IDENTIFYING FEATURES ELIGIBILITY FOR BLOOD DONORS' PREFERENCES USING ARTIFICIAL NEURAL NETWORKS PREDICTION PERFORMANCES

NOR SYUHADA BINTI CHE KHALID

MASTER OF SCIENCE IN INFORMATION AND COMMUNICATION TECHNOLOGY

2017

Faculty of Information and Communication Technology

IDENTIFYING FEATURES ELIGIBLITY FOR BLOOD DONORS' PREFERENCES USING ARTIFICIAL NEURAL NETWORKS PREDICTION PERFORMANCES

Nor Syuhada binti Che Khalid

Master of Science in Information and Communication Technology

2017

IDENTIFYING FEATURES ELIGIBILITY FOR BLOOD DONORS' PREFERENCES USING ARTIFICIAL NEURAL NETWORKS PREDICTION PERFORMANCES

NOR SYUHADA BINTI CHE KHALID

A thesis submitted
in fulfillment of the requirements for the degree of Master of Science in Information and Communication Technology

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2017

DECLARATION

I declare that this thesis entitled "Identifying Features Eligibility for Blood Donors' Preferences using Artificial Neural Networks Prediction Performances" is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : .....

Name : Nor Syuhada binti Che Khalid

Date : 7th August 2017

APPROVAL

I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in term of scope and quality for the award of Master of Science in Information and Communication Technology.

Signature	:	.....
Supervisor Name	:	Assoc. Prof. Dr. Burhanuddin bin Mohd Aboobaider
Date	:	7th August 2017

DEDICATION

Thanks for your kind supports.

To my noble mother, enduring father, gentle sister, adorable brother, quick-witted brother, and happy sister.

To my kind hearted supervisors, wise advisor, generous peers, and acquaintances.

To my strongest allies from North and heart of country, and my special heroes from Land Below the Wind.

This is just a beginning of another long academic journey…

ABSTRACT

Blood donation is an activity that has required people to contribute blood to help others during need for critical and near with fatal conditions such as organ transplant, post-partum haemorrhage, thalassemia, bowel operation, and orthopaedic surgery. Blood supplies extremely needed without fail. Therefore, blood donation service should retrieved useful information, especially to attract specific target groups of donors. However, this information required must be up-to-date, prepared systematically as prediction components, needed to scale down based on specifics target groups to make information extraction become better, and prediction algorithm has to adaptable with several sample sizes and features types because data may origin from different sources will have variety of datasets. Furthermore, blood donors' preferences are based on human opinions, which could cause different priority, conditions, and data retrieval method based on different communities, organisation, or places. As a solution, these research focuses are to collect new data on blood donors' preferences, construct Features Arrangement (FA) as dataset preparation for prediction model and criteria to distinguish between leading features (LF), features, and main leading features as main targets, and apply prediction algorithm which is artificial neural network. There is main dataset has collected from survey questionnaires. Features Arrangement has applied Pearson correlation between potential leading features and features as measurement to main leading features' criteria. This study has found out about main leading features which have influenced directly by less number of positive significant relationships with their attributes or features that have known as member features (MF). Therefore, decreasing number of positive significant relationships, regardless numbers of significant relationships, have yield better performance of blood donors' preferences predictor on main leading features as priority groups of respondents. FA has been implemented to select most and least associated features sets, from LFs and MFs. As summary, main blood donors' preference in Malaysia at 2015 is gender; meanwhile least preference is donation fear. Another recommended main preferences besides than gender as additional information are donating as religious purpose, donated more than once per year experience, health self-awareness and save another people, longer donation experience, overcome donation fear, high overall donation volume, tend to donate for family or acquaintances, donate frequently, up to date donation, information announcement medium such as social media, donation experience, and favourite donation center. Another least preferences recommended by FA are donation fear, favourite donation center, marriage status, up to date donation, interested to overcome donation fear, high overall donation volume, and donation motivation by celebrities. These findings of this study contribute as beneficial information to improve blood donation or healthcare service, as guide to collect and arrange data into prediction or another data mining problems, and extend another study for flexible algorithms with various datasets.

ABSTRAK

Derma darah adalah aktiviti yang dilakukan orang untuk menyumbang darah demi kepentingan orang lain yang memerlukan, kritikal dan hampir maut seperti pemindahan organ, pendarahan selepas bersalin, talasemia, pembedahan usus, dan pembedahan ortopedik. Bekalan darah sentiasa diperlukan. Jadi, perkhidmatan derma darah perlu memperoleh maklumat yang berguna, terutamanya untuk menarik kumpulan sasaran penderma tertentu. Walaubagaimanapun, maklumat diperlukan mestilah terkini, disediakan secara sistematik sebagai komponen ramalan, perlu dikurangkan dengan memfokuskan kumpulan tertentu untuk mengekstrak maklumat yang lebih baik, dan algoritma ramalan perlu sesuai dengan pelbagai saiz sampel dan jenis ciri-ciri kerana data mungkin berasal daripada bermacam jenis mengikut pelbagai set data. Tambahan pula, pemilihan penderma darah berdasarkan kemampuan manusia mengikut keutamaan yang berbeza berdasarkan keadaan dan kaedah mendapatkan semula data, seperti komuniti, organisasi, atau tempat. Oleh itu, objektif kajian adalah untuk mengumpul data baru kepada pilihan penderma darah, membina ciri-ciri mengetuai utama sebagai persediaan set data untuk model ramalan dan kriteria untuk membezakan antara ciri-ciri mengetuai (LF), ciri-ciri, dan ciri-ciri mengetuai utama sebagai sasaran utama, serta mempraktikkan algoritma klasifikasi yakni rangkaian neural buatan. Penyusunan Ciri-ciri (FA) telah mengaplikasikan korelasi Pearson antara kriteria mengetuai yang berpotensi dan ciri-ciri atau ciri-ciri ahli (MF) sebagai pengukuran ciri-ciri mengetuai utama. Kajian ini telah mengenal pasti tentang ciri-ciri mengetuai utama telah mempengaruhi secara langsung dengan kurangnya jumlah hubungan yang signifikan positif dengan sifat-sifat atau ciri-ciri mereka. Oleh itu, bilangan semakin berkurangan bagi hubungan yang signifikan positif, tanpa mengira bilangan hubungan yang signifikan, menghasilkan prestasi yang lebih baik. FA telah dilaksanakan untuk memilih ciri-ciri yang bersesuaian melalui beberapa set ciri-ciri yang paling berkaitan dan paling kurang berkaitan, yakni LF dan MF. Ringkasannya, pilihan penderma darah yang paling penting di Malaysia pada tahun 2015 ialah jantina, dan yang paling kurang penting ialah ketakutan pendermaan darah. Cadangan pilihan penderma darah yang lain-lain termasuklah menderma atas tujuan agama, menderma lebih daripada sekali dalam setahun, kesedaran kesihatan sendiri dan menyelamatkan orang lain, pengalaman menderma yang lebih lama, mengatasi ketakutan pendermaan darah, menderma darah dalam isipadu keseluruhan yang tinggi, cenderung untuk menderma demi keluarga atau kenalan, kerap menderma, menderma pada masa terkini, medium pengumuman maklumat seperti media sosial, pengalaman menderma, dan pusat menderma kegemaran. Manakala, pilihan yang paling kurang dicadangkan ialah ketakutan pendermaan darah, pusat menderma kegemaran, status perkahwinan, menderma pada masa terkini, berminat untuk mengatasi ketakutan pendermaan darah, menderma darah dalam isipadu keseluruhan yang tinggi, dan motivasi menderma oleh selebriti. Penemuan ini menyumbang sebagai maklumat yang berguna untuk memperbaiki perkhidmatan pendermaan darah atau penjagaan kesihatan, melalui panduan untuk mengumpul dan mengatur data dalam ramalan atau masalah perlombongan data lain, lalu mengembangkan kajian untuk menghasilkan algoritma mesra pelbagai set data.

ACKNOWLEDGEMENTS

First and foremost, I would like to take this opportunity to express my sincere acknowledgement to my supervisor Associate Professor Dr. Burhanuddin bin Mohd Aboobaider from the Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM) for his essential supervision, support and encouragement towards the completion of this thesis.

I would also like to express my greatest gratitude to Miss Nuzulha Khilwani binti Ibrahim, co-supervisor of this project for her advice and encouragements in this project. Special thanks to UTeM short term grant funding for the financial support throughout this project.

Particularly, I would also like to express my deepest gratitude to my strongest allies, postgraduate peers, lecturers, technicians, and staffs from Biocore Applied Research Group lab, and other graduate laboratories from Faculty of Information and Communication Technology, UTeM.

Special thanks to all my peers, my mother, father and siblings for their moral support in completing this degree. Lastly, thank you to everyone who had been to the crucial parts of realization of this project.

TABLE OF CONTENTS

	PAGE
DECLARATION	
APPROVAL	
DEDICATION	
ABSTRACT	**i**
ABSTRAK	**ii**
ACKNOWLEDGEMENTS	**iii**
TABLE OF CONTENTS	**iv**
LIST OF TABLES	**vi**
LIST OF FIGURES	**x**
LIST OF APPENDICES	**xiii**
LIST OF ABBREVIATIONS	**xiv**
LIST OF PUBLICATIONS	**xv**
CHAPTER	
1. INTRODUCTION	**1**
1.1 Research Background	1
1.2 Problem Statements	3
1.3 Research Objectives	4
1.4 Research Hypothesis	5
1.5 Research Contribution	6
1.6 Research Scopes	7
1.7 Research Significance	9
1.8 Thesis Organisation	11
2. LITERATURE REVIEW	**12**
2.1 Introduction	12
2.2 Blood Donors' Preferences	13
2.3 Features Eligibility	43
2.4 Neural Network Techniques	43
2.5 Summary	45
3. METHODOLOGY	**46**
3.1 Introduction	46
3.2 Research Design	46
3.3 Initial Stages	49
3.4 Implemetation Stages	51
3.5 Evaluation Stages	56
3.6 Summary	58
4. DATA COLLECTION AND ANALYSIS	**59**
4.1 Introduction	59
4.2 Data Collection	60
4.3 Data Preparation	60
4.4 Data Analysis	60
4.5 Summary	85

5. FEATURES ARRANGEMENT AS EMBEDDED BASED ON CORRELATION **86**
5.1 Introduction 86
5.2 Features Arrangement 87
5.3 Results 88
5.4 Summary 128

6. CONCLUSIONS AND FUTURE WORKS **129**
6.1 Introduction 129
6.2 Conclusions 129
6.3 Future Works 131
6.4 Summary 132

REFERENCES **133**
APPENDICES **144**

LIST OF TABLES

TABLE	TITLE	PAGE
2.1	Prediction Task Components Description	16
2.2	Prediction Algorithm	17
2.3	Past Study Summary on Blood Donors' Preferences Prediction	28
2.4	Overview on blood donors' preferences prediction researches	30
2.5	Concerned suggestion for improvement of blood donors' preferences prediction researches	33
2.6	Advantages and drawbacks of blood donors' preferences prediction algorithms applied on past studies	36
2.7	Blood Donors' Preferences and Their Definition	39
3.1	Research Design Tasks	48
4.1	Glossary of numerical substitutions in variables of blood donors' preferences in Malaysia, 2015 survey data	61
4.2	General descriptions of all variables (yellow highlight means that mean<median≤mode, noncolored means mean>median≥mode, red means median=mode)	63
4.3	Frequencies and percentages of variable values	66
4.4(a)	Percentage comparison of a variable values distribution in other variables for Q1-Q3	70

4.4(b)	Percentage comparison of a variable values distribution in other variables for Q4	71
4.4(c)	Percentage comparison of a variable values distribution in other variables for Q5-Q7	72
4.4(d)	Percentage comparison of a variable values distribution in other variables for Q8-Q10	73
4.4(e)	Percentage comparison of a variable values distribution in other variables for Q11A,Q11B	74
4.4(f)	Percentage comparison of a variable values distribution in other variables for Q11C,Q11D	75
4.4(g)	Percentage comparison of a variable values distribution in other variables for Q12A-Q12C	76
4.4(h)	Percentage comparison of a variable values distribution in other variables for Q12D-Q12F	77
4.4(i)	Percentage comparison of a variable values distribution in other variables for Q12G	78
4.4(j)	Percentage comparison of a variable values distribution in other variables for Q12J, Q12K	79
4.4(k)	Percentage comparison of a variable values distribution in other variables for Q12L,Q12M	80
4.5	Correlation matrix of 27 variables	81
4.6	Counts of associated variables for each variable	82
4.7	Counts of positive associated variables for each variable	84
5.1	Step 1: Assignation of Initial LFs and MFs (similar with correlation matrix match up in Chapter 4)	89

5.2	Step 2-1: Frequency of Associations, n_{FA} between LF and chosen MFs	91
5.3	Comparison n_{FA} between LFs	92
5.4	Comparison n_{FA} between experiments	95
5.5	Step 2-2: Positive association ratio between LFs and chosen MFs (Red is for maximum, blue is for minimum)	97
5.6	Step 3-1: Epochs of paired LFs and MFs	100
5.7	Comparison epochs between LFs	102
5.8	Comparison of epochs between experiments	104
5.9	Step 3-2: Training performance (CE) of paired LFs and MFs, smaller is better	106
5.10	Comparison of training performance (CE) between LFs	108
5.11	Comparison of training performance (CE) between experiments	109
5.12	Step 3-3: Validating Performance (CE) of Paired LFs and MFs, smaller is better	111
5.13	Comparison of validating performance (CE) between LFs	113
5.14	Comparison of validating performance (CE) between experiments	115
5.15	Step 3-4: Testing performance (CE) of paired LFs and MFs, smaller is better	117
5.16	Comparison of testing performance (CE) between LFs	119
5.17	Comparison of testing performances maximum and minimum between LFs	120
5.18	Step 4 & 5: Performance (CE) ranking for maximum and minimum cutter for training, validating, and testing, using feature arrangement (FA)	123
5.19	Comparison descriptive between MaxCut and MinCut	125

5.20 Occurrence of simplified leading features from MaxCut by 30 experiments 126
5.21 Step 5-2: Distributions of simplified leading features from MinCut by 30 experiments 127
5.22 Final LFs by MaxCut and MinCut during 30 experiment runs based on ranking 128
6.1 Achievements of research stages based on research objectives 130
6.2 Strengths and weaknesses of current research with Suggestions for future works 132

LIST OF FIGURES

FIGURE	TITLE	PAGE
1.1	More similar patterns on a class	2
1.2	Features relationships with target class, where r is Pearson correlation	3
1.3	Research Scopes	8
2.1	Prediction task illustration (Tan et al., 2004)	15
2.2	Portrayal of blood donors' preferences on prediction	20
2.3	Summary of possible blood donors' preferences prediction from various representations and class types	25
2.4	Simple flowchart of various blood donors' prediction processes from past studies	27
2.5	Blood Donors Data Representation Structure	42
3.1	Research design idea	47
3.2	Arrangement of Dataset 1 based on research objectives	59
3.3	Features Arrangement	51
3.4	Feedforward prediction	54
3.5	Data distribution among training, validation, and testing phases during prediction	55
3.6	Most performed leading features evaluation	56
4.1	Glossary of numerical substitutions in variables of blood donors' preferences in Malaysia, 2015 survey data	63

4.2	Comparison of mean, median, and mode for each variable	64
4.3	Comparison of coefficient of variances between variables	65
4.4(a)	Percentage distrubutions of variables values for Q1-Q11B	68
4.4(b)	Percentage distrubutions of variables values fror Q11C-Q12M	69
4.5	Comparison of associated variables count and positive associated variables count	85
5.1	Steps Flow of FA	87
5.2	Comparison of n_{FA} maximum and minimum between LFs	93
5.3	Comparison of n_{FA} mode, average, and standard deviation between LFs	94
5.4	Comparison of n_{FA} coefficient of variance between LFs – Low variances	94
5.5	Comparison of n_{FA} maximum and minimum between experiments	95
5.6	Comparison of n_{FA} mode, average, and standard deviation between experiments	96
5.7	Comparison of n_{FA} coefficient of variance between experiments – Low variances	96
5.8	Comparison of epochs maximum and minimum between LFs	102
5.9	Comparison of epochs mode, average, and standard deviation between LFs	103
5.10	Comparison of epochs coefficient of variance between LFs – Low variances	103
5.11	Comparison of epochs maximum and minimum between experiments	105
5.12	Comparison of epochs average and mode between experiments	105
5.13	Comparison of epochs coefficient of variance between experiments – Majority LFs are in high variances	105
5.14	Comparison of training performances maximum and minimum between LFs	109

5.15	Comparison of training performances coefficient of variance between LFs – Majority LFs are in low variances	110
5.16	Comparison of training performances maximum and minimum between experiments	110
5.17	Comparison of training performances coefficient of variance between experiments – Majority LFs are in high variances	110
5.18	Comparison of validating performances maximum and minimum between LFs	113
5.19	Comparison of validating performances coefficient of variance between LFs – Majority LFs are in low variances	114
5.20	Comparison of validating performances maximum and minimum between experiments	116
5.21	Comparison of validating performances coefficient of variance between experiments –LFs are in high variances	116
5.22	Comparison of testing performances coefficient of variance between LFs –LFs are in high variances	119
5.23	Comparison of testing performances maximum and minimum between LFs	120
5.24	Comparison of testing performances maximum and minimum between experiments	121
5.25	Comparison of testing performances coefficient of variance between experiments –LFs are in high variances	121
5.26	Comparison of MinCut and MaxCut frequencies, n_{LF}	124

LIST OF APPENDICES

APPENDIX	TITLE	PAGE
A	Approval Letter To Conduct Research For Finishing Study	145
B	Survey Form Of Blood Donors' Preferences In Malaysia 2015	147
C	Glossary For Survey Data Of Blood Donors' Preferences In Malaysia 2015	152
D	Table Of Numerical Substitutions For 27 Variables Of Blood Donors' Preferences In Malaysia 2015	154
E	Descriptive Statistic Of Blood Donors' Preferences In Malaysia 2015	171
F	Correlation Matrix Of Blood Donors' Preferences In Malaysia 2015	179
G	Code Of Blood Donors' Preferences In Malaysia 2015 Prediction Using Scg Functions	191

LIST OF ABBREVIATIONS

FA	Features Arrangements
SCG	Scalar Conjugate Gradients
ANN	Artificial Neural Network
CV	Coefficient Of Variance
SD	Standard Deviation
CE	Cross Entropy
LF	Leading Feature
MF	Member Feature
MaxCut	Simplification of LFs from Cutting by Maximum Values
MinCut	Simplification of LFs from Cutting by Minimum Values

LIST OF PUBLICATIONS

Nor Syuhada Che Khalid, Burhanuddin Mohd Aboobaider, Ahmad, A., and M.K.A, G., 2013. Classification Techniques in Blood Donors Sector – A Survey. *e-Proceeding of Software Engineering Postgraduates Workshop (SEPoW) Theme : Innovative Software Engineering for Creative*. Melaka: UTeM, pp.114–118.

Nor Syuhada Che Khalid, M.A. Burhanuddin, Nuzulha Khilwani Ibrahim, Zahriah Sahri, M.K.A.. Ghani., 2017. Data Mining Variables and Features Selection for Malaysia Blood Donors' Preferences using Correlation Technique. *Journal of Engineering and Applied Sciences*, vol. 12, pp. 3638-3643. **(Published) Scopus**

CHAPTER 1

INTRODUCTION

1.1 Research Background

Social studies and marketing analysis are introduced mainly on commercial factors such as beauty, price, award, altruism, privilege, advertise, empathy and service. In other hand, risk, fear, problem, difficulties, hurdle, and dissatisfaction, would become a huge drawback for potential targets.

As main concern for blood donation and collection service provider, blood bank and blood transfusion centre always have to ensure that blood collection always continue to operate and collect blood without fail. Blood stocks are highly required by hospitals for critical or near death situations such as accident, giving birth, surgeries that may need blood transfusion, for example in case of post-partum haemorrhage, thalassemia, bowel operation, and orthopaedic surgery. Blood is necessary to save life. Proper information on attraction points and drawback of blood collection and donation service, blood donors and non-donors feedback for their background, donation habit, experience, fear, readiness to overcome donation fear, and incentives.

Furthermore, this information would call for preparation, extraction, and interpretation stages. Existence of various data collection such as questionnaires, medical records, interview, and donation records should lead to distinct preparation and extraction. General purpose of questionnaires is to get latest information from respondents to verify and increase past information. Questionnaires can be prepared as some variables before automate analysis by

computer.

Based on background and information attained from respondents, data will be appointed into possible sets of dependent and independent variables based on random suggestion because they do not have necessary fixed roles to influence another or not directly unlike variables on scientific experiment such as engineering, chemistry, and medicine. Every human behaviour, thinking, and choice is not that simple to distinguish from one to another. Nevertheless, certain similarities from a specific target group may exhibit resemblance in opinion or preferences. These sets of similarities or patterns from different groups should be examined and compared based on their potentials, whether they are significant to influence other variables more or less, as imagined in Figure 1.1.

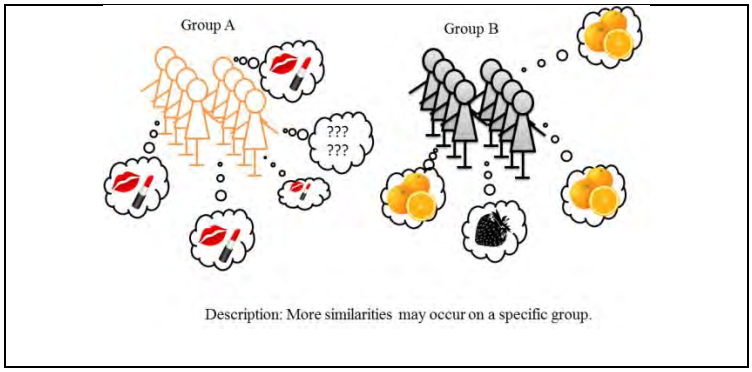

Figure 1.1: More similar patterns on a class

Therefore, similarities on blood donors preferences may reveal more after comparing many variables altogether after prediction. Similarities or patterns influences are significant and certain direction by specific classes will yield better analysis results. Relationship between target classes and features are visualised like Figure 1.2.

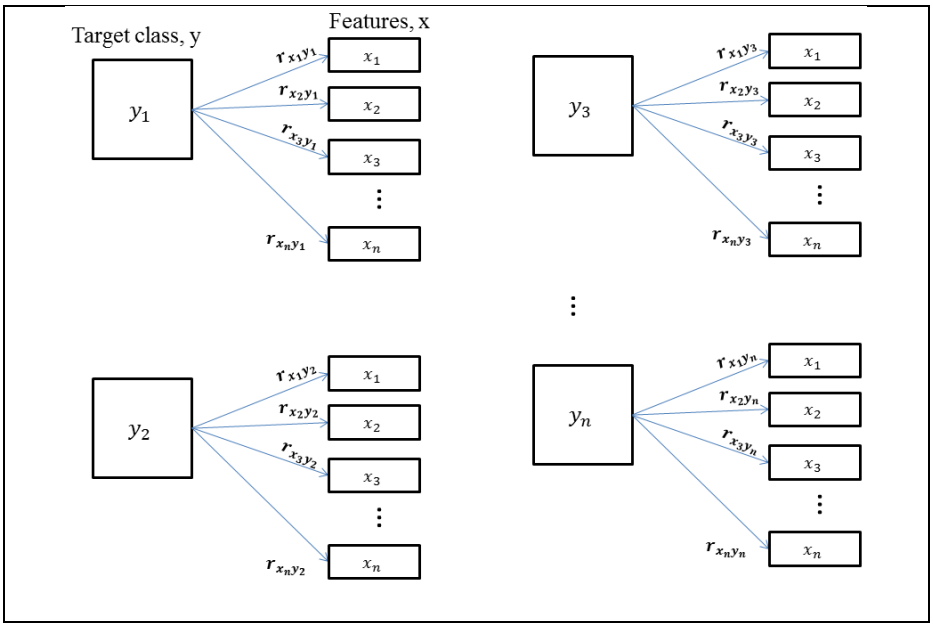

Figure 1.2: Features relationships with target class, where r is Pearson correlation

As conclusion, features relationship with target classes would produce better prediction. Blood donors' preferences prediction objective is to predict blood donors' habit and motivation to know further information that may encourage them to donate blood. This research would like to express some preparation methods on past and current datasets to prepare for prediction purpose.

1.2 Problem Statements

Regarding to past studies, data preparation for blood donors' preferences prediction has not mentioned clearly as guide or tips for another prediction research. Besides that, old datasets from past studies were available. However, latest blood donors' preferences are useful to update information and verify relevancy of past information too.

Additionally, information were limited not until to prepare blood donors dataset to become prediction model when useful information usually come from various datasets. Some