)
Figure 41 Regression: donations number vs donor gender

Figure 42 Regression: donations number vs. length of intervals between first two donations
Figure 43 Regression: years of donating vs. number of donations

Figure 44 Regression: first donation age vs. months between first two donations
Figure 45 Regression: next donation vs. time since the last donation

Figure 46 Regression: average donation interval vs. donating period
Figure 47 Regression: gender vs. average donation interval

Figure 48 Regression: donor gender vs. donating period
Figure 49 Regression: average deferral length vs. donating period

Figure 50 Regression: average deferral length vs. number of donations
Figure 51 Regression: number of deferrals vs. number of donations

Figure 52 Regression: average deferral length vs. donor age
Figure 53 Regression; average deferral length vs. donor gender

Figure 54 Regression: first donation age vs. donating period

<table>
<thead>
<tr>
<th>Reference</th>
<th>Prediction</th>
<th>Y</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>42</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>19</td>
<td>1047</td>
<td></td>
</tr>
</tbody>
</table>

Table 25 Donation prediction with LDA (three-month)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Prediction</th>
<th>Y</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>53</td>
<td>95</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>8</td>
<td>969</td>
<td></td>
</tr>
</tbody>
</table>

Table 26 Donation prediction with QDA (three-month)
<table>
<thead>
<tr>
<th>Reference</th>
<th>Prediction</th>
<th>Y</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>32</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>29</td>
<td>1056</td>
<td></td>
</tr>
</tbody>
</table>

*Table 27 Donation prediction with CART (three-month)*

<table>
<thead>
<tr>
<th>Reference</th>
<th>Prediction</th>
<th>Y</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>38</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>23</td>
<td>1047</td>
<td></td>
</tr>
</tbody>
</table>

*Table 28 Donation prediction with MLP (three-month)*

<table>
<thead>
<tr>
<th>Reference</th>
<th>Prediction</th>
<th>Y</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>25</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>36</td>
<td>1060</td>
<td></td>
</tr>
</tbody>
</table>

*Table 29 Three-month donation prediction with Feed-Froward network (NNET)*

<table>
<thead>
<tr>
<th>Reference</th>
<th>Prediction</th>
<th>Y</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>36</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>25</td>
<td>1030</td>
<td></td>
</tr>
</tbody>
</table>

*Table 30 Three-month donation prediction with PNN*

<table>
<thead>
<tr>
<th>Reference</th>
<th>Prediction</th>
<th>Y</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>92</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>27</td>
<td>958</td>
<td></td>
</tr>
</tbody>
</table>

*Table 31 One-year donation prediction with CART*

<table>
<thead>
<tr>
<th>Reference</th>
<th>Prediction</th>
<th>Y</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>79</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>40</td>
<td>980</td>
<td></td>
</tr>
</tbody>
</table>

*Table 32 One-year donation prediction with LDA*
<table>
<thead>
<tr>
<th>Reference</th>
<th>Prediction</th>
<th>Y</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Y</td>
<td>79</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>40</td>
<td>983</td>
</tr>
</tbody>
</table>

Table 33 One-year donation prediction with MLP

<table>
<thead>
<tr>
<th>Reference</th>
<th>Prediction</th>
<th>Y</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Y</td>
<td>34</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>21</td>
<td>101</td>
</tr>
</tbody>
</table>

Table 34 One-year donation prediction with LDA (active donors only)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Prediction</th>
<th>Y</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Y</td>
<td>43</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>12</td>
<td>70</td>
</tr>
</tbody>
</table>

Table 35 Three-month donation prediction with QDA (active donors only)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Prediction</th>
<th>Y</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Y</td>
<td>35</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>20</td>
<td>96</td>
</tr>
</tbody>
</table>

Table 36 Donation prediction results with CART (active donors only)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Prediction</th>
<th>Y</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Y</td>
<td>32</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>23</td>
<td>90</td>
</tr>
</tbody>
</table>

Table 37 Donation prediction results with MLP (active donors only)
Figure 5.5: CART tree generated for predicting three-month donation.
Figure 56: Tree generated for predicting one-year donations (active donors only)
LIST OF REFERENCES


Ferguson, E., France, C. & Abraham, C., 2007. Improving blood donor recruitment and


180