A HYBRID ENSEMBLE BOOSTING MODEL FOR ENHANCED BLOOD DONOR RETENTION

NAHASHON KIARIE

A Thesis Submitted in Partial Fulfilment of Requirements for the conferment of the Degree of Master of Science in Information Technology of Meru University of Science and Technology

2025

DECLARATION

This thesis is my original work and to the best of my knowledge it has not been presented for the award of a degree in any other institution.

Signature.....Date: .....

NAHASHON KIARIE

CT402/201267/20

This thesis has been presented with our authority as University Supervisors.

..... Date: .....

Dr. Mary Mwadulo, PhD

Meru University of Science and Technology, Kenya

..... Date: .....

Dr. Amos Chege Kirongo, PhD.

Meru University of Science and Technology, Kenya

DEDICATION

This work is dedicated to my dear wife Josphine for her unwavering encouragement and support throughout this academic journey. To my two wonderful children, Lexie and Ethan who have been a constant source of inspiration and motivation. This work is for you, I hope that this work inspires you to pursue your dreams with passion and perseverance.

ACKNOWLEDGEMENT

First and foremost, I would like to thank the Almighty God for providing me with good health, wisdom, and strength throughout the journey to be able to undertake this study.

I am deeply grateful to my Supervisors, Dr. Amos Kirongo Chege and Dr. Mary Mwadulo for their academic counsel and guidance, insightful feedback, and unreserved support throughout this work.

To my fellow classmates and students at the School of Computing and Informatics who assisted me in any way. It is through your constructive discussions, technical support and moral encouragement that this thesis was able to be realized.

Thank you all for your support and contributions.

ABSTRACT

Blood donor retention is critical for maintaining a stable and reliable blood supply, yet predicting donor retention remains a complex challenge. Previous attempts to develop blood donor retention models relied on single algorithms and achieved relatively low prediction accuracy limiting their practical application for donor retention. The Light Gradient Boosting Machine (Light GBM) algorithm employs leaf-wise growth strategy, excels in loss reduction and hence improves accuracy. However, this may lead to potential overfitting, on the other hand, the Extreme Gradient Boosting(XGBoost) algorithm incorporates a robust mechanism for combating overfitting, such as the regularization parameter, column sampling, and weight reduction on new trees but employs a level-wise growth strategy, which is sometimes computationally intensive. This study developed a hybrid ensemble gradient boosting model based on XGBoost and Light GBM. The ensemble model leverages on the high accuracy of Light GBM while mitigating overfitting through and the overfitting prevention strategies of XGBoost. The data was obtained from the Kenya blood banks with 5000 records and nine features. The base models were trained in parallel, a weighted ensemble model was created by assigning weights to the respective prediction results of each model, the ensemble model was then evaluated and the accuracy compared with the accuracy achieved by the individual algorithms. Bayesian hyperparameter optimization was implemented on the base learners in order to find the best combination of hyperparameters and further improve the performance of the model. The ensemble model achieved a performance accuracy of 99.00% and F1 score of 99.00%. This study enables blood agencies to accurately predict blood donor retention, thereby reducing the need for constant donor recruitment efforts and saving both time and costs. Additionally, it will provide insights for targeted retention strategies, ensuring a steady blood supply, ultimately saving lives and improving healthcare systems.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS AND ACRONYMS

ANN	Artificial Neural Network
AUC	Area Under the Curve
CRISP-DM	Cross Industry Standard Process for Data Mining
DM	Data Mining
DNN	Deep neural network
DT	Decision Tree
FN	False Negatives
FP	False Positives
KBTTS	Kenya Blood Transfusion and Transplant Service
KNBTS	Kenya National Blood Transfusion Services
KNBTS	Kenya National Blood Transfusion Services
KNN	K-nearest neighbor
Light GBM	Light Gradient Boosting Machines
LSTM	Long short-term memory
MAE	Mean Absolute Error
NB	Naïve Bayes
NN	Neural network
RF	Random Forest
RMSE	Root Mean Squared Error
TN	True Negatives
TP	True positives
TPD	Theory of Planned Behavior
TT	Training Time
UK	United Kingdom
VNRBD	Voluntary Non-Remunerated Blood Donors
WHO	World health Organization
XGBoost	Extreme Gradient Bosting

DEFINITION OF OPERATIONAL TERMS

Blood- the red fluid that circulates in the arteries and veins of human beings and other vertebrate animals and transports oxygen to the tissues.

Blood donation- Voluntary act of giving whole blood or blood components to a person in need.

Blood donor retention - ability of blood donation centers to keep donors active in their blood donation and prevent them from lapsing.

Machine Learning Using statistical models and algorithms to analyze and draw conclusions from data patterns to construct computer systems that learn and adapt without explicit instructions

Ensemble learning is a meta-approach to machine learning that combines the predictions of multiple models improve predictive performance

Gradient Boosting is a robust boosting algorithm that converts numerous weak learners into strong learners by employing gradient descent to train each new model in order to minimize the loss function of the previously trained model.

LightGBM (Light Gradient Boosting Machine) is an open-source, fast, and efficient gradient boosting framework that is designed for machine learning tasks like classification and regression.

XGBoost (eXtreme Gradient Boosting) is a popular and powerful open-source machine learning library that implements the gradient boosting framework. It is known for its high performance, flexibility, and efficiency in handling large datasets

Model optimization refers to the process of improving a machine learning model's performance in order to produce a model that generalizes well to unseen data, resulting in higher accuracy.

Hyperparameter tuning is the process of optimizing a machine learning model in order to find the best set of hyperparameters that maximizes model performance.

CHAPTER ONE
INTRODUCTION

1.1. Overview

This chapter serves to introduce the study. It explains the concept of blood and blood donation, and describes blood donation supply chain and the blood donation strategies in Kenya and the world at large. Additionally, it introduces machine learning and its applications in healthcare and the ensemble gradient boosting models. Further, it outlines the problem statement, the objectives of the study, the justification and significance of the study, and the scope and limitations of the study.

1.2. Background of the Study

1.2.1. Blood Donations

Blood is an essential component of the human body responsible for transportation of nutrients, oxygen, hormones and other elements necessary for proper functioning of the body. A healthy blood supply is essential for effective body performance. Red blood cells, white blood cells, plasma, and platelets comprise the blood. While white blood cells fight infection, red blood cells carry oxygen(Okuthe et al., 2022). The platelets aid in blood clotting, and plasma is the liquid component of blood. Combined, these elements of the blood maintain the body health and any abnormality in any one of them can result in life-threatening medical issues(Medvedev, 2021). Although there are more than 30 major blood group systems, the ABO and the RhD blood group systems are the most essential when discussing blood donation and blood transfusion. Donated blood is used for a variety of purposes, including transfusions for patients with blood disorders, surgery, and trauma cases(Mouncif & Bellabdaoui, 2022). Blood donation plays a vital role in saving lives and improving patient outcomes.

The process of blood donation involves donor recruitment, blood collection, testing, distribution and transfusion of blood and its components to meet the needs of the patients (Delaney et al., 2022). For one to be eligible for blood donation, they must meet certain eligibility requirements. Generally, potential donors must be between the ages of 17 and 65, and in good health. They must weigh at least 50 kilograms and should not be having any health issues(WHO, 2018). Although the activities in the blood donation cycle are similar, different countries and organizations have adopted different ways of managing their activities from recruitment, donation, inventory, and transfusion.

1.2.2. Global Blood Donation Management Strategies

The World Health Organization (WHO) advocates for the centralization of blood collection, testing, storage, and distribution activities within national frameworks, emphasizing the importance of efficient and integrated blood supply networks. Additionally, the national blood system should be regulated by a national blood policy and laws that assure the safety and quality of blood and blood-related products. Many countries have established national bodies that are mandated to oversee the implementation of WHO recommendation(McElfresh et al., 2021).

The Food and Drug Administration (FDA) regulates blood safety in the United States. Additionally, the American Red Cross and America's Blood Centres manage the national blood supply in the United States(Baron et al., 2020). The Chinese Red Cross Society (CRCS) is the main organization responsible for blood donation in China. The CRCS has established a comprehensive blood donation system, which includes donor screening, donor education and monitoring, and quality assurance. Additionally, the Chinese government has implemented various initiatives to promote safe and efficient blood management(Weidmann et al., 2022).

In Africa various state-owned agencies have been established to manage and coordinate blood donation, storage and distribution. In Zambia, Zambian Integrated Blood Donor Database Management System has been established to track, record and manage blood donor data for blood donor retention purposes(Loua et al., 2021).

World health organization (WHO) recommends that the minimum country blood stocks at any time should be at least 1% of the population, However, many countries in Africa are unable to meet even half of this requirement, resulting in critical blood shortages. (WHO, 2023). Half of the world's maternal deaths from severe bleeding occur in sub-Saharan Africa. Approximately 65% of these deaths are reported to occur during the postpartum period, 80-90 percent of these maternal deaths are due to bleeding complications which are compounded by lack of adequate blood. The low blood supply has also been worsened by the prevalence of infectious diseases in Africa, including Covid 19, HIV, malaria, and tuberculosis (Murtagh & Katulamu, 2021). These illnesses frequently result in the death of red blood cells, responsible for carrying oxygen throughout the body. Up to 65 percent of blood transfusions in many developing nations are administered to children under the age of five(WHO, 2023).

1.2.3. Blood donations in Kenya

The Kenya Tissue and Transplant Authority (KTTA) is the government agency that is mandated to collect, test, process, store, and distribute blood to all hospitals in Kenya(Kanagasabai et al., 2021). To ensure proper management and coordination of blood donation activities. The agency has established six Regional Blood Transfusion Centers (RBTCs), as well as 14 satellite centres countrywide(Moore et al., 2020). Any willing blood donors can walk to any of these blood centres or any government hospital-based blood banks and donate blood voluntarily(BM et al., 2022).

The KTTA in collaboration with the Ministry of health has established a Kenya blood bank management system named DamuKe. The system was established in 2022 and tracks the process of blood donation, from donation to transfusion. The system integrates various functions including blood donor registration, inventory management, as well as tracking of the donated blood. The system aims to achieve efficiency in the blood donation supply chain and help blood banks to enhance decision making and strategic planning for blood donations drives. The agency conducts planned blood donation camps in various towns across the country, as well as in secondary and tertiary institutions. The World Bank reports that in every 10 minutes, seven people need a blood transfusion(World Bank, 2022). However, the country suffers from acute blood shortage as only 16% of the blood needed in the country is collected(World Bank, 2022). One of the primary causes of the blood supply shortage is insufficient donor retention strategies.

1.2.4. The Challenge of Donor Retention

Donor retention is the process of retaining blood donors to give blood regularly(Delaney et al., 2022). However, blood donation centres face challenges due to declining blood donor retention rates, hence reducing their ability to provide sufficient blood(Dei-Adomakoh et al., 2021). One of the major factors that contribute to blood shortages is the high rate of blood donor attrition, with many first-time blood donors failing to return for subsequent blood donations. Recruiting new donors is often expensive and time-consuming (Ou-Yang et al., 2017). Regular donors are very valuable since they provide a reliable source of safe blood, and reduce the cost and time to constantly keep recruiting new blood donors.

Donor retention is also crucial because the returning blood donors already have a history of safe blood donations and hence reduce the risk of transfusion transmitted infections(Delaney et al., 2022). They are also already familiar with the blood donation

process and therefore reduce the time and resources needed for training and registration. Additionally, they also become advocates for blood donation, encouraging many other people to become blood donors, and hence contributing significantly to the stability and safety of the blood supply system(Mouncif & Bellabdaoui, 2022). Enhancing donor retention is very crucial in ensuring a stable supply of blood, reducing costs associated with recruitment, and improving the overall healthcare outcomes.

1.2.5. Machine learning application in healthcare

Machine learning has become a transformative force in healthcare enabling transformations and unlocking limitless possibilities (Panesar, 2019). One of the key benefits of machine learning in healthcare is its ability to analyse huge amounts of complex data, including electronic health records, medical images, genetic information, real-time patient monitoring data among other medical data(Golas et al., 2018). By uncovering hidden patterns and insights within the data, machine learning algorithms can be able to assist healthcare professionals to make more informed, data-driven decisions.

Machine learning has been applied in the medical realm to perform diagnostics such and to analyse medical images, such as X-rays, MRIs, and CT scans(Nagassou et al., 2023). They can analyse images with a high level of precision, often surpassing human radiologists (Xiao et al., 2018). By analysing historical patient data, ML models can be able to predict the likelihood of disease outbreaks, patient admissions and readmissions, and any adverse drug reactions to the patients. Additionally, ML models are used to optimize scheduling, manage medical supply chains as well as streamline administrative and medical tasks, thereby reducing costs and improving the overall efficiency of healthcare delivery (Marade et al., 2019).

ML models can be used in the pre-donation stage of the blood donation supply chain to analyse donor data and identify patterns associated with return donors, predict the risk of

adverse reactions and complications, and identify donors who may need special attention during donation at screening stage.

In the donation process, the models can be used in real time monitoring to analyse the blood donors' vital signs and reactions during the donation. This can help in detecting any signs of discomfort, fatigue or any other adverse reaction and allow prompt interventions during the donation process(Suessner et al., 2022).

In the post donation stage, ML models can be used to analyse donor feedback and understand their levels of satisfaction, as well as factors that may influence their return for blood donation. Such information is critical in developing personalized strategies for donor retention, and hence improve long term engagement with donors(Shashikala et al., 2019). By utilizing the power of technology, the machine learning predictive models can help to identify key factors that influence blood donor retention and hence enable personalized donor retention strategies tailored towards individual donor's needs and preferences(Kauten et al., 2022)

1.2.6. Ensemble Boosting Models

Ensemble models are powerful machine learning models that combine multiple weak learners to create a strong predictive model(Malek et al., 2022). They leverage on the strengths of multiple algorithms to enhance the predictive performance of the model and provide more accurate and robust predictions as compared to single models(Guo et al., 2022). The ensemble boosting models sequentially train models focusing on correcting errors in the previous models. This iterative process ensures that the models are able to achieve high accuracies and robustness. The adaptability and accuracy of these models make them invaluable tools, particularly in predicting blood donor retention where high accuracy is crucial. The popular gradient boosting ensemble models include, XGBoost, Light GBM, AdaBost Catboost(Bentéjac et al., 2021). Despite their huge potential and

improved prediction accuracies the ensemble and specifically gradient bosting ensembles have not found much application in blood donor retention, most existing studies have relied on single algorithms and achieved relatively low prediction accuracy limiting their practical application for blood donor retention.

1.3. Statement of the Problem

Blood donation plays a vital role in healthcare, saving countless lives annually. However, the high turnover rate of blood donors poses a significant challenge for blood banks, hindering their ability to maintain a consistent and stable blood supply.

Previous attempts to develop blood donor retention models, such as those by Shashikala et al. (2019), Marade et al. (2019), Pabreja and Bhasin (2021), and Selvaraj et al. (2022), relied on single algorithms and achieved relatively low prediction accuracies limiting their practical application for blood donor retention. Other studies by Cloutier et al. (2021) and Saad Alkahtani and Jilani (2019) utilized random forest as a single classifier. While Random Forests offer robustness and ease of interpretation, they struggle with imbalanced datasets, a common characteristic of blood donation data, which typically has more non-donors than donors. This imbalance can lead to biased predictions that favor the majority class, thereby reducing the model's effectiveness(Fife & D'Onofrio, 2022).

On the other hand, gradient boosting algorithms like XGBoost and Light GBM offer promising alternatives. Light GBM which is known for its fast training speeds, employs leaf-wise growth strategy which reduces memory usage and enhances efficiency(Nagassou et al., 2023). It also adopts gradient-based one side sampling and exclusive feature bundling and excels in loss reduction, thereby achieving high accuracies. However, this may lead to potential overfitting. Conversely, XGBoost has built-in mechanisms for handling imbalanced datasets, incorporates a robust mechanism for combating overfitting, such as the regularization parameter, column sampling, and weight reduction on new trees.

Nonetheless, since it employs a level-wise growth strategy, it can sometimes be computationally intensive(Liang et al., 2019).

This study developed and evaluated a hybrid ensemble model based on XGBoost and LightGBM for predicting blood donor retention. The study aimed at leveraging on the complementary strengths of XGBoost and LightGBM to enhance accuracy, improve robustness, handle imbalanced data, as well as reduce overfitting. Bayesian hyperparameter optimization was implemented in order to find the best combination of hyperparameters and further improve the performance.

1.4. Objectives

1.4.1. General Objective

The study general objective was to develop, optimize and validate a novel hybrid ensemble gradient boosting model for enhanced blood donor retention.

1.4.2. Specific Objectives

The specific objectives of the study were to:

i. Conduct a baseline survey on the existing blood donor retention models
ii. Develop a hybrid ensemble gradient boosting model for blood donor retention.
iii. Optimize the ensemble gradient boosting model by fine-tuning hyperparameters.
iv. Validate the performance of the developed ensemble gradient boosting model for blood donor retention.

1.5. Research questions

i. What are the existing blood donor retention models and how are they effective in blood donor retention?

ii. How can a hybrid ensemble gradient boosting model for enhanced blood donor retention be developed?

iii. How can the developed hybrid ensemble blood donor retention model be optimized?

iv. How can the performance of the developed ensemble gradient boosting model be validated?

1.6. Justification for the Study

This study is justified by the urgent need to address the issue of blood donor retention. Existing approaches are labor-intensive and time-consuming, and they usually fail to provide precise and timely insights. Furthermore, weaknesses in current donor management systems require intelligent-based solutions. Advances in machine learning, specifically ensemble gradient boosting, offer an exciting opportunity to develop predictive models capable of analyzing large amounts of data(Z. Zhang et al., 2019).

This study is important as it has the potential to improve blood supply stability, optimize resource allocation, reduce costs, improve blood donor experience and enable personalized communication strategies. By accurately predicting donor retention, blood banks and organizations can take targeted action to improve retention rates, ensuring a steady and sustainable blood supply to those in need ultimately saving lives.

1.7. Significance of Study

This study will benefit the blood donation institutions in several ways, they can utilize this model to increase their efficiency by identifying donors who are at risk of not donating again, they can concentrate their efforts on retaining them. This can contribute to a

substantial increase in the number of available blood donors who can donate blood and help to save lives.

Additionally, the model will assist blood centers in developing focused retention strategies that are more effective in maintaining donors over a long time by determining factors that influence donor behavior(Shehu et al., 2024). This can eliminate the need for continued donor recruitment, which is both costly and time-consuming.

The application of ensemble gradient boosting in this context demonstrates application of advanced machine learning techniques to the solving of real-world problems. This contributes to the field of ICT by demonstrating how modern algorithms can be employed to address complex problems in healthcare, moreover the methodology employed in this work has the potential to be utilized in the development of machine learning models for the prediction of various health-related outcomes.

The study anticipates to make clear the importance of Artificial Intelligence and machine learning in blood donor management and predictions to policymakers, and help them to understand and appreciate the significance of AI. It is anticipated that appreciation of machine-learning in blood donor management will result in advocacy for changes in the regulatory environment in order to enable a blood donation cycle that is more technologically focused. Additionally, improvement of blood donor retention rate leads to a subsequent reduction in the amount of time needed to recruit new blood donors and results in cost saving.

In addition, the research will provide researchers and academics with a point of reference on application of ICT in healthcare, and provide knowledge for future studies in blood donor predictions and other fields related to AI and ML, and the development of improved retention and churn prediction models. Moreover, the research will also facilitate identification of research gaps for future studies.

1.8. Scope of the study

The main focus of the work was the development of a predictive blood donor retention model based on ensemble gradient boosting. This includes data collection, feature selection, model training, model optimization and validation using machine learning techniques, particularly ensemble gradient boosting.

The study utilizes data from blood banks in Kenya. While the model is developed and validated using this regional data, the methodologies and findings are intended to be adaptable to other geographical contexts in Africa and in the world.

1.9. Limitations of the Study

While the study of predictive blood donor retention rate models indicates greater prospects, it also had certain limitations that were considered. First, the accuracy and effectiveness of predictive models is highly dependent on data availability and quality. Limited or incomplete data can lead to biased or inaccurate predictions and affect the reliability of the model. The researcher was able to obtain data directly from the blood banks in Kenya. The data undergone extensive and thorough pre-processing to ensure data quality, consistency and completeness.

Secondly collection and utilization of blood donor information for research purposes could raise ethical and privacy concerns. The researcher strictly adhered to ethical guidelines and ensured that all the donor information was anonymized and handled in accordance with data protection laws. All personal identifiable donor information was omitted from the extracted data to protect the identities of the donors. Additionally, workbooks and databases remained password-protected as an additional measure of data security.

1.10. Assumptions of the Study

This study makes the following assumptions:

First, the study assumed that the data collected is a true representative of the actual blood donor population in Kenya. Secondly, the study assumed that the features selected for the development of the model are sufficient and relevant to capture the factors that influence blood donor retention.

Thirdly, the study assumed that blood donor behavior remains relatively consistent over time and that the historical data used in the study accurately reflects future behavior and do not significantly change over time.

Finally, the study assumed that the findings and insights which are derived from the Kenyan context can be generalized to other regions across the world with related socio-demographic characteristics as Kenya.

1.11. Thesis Organization

The first chapter of this study presents the introduction which includes the background of the study, the problem statement, the objectives of the study, significance of the study, the scope of the study, limitations and the assumptions of the study.

The second chapter presents a thorough review of the literature related to blood donor retention from the blood and blood donation to the machine learning models, the ensemble models, datasets used, methods of validation, optimization methods and finally a summary of the reviewed work is presented in a table. The chapter also provides the theoretical and conceptual framework.

The third chapter is the research methodology and describes the research design employed, the research process, data collection methods, the experimental set-up and the models adopted for the experiment.

The fourth chapter describes how the experiment was carried out, the model implementations and the discussion of the results at each step.

The fifth chapter presents the summary of the results, the findings, conclusions and recommendations.

CHAPTER TWO
LITERATURE REVIEW

2.1. Overview

This chapter provides an overview of the literature relevant to the research questions outlined in chapter one. The aim of the review is to identify factors that influence blood donation and highlight the gaps which the study aims to address. This section explains the concept of blood and blood donation, machine learning models used by various researchers to predict blood donations, with their varying degrees of accuracy, and discusses the theoretical framework underpinning the research. Further, it presents, in form of conceptual framework the relationships between the various factors that influence blood donation, and summarizes the reviewed literature in a table

2.2. The Concept of Blood and Blood Donations

2.2.1. Importance of Blood

Blood is an essential component of the human body which transports nutrients, oxygen, hormones, and other elements necessary for the body's cells to function. A healthy blood supply is essential for the body's effective functioning. Blood is comprised of Red blood cells, white blood cells, plasma, and platelets. White blood cells fight infection while red blood cells carry oxygen(Okuthe et al., 2022). Platelets aid in blood clotting and plasma is the liquid component of blood. Together, these elements maintain the body's health; any abnormality can result in life-threatening medical issues.

The transportation of oxygen to the body's cells is one of the blood's most crucial roles. Haemoglobin is a protein found in red blood cells that bind to oxygen molecules and carries them around the body. Without oxygen, the body would not be able to function correctly because cells wouldn't be able to produce energy. Hormones, which are chemicals released by the endocrine system to control various bodily processes, are also

carried by the blood. Hormones regulate growth, development, reproduction, metabolism, and reproduction. White blood cells, which are carried by blood, aid in the fight against infection(Medvedev, 2021). These cells can identify and combat viruses and bacteria.

The classification of blood groups is established based on the presence of a protein, known as an antigen, located on the surface of red blood cells. The ABO system is characterized by the presence of A and B antigens, while the RhD system is distinguished by the presence of the D antigen. An individual's blood group is determined by the combination of their ABO group and their RhD group(Ingle & Patil, 2022).

Blood donation is the voluntary and altruistic act of offering one's blood in order to assist others in dire need of blood transfusions or medical treatment. Donated blood is used for a variety of purposes, including transfusions for patients with blood disorders, surgery, and trauma cases(Mouncif & Bellabdaoui, 2022). The importance of blood donation cannot be overstated as it plays a vital role in saving lives and improving patient outcomes. The process involves the collection, testing, and distribution of blood and its components to meet the needs of patients suffering from acute and chronic illnesses, accidents, and surgeries(Delaney et al., 2022).

2.2.2. Blood Donation Supply Chain

For one to be eligible to donate blood, individuals must meet certain eligibility requirements. Generally, potential donors must be between the ages of 17 and 65 and in good health. They must weigh at least 50 kilograms and not have any current health issues(WHO, 2018).

After meeting the eligibility requirements, donors undergo a blood donation process. Generally, the process begins with registration, where the donor provides their contact information and answers medical questions. The donor will then be asked to provide a small sample of their blood, which undergoes a health screening test for infectious

diseases(Nzoka & Anande, 2014). If the test results are satisfactory, the donor will be asked to sit in a comfortable chair and have their arms prepared for donation. A sterile needle will be inserted into their arm and the blood collected. The entire process usually takes about 8-10 minutes. After the donation, the donor should rest for several minutes and drink plenty of fluids. They may experience some minor side effects such as dizziness, light-headedness or fatigue.

After blood is donated it undergoes rigorous testing for HIV, hepatitis, syphilis, and other diseases. Additionally, the blood is typed and cross-matched to determine recipient compatibility. After testing, blood is then stored in refrigerated blood storage facilities to maintain its freshness(Moore et al., 2020). Different blood components have different storage requirements, so they are kept in different containers in order to preserve them. Blood is requested by hospitals and healthcare facilities when needed for patients undergoing surgery and other medical treatments, it is delivered under regulated conditions to these facilities, ensuring its safety and quality. Medical personnel deliver transfusions to patients in need after which the patients are monitored to ensure they are stable and feedback is given in case of any anomaly.

Although the activities in the blood donation cycle are similar, many countries in the world have established national bodies and organizations that coordinate and manage blood donation activities from recruitment, donation, inventory, and transfusion.

2.2.3. Blood Donations in Kenya

Blood donation in Kenya is coordinated through various institutions. Which includes government institutions such as the ministry of health and the Kenya Tissue and Transplants authority. The agencies also work with non-governmental organizations such as the Kenya Red Cross, universities, secondary schools and corporate entities to facilitate donor recruitment, coordinate blood drives and outreach programs. The world bank reports

that seven people need a blood transfusion every 10 minutes, yet the country suffers from acute blood shortages as less than 30% of the blood needed in the country is collected(World Bank, 2022). One of the primary reasons for low blood collections is over-reliance on school-age donors to maintain nationwide blood donation volumes.

Data from the Kenya Tissue and transplants authority shows that there is usually a drastic drop in blood donations during the school holidays in April, August and December of a year. The largest donor group is the group of 17 to 20 year olds with 57.7%, followed by 21 to 25 year olds (27.8%). This information clearly demonstrates the need to work towards expanding the blood donor population in the 25-65 age group and ensure they are retained to continue donating blood through their entire donation lifetime.

(i) Kenya Tissue and Transplants Authority (KTTA)

The Kenya Tissue and Transplant Authority(KTTA) is the government agency that is mandated to collect, test, process, store and distribute blood to all health facilities in Kenya. The agency formerly known as "The Kenya national Blood Transfusion services" was established in the year 2000 within the ministry of health. The organization was renamed through a gazette notice number 129 of August 2022 and has been given expanded mandate to assure safety and quality blood transfusion, tissue and human organ transplant services through oversight, regulation and support(Kanagasabai et al., 2021).

The agency has established six Regional Blood Transfusion Centres(RBTCs), namely Nairobi, Embu, Nakuru, Eldoret, Kisumu and Mombasa and 14 satellite centers located in Machakos, Kisii, Voi, Meru, Naivasha, Kakamega, Kericho, Nyeri, Garissa, Malindi, Thika, Lodwar, Bungoma and Kitale (Moore et al., 2020). The RBTCs are strategically located across the country and serve as centralized blood banks. The reason for regional blood repositories is to ensure that dozens of neighboring local hospitals are always

stocked with adequate amount of blood to meet the transfusion needs especially in the event of an emergency.

Any willing blood donors can walk to any of these RBTCs and donate blood. They can also choose to walk to any of the government hospital-based blood banks and donate blood voluntarily. Most blood donations in Kenya are collected during blood drives held in communities, schools and planned open air blood donation camps(BM et al., 2022).

(ii) The Kenya Blood Management System

Before a blood donor donates blood they are required to fill a blood donor questionnaire. The information contained in the blood donor questionnaire include: Donation site information, donor personal and demographic information, questions on the medical condition and history of the blood donor. The information collected on this questionnaire is keyed in to the Kenya blood bank management system.

The Kenya DamuKe is a blood bank management system established by the Ministry of Health in Kenya in collaboration with the Ministry of Information and Communication Technology(ICT). It is a track and trace system for donated blood from the donor to the recipient. The DamuKe system was established in 2022 and comprises of a mobile and web-based application that enables Kenyan blood banks to carry out their operations more effectively(Ministry of health Kenya, 2023). It encompasses all aspects of blood banking, such as blood donation, inventory management, and transfusion management(Kenya Blood Transfusion and Transplant Service, 2023).

The Kenya blood bank management system operates by providing a central database for all Kenyan blood banks to store their information. Blood banks can then monitor their blood supply, track the inventory, and coordinate blood transfusions using that data, which is accessible from anywhere in the country. The system is still in its early stages of adoption and development and most modules are still not yet being used(Ministry of health

Kenya, 2023). Although the information being collected by this blood donor questionnaire is quite vast. There has been little or no use of this data to identify trends, patterns, and insights that can influence blood donation and retention.

2.2.4. Challenges in Blood Donor Retention

Blood donor retention is the ability of blood centers to keep donors active and prevent them from lapses in their blood donation. It is a more cost-effective way of retaining active blood donors as opposed to recruiting new donors, this strategy is essential for ensuring the continuity and safety of the blood supply(Van Dongen, 2015). Blood has a short life span and cannot be manufactured in laboratories, its demand is very high and therefore its supply should always remain constant. Emergency situations such as accidents, medical operations and diseases necessitate regular blood transfusion(Dei-Adomakoh et al., 2021). The demand for blood and blood products is constantly increasing due to population growth, advancements in medical procedures, and rising incidence of diseases such as cancer and chronic conditions that require regular transfusions. However, this increasing demand is not being met adequately, resulting in blood shortages and their subsequent impact on healthcare systems worldwide(World Health Organization, 2017).

The age profile of blood donors indicates that in low- and middle-income countries, more young donors donate blood that in high-income countries(WHO, 2023). In Kenya for example most of the blood is collected in secondary schools and colleges. This over reliance on school-going students results in a significant shortage of blood donations when schools are closed and since most of these students are not retained to continue donating blood even after they are through with high school and college. It is therefore critical to put in measures to ensure that these donors are retained and are able to donate blood frequently through their lifetime. Returning blood donors are better as they ensure a regular blood supply and reduce the hassle of finding first-time donors. Patients' lack of timely

access to safe blood and the large gap between blood demand and supply are issues facing many nations, particularly those in the developing world(Kanagasabai et al., 2021). Insufficient recruitment and retention of blood donors is one of the primary causes of the blood supply shortage.

2.2.5. Factors That Influence Blood Donor Retention

There are numerous studies that have explored factors that motivate people to donate blood, and become regular donors after the first donation. Majority of these studies have found altruism, that is the desire to save lives, to be the main factor that influence blood donors to donate blood. In Africa, family or replacement donors who donate blood in response to a patient or family member who is in need of blood account to over 60% of donations(Asamoah-Akuoko et al., 2017).

Age has been found to be an important demographic factor affecting the return of blood donors. Studies have shown that older people are more likely to become regular donors and have a higher donor retention rate(Shama et al., 2022). Older donors often have a greater sense of social responsibility and are more interested in making a positive impact in their community by donating blood. Additionally, as people age, they may become more aware of their own health needs and the importance of maintaining a stable blood supply. The time lapse between the last blood donation made by a donor and return has also been found to be a very key factor that influence the return of a blood donor to donate blood. Blood donors with a shorter interval between their last two donations have higher chances of becoming regular donors as compared to blood donors with longer intervals. Moreover, a short blood donation interval indicates a higher donation frequency in previous years(Shehu et al., 2024).

The number of previous donations made by a donor have also been found to greatly impact donor return. Blood donors with a higher number of previous donation have been found to

return for more donations, this is because these donors are already familiar with the donation process and experience they also act as ambassadors for blood donations often bringing on board other blood donors(Van Dongen, 2015).

The longer the duration since the first donation may also signify a donor commitment and an established habit of donating blood. This has been associated with a higher likelihood of the blood donor returning and continuing to donate(Cloutier et al., 2021).

Gender has also been found to influence blood donation. Research by (Kanagasabai et al., 2021) shows that men account for approximately 70% of blood donors in various countries while women account for 30%, women are more likely to be differed from donating blood due to different factors such as low levels of haemoglobin or a body weight below 50 kg, pregnancy and breastfeeding, however women are more likely to be regular donors as compared to men (Shama et al., 2022). This could be due to various factors, including higher levels of awareness about blood donation, a greater sense of empathy and altruism. Socioeconomic status may play a role in blood donor return and retention(Asamoah-Akuoko et al., 2017). Those with higher socioeconomic status may have better access to information and resources, making them more likely to donate blood on a regular basis. Additionally, those with higher incomes may be more willing and able to take time off from work or engage in philanthropic activities, including blood donations.

Studies indicate that people with a higher level of education tend to be more aware of the importance of donating blood and are more likely to donate regularly (Ou-Yang et al., 2017). Education can lead to better knowledge of blood needs, better understanding of eligibility criteria, and awareness of the impact of donation on patients' lives. Ethnicity and cultural background can also impact blood donor return and retention(Asamoah-Akuoko et al., 2017). Different ethnic and cultural groups may have different levels of awareness, beliefs and attitudes towards blood donation.

2.2.6. Strategies Used for Blood Donor Retention

Blood donation organizations often face the challenge of retaining existing blood donors to ensure they continue to donate and provide a stable and reliable supply. Several methods have been used try and improve blood donor retention.

Donor appreciation and recognition has been used to recognize blood donors for their contributions to blood donations. Appreciation methods may include thank you notes, publicly acknowledging the donor's dedication and efforts as well as offering small incentives and gifts(Ferguson, 2021). This may make the donor to feel valued and encourage the donor to continue donating blood. However, this method is often costly and may not be sustainable especially developing countries where blood donation organizations are resource constrained, moreover using incentives to motivate donors may often raise ethical issues as WHO guidelines require that blood donations should be voluntary.

Improving blood donation experience has also been use to improve the donor retention. Enhancing donor activities such as reducing the waiting time, providing clean and comfortable facilities, friendly handling of donors is used to make the donation process more pleasant, improve overall satisfaction of the blood donor and encourage them to return for future blood donations. However, these methods may also come with additional costs and may also be difficulty to tailor the experiences to individual donor preferences(Shehu et al., 2024).

Targeted awareness and educational campaigns programs are also used to retain blood donors. They are used to educate blood donors on the donation process and eliminate the misconceptions and fears about the blood donation process. These activities can foster a sense of altruism and motivate donors to repeat donations(Ferguson, 2021). However, it may require a lot of time, effort and investment to keep the campaigns effective and may

not immediately translate to increased donations, additionally they can be difficult to measure and may also vary depending on the targeted audience.

Technological solutions such as mobile applications and online portals are also used to improve blood donor retention. These solutions can be used to provide easy access to donation information and donors personal information, to book and schedule donations as well as provide detailed donation history for the donor(Nagassou et al., 2023).They can also facilitate efficient communication with the donor such as feedback and reminders moreover they can be intergraded with other blood bank systems to streamline the donation process.

Advanced technological solutions such as predictive models can be used to predict and analyze the blood donor history, analyze demographic information and donation patterns so as to identify the donors likely to return for donations(Kauten et al., 2022). Additionally they can be used to segment donors based on their preferences and behavior and enable tailored experiences which can improve the conversion rates of new donors to become regular blood donors(Shashikala et al., 2019).

2.3. Machine Learning

Machine learning is the application of artificial intelligence (AI) which enables systems to automatically learn from experience and improve without being explicitly programmed(Fei et al., 2023). It is primarily utilized for classification, forecasting, and prediction applications. Machine learning functions by learning data and generating prediction rules by recognizing patterns in the data, as opposed to following a predefined and hard-coded algorithm. The recursive nature of machine learning allows it to adapt and evolve in response to new data changes(Panesar, 2019). Machine learning can be divided into un-supervised machine learning and supervised machine learning.

2.3.1. Un-Supervised Machine Learning

Un-supervised machine learning deal with unlabeled data, the algorithms aim to discover patterns, structure, or relationships within the data (Bi et al., 2019). This method is particularly useful when the objective is to examine and obtain insights from data(Lestandy et al., 2020). There are two main methods of un-supervised learning which namely clustering and association.

(a) Clustering

Clustering involves dividing the data into groups or clusters of similar items and discovering underlying groups of data with similar behavior. Examples of clustering include K-Means clustering which divides data into k clusters and each data point is assigned to the nearest cluster center(Kauten et al., 2022). Hierarchical clustering works by building a hierarchy of clusters. The clusters are often formed based on the proximity of data points. In the context of blood donation clustering has been used to group donors into clusters such as donors and non-donors, first time donors or returning donors, young or old donors and discover the pattern in each group.

(b) Association

Association involves discovering interesting relationships between the data. The models use association rules to find relevant associations amongst the data(Xiao et al., 2018). In blood donations associations can be used to predict blood donors who are more likely to return for donations based on their past behavior and interactions, blood banks can leverage on these insights to develop more effective blood donor retention strategies

2.3.2. Supervised Machine Learning

Supervised learning is the process of training a machine using labeled data, which indicates that some data is already labeled with the correct response(Jiang et al., 2020). The machine is then provided with a new set of examples (data) in order for the supervised learning

algorithm to analyze the training data and generate the correct result from labeled data(Panesar, 2019). The primary benefit of supervised learning is that it permits machines to generalize patterns from known data and apply them to new, unobserved data. Supervised machine learning algorithms include, linear and logistics regression, decision trees, support vector machines, neural networks, naïve bayes and the ensemble learning models such as bagging and boosting algorithms.

(a) Linear and Logistic Regression

Linear regression is a simple and widely used prediction algorithm (Marade et al., 2019; Pabreja & Bhasin, 2021; Selvaraj et al., 2018; Wu et al., 2022). The algorithm was used in blood donation predictions in the study by Saad Alkahtani and Jilani (2019), achieving an accuracy of 93%, and in the study by Pabreja and Bhasin (2021), where it achieved an accuracy of 68%. Linear regression seeks to establish a linear relationship between the input characteristics and the target variable. The algorithm estimates the coefficients of the linear equation that best fit the data, minimizing the difference between the predicted and the actual values (Prion & Haerling, 2020).

During prediction, the algorithm applies these coefficients to the input features to generate the predicted result. Linear regression is simple and easy to interpret and provides high efficiency with large datasets. However since it assumes a linear relationship it may struggle with complex and nonlinear patterns(Leipnitz et al., 2018). Logistic Regression assumes a linear relationship between the predictors and the outcome, making understanding the relationship between variables easier. However, the linear assumption may limit its ability to capture complex patterns and interactions between variables.

(b) Decision Trees

Decision trees are a versatile algorithm that uses a hierarchical structure of nodes to make predictions(Wu et al., 2022). The tree splits the data based on characteristics and creates decision rules that lead to the final prediction. Each internal node represents a decision based on a feature and each leaf node represents the predicted outcome. The tree is built by recursively choosing the best feature to partition the data, optimizing criteria such as information gain or Gini impunity(Zulfikar et al., 2018).

The CART (Classification and Regression Trees) algorithm is also another a frequently employed decision tree in prediction of blood donors. The CART algorithm adopts a binary tree configuration and conducts recursive data partitioning using a selected attribute and threshold(Rivera-Lopez et al., 2022). The objective is to reduce a particular measure, such as Gini impurity or mean squared error while performing the partitioning procedure. Like C4.5, CART provides interpretability, enabling researchers to comprehend and elucidate the decision-making mechanism of the decision trees produced.

The algorithm exhibits versatility in accommodating categorical and numerical variables, rendering it suitable for accommodating the heterogeneous attributes typically encountered in blood donor retention datasets. Nevertheless, the CART algorithm is prone to overfitting, particularly when the decision tree's complexity rises and assimilates irrelevant data from the training set(Panesar, 2019). Furthermore, like the C4.5 algorithms, the CART algorithm must be revised to handle outliers and noisy data. Slight deviations or disturbances within the training dataset have the potential to result in distinct tree structures and forecasts, thereby influencing the dependability and consistency of the model(X. Li et al., 2022).

(c) Support Vector Machines (SVM)

Support Vector Machine (SVM) is a powerful algorithm used for both classification and regression tasks, and it has demonstrated higher accuracy in blood donor predictions. In the study by Wu et al. (2022), SVM achieved an accuracy of 95%, compared to 93% in the study by Saad Alkahtani and Jilani (2019) and 78.4% in the study by Selvaraj et al. (2022). SVM maps the input features into a high-dimensional feature space and constructs the optimal hyperplane by maximizing the distance between support vectors (data points closest to the hyperplane)(Suessner et al., 2022). During prediction, SVM determines the class or regression value based on the position of the new data point relative to the hyperplane.

SVM are highly effective in high dimensional data and works well with both linear and nonlinear relationships, it is less prone to overfitting than other algorithms, making it a robust choice. However, SVM's performance is sensitive to the choice of kernel function and hyper parameters, requiring careful tuning(Lukmanto et al., 2019). The complexity of the decision boundaries generated by SVM can also hinder interpretability.

(d) Artificial Neural Networks

Neural Networks, particularly deep learning models, have gained significant attention due to their ability to learn complex patterns from large amounts of data(Pabreja & Bhasin, 2021). Neural networks consist of interconnected nodes (neurons) organized in layers. Each neuron applies a mathematical transformation to the input, and the network learns to adjust the weights between neurons during training to minimize the prediction error. The network's architecture, including the number of layers and neurons, determines its complexity and capacity to capture intricate relationships(Ingle & Patil, 2022).

Neural network’s ability to apprehend intricate patterns and correlations within the data enables a more exhaustive comprehension of the determinants that impact donor retention.

Neural Networks exhibit a high degree of adaptability to diverse data types, successfully incorporating numerical and categorical variables frequently encountered in donor datasets(Awwalu et al., 2019). In the study by Shashikala et al. (2019), the decision tree algorithm achieved an accuracy of 85%, compared to 70% in the study by Pabreja and Bhasin (2021). Even with their strengths, Neural Networks exhibit certain limitations that warrant attention. A significant obstacle in this context is the inclination towards overfitting, especially when the models possess a considerable number of parameters compared to the amount of data at hand.

Overfitting is a phenomenon that arises when a model acquires an excessive level of familiarity with the training data, resulting in suboptimal generalization performance when applied to novel data. Furthermore, Neural Networks are frequently regarded as opaque models owing to their intricate internal architectures and high dimensionality(Dong et al., 2023). This phenomenon results in decreased interpretability, rendering the comprehension and explication of the model's decision-making process more arduous.

(e) Naive Bayes

Naive Bayes is a probabilistic algorithm based on Bayes' theorem. It assumes that the features are conditionally independent given the target variable. Naive Bayes calculates the probabilities of each class given the input features and selects the class with the highest probability as the prediction. It estimates the probabilities using the training data, utilizing the prior probabilities and likelihoods of each feature(Zulfikar et al., 2018).

Despite the "naive" assumption of independence, Naive Bayes often performs well in practice, especially with text classification tasks(Zulfikar et al., 2018). It produces fast training and prediction times especially with high dimensional data. However, it assumes independence between features and may nor not perform well when the independence assumption is violated. It is also less effective with numerical or continuous variables. The

algorithm recorded an accuracy of 97% in the study by Kewat and Sharma (2018), which is the highest accuracy achieved in all the reviewed studies on blood donation predictions. Additionally, the algorithm achieved an accuracy of 81% in the study by Zulfikar et al. (2018).

2.4. Ensemble Machine Learning Techniques

Ensemble models are predictive models that combine the predictions of multiple individual models to create a more powerful and accurate predictive model(Fei et al., 2023). This concept gained more popularity in the late 20th century when researchers realized that by combining several individual models, the overall predictive accuracy of the predictive model can be significantly improved. Using the collective intelligence of the ensemble, it attempts to mitigate errors or biases that may exist in individual models(Malek et al., 2022).

There are two types of ensemble learning: serial integration and parallel integration. Bagging is a parallel integration technique, whereas boosting is a serial integration technique, both involve sequentially generating a set of base learners and using the residual of the current model to construct the learner(H. Liu, 2021). Because of their outstanding prediction performance and flexibility, ensemble models have enjoyed widespread adoption in machine learning. They are utilized in a variety of fields, including banking, healthcare, and natural language processing, where high accuracy and interpretability are critical(Liang et al., 2019). Bagging entails training numerous models using bootstrapped subsets of data and averaging their predictions to reduce variance and boost stability while boosting, focuses on boosting model accuracy by giving greater weights to misclassified cases, resulting in a strong prediction model from a collection of weaker models(Pham & Ho, 2021).

2.4.1. Bagging

Bagging, or Bootstrap Aggregating derives its name from the fact that it incorporates Bootstrapping and Aggregation into a single ensemble model. Given a data sample, multiple bootstrap subsamples are drawn. Each bootstrapped subsample is utilized to develop a decision tree. Following the formation of each subsample Decision Tree, an algorithm is used to aggregate over all decision trees to produce the most accurate predictor(Guo et al., 2022).

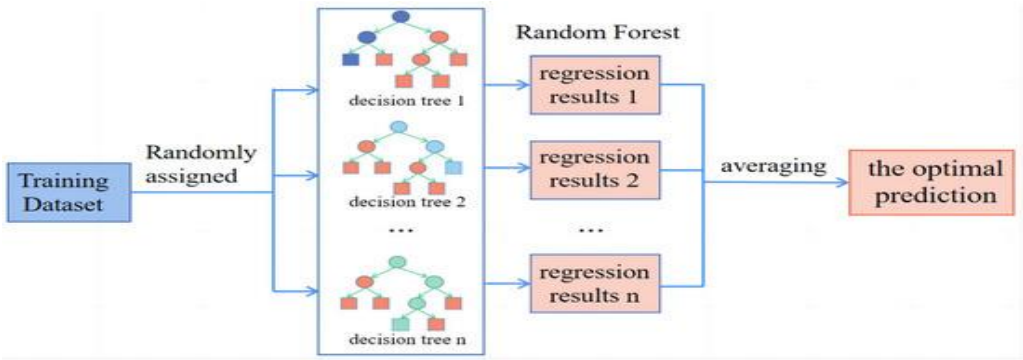
(a) Random Forest

Random Forest is an ensemble algorithm that combines multiple decision trees to make predictions(Wu et al., 2022). Random forest has shown a good performance accuracy in blood donations predictions achieving an accuracy of 93% in the study (Saad Alkahtani & Jilani, 2019). It also achieved an accuracy of 91% in (Cloutier et al., 2021) and 78% in (Selvaraj et al., 2022). RF has the advantage of handling many input variables. This makes it suitable for capturing complex patterns in the data. However, the interpretability of the model may be limited, making it challenging to understand the underlying factors influencing donor retention(Cloutier et al., 2021). During prediction, each tree makes a prediction independently, and the final score is determined by aggregating the predictions (e.g., by majority voting or averaging)(Marade et al., 2019).

Random Forest reduces overfitting and improves generalization compared to a single decision tree, they are also robust and good in handling high dimension data and feature importance estimation. A potential drawback of Random Forest is its computational cost, especially with large datasets which may be expensive especially when large datasets are involved and may over fit if the number of trees is not optimized(Saad Alkahtani & Jilani, 2019). Figure 2.1. below shows random forest modelling flowchart.

Figure 2. 1

RF Modelling Flowchart



2.4.2. Boosting

Boosting is a machine learning technique that has the ability to turn weak learners into strong classifiers(Freund & Schapire, 1999). It is a type of ensemble meta-algorithm used to reduce distortion and variance. The primary concept of boosting is to repeatedly apply the underlying learning algorithm to adjusted iterations of the input data. Boosting techniques are a class of algorithms that utilize the input data to train a weak learner. The weak learner's predictions are computed, and then misclassified training examples are selected. The subsequent weak learner is trained using an adjusted training set that incorporates the incorrectly classified instances from the previous phase of training(B. Zhang et al., 2019). The iterative learning process is repeated until a predetermined quantity of basis learners is attained, and the base learners are assigned weights.

(a) Gradient Boosting

Gradient Boosting is an ensemble algorithm similar to Random Forest, it combines multiple weak prediction models (typically decision trees) in a sequential manner to create a strong predictive model(Bentéjac et al., 2021). Gradient Boosting starts with an initial model and then builds subsequent models to correct the errors of the previous models. Each subsequent model focuses on the samples that were poorly predicted by the previous

models. During prediction, the final outcome is determined by combining the predictions of all the models, giving higher weight to the more accurate models.

Gradient Boosting has demonstrated high accuracy, it can handle various variable types and automatically handles missing data, making it robust in handling real-world datasets(Rodríguez-Tomàs et al., 2022). Gradient boosting models are able to handle complex interactions between variables to produce high predictive accuracy. However they may be prone to overfitting if not properly regularized and can be computationally expensive(McElfresh et al., 2021).

(b) Adaptive Boosting (AdaBoost)

AdaBoost, also referred to as reinforcement learning or promotion method(Mahesh et al., 2022). It is an ensemble method that produces a strong classifier from a collection of weak classifiers. It is based on its ease of use, performance in real-time target detection, feature selection, and fast training time. Each sample prediction is assigned a weight, and the incorrect prediction is identified and given to the subsequent base learner with a high weight, while the weight of the rightly predicted sample is lowered. The procedure continues until the algorithm identifies a classification model that correctly classifies these samples. The final strong learner is not identified until the predetermined maximum number of iterations is reached or the predetermined error rate is low enough(Chengsheng et al., 2017). AdaBoost is sensitive to anomalous samples, and abnormal samples may be given a higher weight during iteration, which may influence the accuracy of the prediction.

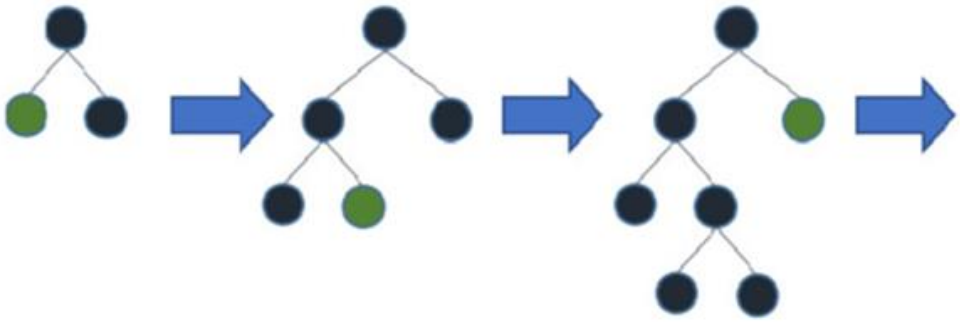
(c) Light Gradient-Boosting Machine (LightGBM)

A team from Microsoft devised the light gradient boosting machine (LightGBM) in April 2017 to reduce implementation time(Guolin Ke et al., 2017). In LightGBM decision trees are grown leaf-wise, instead of checking all prior leaves on every new leaf, all attributes are sorted and classified into bins. This kind of implementation is described as

Histogram. The majority of samples with a minor gradient are excluded using gradient-based one-sided sampling (GOSS), and only the samples that remain are used for calculating the information gain in order to make up for the loss in data volume and maintain accuracy. Exclusive Feature Bundling (EFB) links numerous mutually exclusive features to a given feature in order to reduce dimensionality. LightGBM has a number of advantages, including improved accuracy, faster training speed, the ability to handle large-scale data, and GPU learning support(Guo et al., 2022). Figure 2. 2. shows the light gradient boosting leaf-wise growth

Figure 2. 2

Light Gradient Boosting Leaf-Wise Growth



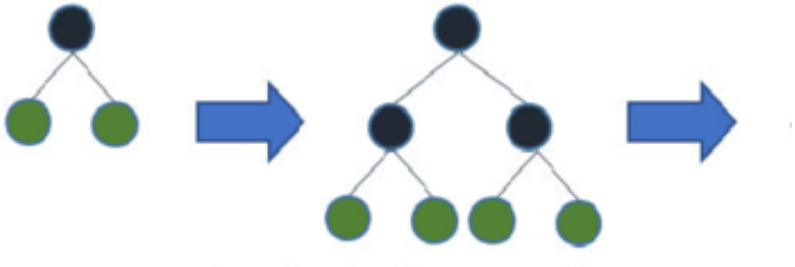
(d) Extreme Gradient-Boosting (XGBoost)

XGBoost is a very scalable decision tree ensemble built on gradient boosting(Chen & Guestrin, 2016). Although XGBoost has several optimizations and is different from light gradient-boosting, it shares the same fundamental concept as GBDT. XGBoost constructs an additive expansion of the objective function by minimizing a loss function, much to gradient boosting. A different loss function is utilized to regulate the complexity of the trees. XGBoost concentrates on lowering computational complexity which is the most time-consuming step in decision tree construction algorithms. The complexity reduction ensures that models are trained faster and require less storage space.

Additionally, XGBoost uses randomization approaches to reduce overfitting and to increase training speed(Bentéjac et al., 2021). XGBoost has become the top algorithm for prediction problems hence winning a lot of Kaggle competitions. Figure 2.3. shows Extreme gradient boosting level-wise growth.

Figure 2. 3

Extreme Gradient Boosting Level-Wise Growth



2.5. Machine Learning Approaches in Blood Donor Retention

Machine learning algorithms have been used in various studies to predict blood donations and blood donor retention. In their study on forecasting blood donor response using predictive modelling approach(Marade et al., 2019) used predictive modeling approach to predict whether a particular donor will donate blood within coming months. The study used existing dataset obtained from the open database of Blood Transfusion Service Centre in Taiwan.

The data was collected using the Recency, Frequency, and Monetary(RFM) model which is a popular model mostly utilized for customer churn predictions. The dataset contained five main variables namely: Recency which denotes the number of months since the person last donated blood, Frequency - total number of donations, Monetary - total blood donated in c.c.). Time in months since first donation and a binary variable representing whether the donor donated blood in March 2007. The study compared various classification algorithms such as K-nearest Support vector machines, Neighbours (KNN), Decision tree, Gaussian

Naive Bayes, and logistic regression. The study utilized holdout validation and divided the data into 70% training and 30% testing segments. The results show that decision tree produced the best accuracy at 0.60. The accuracy achieved in this study was quite low and needs improvement additionally the variables used can be increased to improve on the accuracy.

The study by Zulfikar et al., (2018) classified eligibility of blood donors using decision trees and Naive Bayes classifiers. The study employed a data set of 500 blood donors, obtained from a humanitarian organization in Indonesian. Holdout validation method was utilized of which 400 donors were used for training and 100 for testing. The accuracy of the decision tree classifier was 78.5%, while the accuracy of the Naive Bayes classifier was 81.5%. While the study utilized a real-world blood donor database for the research and compared two distinct machine learning algorithms, the accuracy achieved by the algorithms needs improvement.

The training dataset utilized for the study was also quite minimal which may limit generalization. Although Naïve Bayes achieved a high accuracy in this study. The algorithm can struggle when dealing with imbalanced datasets and does not have inbuilt mechanism for handling missing values. Additionally, the algorithm typically uses simple probabilistic models, making them less suitable for capturing complex relationships in data(Kalcheva et al., 2020). Decision trees utilized in the study are also susceptible to overlap, which can delay decision-making and increase memory consumption.

Using the dataset from Yangzhou Blood Station in China, Wu and his fellow researchers in 2022 gathered information about experienced blood donors recruited via short message service (SMS) and developed seven machine learning-based recruitment models(Wu et al., 2022). The Seven models included Random forest, Support Vector Machines, Decision trees, Linear regression, KNN, XGBoost and Deep neural networks. Thirteen

characteristics were outlined as a method for evaluating and predicting blood donors' intentions to donate. Area under the receiver operating characteristic curve (AUC), accuracy, precision, recall, and F1 score were used to evaluate the performance of the prediction models on the complete dataset. Overall, 95,476 SMS recruitments and their donation outcomes were included in the modelling study. To effectively classify a blood donor, the most relevant features were screened out and ranked by their importance in the trained model. Initial training records were divided into 70% and 30% for the training and test dataset respectively.

The accuracy of three superior models was validated using the tenfold cross-validation method based on blood groups. Blood donation interval, age, and donation frequency were discovered to be the most accurate predictors of the donation for experienced donors. The Extreme Gradient Boosting and Support vector machine models had the highest mean performance accuracy of (95%) among the seven baseline models. The study found that in the training-test dataset, XGBoost had the best parameters and it was capable of learning hidden patterns and accurately matching the characteristics of blood donors. Although the study utilized quite an extensive dataset and a variety of features, the study only dealt with only experienced blood donors leaving out students and new blood donors. Blood donors with missing critical data were also excluded from modelling processes, which affected the profile of the blood donors causing selection bias which may ultimately influence the results and limit generalization.

The study on analytics framework for blood donor classification in 2021 classified students from an Indian state university as potential blood donors or non-donors using data visualization techniques(Pabreja & Bhasin, 2021). Students who were enrolled in a bachelor's degree program at Delhi state university, were surveyed using online questionnaire created with Google forms. Twenty questions regarding the characteristics

of the donor were asked. Respondents were asked a total of 20 questions and were instructed to choose the option that best suited them on a Likert scale. Using convenient sampling technique, a total of 448 participants replied to the study. K-nearest neighbor and logistic regression algorithms were used to test the data on accuracy, precision, sensitivity and specificity.

The study used data visualization bar plots to extract features for input to the machine learning algorithms. Recursive Feature Elimination (RFE) method was employed to rank and select the features. Model training was done using 70% of records and 30% for testing. Tenfold Cross validation was also used and KNN was found to achieve the best classification accuracy when k=8. KNN classifier produced the best results with an Accuracy of 0.7027, Precision of 0.7209, Sensitivity value 0.7949, F1-score equal to 0.7561 and Specificity value of 0.5789. The Results obtained in this study may be skewed since the data belongs to students only, the students also belong to the same university and same education level and hence related socio economic factors. The study utilized quite a minimal dataset which may limit generalizations moreover data provided via online questionnaires may not be completely verifiable, the accuracy achieved by the models also needs improvement.

Soft computing with data mining techniques have also found application for prediction in blood donation domain. The study by Kewat and Sharma, (2018) employed Naive Bayes soft computing algorithm to classify and predict blood donors according to their sex and blood group. The blood donor's data was obtained from the Kota blood bank, comprising a total of 5656 cases and 12 attributes. The results showed that the generated classification rules carried out perfectly with accuracy rate of 97.5588%. Despite the model's high accuracy in this study, Naive Bayes is a relatively simplistic probabilistic model, and this can be a disadvantage when dealing with complex datasets. The underlying premise of

feature independence is one of the major drawbacks of Naive Bayes. Given the class label, it assumes that all features are completely independent. This assumption is extremely simplistic and does not hold in many real-world settings.

Naive Bayes is insensitive to feature relationships and dependencies because of its assumption of feature independence when dealing with structured data, where feature interactions are important, this makes it less effective(Kalcheva et al., 2020). In addition, Naive Bayes is unable to handle missing data well and assigns non-zero probabilities to features that are not relevant, which can have an impact on the accuracy of predictions.

While predicting the return rate in young blood donors the study by Cloutier et al., (2021) extracted data from a blood donation management information system managed by Héma-Québec a non-profit organization that supplies hospitals in the Canadian province of Quebec with blood and other biological products of human origin.

The final dataset analyzed included 81 986 donors aged 18–24 at the time of their most recent donation. The data contained 11 main attributes, Additional information was acquired from the marketing database, that included data pertaining to donor contact details. The study employed Random forest mean decrease accuracy (MDA) method to measure the features impact on the accuracy of the model and cross validation was used to validate the model.

The random forest model accurately predicted over 91% of donation frequencies, with an overall average error rate of 8.16% and specific error rates of 4.6% and 12.3% for the 'unreturned donor and returned donor groups respectively. The model's best predictive variables were found to be the number of marketing department contacts, the age of the donors, the number of adverse reactions during donation, the donors' status, and their ethnicity. Although the model achieved a considerable high level of accuracy, the study included only young donors between the age of 18- 24 years.

Additionally, donors that were contacted by the marketing professionals may result in bias. Random Forest is slow to construct and challenging to interpret, particularly when the ensemble comprises an extensive number of decision trees, as it must independently build and evaluate each decision tree(Fife & D'Onofrio, 2022). Memory consumption can also be substantial when working with enormous datasets or ensembles containing numerous trees. This can restrict their applicability on systems with limited memory.

The study by Selvaraj et al., (2022) aimed at building a forecasting system for donation of blood using SVM Model, obtained data from a Blood Transfusion Service Center in Hsin-Chu City in Taiwan. The dataset included 748 donors with five main variables: R (Recency - months since last donation), F (Frequency - total number of donations), M (Monetary - total blood donated in cc), T (Time - months since first donation), and a binary variable indicating whether a donor donated blood in March 2007 (1 for donating blood; 0 for not donating blood).

The correlation matrix was used in the study to define the relationship between two variables with 10-fold cross validation. At 78.4 percent, Support Vector Classifier obtained the highest accuracy. The research was based on information from an isolated blood bank in Taiwan. This may limit generalization to other blood banks or countries especially in Africa. To aid the prediction model further, additional data tuples may be added. By adding more information, the results could become accurate. SVM can also be compared to other machine learning techniques to access the performance on the same dataset.

The study by Salazar-Concha and Ramírez (2021) aimed to predict the intention to donate blood among blood donors using a Decision Tree Algorithm. The research employed a convenience sampling method to distribute in-person questionnaires to adult blood donors at two health facilities located in Valdivia, Chile. Specifically, a cross-sectional survey was undertaken in April and March 2020, all surveys were administered as the

last step in the blood donation procedure. The anonymity of the respondents was preserved all through the data collection procedure. The study utilized decision tree using C4.5, information gain for feature selection and tenfold cross validation. The questionnaires were modelled based on theory of planned behavior and administered to adult users in two health centers in Valdivia (Chile). 197 participants responded; seven variables were used. The model achieved an accuracy of 84.17%. The sample size utilized in this study was quite minimal, a factor which can greatly limit generalization. The accuracy achieved in the study can also be improved.

Various other studies have attempted to classify blood donors according to their characteristics and predict whether they are likely to donate in future. (Hanieza et al., 2019) uses multiple logistic regression to identify the association between various blood donor characteristics such as the willingness to donate blood, number of months since the last donation, number of donations, total volume donated and the number of months since the first donation. Secondary data was retrieved from UCI Machine Learning Repository which gives information about blood donation by staff and students from university in Hsin-Chu City in Taiwan. The Variance Inflation Factor (VIF) was used to check multi-collinearity where two or more independent variables were highly correlated to each other. The results showed that only three variables contributed to the willingness to donate blood which are: total months since last donation, frequency of donating blood and total volume of donated blood.

Facebook has utilized its large global social network to provide a blood donation tool that connects millions of people in selected countries around the world. Several researchers from Facebook studied the tool, which connects donors to blood donation opportunities in selected countries (McElfresh et al., 2021). Blood donors can find donation opportunities

and opt-in to receive notifications of the opportunities, blood recipients can also get to state their need and availability.

The study developed an online matching model that measured donor action as a proxy for actual donation and developed an online matching model based on automated policies for matching donors with donation opportunities. In simulations, the matching strategy was able to increase the number of donations by 5-10%. Those who are not in Facebook are not able to access and benefit from the tool. Likewise, those who do not have internet access cannot be able to access the tool. The study measured donor action on Facebook as an actual donation but this might not be the case since not all people will turn up to do the actual donation.

2.6. Types of Datasets

A diverse range of datasets have been utilized for prediction of blood donors (S. Liu et al., 2021). These datasets encompassed different sources and characteristics, each offering unique strengths and weaknesses in understanding donor retention.

2.6.1. Hospital/Clinical Datasets

These datasets are derived from healthcare institutions and comprise of detailed information about donors' medical history, demographic factors, donation history, and retention outcomes.

The study by Cloutier et al., (2021) utilized data from a blood donation management information system belonging to a non-profit organization that supplies hospitals in Quebec province of Canada. The dataset had 81986 donors with 11 variables. The strengths of hospital/clinical datasets lie in their comprehensive nature. They capture a wide range of donor characteristics and health conditions variables, enabling researchers to explore the impact of medical factors on donor retention (S. Liu et al., 2021). The richness of these datasets provides valuable insights into the interplay between donor

health and retention outcomes. However, a potential weakness of such datasets is their limited generalizability(Gandin et al., 2021). They often represent a specific group of donors who have interacted with healthcare institutions, potentially leading to selection biases. Consequently, caution must be exercised when extrapolating findings to the broader donor population.

2.6.2. Donor Database Datasets

Another common type of dataset utilized in prediction of blood donors is the donor database datasets(Shashikala et al., 2019) (Zulfikar et al., 2018). Blood banks or organizations managing blood donation programs collect and maintain these datasets.

The study (Kewat & Sharma, 2018) utilized data from Kota Blood Bank in India to evaluate the performance of the Naïve Bayes soft computing algorithm. The dataset has 5656 records with 12 attributes. The model was able to obtain a considerably high performance accuracy of 97.6%.

The studies by Selvaraj et al., (2022) and Hanieza et al., (2019) used online open dataset that was collected from a Blood Transfusion Service Centre's in Hsin-Chu City, Taiwan, which contains 748 records with five variables. Donor database datasets typically include donor demographics, donation history, communication records, and retention outcomes.

The strengths of donor database datasets lie in their large sample sizes and representativeness of the donor population. They provide a broader perspective on donor behavior and retention, hence allowing for more generalizable findings. Furthermore, these datasets often span multiple years, enabling longitudinal analyses and examining trends. However, donor database datasets may have weaknesses(Xiao et al., 2018b). They may contain missing or incomplete information, particularly if data collection processes must be consistent across different blood banks or organizations. Moreover, data quality

and standardization challenges may arise when integrating data from various different sources.

2.6.3. Questionnaire Datasets

Some studies relied on survey or questionnaire datasets, where data were directly collected from blood donors through self-reported responses. The study by Salazar-Concha & Ramírez-Correa, (2021) utilized in-person questionnaires to collect data from adult blood donors at two health facilities located in Valdivia, Chile. Seven variables were used with a total of 197 participants. They aimed to predict the intention to donate blood among blood donors using a Decision Tree Algorithm.

The study by Pabreja & Bhasin, (2021) in 2021 also administered questionnaires to students at an Indian state university when developing an Analytics Framework for Blood Donor Classification. The questionnaire datasets encompassed information about donor motivations, satisfaction, experiences, and attitudes.

The strengths of survey datasets lie in their ability to capture subjective donor perspectives and attitudes, and they provide valuable insights into the psychosocial factors influencing donor retention. Surveys also allow researchers to tailor questions to specific research objectives and explore specific aspects of donor behavior. However, survey-based data may be susceptible to response biases, as participating donors may differ from non-responders(Golas et al., 2018). Additionally, self-reported data may be subject to recall bias and social desirability bias, potentially impacting the accuracy of the information obtained.

2.6.4. Integrated Datasets

Integrated datasets combine information from multiple sources such as donor databases, clinical records, and external sources like social media or administrative data. Researchers from Facebook have utilized the power of social media and their large global social media

network to gather data from its users and develop an online matching model based on automated policies for matching blood donors with donation sites(McElfresh et al., 2021). Integrated datasets offer strengths in capturing various factors influencing donor retention, including donor-specific and external variables. By combining different data sources, these datasets provide a more holistic view of donor behavior and retention, allowing for a comprehensive analysis of the complex dynamics involved(Panesar, 2019). However, integrating disparate datasets presents challenges regarding data harmonization, standardization, and privacy concerns. Rigorous data integration processes must be implemented to ensure data quality, preserve donor privacy, and maintain the integrity of the analysis.

2.6.5. Intervention Datasets

Some studies adopted an experimental approach, utilizing intervention datasets. These datasets were generated by implementing specific interventions, such as targeted communication strategies or incentive programs and measuring their impact on donor retention outcomes(Wu et al., 2022).

The study by Wu et al., (2022) gathered information about experienced blood donors recruited via short message service (SMS) and developed seven machine learning-based recruitment models. Experimental datasets offer strengths in assessing the effectiveness of interventions in a controlled setting, allowing for causal inferences and insights into the mechanisms driving donor retention. By manipulating specific variables, researchers can directly investigate the effects of interventions on donor behavior. However, experimental datasets may have limitations regarding sample size and generalizability to real-world blood donation programs(Xiao et al., 2018). When implementing interventions that may affect donor behavior, ethical considerations must also be considered, to ensure that

interventions are conducted with the highest standards of participant safety and informed consent.

2.7. Feature Selection Techniques

The process of selecting a subset of representative features that can impact a model's performance is known as feature selection. Choosing the most relevant and acceptable features is an important step before getting rid of the extraneous ones (Chen & Guestrin, 2016). It is generally preferred for a classification model to be trained on a small range of characteristics in order to simplify the model and reduce the quantity of data it requires, even when there may be a large number of features available, by building a classifier with particular traits, various benefits can be attained such as minimizing the amount of data that needs to be stored, shortening the training process, reducing the training time; and reducing the dimensionality to improve prediction accuracy.

There are three major methods commonly used for feature selection they include: filter, wrapper and embedded or intrinsic methods(J. Li et al., 2018). Filter methods use statistical measures to score the correlation or dependence between input variables that can be filtered to choose the most relevant features. They are faster and quite useful especially when a number of features are involved. Examples of filter methods include correlation based methods, Information Gain, Chi-Square Test and Fisher's Score. Wrapper methods evaluate a variety of classifiers to find the best combination of features that maximizes performance. This usually takes quite some time and hence computationally expensive(Remeseiro & Bolon-Canedo, 2019). Examples of wrapper methods include forward selection, backward elimination, exhaustive feature selection and recursive feature elimination.

Some machine learning algorithms have inbuilt mechanisms for ranking and selecting features, they usually combine the qualities of both the filter and wrapper methods. This

kind of feature selection technique is known as embedded/intrinsic. Examples of embedded feature selection methods include Lasso regularization and tree-based methods, including various decision trees ensembles like random forest and gradient boosting(J. Li et al., 2018). The studies by Wu et al., (2022) and (Selvaraj et al., (2018) used correlation based feature selection techniques.

Correlation is a straightforward and simple technique for selecting features. It measures the linear relationship between each feature and the dependent variable (blood donor return in this instance). It is appropriate for identifying direct linear relationships between features and the objective(Prion & Haerling, 2020). It can help identify characteristics that may have a direct, straight-forward effect on the target variable. Correlation may not, however, capture more complex, nonlinear relationships between features and the target, and it may neglect significant nonlinear feature interactions or dependencies.

The studies by Zulfikar et al., (2018) and Salazar-Concha & Ramírez-Correa, (2021) employed information gain for their feature selection. Information gain is a statistical method that divides the data according to a characteristic to determine the reduction in entropy (uncertainty) for that feature. The main goal of this method is to determine the most important characteristics for the prediction model by calculating feature importance scores based on information gain(J. Li et al., 2018). By figuring out these scores, it is possible to decide with certainty which characteristics to analyze and which ones to leave out. Information gain is more general and is suitable for both linear and nonlinear relationships between the features and the target. It can recognize complex interactions and dependencies between attributes. However, it can be computationally intensive, particularly when the dataset contains a large number of features, and it may select elements that have a strong mutual information with the desired feature but do not have a straightforward, readily interpretable relationship.

2.8. Validation Methods

Model validation is the process of assessing the performance of a machine learning model using a dataset that was not previously used during the training of the model(Bates et al., 2023). This is done to ensure that the model is able to generalize well on new, unseen data and that the model is not simply memorizing the training data. Various validation methods have been employed in the prediction of blood donations and retention to assess the performance and effectiveness of the developed models(Marade et al., 2019). These validation methods included holdout validation, cross-validation, bootstrapping, and external validation. Each method has its strengths and weaknesses, and understanding them is crucial for accurately interpreting the studies' findings.

2.8.1. Holdout Validation

Holdout validation, also known as train-test split, is a commonly used validation method. This approach divides the dataset into two subsets: a training set and a testing set. The model is trained on the training set and then evaluated on the testing set. The method was utilized by studies by Zulfikar et al., (2018), Wu et al., (2022) and Pabreja & Bhasin, (2021) to divide data into 70% training set and 30% testing set.

The main strength of holdout validation is its simplicity and computational efficiency. It allows researchers to assess the model's performance on unseen data quickly. Additionally, holdout validation provides a clear distinction between the training and testing phases, making it easy to understand and implement. However, holdout validation may suffer from high variance if the training and testing sets do not represent the overall dataset(Awwalu et al., 2019). This can happen if the dataset is imbalanced or if certain patterns in the data are not captured in the split. It only utilizes part of the dataset for training and testing, which may limit the model's performance and generalizability.

2.8.2. Cross-Validation

Cross-validation is another widely used validation method. This method has been applied in various studies on prediction of blood donations including studies by McElfresh et al., (2021), Cloutier et al., (2021), (Saad Alkahtani & Jilani, (2019) and Salazar-Concha & Ramírez-Correa, (2021).

Cross-validation addresses the limitations of holdout validation; it involves dividing the dataset into multiple subsets or folds. The model is trained and tested iteratively, with each fold serving as the testing set while the remaining folds are used for training. The results from each iteration are then averaged to obtain an overall performance measure. Cross-validation provides a more robust estimate of the model's performance as it utilizes the entire dataset for training and testing, it also helps to mitigate the impact of dataset variability and provides a more reliable assessment of the model's generalization capabilities.

Additionally, cross-validation allows for evaluation of different model configurations and hyperparameter settings, providing insights into a model selection and performance optimization(Bates et al., 2023). However, cross-validation can be computationally expensive, especially when dealing with large datasets or complex models. The performance results can also vary depending on the chosen number of folds, and the method may be sensitive to the specific configuration of the folds.

2.8.3. Bootstrapping

Bootstrapping is a resampling technique used for validation. It involves randomly sampling of the dataset with replacement to create multiple bootstrap samples. Each sample is used to train and test the model, and the results are averaged to assess its performance. The study by Brodeur et al., (2020) employed bootstrap aggregation with cross validation to reduce overfitting in reservoir control. Bootstrapping helps to quantify

the uncertainty in the model's performance by generating multiple estimates. It is particularly useful when the dataset is limited and the model's performance needs to be evaluated more reliably.

Additionally, bootstrapping can provide insights into the stability of the model's performance by assessing its variability across different samples. However, bootstrapping can be computationally intensive, especially with large datasets, as it requires repeatedly resampling from the dataset(Brodeur et al., 2020). It may also not be feasible in cases where generating multiple bootstrap samples is impractical due to resource limitations. Furthermore, it is important to note that bootstrapping assumes that the bootstrap samples are representative of the underlying population, which may not always be the case.

2.8.4. External Validation

External validation involves evaluating the model's performance on an independent dataset not used during model development. This validation method assesses the model's generalizability to new and unseen data. External validation offers a more thorough examination of the model's performance using a different dataset in practical situations. It aids in figuring out whether the model's performance is consistent across scenarios outside of the dataset it was trained on.

External validation is crucial in assessing the model's applicability to various populations or circumstances(Eertink et al., 2022). However, finding an external dataset might be difficult and necessitate working with other academic institutes or groups. To establish meaningful comparisons, it is essential to ensure that the external dataset is sufficiently similar to the original dataset in terms of characteristics and distribution. Furthermore, external validation might only be possible if the necessary external data is available or access to such datasets is not prohibited by law or privacy concerns. There is only one blood donations dataset publicly available online which most of the studies have used for

model training and hence not possible to perform external validation for most of the studies (Bui et al., 2021).

2.9. Performance Optimization Strategies

Model performance optimization is the process of refining a machine learning model to improve its accuracy, efficiency, and overall effectiveness (Victoria & Maragatham, 2021). This process ensures that the model generalizes well to new data, minimizes errors, and operates within desired constraints such as time or computational resources. There are various techniques used for model optimization, each with specific goals and methods.

2.9.1. Regularization

Regularization is a machine learning technique that is used to prevent model overfitting. It works by reducing the model complexity by adding a penalty to the loss function, thereby improving generalization to unseen and helps the model to perform better(Alibrahim & Ludwig, 2021). Regularization techniques include L1 (Lasso) and L2 (Ridge) regularization. L1 Regularization encourages sparsity by shrinking some coefficients to zero. This enables the model to effectively perform feature selection by removing irrelevant features from the model. L2 Regularization Shrinks the coefficients of all features, but not necessarily to zero which creates a smoother model that is less sensitive to noise.

2.9.2. Transfer Learning

Transfer learning is the process of taking pre-trained models and using them on new and unrelated problems thereby reducing the need for large datasets and extensive training This is mostly applied in image recognition where the pre-trained model has already learned low-level features like edges, shapes, and colors, and apply it to a new task(Zhuang et al., 2021). Transfer learning significantly reduces training time and improves performance. However, if the source and the target tasks are not related the pre-trained model may hinder

the performance of the new model. Additionally, the pre-trained models may inherit bias from the training data they used and pass this bias to the new model hence impacting the model performance(Nanni et al., 2020).

2.9.3. Hyperparameter Tuning

Hyperparameters are configuration settings that are used to control how a machine learning model learns(Yu & Zhu, 2020). They can be preset before the model begins training and they remain constant during the training process. Hyperparameter tuning is the process of adjusting the parameters that govern the learning process in a machine learning model so as to enhance their performance. Different machine learning models have different parameters that impact model accuracy and training speed, tuning these parameters is paramount to achieving the optimal performance of the model (Bischl et al., 2023). The three main methods for hyperparmeter tuning include random search, grid search and Bayesian search.

Random search involves randomly searching for the combinations of hyperparameters from a specified range. Although this method enables the model to explore a wide range of hyperparameters it may take many iterations to find the best combination of hyperparameters and this may increase the computational time and resources with no guarantee of finding the best combinations(Kervanci et al., 2024).

Grid search involves evaluating a set of hyperparameter combinations that are defined by a grid. This is done by manually searching through a specified subset of the hyperparameter space to find the best combinations. This process may take time and therefore become computationally expensive, especially for high-dimensional hyperparameter spaces(Alibrahim & Ludwig, 2021).

Bayesian search is a statistical modelling technique which employs Gaussian process to represent the relationship between combinations of hyperparameter and the model

performance. It sequentially evaluates new hyperparameter combinations based on previous results, while prioritizing areas with a higher likelihood of good performance(Alibrahim & Ludwig, 2021). This makes it more efficient than random and grid search since it focuses the evaluations on promising areas based on past results, reducing the total number of combinations needed to explore hence reducing the computational time.

2.10. Theoretical Framework

This section analyzes and examines the theory related to the area of study. A theoretical review contextualizes the research by outlining the intellectual history and larger theoretical landscape within which the study is placed.

2.10.1. Theory of Planned Behavior (TPB)

The Theory of Planned Behavior (TPB) is a popular psycho-social theory that helps to explain and predict behavior. According to the theory of planned behavior (TPB), an individual's behavior is primarily influenced by attitude, subjective norms, and perceived behavioral control.

Attitude is a person's assessment of their behavior, subjective norms are their perception of social pressure from others, and perceived behavioral control is their confidence in their capacity to carry out the behavior(Rajeh, 2022). TPB describes how people's attitudes, subjective standards, as well as perceived control over their behavior influence their motives and subsequent behaviors(Ajzen, 1985). It has been used to describe and forecast people's behavior in various contexts, including health, education, leisure, and other topics. This theory has also been applied to research and forecast people's attitudes toward blood donation(Giles, 2004).

A modified and expanded version of the TPB was employed to explain the factors that influence blood donation intention in a study done among Chinese university students.

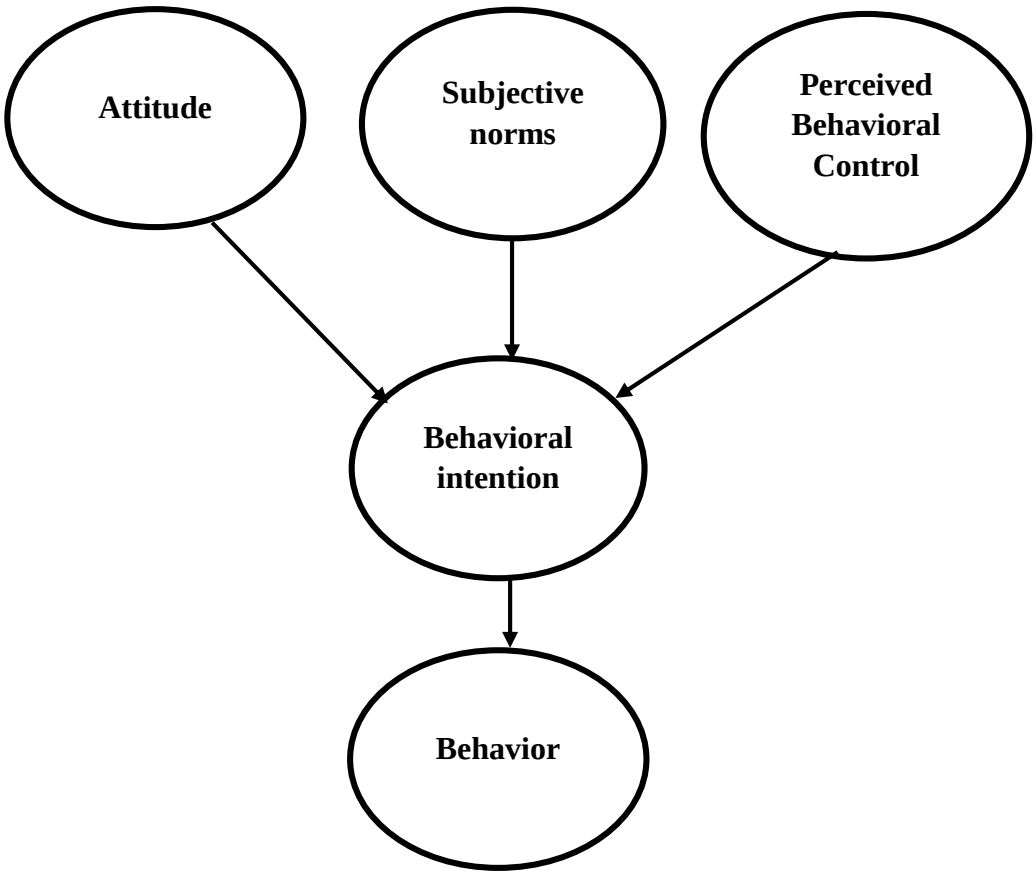
Attitudes, subjective norms, and perceived control over behavior were found to be important determinants of blood donation intention in the study by J. Liu & Han, (2023). TPB can be utilized in this study in data collection to gather data related to past blood donation behavior. This includes information on donors' attitudes, subjective norms, and perceived behavioral control, as well as demographic and behavioral data. We can also utilize TPB constructs in feature engineering to create features for the predictive model. Predictive machine learning models can utilize the insights from the TPB to produce more precise predictions of blood donation behavior. For instance, a person's attitude toward giving blood may be influenced by how well they comprehend the need for blood, how safe they believe the donation process to be, how much they value giving blood among other factors. The opinions of one's family, friends, religious leaders, and medical experts may impact one's subjective norms and how that person feels under social pressure to donate. The individual's comprehension of the procedure, the accessibility of donation sites, and confidence in their capacity to donate may all impact how much behavioral control they perceive themselves to possess(Rajeh, 2022).

By integrating the constructs of attitudes, subjective norms, and perceived behavioral control from the Theory of Planned Behavior into the predictive model for blood donor retention, the developed model can provide insights into the factors that drive donor behavior. This can help blood banks to identify potential areas of intervention and develop strategies to enhance donor retention, by addressing attitudinal barriers, leveraging social influences, and improving perceived behavioral control.

Figure 2. 4. shows the TPB Model.

Figure 2. 4

Theory of Planned Behavior (TPB) Model

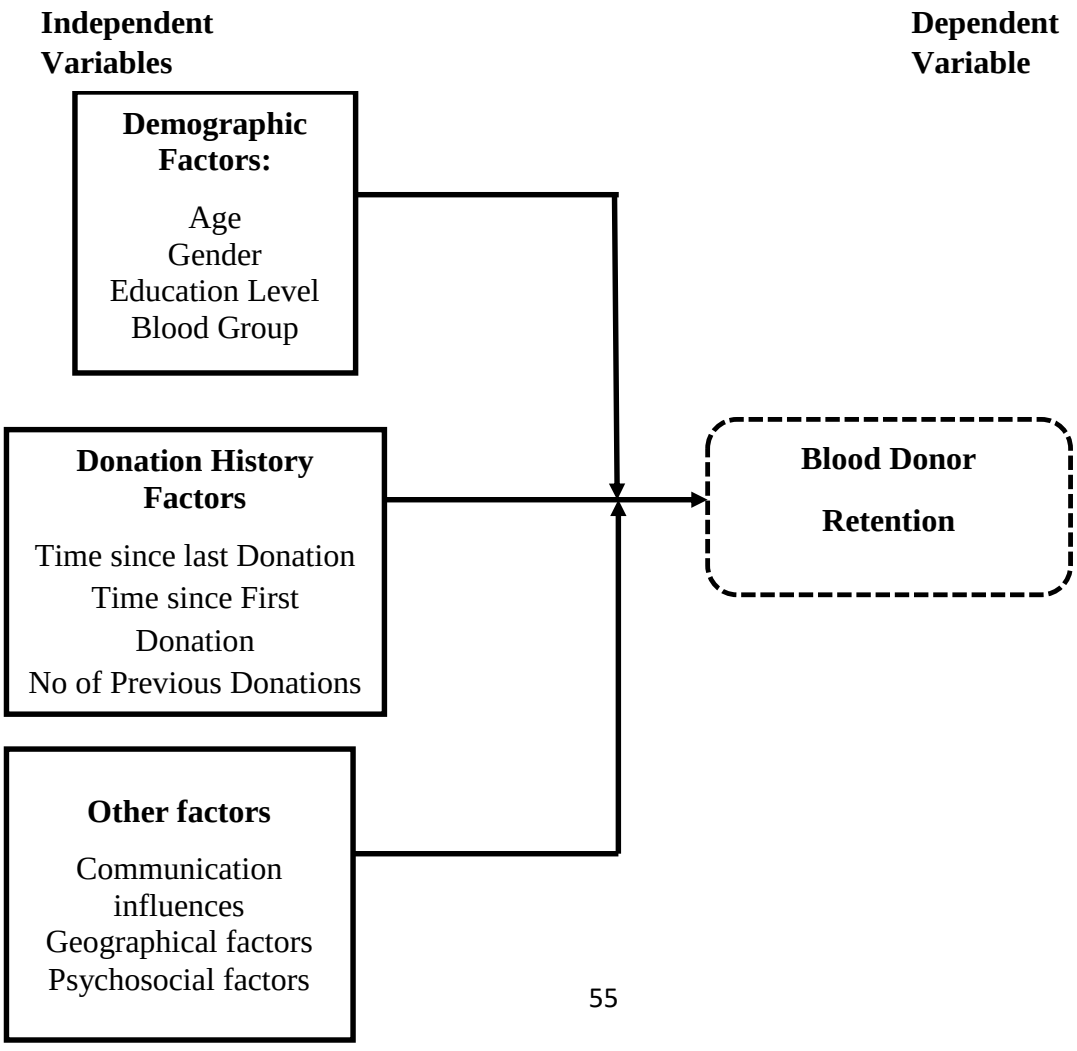


2.11. Conceptual Framework

The conceptual framework visually organizes and presents the variables of the study, helping to clarify the relationships being investigated. It serves as a guide for data collection, analysis, and interpretation, ensuring that the study remains focused on how the independent variables affect donor retention. In this study the independent variables that influence blood donor retention include donor demographics such as age, gender, education level and blood group. Donation history variables which include time since last donation, time since first donation, number of previous donations, total volume of blood donated and the blood donor recent activities such as whether they donated blood in a certain month or year. The conceptual framework of this study is show in the figure 2.5

Figure 2. 5

Conceptual Framework



2.12. Summary of reviewed literature

Table 2. 1

Summary of Reviewed Studies

Study	Algorithms	Feature Selection	Dataset & Attributes	Validation Methods	Results	Gap
(Zulfikar et al., 2018)	Decision trees and Naïve Bayes	Information gain	humanitarian blood donation organization in Indonesia with 5 attributes 400 for training and 100 testing	Validation method not mentioned	Decision tree obtained an accuracy value of 78.5% and naive bayes classifier 81.5%	Training data used was quite minimal. Accuracy can be improved.
(Wu et al., 2022a)	XGBoost,RF, DNN, SVM, KNN, decision tree, and linear regression models.	Not Mentioned	Data from experienced blood donors through message recruitment. 95,476 donors with 13 features	Tenfold cross-validation	eXtreme Gradient Boosting (XGBoost) and Support vector machine models (SVM) achieved the best performance accuracy (95%). F1 score 84.3% AUC 80.9%	Students were excluded from the donor recruitments, thus affecting the profile of donors. Focused on experienced donors only hence causing selection bias
(Shashikala et al., 2019)	KNN, naïve Bayes, and neural network methods	-Not mentioned	Data from an online questionnaire using google form. 246 records with more	Not mentioned	Naïve Bayes 86.99 k-nearest neighbours 85.54	Training data used was quite minimal, Data provided through online questionnaire may not be

			than 25 features			completely verifiable. Accuracy can be improved.
(McElfresh et al., 2021)	gradient boosted decision tree (GBDT).	Not mentioned	Data from Facebook blood donation tool	10-fold cross validation	Result showed a slight increase: 5% in Meaningful actions (a proxy for donations).	Measures donor action as proxy to an actual donation. No statistics for actual donations. Users privacy may not be guaranteed. Does not directly address donor returning rate after the first donation
(Kewat & Sharma, 2018)	Naïve Bayes.	Not mentioned	Dataset used was collected from Kota blood bank having 5656 instances with 12 attributes	10-fold cross-validation. T	The results showed that the generated classification rules carried out perfectly with accuracy rate 97.5588%.	Naïve Bayes needs to be compared with other algorithms to evaluate the performance on the same dataset.
(Cloutier et al., 2021)	Random forest	mean decrease accuracy	blood donors aged between 18 and 24 years from the province of Quebec, Canada, 81986 donors with	cross validation	The model correctly predicted more than 91% of the donation frequencies,	Donors contacted by the marketing personnel could potentially lead to bias RF needs to be compared with other algorithms to evaluate the performance

			11 variables			on the same dataset Accuracy can be improved
(Marade et al., 2019)	K-NN, Naive Bayes, Support vector machines, Decision tree, and logistic regression.	Not Mentioned	Blood Transfusion Service Centre in Taiwan. The data set comprises of 748 donors with five variables	Hold out validation	Decision tree produced the best accuracy at 0.6.	This accuracy obtained was quite low and needs improvement. More attributes that influence blood donations can be included to improve the prediction accuracy
(Saad Alkahtani & Jilani, 2019)	logistic regression (LG), random forest (RF) and support vector classifier (SVC).	-	data was collected from a Saudi blood bank containing 8 attributes	5-fold cross-validation	LG had the best accuracy at 93.49%. Followed by RF with 93.31% and SVM with 93.13%	LR is limited in its ability to manage interactions and complex feature interactions, and this may be critical for certain datasets. RF may over-fitting the data, particularly when the number of trees is large.
(Pabreja & Bhasin, 2021)	K-Nearest Neighbor) and logistic regression.	Recursive feature Elimination	Questionnaire administered to undergraduate students in Delhi state university. 488 participants	Not mentioned	K-NN classifier with an Accuracy of 0.7027 than logistic regression at 0.68 F1 score 75.6	Results may be skewed since the data belongs to students in the same university hence same education level and related

			with contains 19 features		Precision 72.09 Recall 71.4	socio economic factors. Accuracy is quite low and can be improved.
(Salazar-Concha & Ramírez-Correa, 2021)	decision tree using C4.5.	information gain	Questionnaire administered to adult users in two health centres in Valdivia (Chile). 197 participants with Seven variables were used.	tenfold cross validation with Grid optimisation	Accuracy achieved 84.17% Recall 87.8	The sample size was quite minimal. The sample size limitation does not allow generalizing this result to the whole population. Accuracy can be improved
(Selvaraj et al., 2022)	Support Vector Machines	Correlation	Blood Transfusion Service Centre's data from Hsin-Chu City, Taiwan 748 donors with five variables five primary variables	Cross validation with grid search	Support Vector Classifier, achieved an accuracy of 78.4 percent	Data was obtained from an isolated blood bank in India this may limit generalization to other blood banks or countries. Accuracy could be improved

2.13. Research gaps

The reviewed studies employed various models, including logistic regression, decision trees, random forests, support vector machines, and artificial neural networks, to predict blood donor retention. These models utilized a range of variables, including donor demographics, donation history, psychosocial factors, communication and engagement, external influences, health-related variables, and geographical factors. The validation methods employed included holdout validation, cross-validation, bootstrapping, and external validation.

Several gaps and opportunities for further research directions have been identified through the review. Although machine learning models have been utilized for blood donor prediction, the accuracies achieved in most research are rather low; hence, there is a need to increase model prediction accuracy through advanced algorithms and feature engineering techniques (Pabreja & Bhasin, 2021a), (Salazar-Concha & Ramírez-Correa, 2021).

Several research studies have utilized relatively small training datasets the studies include the study by Pabreja & Bhasin, (2021), Shashikala et al., (2019) and Zulfikar et al., (2018). This constraint limits the model's ability to generalize and generate accurate predictions indicating the necessity for larger and more diverse training datasets to increase model accuracy and generalizability of findings.

Some studies have employed different algorithms without evaluating their performance with other algorithms(Cloutier et al., 2021), (Salazar-Concha & Ramírez-Correa, 2021). There is need for a systematic evaluation and comparison of several machine learning algorithms to discover the best one(s) for blood donor retention prediction.

Several studies have utilized acquisition of data via online questionnaires and online forms(Shashikala et al., 2019). These sources may lack data verifiability, and there is a need to investigate more reliable and verifiable data sources and data collection techniques.

While machine learning models were employed in the reviewed studies, there is a need for more advanced modeling techniques, such as ensemble methods and deep learning algorithms. Ensemble methods, such as stacking or boosting, have the potential to overcome the weaknesses of individual models and improve predictive accuracy by combining multiple models(Golas et al., 2018).

CHAPTER THREE
RESEARCH METHODOLOGY

3.1. Overview

This chapter provides a detailed description of the research design, data collection procedures, data pre-processing, model development, and evaluation techniques used to achieve the research goals.

3.2. Research design

The study adopts a mixed research design involving exploratory, explanatory and experimental design. The exploratory design was used to study and understand the problem and how the problem can be solved. It was also used to gain a deeper understanding of the data, to identify the various patterns in the data and to explore the potential variables that influence blood donor retention. Explanatory research design was used to understand and explain the relationships between variables and to quantitatively analyze the data to provide explanations for the observed relationships. Experimental research was applied to build test and validate the blood donor retention model based on ensemble gradient boosting. The study adopted the Cross-Industry-Standard Process for Data Mining (CRISP-DM) process model to develop the model.

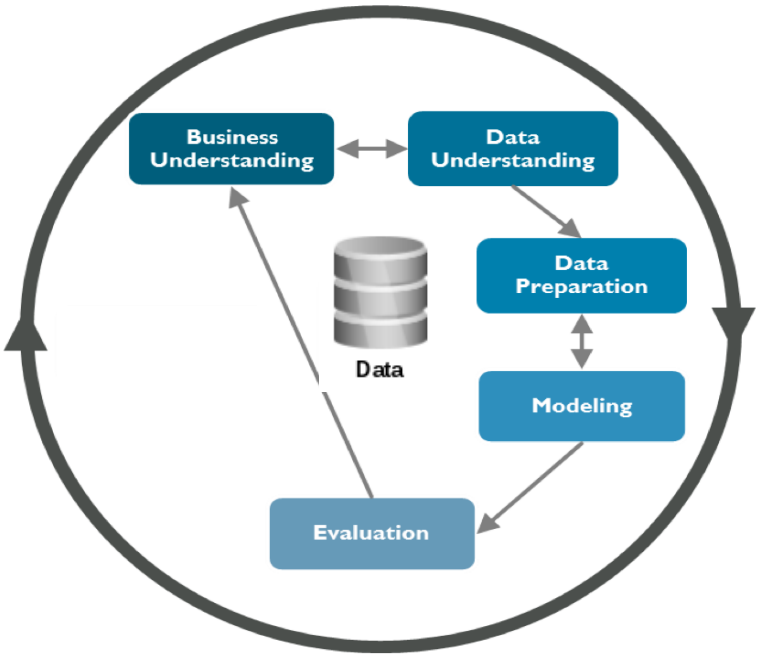
3.2.1. Overview of CRISP-DM

The CRISP-DM open standard data mining process model was conceived in 1996 and its development began in 1997 as part of an ESPRIT-funded EU project. Since then, the model has evolved into a technology-agnostic industry standard for data mining processes(Shearer, 2000). The CRISP-DM provides a structured and flexible approach to handling complex data mining tasks. It breaks the project into clear, concise and distinct phases. In the context of this

research, the methodology provides a number of significant benefits. First, it presents a structured framework that ensures all essential steps are executed meticulously, thus encouraging a systematic and organized analysis. This structured approach is highly effective and enables researchers to save precious time and resources by eliminating unnecessary tasks and concentrating on the most essential aspects of the project. In addition, the methodology is highly adaptable, promotes transparency through extensive documentation of each analysis stage. This enhances reproducibility, and permits the detection and correction of possible biases or errors(Nadali et al., 2011). This analysis improves prediction accuracy and reduces the likelihood of overfitting and other modeling errors. By adhering to this structured methodology, researchers are able to navigate the complexities of data mining with greater clarity, resulting in a more robust research process and more accurate predictions. CRISP-DM Methodology consists of six steps as outlined in figure 3.1 below

Figure 3. 1

CRISP-DM Methodology



3.2.2. Business Understanding

Business understanding phase focused on understanding the objectives and requirements of the project. This stage is very crucial as it built the foundation for the project. The researcher identified the key stakeholders including blood donation organizations, and sought to understand their specific needs, requirements and challenges and determined the importance of blood donor retention and how a predictive model can help optimize donor retention strategies. Both primary and secondary sources of literature review were used to understand the problem. The researcher also determined the availability of the required resources such as computing requirements, availability of training and testing data, assessment of risks and contingencies, as well as conduct a cost-benefit analysis.

3.2.3. Data understanding

During this phase relevant data sources were identified, data was collected and examined(Nadali et al., 2011). Data used in this study was obtained from Kenya blood bank management system it consists of blood donors registered in the system from the year 2022. This data was derived from the blood donation questionnaire which was administered to blood donors before donating blood and other sections are filled by medical officers after the blood donation. The data was queried and fetched from the blood bank management database. The provided data consists of 5000 records with nine (9) features.

To gain a preliminary understanding of the data and its structure and quality the following activities were carried out, basic quantitative analysis of the data was done using descriptive statistics. The general measures of central tendency such as mean, median, standard deviation, 25% percentile, 75^{th} percentile, minimum and maximum were used. Missing values and duplicate/redundant records were checked using the filter functions of Microsoft excel. The

data was then visualized using histograms, scatterplots, and bar plots in order to determine the overall nature of the data and identify outliers. This was done using the matplotlib and seaborn python libraries. The correlation between the independent variables themselves as well as the correlation between the independent variable and the target variable, was examined using the correlation matrix.

3.2.4. Data preparation

Data preparation is the process of cleaning, converting, and organizing raw data into a format that can be analyzed and modeled. It is a critical step in the machine learning pipeline to ensure reliable results in subsequent phases since data quality has a direct impact on model performance and accuracy (Shearer, 2000).

(a) Data cleaning and preprocessing

This stage converts raw data into a cleansed dataset that can be used in machine learning modelling. The process involved checking for missing or null values, checking for duplicate or redundant records, checking and dealing with outliers, as well as data sampling and filtering.

(b) Encoding

The dataset includes both numerical and categorical variables. Some machine learning models, may not function directly with categorical data. They presume that the variables are numerical. As a result, the data's categorical variables must be transformed to numeric values before being fed into the classifiers. One popular method of encoding categorical variables is label encoding(B. Zhang et al., 2019).

Label encoding uses alphabetical ordering to assign a unique integer to each label. One-Hot Encoding is another popular technique which represents each category as a binary vector. A

new binary column is created for each category in a categorical column. Label encoding is therefore computationally efficient as compared to one hot encoding.

Another method used for encoding is ordinal encoding which is used when there is clear and predefined ordinal relationship. The dataset used in this study had three categorical variables. Gender, education level and blood group. The gender variable was converted to the numerical values '1' and '0' using label encoding. While Education level was encoded using ordinal encoding. The items were converted to numerical values of 0,1,2 and 3 representing, none, primary, Secondary and tertiary education respectively.

The blood group was encoded using one hot encoding. For each blood group, a new binary (0 or 1) column was created. Each donor was assigned a value of 1 in the column corresponding to their blood group and 0 in the others. The original blood group category was replaced with the corresponding binary values in the newly created columns.

(c) Data Sampling

Highly imbalanced data may impede the performance of machine learning models, Data Sampling is used to minimize the data imbalance. Primarily, three sampling methods are employed: Under sampling which decreases the items in the majority class to match the items in the minority class this may result in loss of vital information and may also lead to a reduced model performance(Chaipanha & Kaewwichian, 2022).

Over Sampling replicates, the instances of the minority class which may result to overfitting since the developed model may learn specific details of the replicated cases rather than the general data patterns. Additionally, it increases the size of the dataset, which may lead to higher computational costs. An example of over sampling is Synthetic Minority Oversampling Technique (SMOTE) which generates new synthetic instances of the minority class. SMOTE

can introduce noise if the generated synthetic samples are not representative of the actual data distribution.(Wongvorachan et al., 2023).

Some algorithms allow specifying of class weights directly in order to balance the influence of each class in the data. Class weighting techniques are used to address class imbalance by adjusting the loss function to incorporate weights rather than altering the dataset itself through oversampling or under sampling(Anand et al., 2010). This is a more efficient method and it does not increase the size of the dataset moreover it ensures that the patterns in the dataset are maintained which directly impacts the model learning(Yang et al., 2022). Therefore, blood donation dataset used in this study was balanced using class weighing, with the scale-pos-weight parameter in both XGBoost and Light GBM.

(d) Feature selection

Both XGBoost and LightGBM provide methods for determining the significance and importance of the features in a dataset. XGBoost calculates a feature significance score based on how frequently each feature is used for splitting the data among all boosting rounds (trees) in the model. It does this by summing the gains provided by each feature when employed in tree construction(Chen & Guestrin, 2016).

LightGBM also provides a feature importance score, however it calculates the scores differently. In LightGBM, feature relevance is calculated based on the number of times a feature appears in tree splits, weighted by the mean gain of the splits that use the feature further LightGBM employs a histogram-based technique for splitting, enabling it to effectively calculate feature importance during training(Guolin Ke et al., 2017). The scores serve as an indication of the extent to which each feature contributes to the predictions made by the model(Dorogush et al., 2018). The provided scores were utilized to manually determine

what features are most important by employing either a specific threshold or an established number of top attributes. Light GBM and XGBoost embedded feature selection methods were utilized to select features. Correlation coefficient was also used as an independent method from the two embedded methods.

3.2.5. Modelling

In this phase, the preprocessed data was utilized to construct machine learning models that predict blood donor retention. The resulting data set after the completion of the preceding operations was divided into training and testing sets. The dataset was divided into 80% training data and 20% testing data. This gives the model enough data for training allowing it to learn the underlying trends and patterns while reserving a significant portion for testing. Additionally, this method ensures a robust evaluation of the model performance and reduces the risk of overfitting. Given the availability of labeled data for training, supervised machine learning models are used in this research for modeling.

(a) Model Selection

Model selection involves selecting appropriate algorithms that can effectively fit the problem at hand that is the prediction of blood donor retention, Extreme gradient boosting(XGBoost) and Light gradient boosting algorithms were selected based on available data. Light gradient boosting (Light GBM) is a gradient boosting method that employs tree-based learning techniques. It employs a technique known as Gradient-based One-Side Sampling (GOSS) to reduce the total quantity of samples used for training while maintaining the model accuracy, hence increasing convergence speed(Guolin Ke et al., 2017).

Light GBM constructs decision trees using a split finding algorithm that considers the histogram of the gradient values instead of the actual values, resulting in faster training times.

It also allows for categorical features to be expressed as integers or one-hot encoded. Its benefits include faster training efficiency, lower memory utilization, and improved accuracy. Its downside is that it may grow deeper decision trees and hence overfitting(Liang et al., 2019).

XGBoost is a distributed, scalable gradient-boosted decision tree (GBDT) machine learning library(Chen & Guestrin, 2016). It is also called regularized GBM because it features built-in regularization to prevent overfitting. Other strengths of XGBoost include: ability to allow parallel processing by employing several CPU cores hence making it faster than GBM; It has in-built capability to handle missing values. It runs cross-validation at each boosting iteration, making it easy to find an ideal number of iterations in a single run and offers excellent tree pruning(Chang et al., 2018). It was chosen for this study because it is an ensemble machine learning algorithm that has demonstrated superior performance accuracy and robustness making it the leading machine learning library for regression, classification, and ranking problems.

(b) Model Training

A series of experiments were carried out in this phase to build a well performing model that can accurately predict the blood donor retention. The model was developed using python packages. The two algorithms, XGBoost and LightGBM were trained in parallel. The ensemble model is created using soft weighted voting, with the weights for each individual model automatically calculated based on their predictive performance. The soft weighted voting mechanism combines the predicted probabilities from each base model and then calculates the weights depending on the prediction. The weights are then multiplied with the

individual model prediction to create the ensemble model. This enables the ensemble model to leverage the strengths of each individual model while mitigating their weaknesses.

(c) Hyper parameter Tuning

A model hyper parameter is an external characteristic of a model whose value cannot be estimated from the data. The hyper parameter's value must be set prior to commencement of the learning process. There are three main hyperparameter optimizing techniques: grid search, random search and Bayesian optimization(Claesen & De Moor, 2015). Bayesian optimization was applied to determine the best set of hyperparameters of the base models that produce the most accurate predictions. This method offers a more constrained set of possibilities to adjust particular model hyper parameters with better precision when unsure of the appropriate model hyper parameters and need to narrow the options. Additionally, it requires fewer evaluations to find the best hyperparameters hence making it faster and computationally efficient.

3.2.6. Model Validation

Model validation refers to the process carried out after or during the model training to confirm that a model achieves its intended prediction purpose. This study adopted a mix of both hold-out and cross validation methods. After feature selection the data was divided into training and testing sets. Using the training set the models were trained in parallel using cross validation. K-Fold cross validation was used in the study(Chen & Guestrin, 2016). Cross-validation provides a more reliable and robust estimate of model performance. By using multiple subsets of the data, it helps ensure that the model's performance is consistent and not dependent on a particular random split of the data thereby preventing overfitting and allowing for maximum data utilization(Brodeur et al., 2020). Training data was divided in to K folds

with k=5, trained iteratively on fold K-1, the performance was then assessed on fold k. This process was repeated K times in order to test on every example in the training set. The recorded performance was then utilized to evaluate various hyper-parameter combinations. The hybrid ensemble model was created using weighted voting and then evaluated using the testing set that was set apart in the initial step.

3.2.7. Model Performance Evaluation

In this phase, performance of the predictive model developed was evaluated. The performance was evaluated using the confusion matrix with accuracy, precision, recall and F1 score as the main metrics.

(a) Confusion Matrix

Confusion Matrix is a table used in classification problems to determine where model errors occurred(Naveen et al., 2019). It is an incredibly useful tool for calculating Accuracy, Precision, Recall, Specificity, and most importantly AUC-ROC curves. The matrix displays the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) that the model generated from the test data(Vujovic, 2021). Other metrics such as, precision, recall, F1 score and area under curve value(AUC) was calculated from the confusion matrix. The model's ability to predict blood donor retention rates was assessed and its strengths and weaknesses identified. Table 3.1.and table 3.2 shows the confusion matrix and the confusion matrix evaluations respectively.

Table 3. 1

Confusion Matrix

		Predicted	
		Positive **(P)**	Negative **(N)**
Actual	Positive +	True Positive (TP)	False Negative (FN) **Type II Error**
	Negative -	False Positive (FP) **Type I Error**	True Negative(TN)

Table 3. 2

Confusion Matrix Evaluations

	Metric	**Formula**	**Definition**
1	Accuracy	$= \frac{TP + TN}{TP + TN + FP + FN}$	Proportion of the Total number of predictions that turn out to be correct
2	Precision	$= \frac{TP}{TP + FP}$	Number of the correctly predicted cases that actuallyturn out to be positive
3	Sensitivity/ Recall	$= \frac{TP}{TP + FN}$	Number of the actual positive cases that are predicted correctly with the model
4	F1 Score	$= 2 * \frac{Precision * Recall}{Precision + Recall}$	Provides the balance between Precision and Recall

3.2.8. Deployment

Once a satisfactory model is developed, the model may be deployed to a real-world environment. The model can then be integrated into an easy-to-use application that allows blood donation organizations to predict donor retention return rates. The model can also be integrated to the existing blood banks systems so as to seamlessly predict blood donor retention.

3.2.9. Monitoring

After deployment, the model will be monitored continuously to establish the performance and to collect feedback from stakeholders. Tracking of the accuracy of the predictions will be done to validate the effectiveness of the model over time based on feedback and the evolving needs, the necessary adjustments will be done.

3.3. Data Analysis

This research employed both comparative and quantitative data analysis methods to find patterns within the data and draw meaningful conclusions. The model was implemented using Python programming. Python is a free and open source programming which is easy to learn and use. Jupiter notebook is the platform where the python program was coded. The python libraries utilized include. Numpy for mathematical and statistical operations. Pandas for loading, manipulating and exploring the data, matplolib and seaborn for data visualizations, and scikit-learn which provided the data mining tools for gradient boosting implementations.

3.4. Ethical Considerations

Predictive models raise ethical concerns, particularly with regard to donor privacy and data usage. Extreme care was taken to ensure that the data collected and analyzed for the model was compliant with data protection regulations and policies. The researcher worked with all stakeholders and ensured all the approvals were obtained. The researcher obtained approvals from Meru University of Science and Technology Institutional Research Ethics Review Committee (MIRERC), Kenya National Commission for Science, Technology and Innovation (NACOSTI) and the Meru Teaching and Referral Hospital under the County Government of Meru, Department of Health. All personal identifiable donor information was omitted from the extracted data to protect the identities of the donors. Due care was taken to ensure

confidentiality of the data collected from the organizations. Sharing of any data collected shall be done with relevant permissions.

CHAPTER FOUR
RESULTS AND DISCUSSION

4.1. Overview

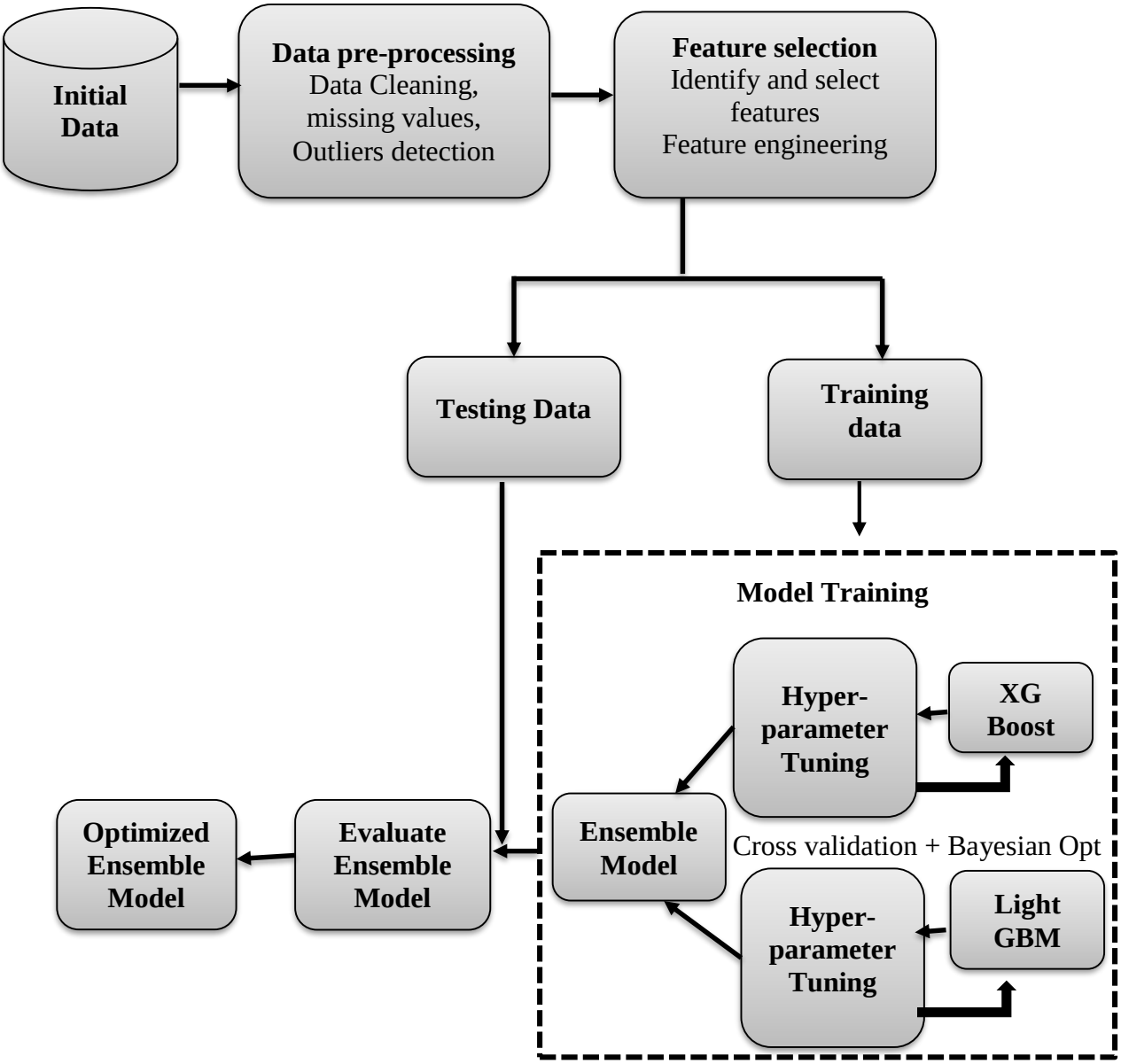
This chapter provides a detailed presentation of the model development, implementation, and testing processes, along with the results obtained. It begins by explaining the data sources in detail and the data preprocessing methods employed, followed by an explanation of the feature selection process. The chapter then outlines the stages of the model training, testing, and evaluation, it further describes the techniques and the process used to optimize the model's performance. In addition, the chapter presents the results, findings, and analysis from each stage of the implementation, providing insights into the effectiveness of the developed model in predicting blood donor retention. The chapter concludes by discussing the overall performance of the model, comparing it with existing approaches

4.1.1. Experimental Setup

The experimental set up followed the following steps: The dataset was extracted, Data preprocessing was done to clean the data, check missing values and detect outliers, feature selection was done to select the best features for model training, data was split into training and testing. The base models were trained using cross validation, performance evaluation for the models was done, model optimization was conducted through hyperparameter tuning to improve performance, and finally the performance evaluation of the optimized hybrid ensemble model was conducted. The experimental set up stages are summarized in figure 4.1. below.

Figure 4. 1

Model Design



4.1.2. Data Source

Data used in this study was obtained from Kenya blood bank management system it consists of blood donors registered in the system from July 2022 to May 2024. This data is derived from the blood donation questionnaire which is administered to blood donors before donating

blood and other sections are filled by medical officers after the blood donation. The data was queried and fetched from the blood bank management database.

4.1.3. Description of The Dataset

The provided data consists of 5000 records with nine (9) features. The features in the dataset are described in table 4.1 below.

Table 4. 1

Data Description

Attribute	Description	Values
Gender	Gender of the door	Male, Female
Age	Age of the blood donor	Number 17-65
Education Level	The highest education level achieved.	None, Primary, Secondary., Tertiary.
Blood group.	The blood group of the blood donor.	A+, A-, B+, B-, AB+, AB-, O+, and O
Months since last donation	Total number of months since last donation Number	0 and above
No of Previous donations	Total number of donations made by the donor including the current donation Number.	0 and above
Months since First Donation	Total number of months since the donor made the first donation Number	0 and above
Total Volume donated	Total volume of blood that the donor has donated since they started donating blood. Number	0 and above
Donated Blood in 2024	Binary variable indicating whether a donor has donated blood in 2024	0 or 1

4.1.4. Exploratory Data Analysis

A structural analysis of the data was done to examine the structure of the dataset based on the variables contained. All variables were either categorical, integers or numeric. Figure 4.2 below shows the first 10 records of the data as extracted from the dataset.

Figure 4. 2

Dataset Head

	A	B	C	D	E	F	G	H	I
1	Gender	Age	Education Level	Months Since Last Donation	Previous Donations	Total Volume Donated	Months since First Donation	Blood Group	Donated Blood in 2024
2	Male	38	Tertiary	50	2	900	56	A+	0
3	Female	22	Secondary	13	1	450	16	A+	0
4	Female	48	Secondary	16	13	5850	79	B+	0
5	Male	31	Secondary	20	7	3150	41	O-	0
6	Female	19	Primary	12	1	450	16	O-	0
7	Male	18	Secondary	4	1	450	10	O+	1
8	Female	20	Secondary	7	2	900	13	AB-	0
9	Male	21	Secondary	12	1	450	15	A-	0
10	Male	25	Primary	9	2	900	15	B+	0
11	Female	48	Secondary	46	5	2250	61	O-	0

Statistical summary of the data of the numerical values was done using python describe code. The structural analysis of the data shows the mean, standard deviation, minimum, maximum, 25th percentile and 75th percentile for each of the numerical variables. The mean shows the average of the values while the standard deviation measures how the numerical values were spread. Figure 4.3. below shows the statistical summary of the numerical variables.

Figure 4. 3

Statistical Summary of the Numerical Variables

	Age	Months Since Last Donation	Previous Donations	Months since First Donation	Total Volume Donated	Donated Blood in 2024
Count	5000.000	5000.000	5000.000	5000.000	5000.000	5000.000
Mean	23.945	13.398	2.366	20.396	1070.100	0.397
Std	8.162	14.676	3.382	19.784	1528.240	0.489
Min	18.000	0.000	1.000	0.000	0.000	0.000
25%	19.000	3.000	1.000	6.000	450.000	0.000
50%	21.500	8.000	1.000	13.000	450.000	0.000
75%	24.000	20.000	2.000	900.000	900.000	1.000
Max	64.000	80.000	54.000	163.000	24300.000	1.000

4.1.5. Univariate Analysis

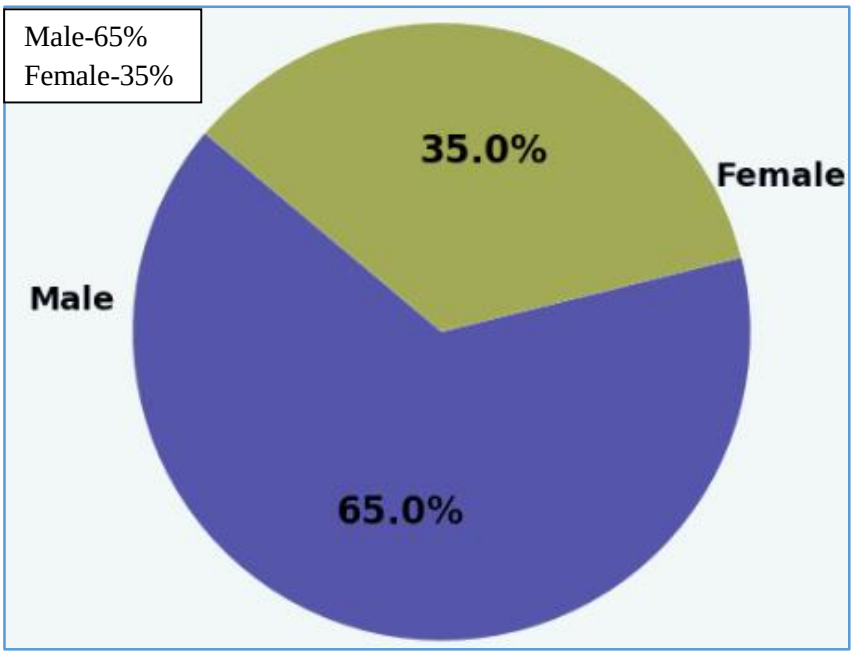
Univariate analysis refers to discovering data patterns within a single attribute. The goal is to gain a comprehensive understanding of the individual variable and examine each variable without considering the relationship with other variables(Hamid et al., 2013). Univariate analysis was done on each of the variables in the data.

a). Gender

The gender distribution of the blood donors shows that male donors account for a majority of blood donors representing 65% while the female blood donors account for 35%. Gender of the donors is an important variable in prediction of blood donors since the psychological differences between the male and female can influence eligibility and frequency of blood donation. Women are also more likely to face donation restrictions due to lower hemoglobin levels pregnancies especially during menstruation and during pregnancy(Murtagh & Katulamu, 2021) . Figure 4.4. shows the gender distribution of the blood donors.

Figure 4. 4

Gender Distribution of the Blood Donors

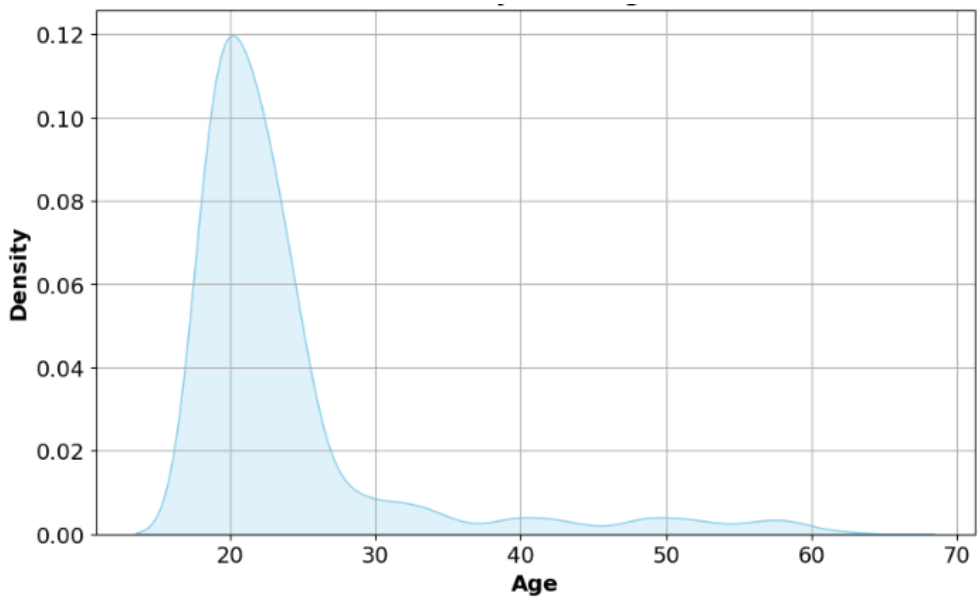


b). Age

The age of the lowest blood donor was 17 years while the maximum was 64 years. The mean age of the blood donors is 23 years, the 25th percentile falls at 19 while the 75th percentile falls at 24 years. This agrees with the studies (Adepoju, 2019), (Okuthe et al., 2022)and (World Bank, 2022) which found that most of the blood in Kenya is collected from high school and college students who fall between the age of 17-24 and hence there is normally acute shortage of blood when the schools were closed especially during Covid 19. Figure 4.5 below shows the age density plot.

Figure 4. 5

Age Density Plot

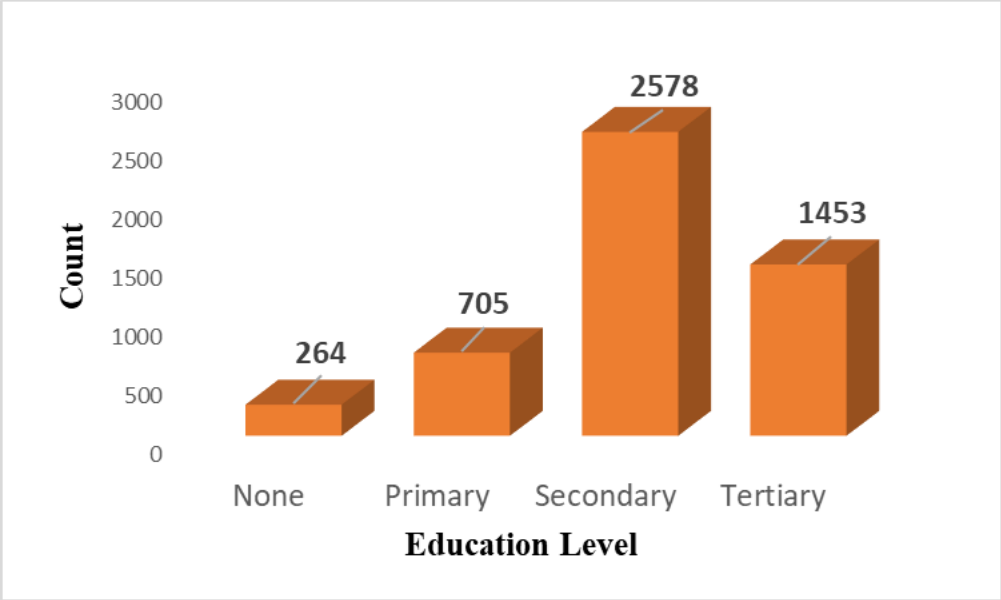


c). Education Level

This indicates the highest education level achieved by the donor. Education level is a good predictor of blood donation since education often correlates with increased awareness and increased understanding of the blood donation and safety. Educated individuals are likely to be more informed on the need to donate blood and its benefits to people and community as well as the implications on health for the blood donors(Lourençon et al., 2011). Additionally, education influences socioeconomic status which can affect the level of accessibility to blood donation centers and availability of time which can facilitate easier participation in blood donation drives. Most of the blood donors had secondary school as their highest level of education achieved followed by tertiary then primary and the lowest was those without basic education(none). Figure 4.6. below shows the distribution education level.

Figure 4. 6

Distribution of the Education Level

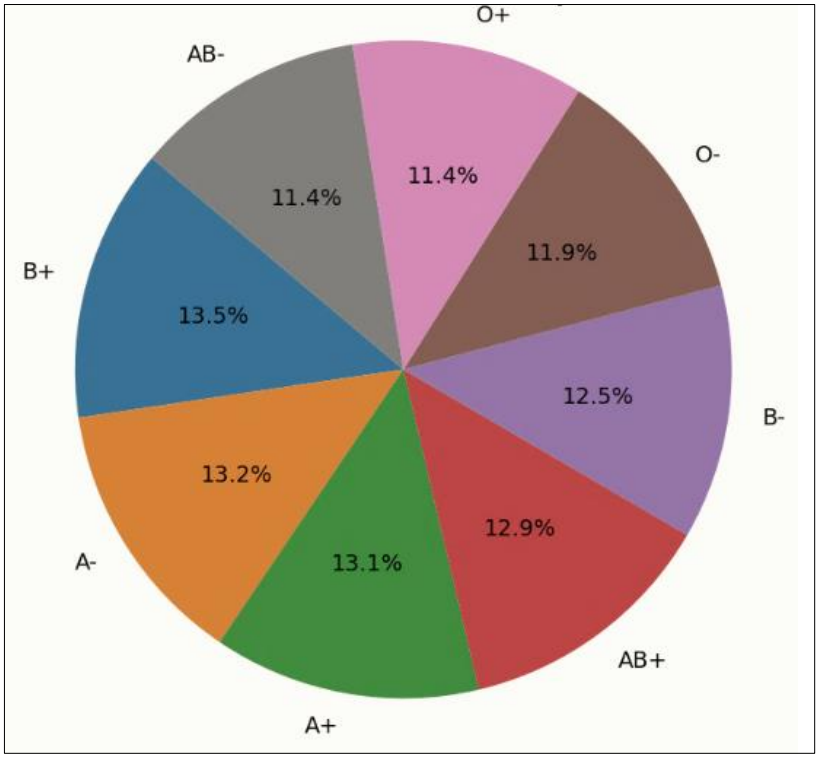


d) Blood Group

A blood group is a classification of blood based on the presence or absence of specific antigens on the surface of red blood cells, the most common types being A, B, AB, and O, and the Rh factor being either positive or negative(Mbuthia et al., 2019). Blood transfusions require accurate matching of the donor and recipient blood groups to avoid any adverse reactions. Proper understanding of the distribution of blood groups among donors is important to predict the availability of compatible blood donors to recipients who are in need. The blood group was evenly distributed among the blood donors with the B+ blood group having a slightly higher percentage (13.5%) and the AB- and O+ with the lowest (11.4%). As shown in figure 4.7. below.

Figure 4. 7

Distribution of Blood Groups

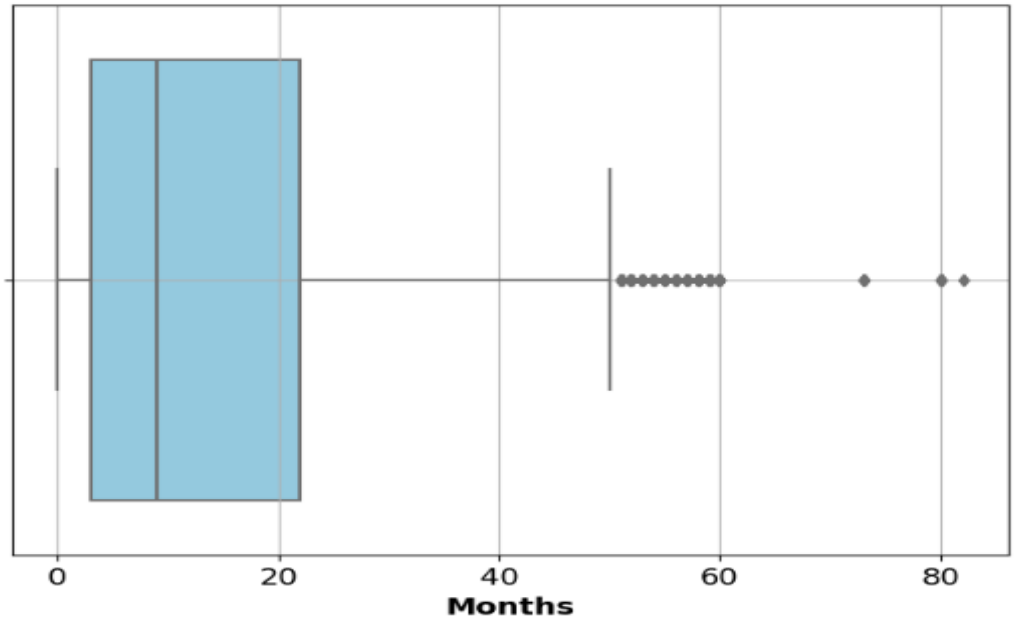


e) Number of Months Since Last Donation

This is the period in months that has elapsed since a donor last donated blood. This variable is very important since it provides important information on the blood donors recent behavior which can influence their likelihood of donating blood again. Figure 4.8 below shows a Box plot of number of months since last donation

Figure 4. 8

Box Plot of Number of Months Since Last Donation



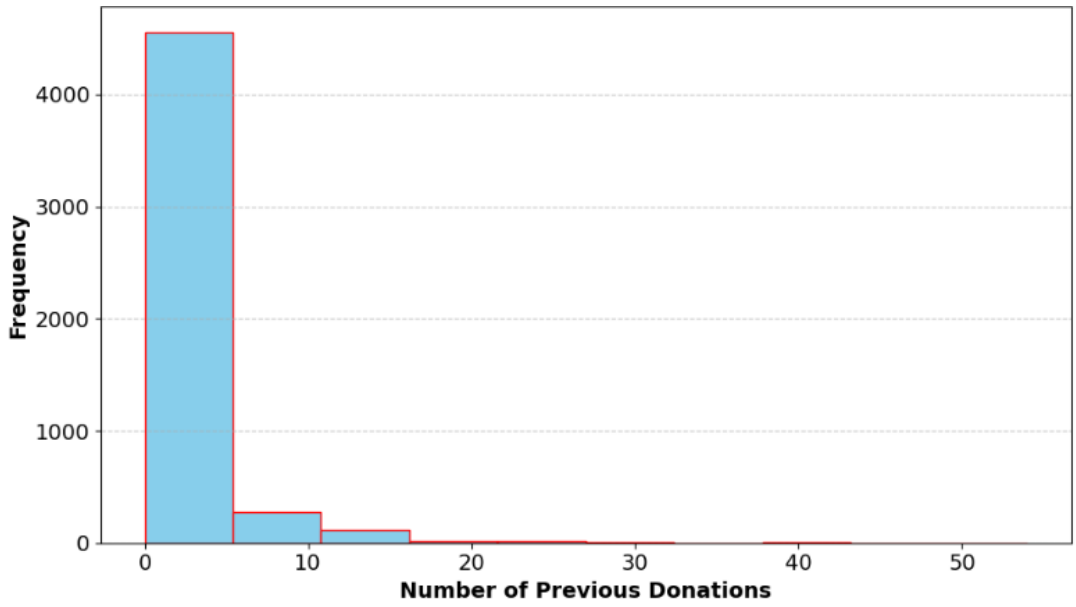
Most of the blood donors had less than 20 months since last donation with most of them being first time blood donors. The average months since last donation is fourteen (14) months while the 50th percentile falls on nine (9) months and the 75th percentile falls at twenty-two (22) months with a few donors with over 40 months since the last donation.

f) Total Number of Previous Donations

This is the total number of blood donations that a blood donor has made in their lifetime. The total number of donations is an indicator of the blood donor commitment and the likelihood of continuing to donate blood. Blood donors with higher number of previous blood donations are regarded as committed and reliable. By analyzing this variable we can identify donor patterns and behavior such as how frequent they donate and changes in their blood donation frequency over a period of time. Figure 4.9. below shows a histogram of the number of previous donations

Figure 4. 9

Histogram of Number of Previous Donations



Most of the blood donors are first time blood donors and hence have only one donation. The median donation is 1 while the mean donation is 2 donations. The 75th percentile falls on two donations while the highest donor recorded 54 number of donations. This implies the need to ensure that the many first time donors return and become regular blood donors hence increasing the number of donations.

g) Total Volume Donated

This is the total amount of blood that donors have donated throughout their blood donation lifetime. High donation volumes indicate long term commitment and consistent involvement and are likely to suggest reliability of a donor and the likelihood of them to continue donating blood. This variable can help to identify blood donor patterns based on past activity and ensure that donors are donating within the recommended guidelines. In Kenya donors are required to donate only one pint of blood per donation, one pint is equivalent to 450 milliliters of blood. Most first time donors have therefore donated a total 450 Ml. The median donation is 450ml,

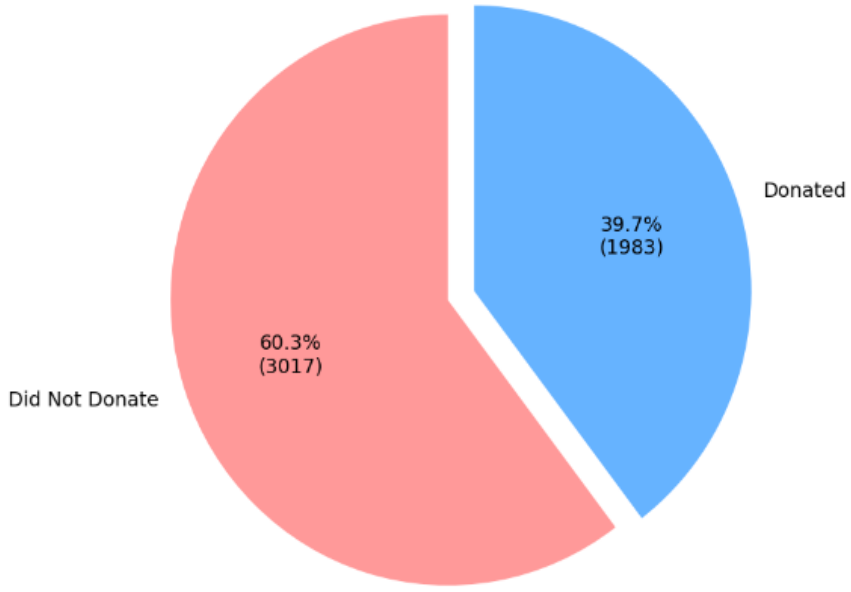
the 75th percentile is 900ml. This is because over 75% of the donors have donated only two times in their lifetime.

h) Donated Blood in 2024

This variable indicates whether a blood donor had donated blood from January to May 2024. It is very important since it measures how recent the blood donor has donated blood. The variable also acts as our target variable. 39.7% of the blood donors have donated blood in 2024. Figure 4.10 shows the proportion of donors who had donated blood in 2024.

Figure 4. 10

Proportion of Donors Who Had Donated Blood in 2024



4.1.6. Bivariate Analysis

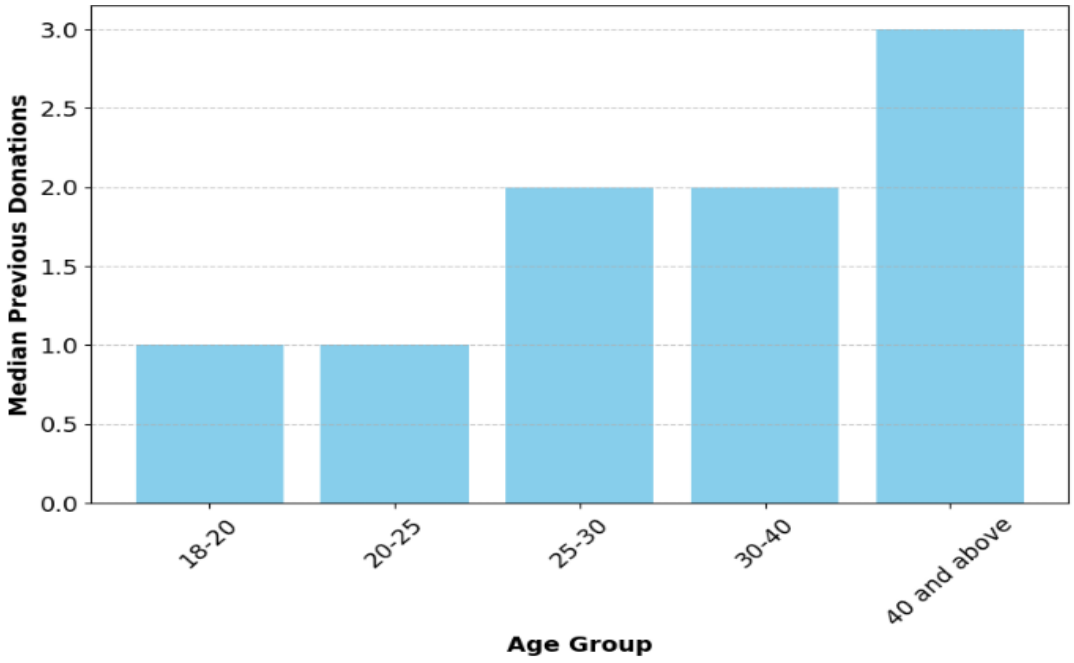
Bivariate analysis is the statistical examination of two variables to determine the relationship between them(Hamid et al., 2013). It is crucial in understanding how one variable influences or is related to another variable. Additionally, bivariate analysis is important to identify associations and potential relationships in the variables.

a) Age and Number of Previous donations

Most young blood donors below the age of 25 years were first time blood donors and hence had only made one blood donation. Most donors above the age of 40 have an average of three donations. This shows that although the number of donations increase with age, there is need to increase the number of repeat donations especially from the age group above 20 years. Figure 4.11. shows the distribution of number of previous blood donations on different ages

Figure 4. 11

Bar Plot of Age Vs Number of Previous Ponations



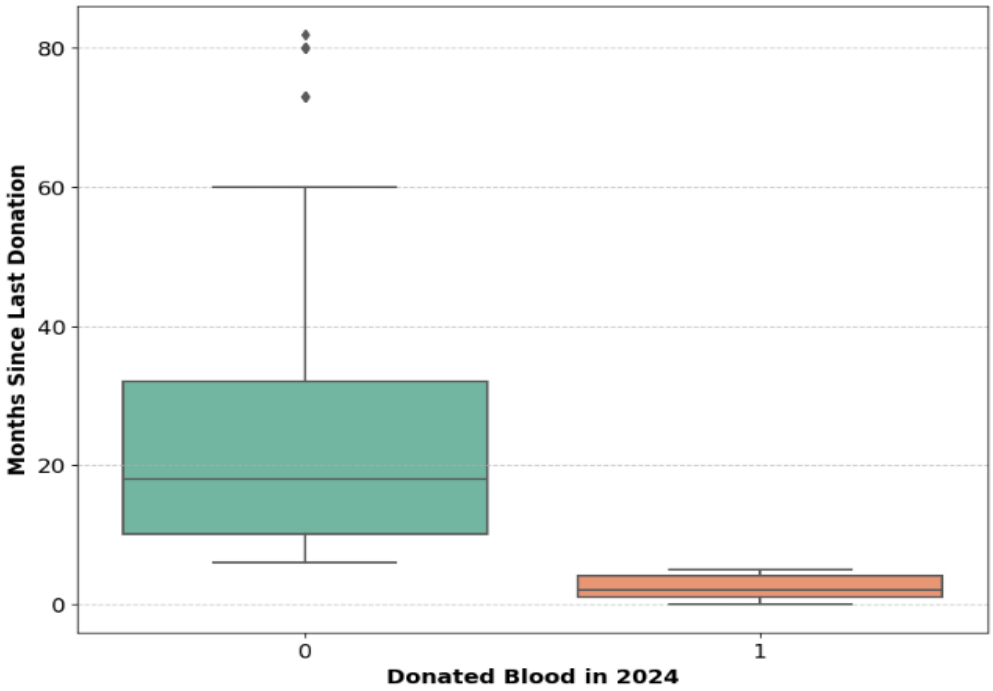
b). Donation Status and Months Since Last Donations

A comparison of Months since last donation and donation status shows that most donors with fewer number of months since last donation are more likely to donate blood while those with a higher number of months since last donation may not likely to donate. This suggests that blood donors with shorter intervals between their blood donations are more likely to return for future donations therefore maintaining regular contact and engagement with recent donors

could be a key strategy for improving donor retention rates. Figure 4.12 shows a Box plot of months since last donation and donation status.

Figure 4. 12

Box Plot of Months Since Last Donation and Donation Status



4.1.7. Data Pre-Processing

Data preprocessing is a very important step in machine learning model development. It involves cleaning, transforming, and preparing the data for further analysis and modeling. The dataset was checked for missing values and duplicates within the data as well as identification of incorrect entries within categorical variables. There were no missing values in the data since the data was extracted from a database and had already been checked for accuracy, completeness and validity.

4.1.8. Outliers Detection

Outliers are data values that differ significantly from the other data in the dataset. They may be extremely small or extremely large data values relative to the rest of the dataset. Outliers may be as a result of errors in the data, in this case they should be removed or may be correct readings but different from the other data points. There were a few outliers in age, months since donation, months since first donation and total volume donated. The outliers were correct values but a bit extreme in those fields. They were not removed since we believe they contain important information and patterns for the study. Figure 4.13 and 4.14 shows an outlier boxplot of age and months since last donation respectively.

Figure 4. 13

Outlier Boxplot of Age

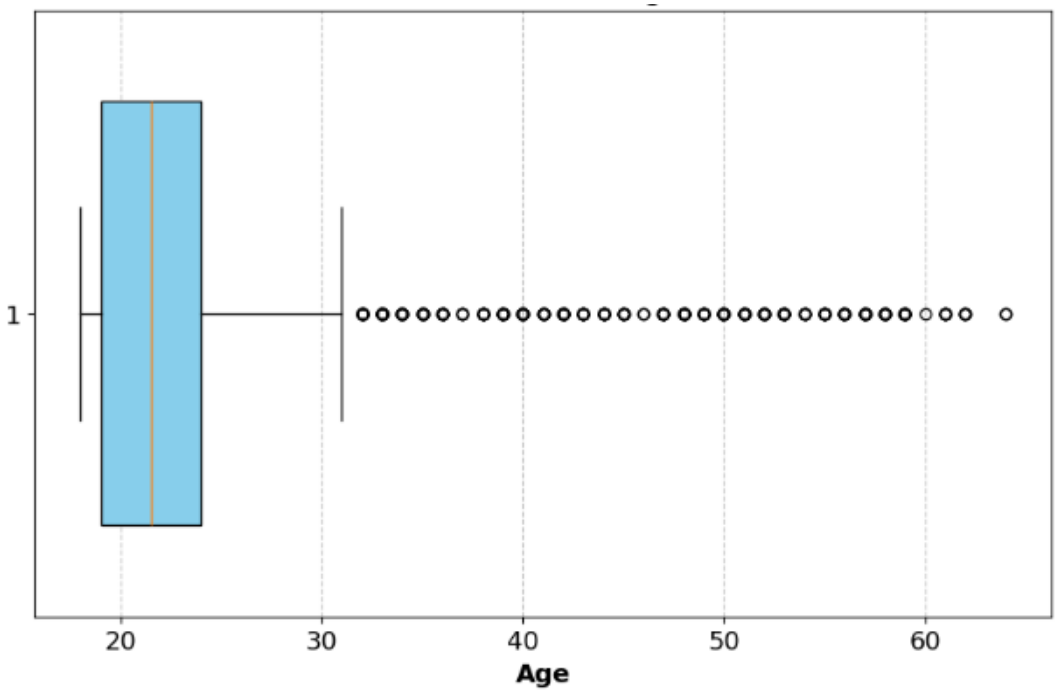
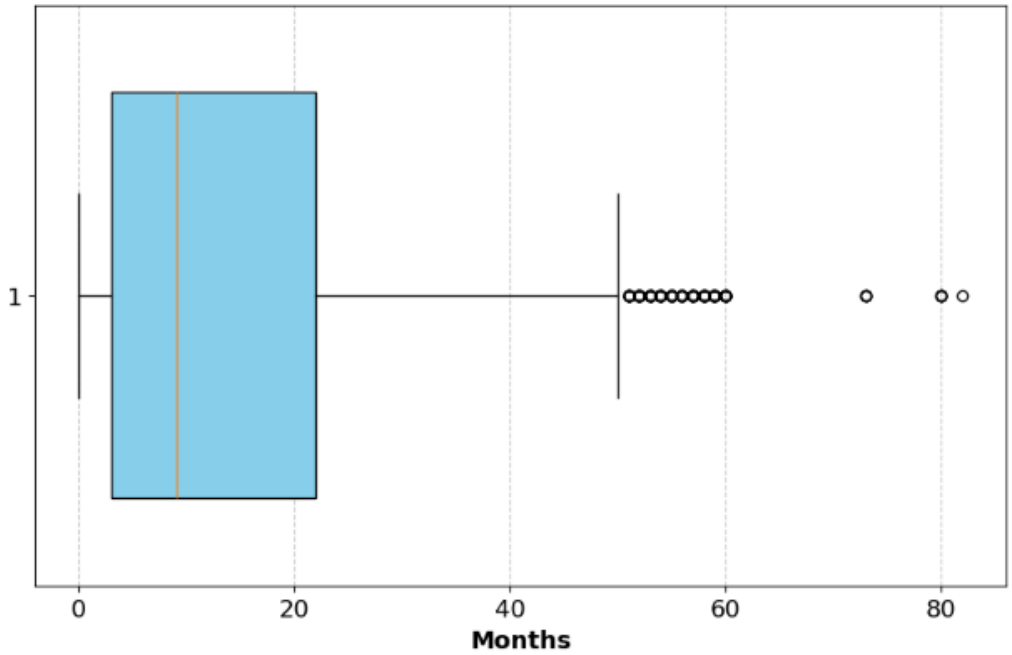


Figure 4. 14

Outliers Boxplot of Months Since Last Donation.



4.1.9. Encoding and Standardization

After dealing with outliers the data was encoded to convert categorical variables to numerical format. Gender was encoded using label encoding which converted the categorical variable male to 1 and female to zero. Nominal encoder was used to encode the education level with values 0,1,2,3 and 4 representing none, primary secondary and tertiary respectively. One hot encoding was used to encode blood group which converted the blood group column into multiple columns for each blood group with values 0 or 1. The data was then standardized using the min-max scaling to ensure that all the values lie within the range of 1-0.

4.1.10. Feature Selection

Three methods were used to determine feature importance in this study. Both the XGBoost and Light GBM library provide an inbuilt function to plot features which are ordered by their

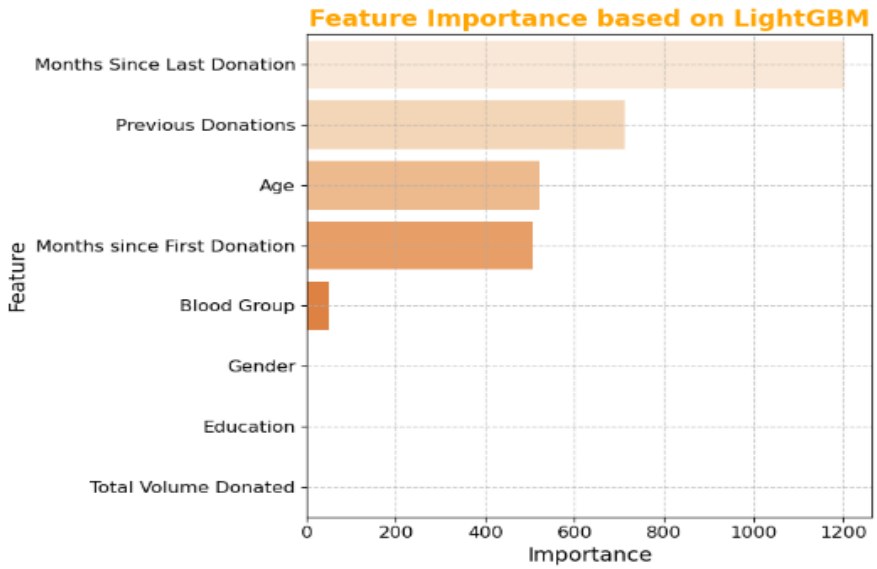
importance(Liang et al., 2019). The XGBoost and Light GBM embedded mechanisms were used to determine feature importance in addition correlation was used to provide an independent oversight and determine feature importance for the variables.

(a) Feature Importance based on Light GBM

The feature importance scores generated by Light GBM ranked months since last donation as the most important factor in predicting whether a donor will return to donate, it was followed by previous donations, age and months since first donation respectively. Blood group, gender, education and total volume donated had the least feature importance scores as shown in Figure 4.15 below.

Figure 4. 15

Feature Importance Based on Light GBM



(b) Feature importance based XGBoost

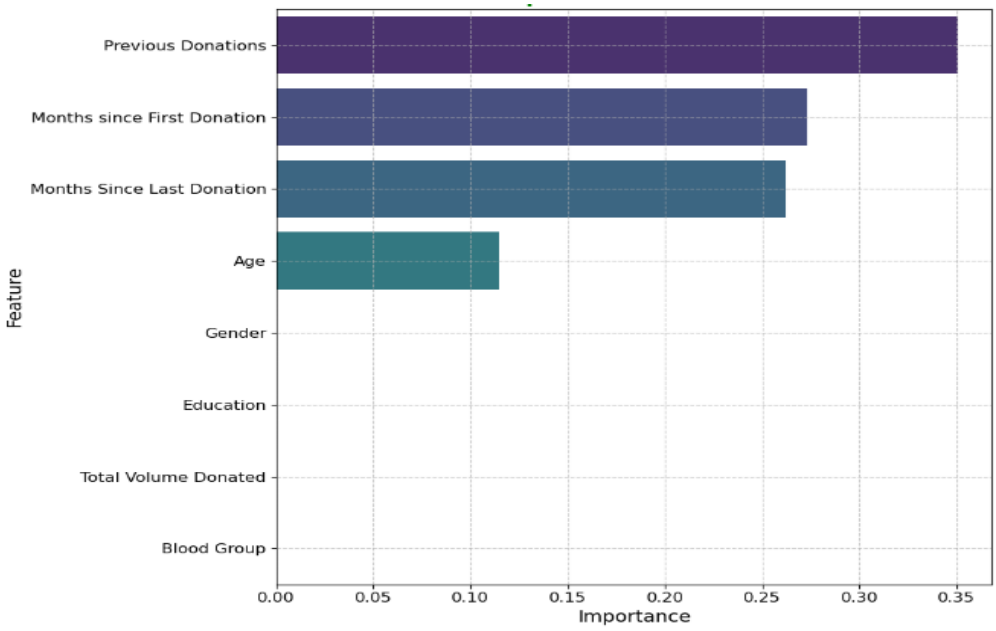
The XGBoost ranked the previous donations as the most important factor followed by months since last donation, months since first donation and age followed in the respective manner.

Gender, education, Total volume donated and blood group were ranked lowest in their order of importance.

Overall, the XGBoost and LightGBM models agreed on the most important features for predicting donor return, months since last donation, months since first donation, number of previous donations and age being the most important factors. Blood group, gender and education level and got the lowest scores indicating their low feature importance. Figure 4.16 below shows the feature importance based on XGBoost.

Figure 4. 16

Feature Importance Scores on Xgboost



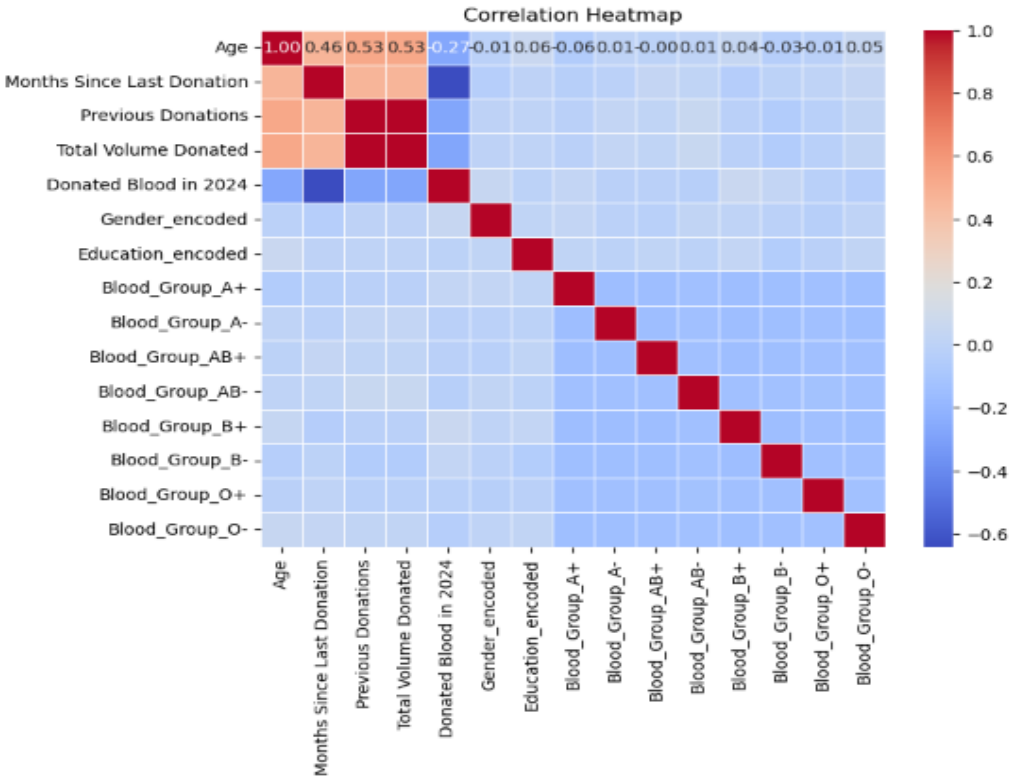
(c) Feature importance based on correlation coefficient analysis

To obtain an independent assessment of feature importance separate from XGBoost and LightGBM inbuilt feature ranking mechanisms, correlation was used. Correlation which is a filter-based technique was used to calculate the correlation coefficient between each feature and the target variable and to identify the best attributes for predicting blood donor retention.

Pearson's correlation was used to measure linear correlation between the variables and the target variable, providing a value between -1 and 1 while Spearman's correlation was used to rank the variables depending on the absolute values of their correlation coefficients. Figure 4.17 below shows the correlation heat map and correlation values.

Figure 4. 17

Correlation Heat Map and Values




```
Correlation with 'Donated Blood in 2024':
Donated Blood in 2024          1.000000
Blood Group_B+                 0.026817
Blood Group_O+                 0.015463
Blood Group_A+                 0.010359
Blood Group_A-                -0.001258
Gender                        -0.002937
Blood Group_O-                -0.004307
Blood Group_AB-               -0.004680
Blood Group_B-                -0.016935
Education                     -0.024313
Blood Group_AB+               -0.026244
Previous Donations            -0.133123
Total Volume Donated          -0.133123
Age                           -0.182847
Months since First Donation   -0.552947
Months Since Last Donation    -0.640210
Name: Donated Blood in 2024, dtype: float6
```

The feature importance based on correlation ranked months since last donation, Months since first donation, age, number of previous donations and total volume donated respectively as the most important features. It also showed that the number of previous donations and the total volume donated were highly correlated.

From the three feature importance determination methods that we considered we can conclude that the months since last donation, Months since first donation, number of previous donations and age are the key features that influence blood donor retention. The other factors received quite low scores highlighting their low significance in predicting blood donor return. They were therefore dropped and the key feature determined were used to train the models.

4.2. Development of the ensemble gradient boosting model for blood donor retention

This section provides a step by step process of how the ensemble gradient boosting model for blood donor retention was developed. The model was developed using Jupiter notebook which is an open source software that is used for implementing machine learning models in Python. Python 3.11.7 was used with libraries including pandas, Seaborn, Scikit-learn, Shap, Lightgbm, and XGBoost. Matplotlib and Optuna were used to create the visualizations so as to better understand the results. The computer used is a Dell Intel Core i5 6th Generation Processor with 2.5 Ghz clock speed, 8gb random access memory and 256 solid state disk.

The development of the model followed the following steps

Step 1. Data is loaded and preprocessed

Step 2. Light GBM and XGBoost Algorithms are trained using their default parameters.

Step 3. Bayesian Optimization is used to find the best set of hyper parameters that minimize the objective function while maximizing accuracy.

Step 4. Models are trained with their best hyper parameters to create the ensemble model.

Step 5. The Performance of the hybrid ensemble model was compared with the performance the individual models and other existing models

4.2.1. Data Loading

Data was loaded into python as a CSV file. The necessary python libraries were imported such as pandas numpy, matplot, and scikit-learn. The dataset was split into features (X) and target (y). The features were standardized using StandardScaler to ensure that each feature has a mean of 0 and a standard deviation of 1. The figure 4.18 shows the python code that was used to load the data, encode the data, preprocess and split the data.

Figure 4. 18

Code Extract For Data Loading and Preprocessing

```
[11]: import pandas as pd
      import numpy as np
      import xgboost as xgb
      import matplotlib.pyplot as plt
      from sklearn.model_selection import train_test_split
      from sklearn.metrics import accuracy_score

      # Load the dataset
      data = pd.read_csv('transfusiondone1.csv')

      # Example: Encoding categorical variables (Gender, Education, Blood Group)
      gender_encoder = {'Male': 0, 'Female': 1}
      data['Gender'] = data['Gender'].map(gender_encoder)

      education_mapping = {'None': 0, 'Primary': 1, 'Secondary': 2, 'Tertiary': 3}
      data['Education'] = data['Education'].map(education_mapping)

      blood_group_dummies = pd.get_dummies(data['Blood Group'], prefix='Blood Group')
      data = pd.concat([data, blood_group_dummies], axis=1)
      data.drop(['Blood Group'], axis=1, inplace=True)

      # Separate features and target variable
      X = data.drop(['Donated Blood in 2024'], axis=1)
      y = data['Donated Blood in 2024']

      # Split data into training and testing sets
      X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

4.2.2. Model Selection

Light Gradient Boosting and Extreme gradient boosting algorithms have been selected for this task. This is due to their proven accuracy and robustness in handling structured data specifically in binary classification tasks. The XGBoost which is a decision tree bosting ensemble is an advanced version of gradient boosting machines. It implements regularization enabling it to effectively combat overfitting. Additionally, it has inbuilt mechanisms for handling imbalanced dataset(Chen & Guestrin, 2016). The light GBM is a very fast gradient boosting technique that is also based on decision trees. The Light GBM uses the leaf wise strategy with a maximum depth and histogram based technique to increase training speed and reduce memory consumption. It also uses one side gradient sampling as well as does exclusive feature bundling (EFB), which enables it to handle large datasets and high-dimensional data effectively(Guolin Ke et al., 2017). These strategies make Light GBM to be superior in improving accuracy and computational efficiency which is required in the prediction of blood donor retention. Combining these two strong models enables us to effectively handle the challenges that come with our imbalanced dataset, reduce overfitting. As well as bring robustness, efficiency and the improved performance required for this task.

4.2.3. Training Process

A systematic training process was followed to rigorously train the model. The data was divided into 80% training and 20% testing. The two gradient boosting algorithms light gradient boosting algorithm and Extreme gradient algorithms were imported and trained based on their default parameters. K fold cross validation was utilized with k=5. This means that the training dataset was split into 5 equal-sized folds, and the model was trained and evaluated 5 times, each time using a different fold as the validation set and the remaining folds as the

training set. This process allows for a more reliable estimation of the model's performance compared to a single train-test split. The models were first trained using their default hyperparameters.

After training and k-fold cross validation, The XGBoost achieved an accuracy 0.9840, the Light GBM achieved 0.9830 while the hybrid ensemble model achieved an accuracy of 0.9830. This indicates that the models were effective in predicting the target variable in our dataset. Figures 4.19. and 4.20. show the confusion matrix for XGboost and Light GBM respectively.

Figure 4. 19

Xgboost Confusion Matrix

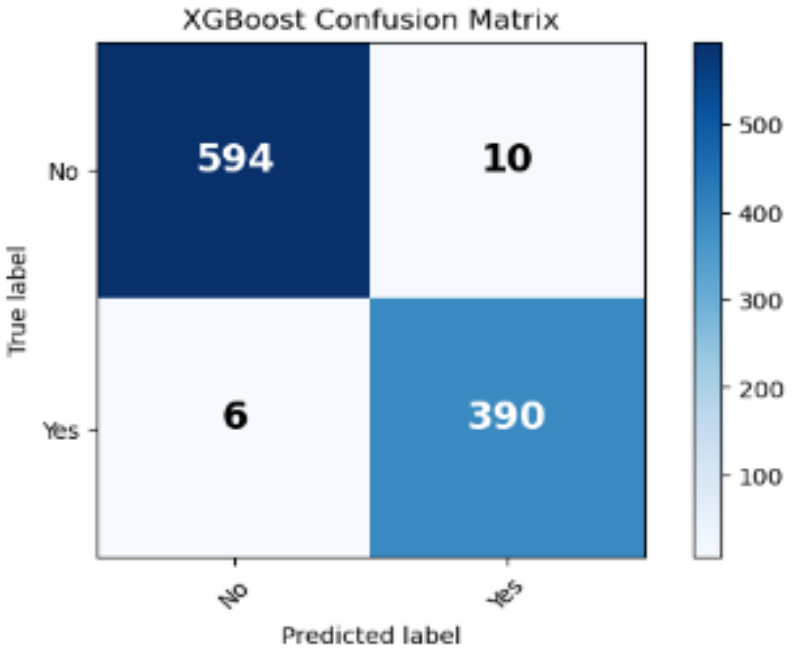
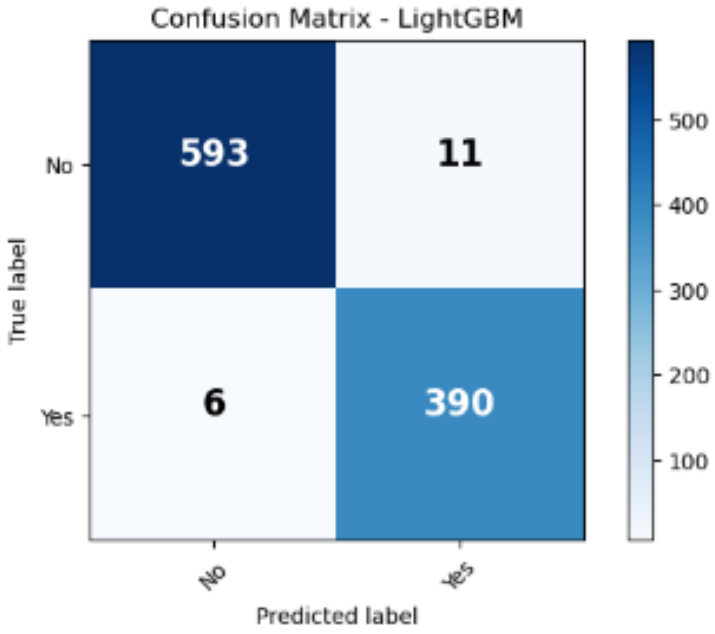


Figure 4. 20

Light GBM Confusion Matrix



The two models also performed well in other metrics with XGBoost achieving a precision 0.9750 of recall 0.9848 and F1 score of 0.9799. while the light GBM achieved a precision 0.9726 of recall 0.9848 and F1 score of 0.9787 as shown of figure 4.21. and 4.22. below.

Figure 4. 21

Xgboost Performance Metrics for Before Optimization.

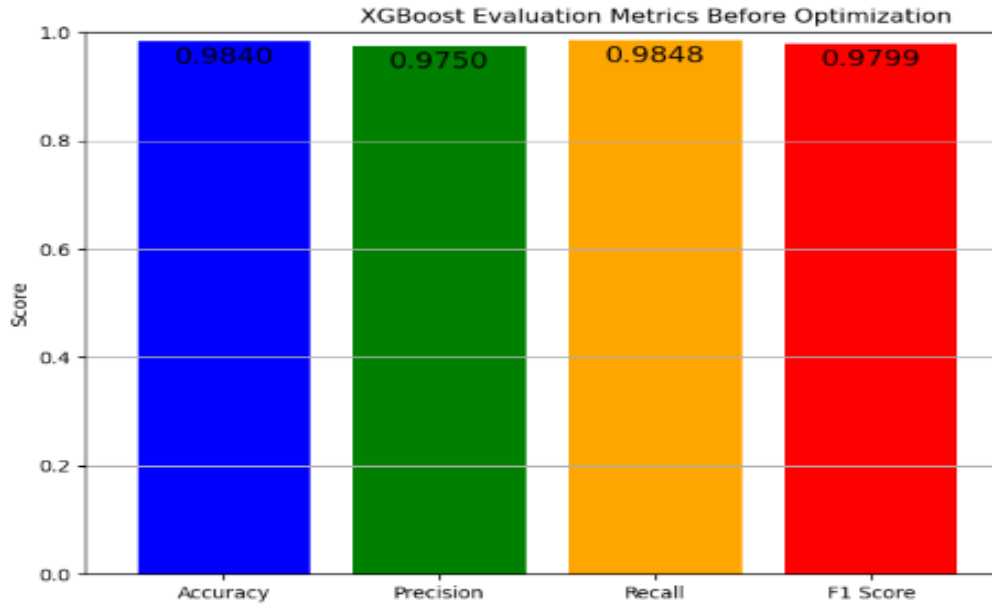
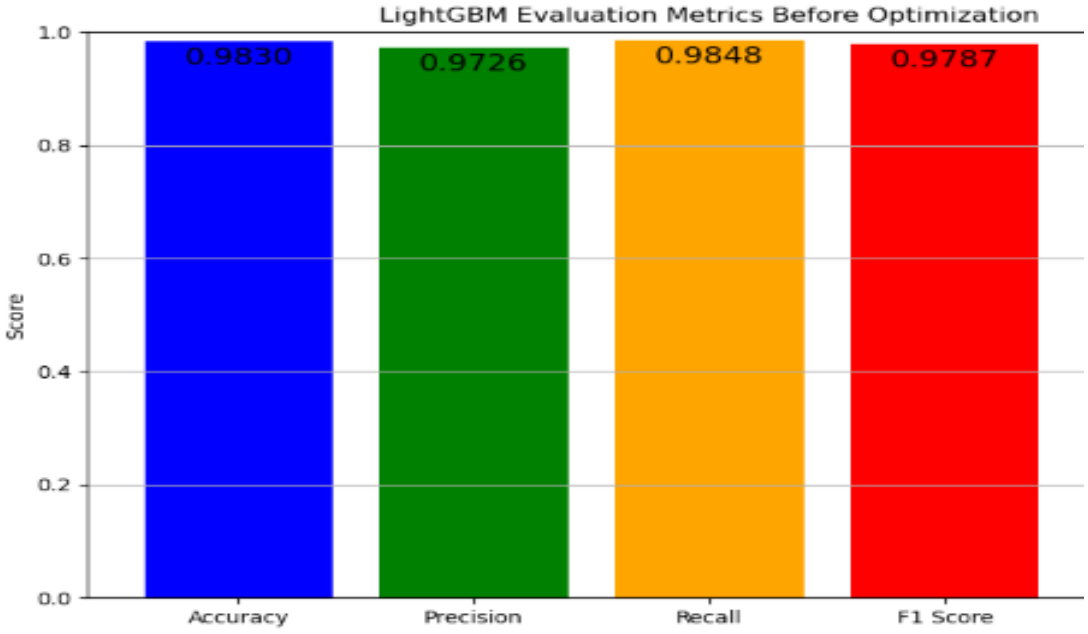


Figure 4. 22

Light GBM Performance Metrics for Before Optimization

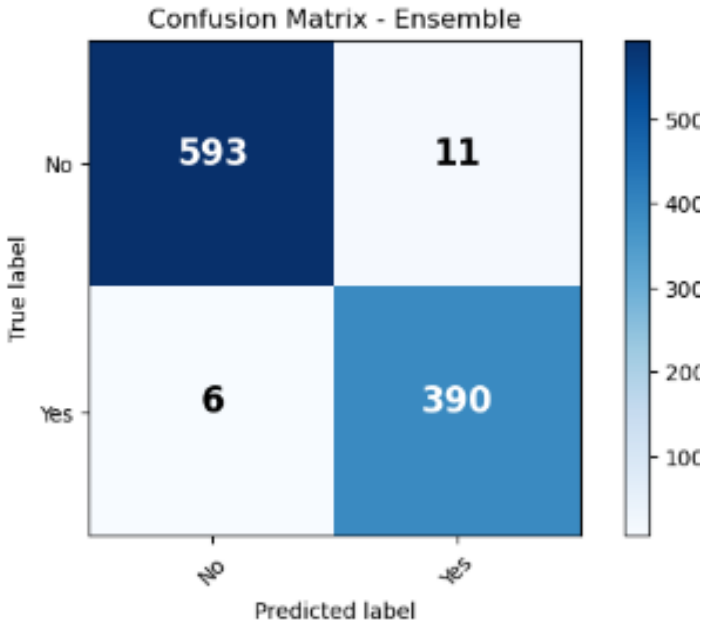


4.2.4. Creation of the ensemble Model

The hybrid Ensemble model was created by training and combining the individual base models in parallel and building a super learner. The XGBoost and the Light GBM were fitted into the dataset as base learners and K-fold cross validation was implemented on the training set with k=5. The out of fold predictions for each of the base models are combined to create a meta data matrix for the meta model. The combinations were then weighed by a weight vector and the optimal value of the weights was selected which reduces the risk on the cross validation for the individual base learners. Finally, the optimal weights obtained were combined in order to construct the hybrid light GBM and XGBoost Model. The ensemble model achieved a performance of accuracy of 0.9900, a precision of 0.9726, recall 0.9848 and F1 0.9787 score of as shown in the confusion matrix in figure 4.23. below. Although these results demonstrated a strong performance, we aimed to further enhance and maximize the models' capabilities by tuning their hyperparameters. Given the critical nature of blood donation, even minor improvements in model performance can significantly impact donor retention rates.

Figure 4. 23

Hybrid Ensemble Model Confusion Matrix Before Optimization



4.3. Optimization of the Hybrid Ensemble model

This section details how the performance of the ensemble gradient boosting model was enhanced. The performance was enhanced by tuning the hyperparameters to get the best set of hyperparameters that reduces the objective function. Both Light GBM and XGBoost parameters were optimized and the ensemble model trained based on the best hyperparameters for the two models.

4.3.1. XGBoost and Light GBM Hyperparameters

Hyper parameters are values and weights that determine the learning process of a model. Both XGBoost and Light GBM provide a wide range of hyper parameters which can be leveraged to maximize performance of the model(Claesen & De Moor, 2015). Table 4.2. shows the Light GBM and XGboost Hyperparameters.

Table 4. 2

Light GBM and XGboost Hyperparameters

Hyper parameter	Description	Effect	Range	Default XGBoost	Default Light GBM
Number of Estimators	The number of trees in the ensemble	Increasing can improve accuracy with large data	100 - 1000	100	100
Learning rate	Controls the pace at which the model learns	Decreasing prevents overfitting	0.01 - 0.5	0.3	0.1
Maximum tree depth	The maximum depth of each decision tree	Decreasing prevents overfitting	3 - 10	6	-1 (no limit)
Subsample/ Bagging fraction	Subsample ratio of the training instances..	Decreasing prevents overfitting	0.5 - 1	1	1
Column sample by tree	subsample ratio of columns when constructing each tree.	Decreasing prevents overfitting	0.5 - 1	1	1
α (L1 Reularization)	L1 (Lasso Regression) regularization on weights	increasing prevents overfitting	0 to 1	0	0
λ (L2 Reularization)	L2 (Ridge Regression) regularization on weights.	increasing prevents overfitting	0 to 1	1	0
min_child_ weight	defines the minimum sum of weights of all observations required in a child	increasing prevents overfitting	1 - 10	1	1e-3
Gamma / min_split_g ain	Minimum loss reduction that is required to make a partition on a leaf node.	increasing reduces overfitting	0 - 1	0	0

scale_pos_ weight	controls the balance of positive and negative weights	adjusts the balance between positive and negative classes	Depends on class imbalan ce	1 for Balanced	1 for Balanced

4.3.2. Bayesian Optimization

Bayesian Optimization was implemented to search and tune the hyper parameters for Light GBM and XGBoost so as to find the best set of optimal parameters for each of the models that minimizes the objective function. Bayesian optimization is a powerful hyperparameter optimization technique that iteratively determines hyper parameters based on the previous iterations results. This saves computational time and reduces the model complexity as compared to random or grid search. The objective function for the base models was to minimize log loss whilst improving accuracy. Since the prediction of blood donor retention is a binary classification problem the objective function was set to binary: logistic. The optimization of the models was conducted by performing different sets of trials. To reduce the computational time taken to search for the hyperparameters. The trials were conducted in sets of between 100 and 205 trials. This was done using cross validation with k=5. A total of over 1000 trials were conducted. The trial no 110 is the one that obtained the minimum log loss of 0.0346 for the XGBoost. The light gradient boosting achieved the lowest log loss of 0.0339 in iteration number 139. A lower log loss suggests that the predicted probabilities are close to the actual class labels, demonstrating better accuracy of the model. The figures 4.24. and 4.25. show the optimization history for XGBoost and light GBM respectively for the set that achieved the minimum log loss.

Figure 4. 24

Hyperparameter Optimization History for XGBoost

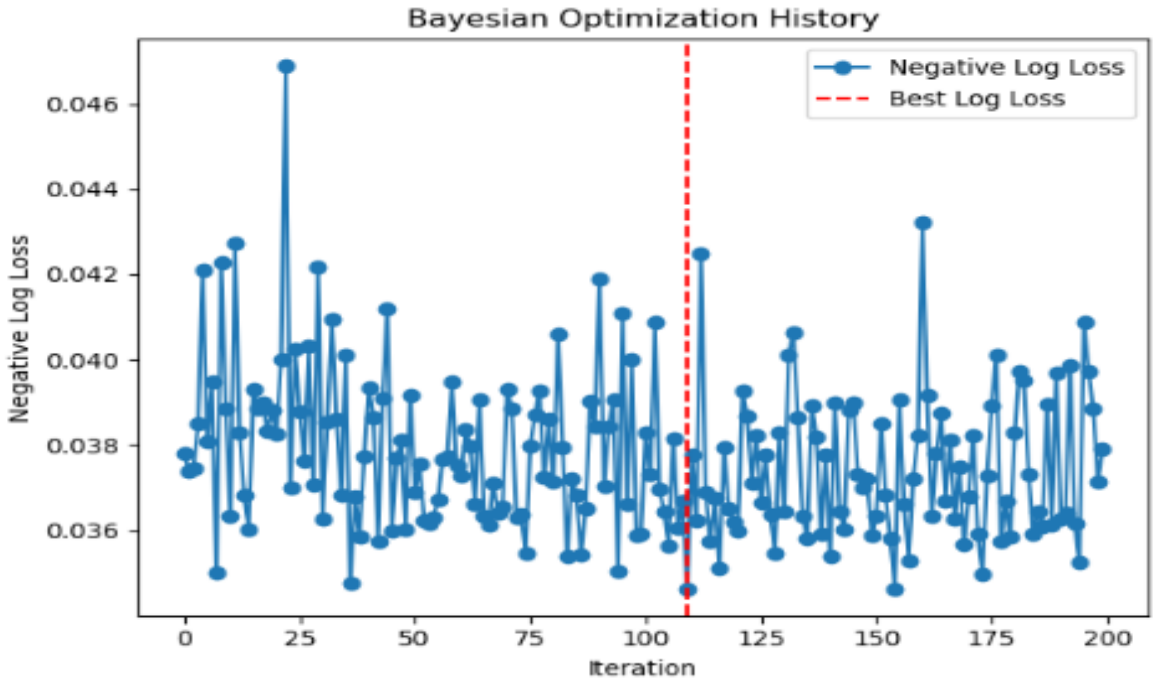
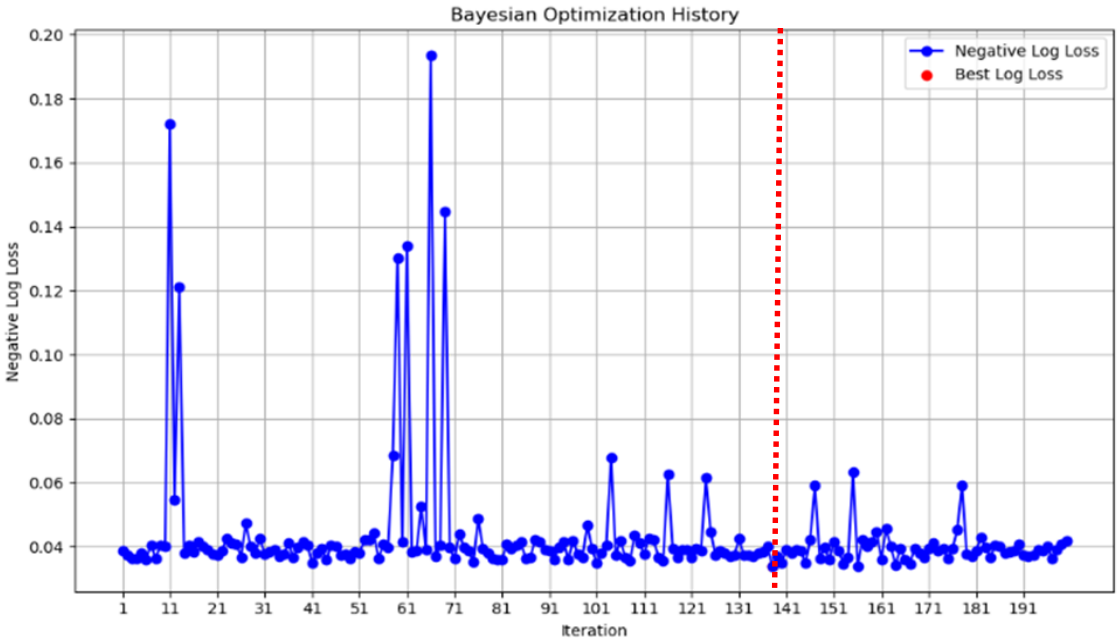


Figure 4. 25

Hyperparameter Optimization History for Light GBM



4.3.3. Summary of base learners hyperparameters

After more than 1000 trials the best optimal hyperparameters were obtained. The table 4.3 shows a summary of the best set of hyperparameters that were able to achieve the minimum error through optimizing the objective function for the base learners.

Table 4. 3

Best Set of Hyper Parameters for Xgboost and Light GBM

Hyper parameter	**XGBoost**	**Light GBM**
Number of Estimators	454	140
Learning rate	0.21526	0.05980
Maximum tree depth	6.00000	9.91125
Subsample/Bagging fraction	0.88669	0.75712
Column sample by tree/Feature fraction	0.69572	0.82064
α (L1 Reularization)	0.33527	0.99250
λ (L2 Reularization)	0.91214	0.54633
Minimum Child weight	1.00000	7.04605

4.3.4. Hyper parameter importance

The set of hyperparameters which achieved the lowest minimized objective function had different impact on the performance of the base models. The number of estimators had the highest impact in XGboost followed by the Maximum depth. Other hyperparameters such as Learning rate, column sampling by tree, sub sample, L1 and L2 regularization had the lowest impact and contributed less than 10% to improving the performance of the model. For the light GBM the number of estimators, maximum depth, column sample by tree and subsample had the highest impact. Although the other hyperparameters has lowest impact on the

performance of the models this does not mean that they did not have any contribution on improving the performance. Their small contributions were very necessary in optimizing the objective function in combination with the other hyperparameters that had a higher impact. The hyperparameter importance scores further guided our hyperparameter tuning focusing on optimizing the parameters that had higher significance to achieve better predictive performance. Figure 4.26. and 4.27. shows the hyperparameter importance for the two base models.

Figure 4. 26

Xgboost Hyperparameter Importance

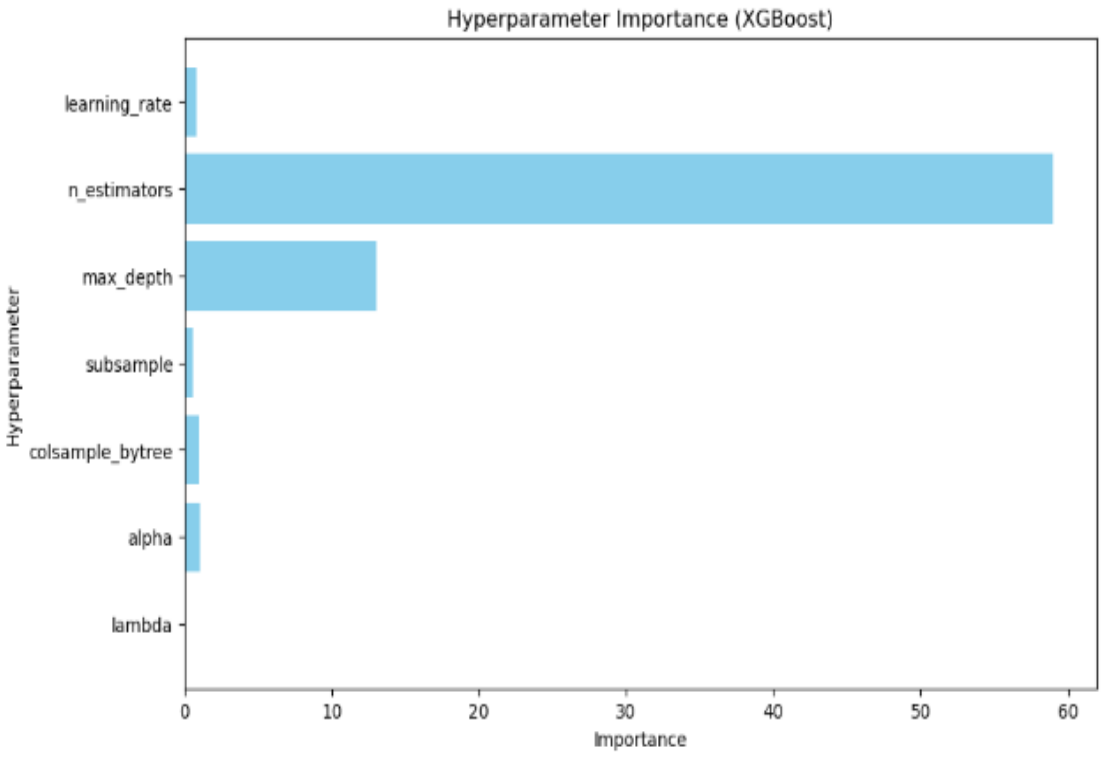
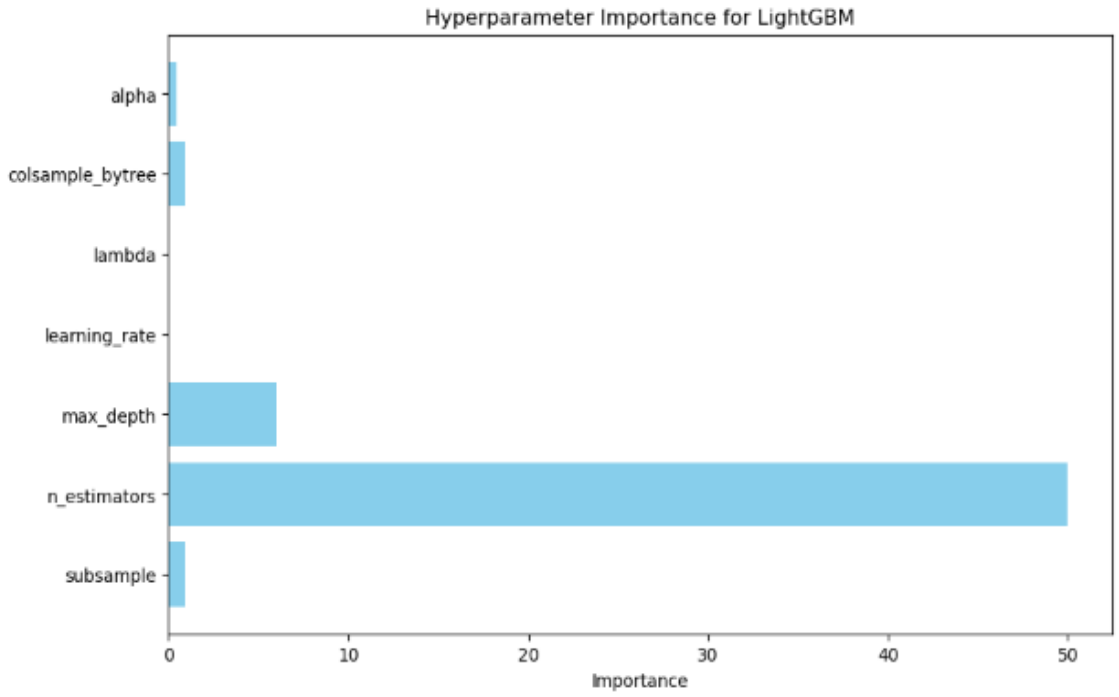


Figure 4. 27

Light GBM Hyperparameter Importance



4.3.5. Learning Curves

Learning curves show the models performance over time. The learning curves below display the performance of the model on both the training and validation sets as a function of the number of training iterations.

(a) Accuracy Learning Curves

We observe that the learning curve for the training accuracy initially surpasses the cross-validation curve for both models. The learning curves for both XGBoost and Light GBM models show considerable increase in accuracy on both training and cross validation sets as the number of iterations increase. This shows that the models were able to learn effectively as the amount of data increases was able to learn the underlying patterns and generalize from the

training data hence able to improve the accuracy. Figure 4.28. and Figure 4.29. shows XGBoost and light GBM accuracy learning curve.

Figure 4. 28

Xgboost Accuracy Learning Curve

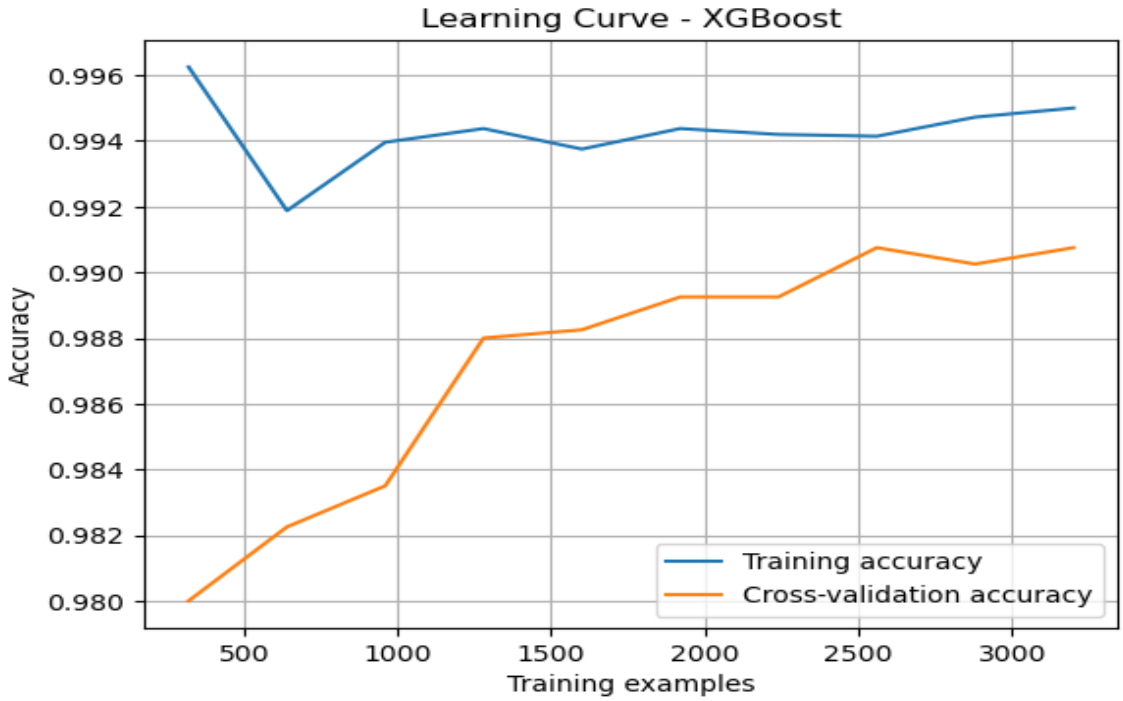
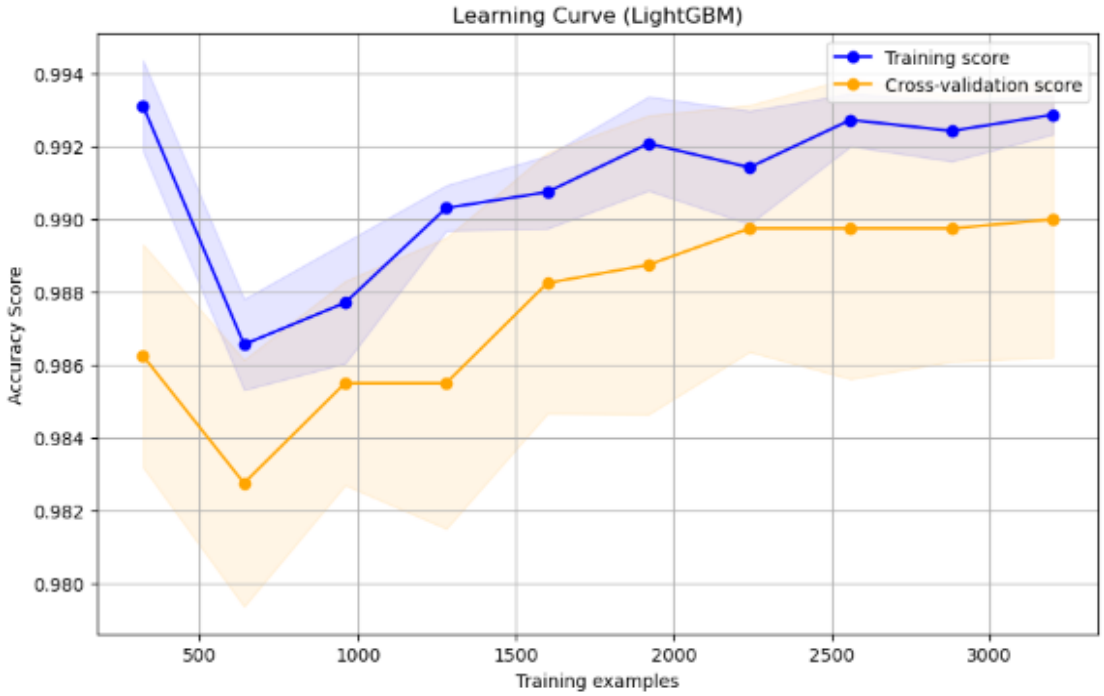


Figure 4. 29

Light GBM accuracy Learning Curve



(b) Log Loss Learning Curves

Both the XGBoost and LightGBM models learning curves show that there is a significant decrease in log loss on training and cross-validation sets as the number of iterations increases. The reduction in log loss signifies an improvement in the models' ability to generalize from the training data to unseen data and make accurate predictions. This enhanced generalization leads to more reliable performance ultimately contributing to the overall robustness and effectiveness of the models. The Log loss learning curves for both XGBoost and light GBM are shown in figure 4.30. and 4.31. respectively.

Figure 4. 30

Xgboost Log Loss Learning Curve

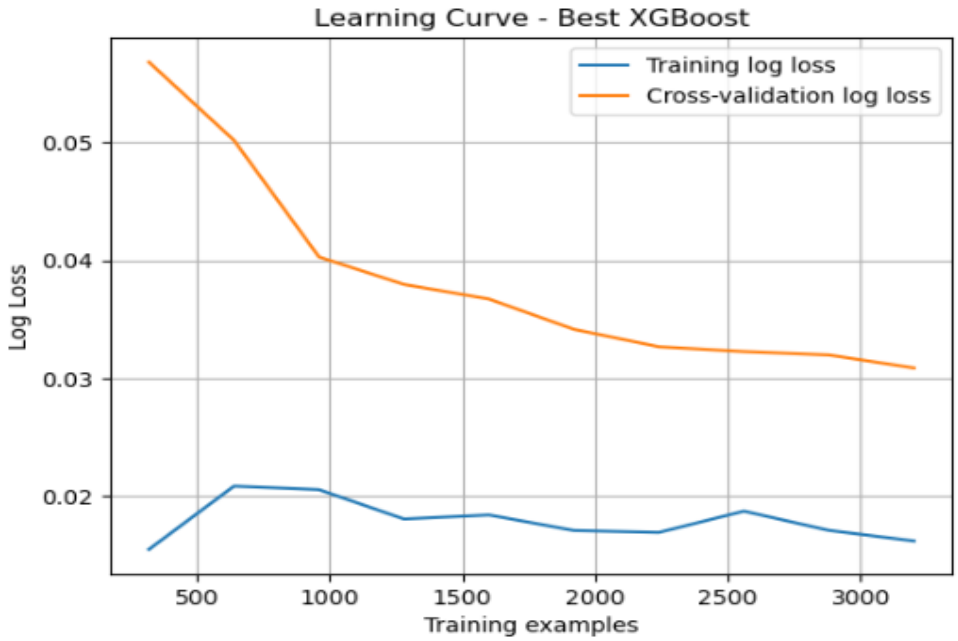
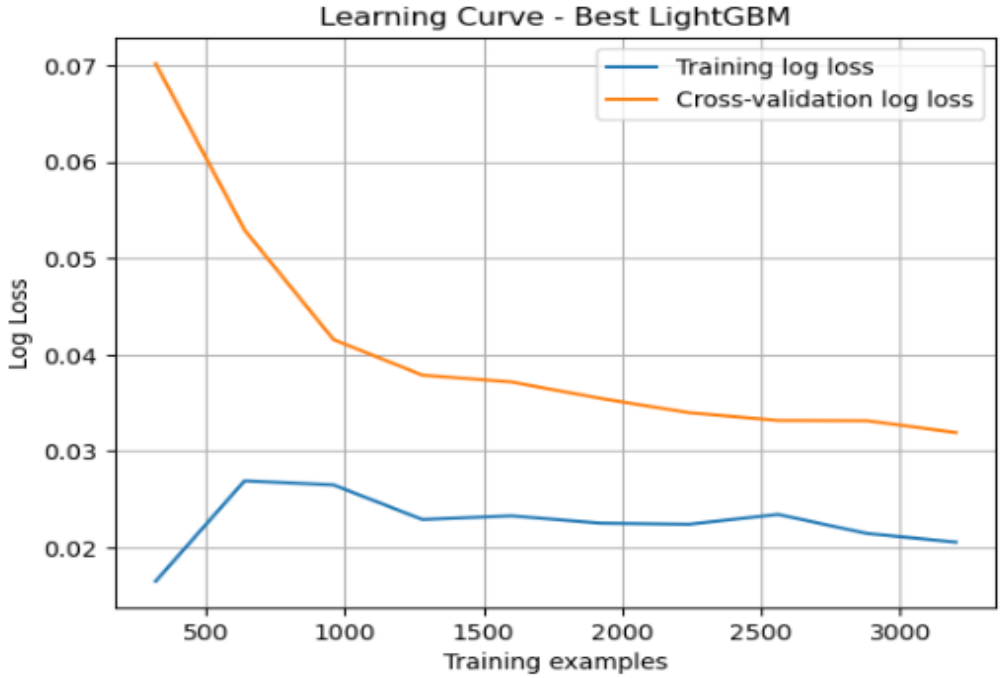


Figure 4. 31

Light GBM log loss Learning curve



4.3.6. Summary of the models before and after optimization

The created hybrid ensemble model showed a better performance in optimizing the objective function which was to minimize the log loss while improving the accuracy. Table 4.4.6. shows the log loss comparison of the model with default parameters and the log loss attained by the hybrid ensemble model after implementing Bayesian optimization to the base learners. The hybrid ensemble model's log loss decreased from 0.0406 to 0.03207 while the accuracy also increased from 0.9830 when using the default XGBoost and Light GBM hyperparameters to an accuracy of 0.99 after Bayesian optimization.

Table 4.4. Performance Evaluation of the hybrid XGBoost and Light GBM before and after optimization

Table 4. 4

Performance Evaluation of the Hybrid Xgboost And Light GBM Before and After Optimization

	Log loss		**Accuracy**	
Model	Without BO	With BO	Without BO	**With BO**
XGBoost	0.0407	0.03461	0.9840	**0.9910**
Light GBM	0.0416	0.03397	0.9830	**0.9880**
LGBM-XGBoost Hybrid Ensemble	0.0406	0.03207	0.9830	**0.9900**

4.4. Performance validation of the ensemble gradient boosting model for blood donor retention

4.4.1. Introduction to model validation

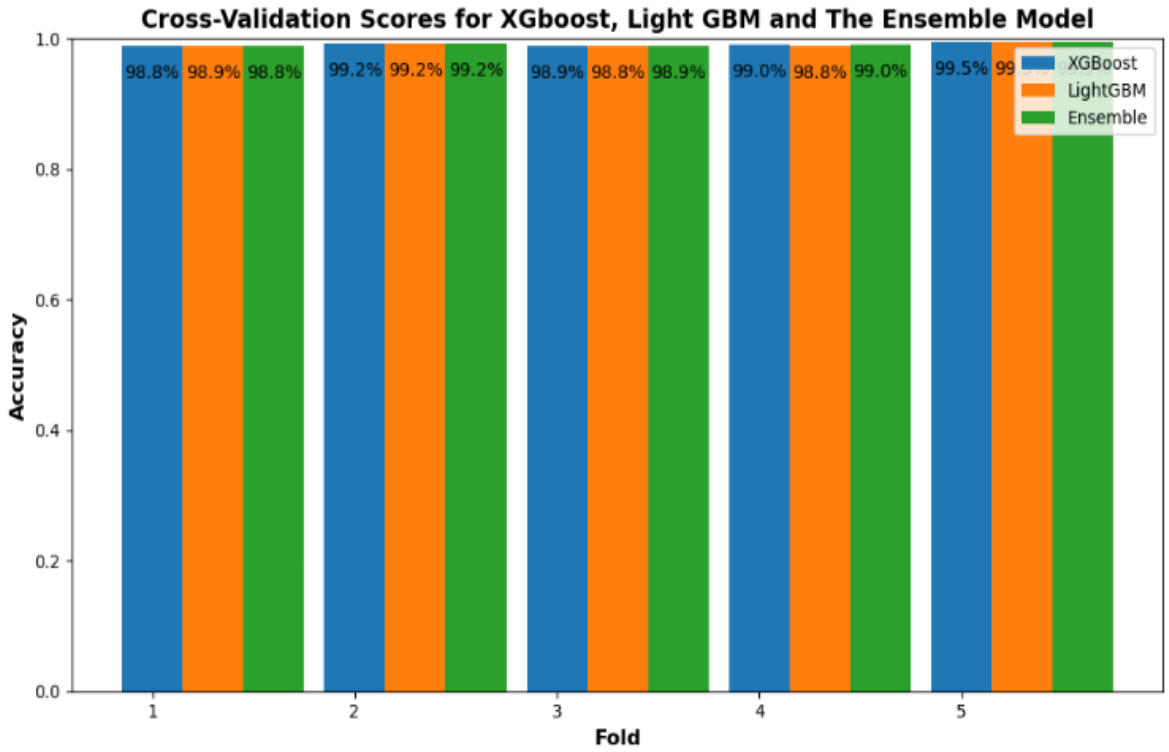
In this step the validation of the model is conducted. Model validation is an important step to ensure that the model is able to generalize well on new data and to ensure that the model does not over fit and that the model is able to effectively predict whether a donor will return for donations.

4.4.2. Cross-Validation Technique

Stratified k-fold Cross validation with K=5 was utilized to train the model so as to maintain class balance across all folds and provide reliable estimation of the model performance. The CVScores visualizer from Yellowbrick library was used to the perform cross-validation and visualize the accuracy scores for each fold. This assisted in understanding how the model's performance varied across different subsets of the data and provided insights into its generalization capabilities. The model performed well in all the folds with a score of over 90%. The cross validation scores are shown in the figure 4.32. below.

Figure 4. 32

Cross Validation Results



4.4.3. Performance Evaluation Metrics

Several performance metrics were used to evaluate the performance of the classifiers. They include, accuracy, precision, recall and F1 score which were calculated from the confusion matrix, Log loss, ROC curve (receiver operating characteristic curve) and AUC: Area Under the ROC Curve were also used to evaluate the performance. Figure 4.33. and 4.34. shows XGBoost and Light GBM Confusion Matrix after Bayesian optimization

Figure 4. 33

Xgboost Confusion Matrix After Optimization

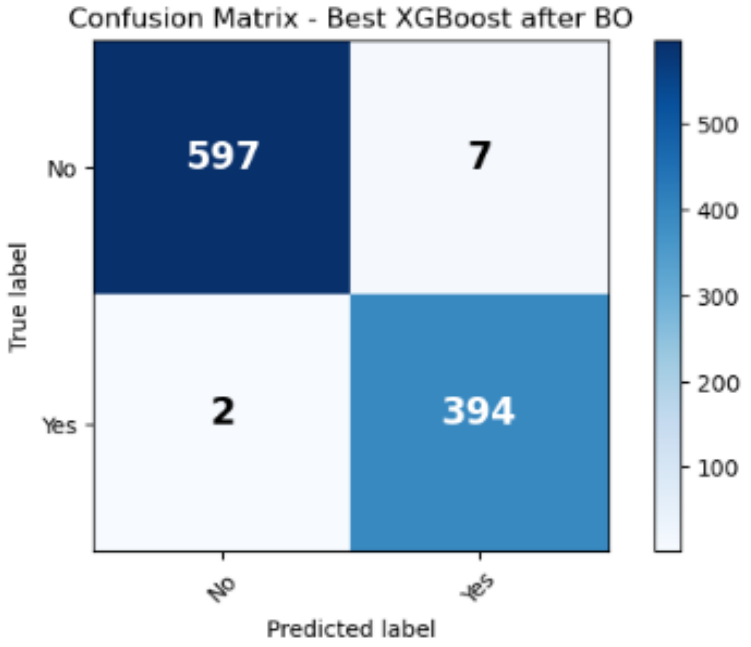
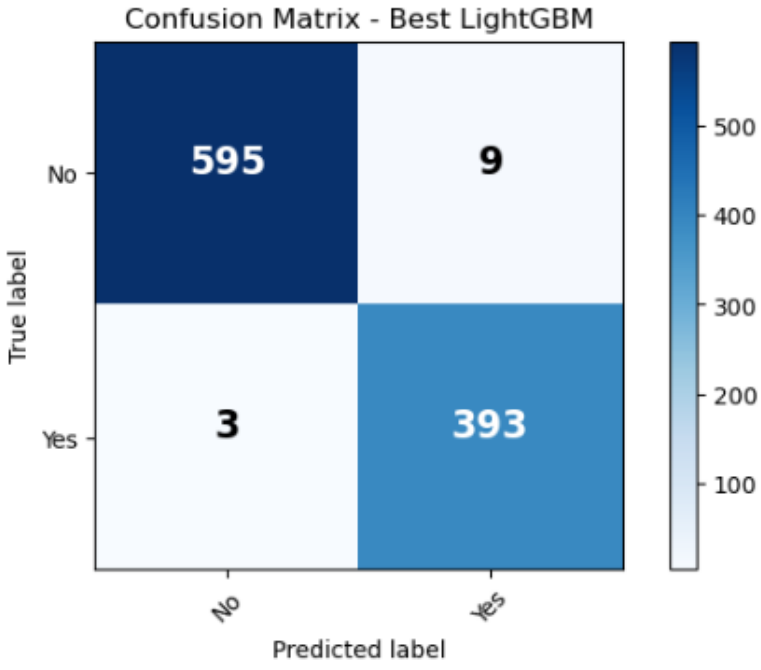


Figure 4. 34

Light GBM Confusion Matrix After Optimization



From the confusion matrix the accuracy achieved by the models was calculated as shown below.

XGBoost Accuracy = $\frac{TP+TN+}{FP+FN+TP+TN} = \frac{597+394}{1000} = 0.991 = 99.1\%$

Light GBM Accuracy $\frac{595+393}{1000} = 0.988 = 98.8\%$

4.4.4. Receiver operating characteristic (ROC) curve and AUC-ROC

The receiver operating characteristic (ROC) curve is a graph that shows the performance of a classification model at all the classification thresholds. ROC is a standard technique that is used for summarizing a classifiers performance over a range of trade-offs between the true positive rate (Sensitivity) against the false positive rate (1 - Specificity) for various threshold values(Vujovic, 2021). The ROC curve shows that the both XGBoost, Light GBM and the Ensemble Model have a high ROC and therefore are able to distinguish between the two classes of donors and non-donors. Figure 4.35. and Figure 4.36. shows ROC Curve with AUC for XGBoost and light GBM after optimization

Figure 4. 35

ROC Curve and AUC for XGBoost after optimization

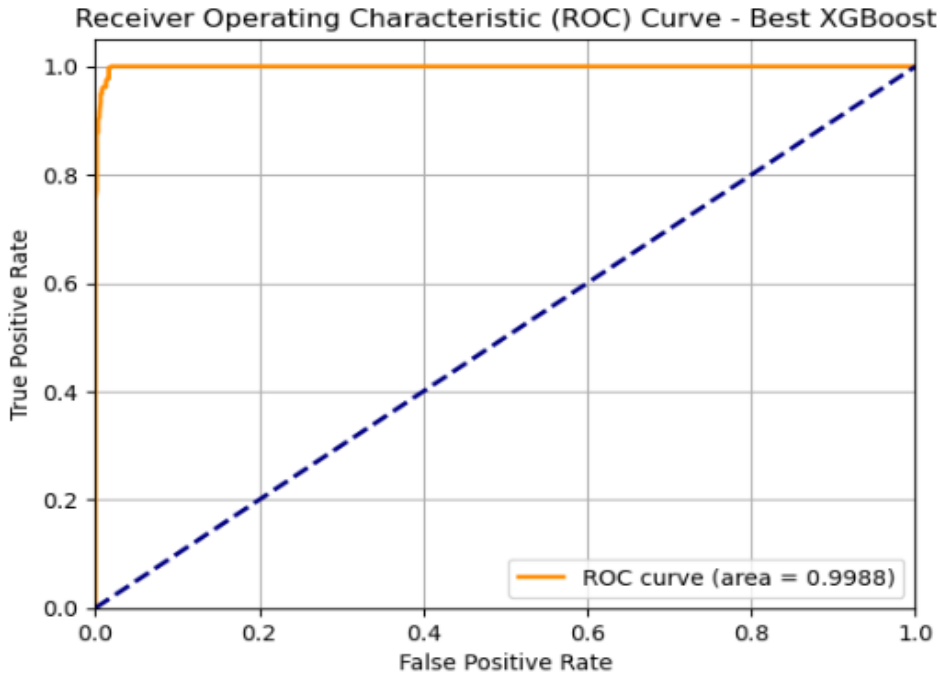
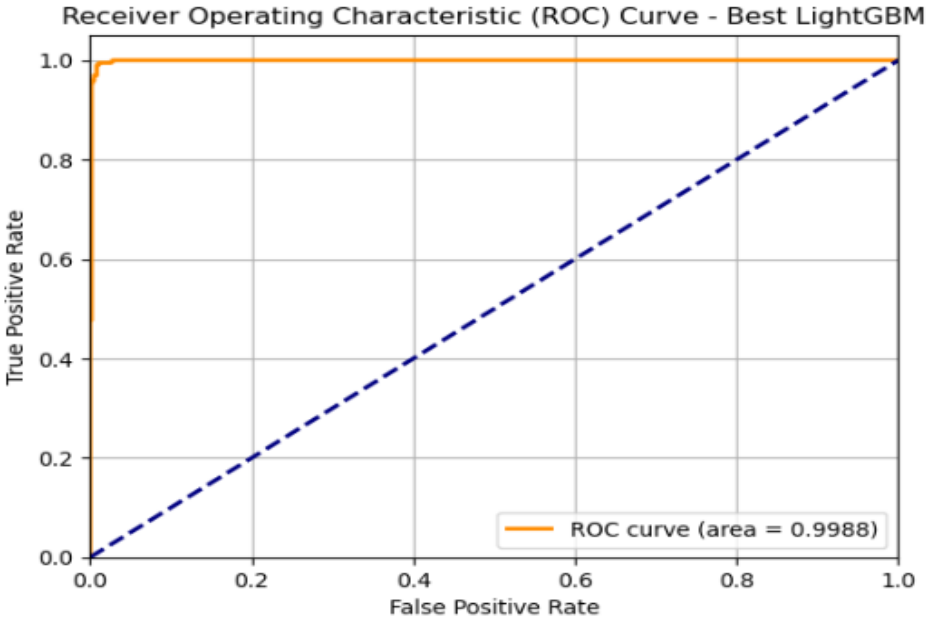


Figure 4. 36

ROC Curve and AUC for Light GBM after optimization



4.4.5. AUC-ROC for the base models after optimization

Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is used to measure the entire two-dimensional area under the entire ROC curve. It is mostly used to measure the ability of a machine learning model to differentiate between two classes typically the positive class and the negative class(Vujovic, 2021). The higher the AUC values the better prediction ability the model. The blood donor data used for this study was imbalanced with more donors being non-returning donors. In the case of such imbalanced data AUC can be a better metric since it is less susceptible to the class imbalance and hence provides a more robust measure of performance. The base models obtained an AUC of 0.9988 and 0.9988 for XGboost and the Light GBM respectively. This shows that the models are extremely good in distinguishing between the positive and the negative classes which is very crucial in blood donor retention.

4.4.6. The Ensemble Model after optimization

The ensemble model was created using the VotingClassifier from scikit-learn's ensemble module. The Two base models that were used for the ensemble are the best LGBM (LightGBM) model and the best XGBoost (Extreme Gradient Boosting) model. These models were previously trained, optimized and saved. The VotingClassifier was initialized. The voting parameter was set to 'soft', which means that the predicted probabilities of the models are weighted across the models. After the ensemble model was initialize, it was trained on the standardized features (X_scaled) and the target variable (y). The soft voting mechanism automatically determines the weights for each base model based on their predictive performance, allowing the ensemble model to leverage the strengths of each individual model while mitigating their weaknesses. The ensemble model was able to achieve an improved

accuracy of 0.9900. Figure 4.37. and 4.38. shows the performance of the ensemble model through the confusion matrix and the ROC- AUC respectively.

Figure 4. 37

Confusion matrix for the ensemble model after optimization

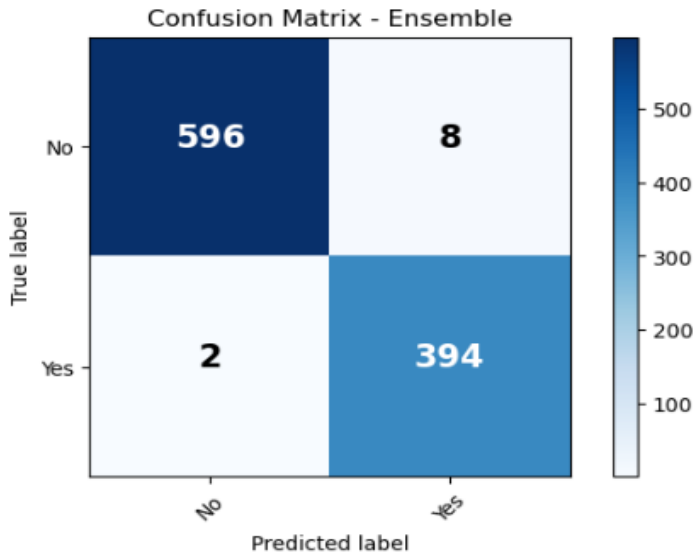
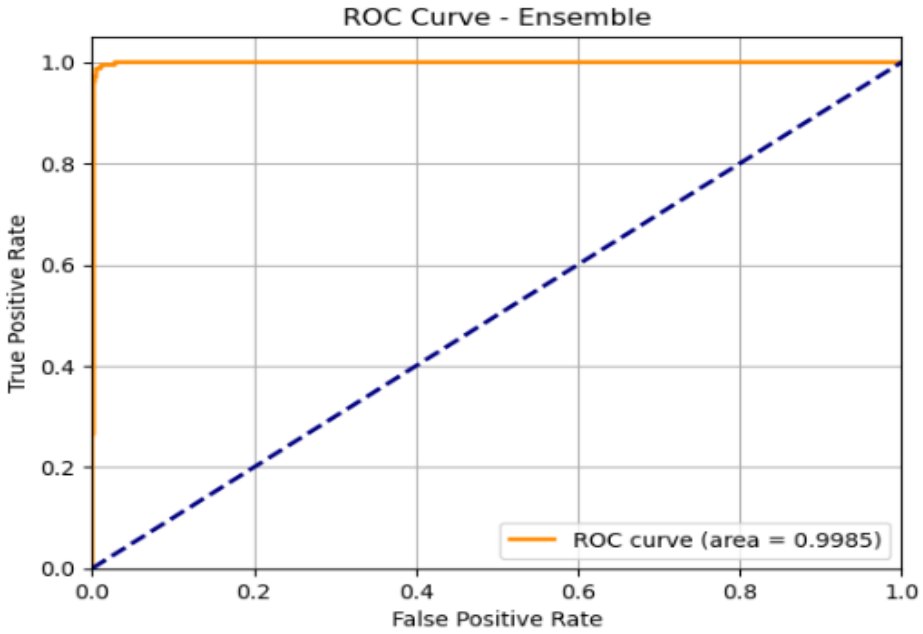


Figure 4. 38

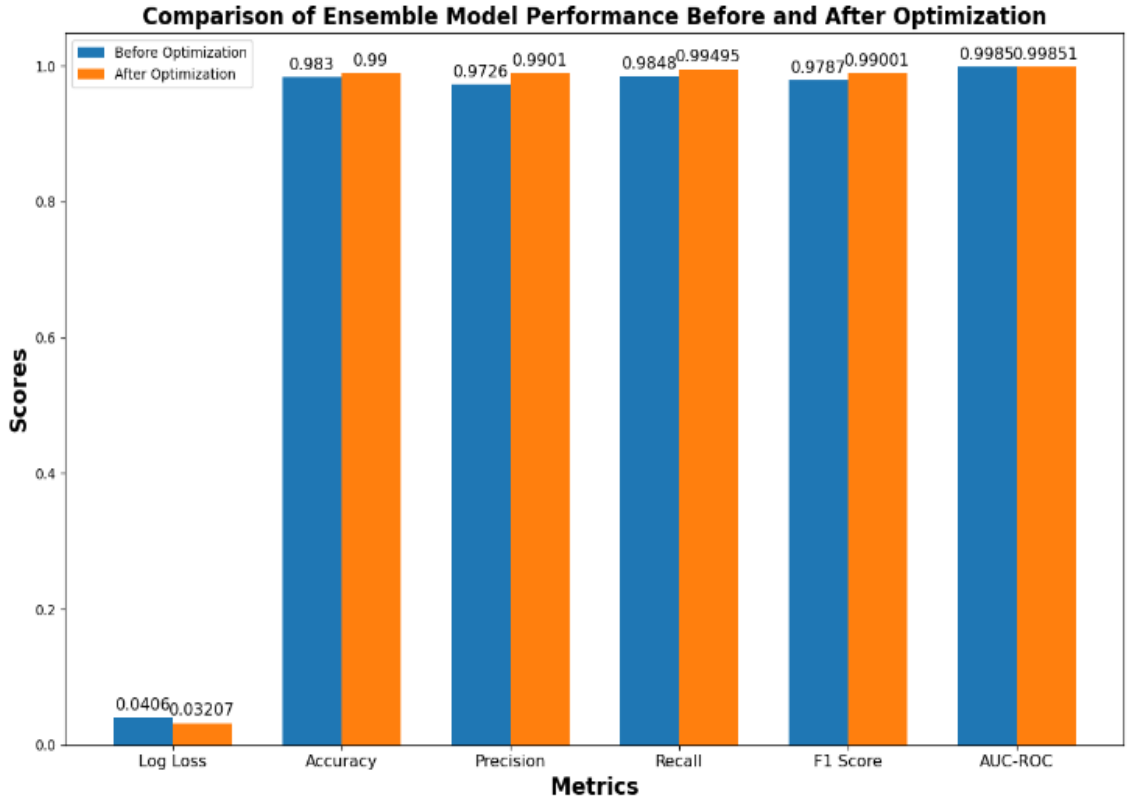
ROC Curve with AUC for the ensemble model.



The ensemble model performed well on all the performance metrics attaining an accuracy of 0.9900, precision of 0.9901, recall of 0.9950, F1 score of 0.9900 and a ROC-AUC of 0.9985. The hybrid ensemble model was able to improve in performance for all the metrics after the optimization. Figure 4.39. shows the comparison of the ensemble model performance for all the metrics before and after the model optimization.

Figure 4. 39

Summary Comparison for the Hybrid Ensemble Model Before and After Model Optimization

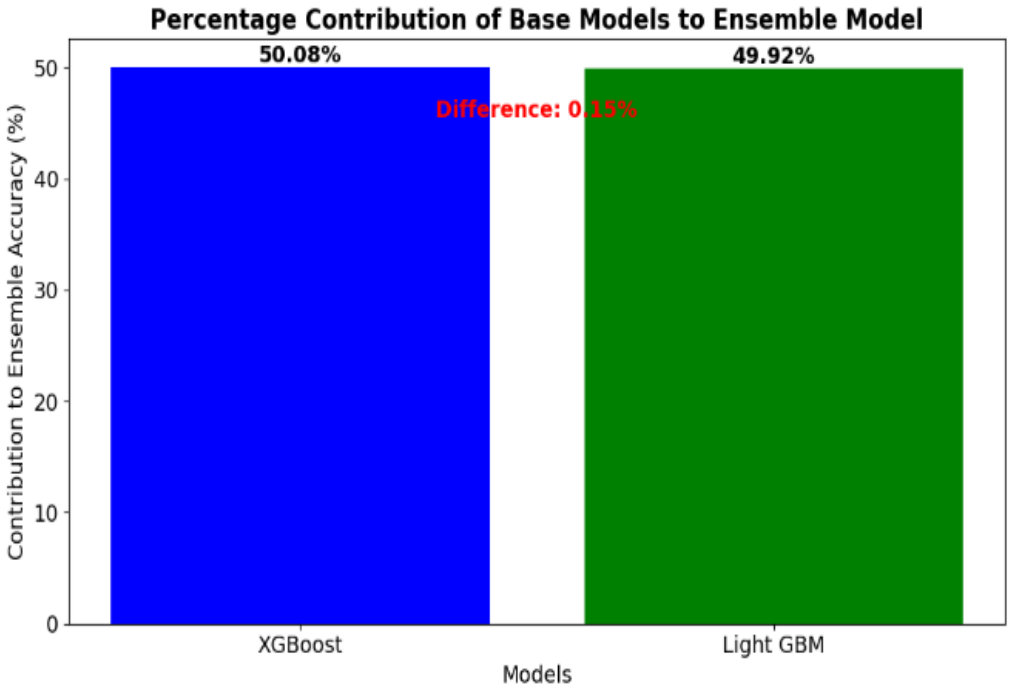


4.4.7. Contribution of the base models

The hybrid ensemble model had an almost equal contribution from both XGBoost and light GBM The percentage contribution for XGBoost was 50.08 while Light GBM contribution was 49.92. there was a slight difference of 0.15. therefore, the base models had almost the same influence and impact on the Hybrid ensemble model output which resulted to improved performance, stability and robustness of the model. Figure 4.40. below shows the contribution of each of the base learners to the final output of the Ensemble hybrid model.

Figure 4. 40

Contribution of Each Base Learner to the Overall Output of the Hybrid Model



4.4.8. Hybrid model interpretation

SHAP plot was used to compute shap values for the contribution of each of the features to the hybrid model performance. Shapley Additive explanation (SHAP) plot is a type of

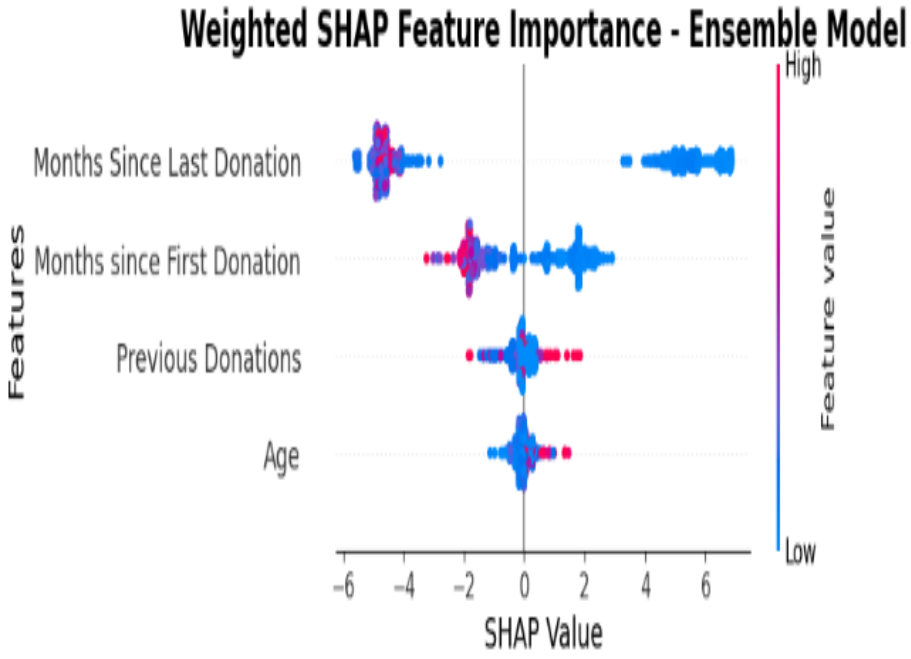
visualization that is used to interpret the output of a machine learning model. It helps to understand the impact of features on model predictions.

Recent donors (low months since last donation) were associated with a higher probability of a blood donor being predicted as likely to return for donation, while a higher number of previous donations also resulted in a higher probability of a blood donor being predicted as likely to return for donation.

The SHAP values for age indicate mixed impact on the model's predictions. This suggests that age is likely to both positively and negatively influence the likelihood of blood donor retention. This is because young donors are the majority of donors but there are also several older people who have done quite a number of repeat donations and therefore have a higher probability of returning for donations. Figure 4.41. shows a weighted SHAP feature importance for the hybrid ensemble model.

Figure 4. 41

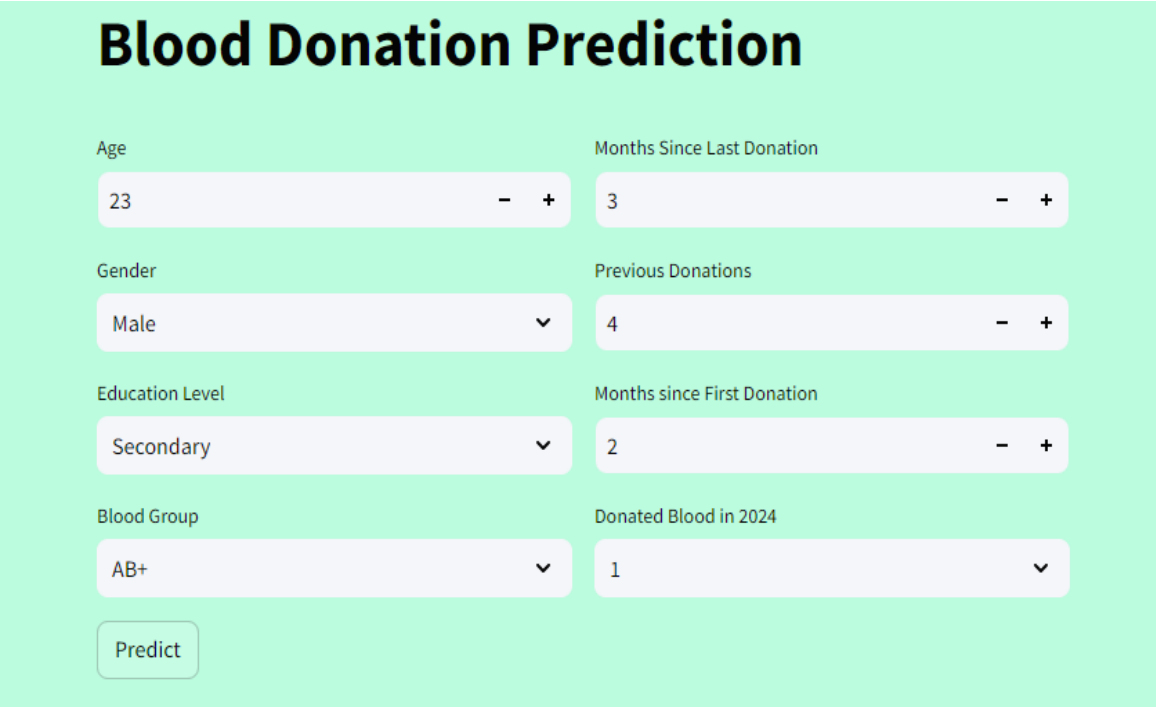
Hybrid Ensemble Model Feature Importance Summary.



4.4.9. Model User Interface

Figure 4. 42

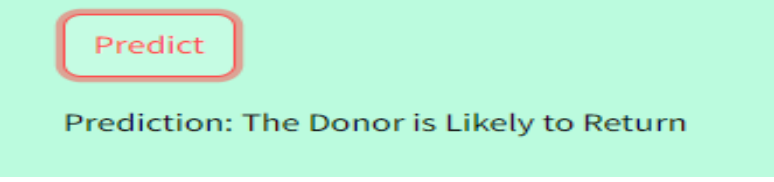
Model User Interface



The figure 4.42. shows the model interface. The interface was created using HTML and python. The ensemble hybrid model was integrated to the interface. The interface provides inputs such as age, gender, months since last donation, number of previous donations, Total volume donated, blood group and education level. After the details are entered the user clicks the predict button. The model loads from the data and returns whether the donor is likely to donate or not likely to return as shown in figure 4.43. below.

Figure 4. 43

Prediction Interface

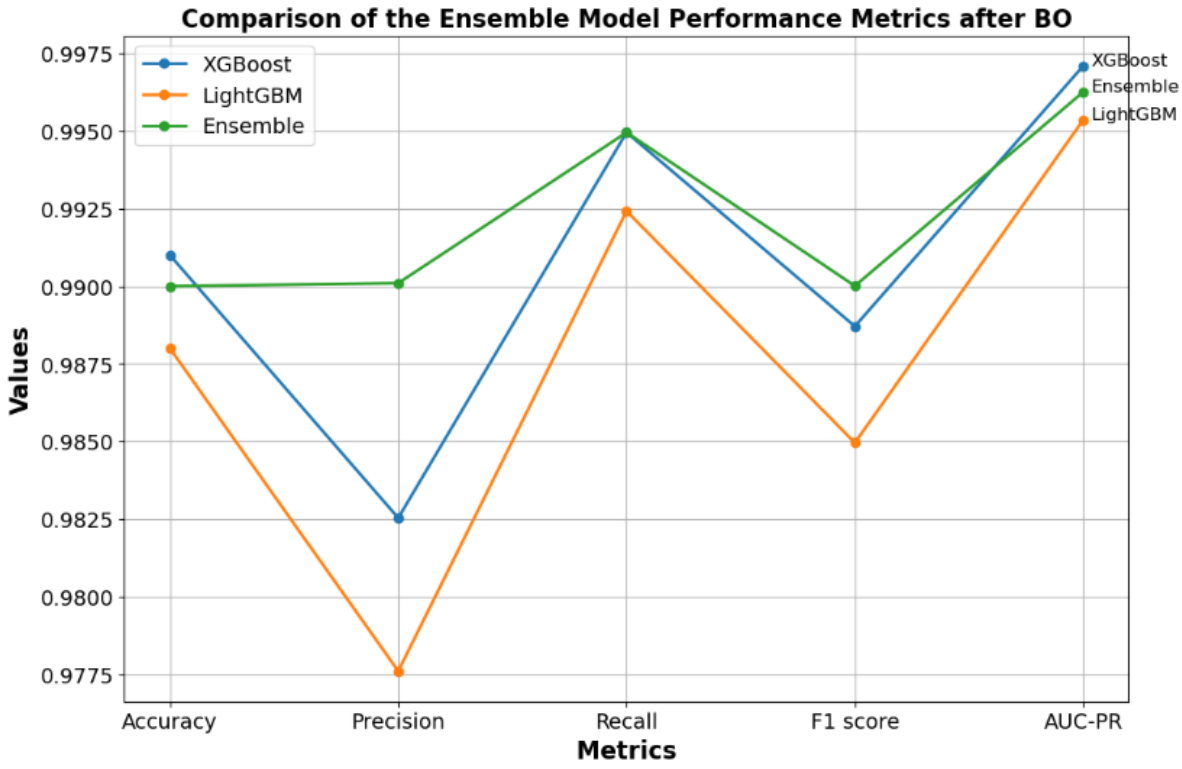


4.4.10. Comparative analysis of the XGBoost, Light GBM and the Ensemble gradient boosting model

The ensemble model performed well across most of the metrics than the individual models. The ensemble model outperformed the base models in precision, Recall and F1 score. This is because the ensemble model leveraged on the strengths of the two models while covering for each other's weaknesses. This shows that the ensemble model is more stable and robust as compared to the base models. Figure 4.44. shows the ensemble model performance as compared to the base models for the five main performance metrics.

Figure 4. 44

Comparative Analysis of the Xgboost and Light GBM and the Ensemble Gradient Boosting Model.



4.4.11. Comparative analysis with the existing models

Table 4. 5

Comparative Analysis of the Ensemble Model with the Existing Model.

Study	Algorithms	Validation	Dataset	Accuracy	F1 Score
(Marade et al., 2019)	Decision tree, and logistic regression.	Hold out validation	Taiwan 748 donors 5 variables	60	-
Pabreja & Bhasin, 2021	KNN	-Not mentioned	488 19 features	70.3	75.6
(Salazar-Concha & Ramírez-Correa, 2021)	Decision Treee C4.5	information gain Tenfold cross validation	197 donors w ith Seven variables	84.17	-
(Selvaraj et al., 2022)	SVM	Correlation and C ross validation	748 donors	78.4	-
(Shashikala et a l., 2019)	Naïve bayes	Not mentioned	246 25 features	86.99	98.8
(Zulfikar et al., 2018)	Decision Tree	Information gain	500 5 attributes	81.5	-
(Cloutier et al., 2021)	RF	MDA Cross validation	Donors aged 18 and 24	91	-
Hybrid Ensemble Boosting model	Xgboost & Light GBM	Cross validation Bayesian Optimization	5000 with 9 features	LGBM- 98.8 XGBoost- 9 9.1 Ensemble Hybrid- 99.0	98.8 98.4 99.0

When compared to the existing blood donor retention models. The hybrid ensemble model achieved the highest accuracy of 99.0. The model also achieved the best F1 score of 0.99 as compared to the existing studies. This highlights its superiority in identifying both the positive and the negative class.

It is important to note that most of the existing studies predominantly relied on accuracy as the performance metric. However, accuracy can be misleading, particularly in imbalanced datasets where the majority class can dominate the results

In contrast, the F1 score provides a more balanced evaluation. By balancing precision and recall, the F1 score provides a more comprehensive assessment of a model's ability to effectively handle both positive and negative instances. This makes it a more appropriate measure of model performance in scenarios where the class distribution is skewed, as is often the case in blood donation datasets. Our study's emphasis on the F1 score allowed us to capture a more accurate representation of model performance, leading to the creation and identification of the superior model. The hybrid ensemble model's achievement of the highest F1 score shows its effectiveness in addressing the challenges associated with imbalanced datasets, reducing the risk of overfitting and thereby improving donor retention predictions.

CHAPTER FIVE
SUMMARY, CONCLUSION, RECOMMENDATIONS AND PUBLICATIONS

5.1. Summary of findings

The goal of this research was to develop, optimize and validate an ensemble gradient boosting for blood donor retention. This was done by addressing the following research objectives.

Objective 1. To conduct a baseline survey on the existing blood donor retention models

The objective was achieved through review of literature to identify the existing studies on blood donor retention and the existing gap in research it was found out that the existing studies focus on individual models and hence the accuracy is limited to the individual model the models ability, also the accuracy achieved by the existing models need improvement. The ensemble models and specifically ensemble gradient boosting have not been experimented on the blood donation problem.

Objective 2. To develop an ensemble gradient boosting model for blood donor retention

The objective was achieved by developing a hybrid ensemble model based on XGBoost and LightGBM for predicting blood donor retention. The model was able to achieve an accuracy of 99.00% and F1 score of 99.0% after optimization which is higher than most of the existing models for blood donor retention. The model is also considered more stable and robust since it is made use of two strong base models and leverages on the strengths of the two models to produce a more accurate and robust model.

Objective 3. To optimize the ensemble model by fine-tuning hyperparameters

The performance of the hybrid ensemble gradient boosting model was optimized by tuning the hyperparameters to find the best set of optimal hyperparameters that minimize the

objective function. Over 1000 Trials were conducted for both Light GBM and XGBoost algorithms to find the best combination of hyperparameters. The trial no 110 is the one that obtained the minimum log loss of 0.03461 for the XGBoost with an improved accuracy of 99.1% while the light gradient boosting achieved the lowest log loss of 0.03397 in iteration number 139 with an improved accuracy of 98.8%. Overall the ensemble model performance improved from a log loss of 0.0406 to 0.03207 and accuracy of 98.3% to 99.0% after the model optimization.

Objective 4. To validate the performance of the developed ensemble gradient boosting model for blood donor retention

The performance of the model was validated using cross validation with k=5. The validation results showed that the model performed well in the prediction of blood donor retention. The hybrid ensemble model achieved an accuracy of 0.9900, precision of 0.9875, recall of 0.994, F1 score of 0.9900 and a ROC-AUC of 0.9979.

5.2. Conclusion

This study developed optimized and validated an ensemble gradient boosting model for blood donor retention, the study specifically utilized XGBoost and LightGBM. These are strong gradient boosting algorithms. The results demonstrate that the ensemble model achieved a high performance across multiple evaluation metrics, including accuracy (99%) and F1 score (99%) demonstrating its effectiveness and potential as a reliable tool for predicting whether a donor is likely to return to donate blood. The ensemble model leveraged on the strengths of the two models while minimizing their weaknesses to produce a more accurate and robust model. The model also successfully integrated various donor-related features to provide robust predictions, offering valuable insights for blood banks to enhance their donor retention strategies. This work demonstrates the potential application of advanced machine learning techniques in addressing critical healthcare related challenges and improving blood donation programs.

5.3. Limitations

Machine learning algorithms are limited to the dataset they employ to train and test the models. Although this study was conducted using dataset obtained from Kenya Blood Donor Management system. The study incorporated rigorous cross-validation techniques to ensure robust model performance and reduce overfitting to the specific dataset. Additionally, future work includes plans to validate and test the model on datasets from different regions and countries to evaluate its generalizability and adapt the model to diverse blood donation systems.

5.4. Contributions

This study contributes to the body of knowledge by conducting an extensive survey of existing blood donor retention models, their strengths, and their limitations. This contribution helps to identify gaps in the existing literature and practice and sets a foundation for further research.

This study introduces a novel hybrid ensemble gradient boosting model for blood donor retention based on XGBoost and Light GBM. By integrating multiple boosting techniques, this model was able to improve prediction accuracy as well as provide a more stable and robust model as compared to the existing single-model approaches. The study therefore contributes to the field of ICT by demonstrating how advanced machine algorithms can be employed to solve real world problems in healthcare.

The hybrid ensemble model was validated using real-world data from blood banks in Kenya. This demonstrates the practical application of the model and its effectiveness in a real-world context, providing evidence for its potential adoption by blood banks to make informed decisions and improve donor retention strategies.

The study also employs Bayesian optimization to fine-tune hyperparameters for the hybrid ensemble model to ensure the model operates at its optimal performance. This highlights the importance of performance optimization in machine learning. The methodology employed in this work has the potential to be utilized in the development of machine learning models for the prediction of other health-related outcomes and the development of other improved retention and churn prediction models. In addition, the research has facilitated identification of research gaps for future studies.

5.5. Future work

This study was conducted using data obtained from blood banks in Kenya. Various other datasets can be used on this model to evaluate its performance. Furthermore, different hyperparameter optimization techniques can be experimented on the base learners to find out which one achieves the best performance.

Future studies could investigate the use of other machine learning algorithms, such as deep learning or reinforcement learning. Other ensemble techniques such as stacking or blending can also be experimented and the performance be assessed to compare their effectiveness with the current XGboost-Light GBM hybrid model for blood donor retention.

Additionally, studies can be conducted using longitudinal data to try and understand blood donor behavior over time, which can aid in developing more dynamic and adaptive retention models.

5.6. Recommendations

The study recommends that this model be adopted and integrated with the blood bank systems in order to predict whether a donor is likely to return to donate blood. The integration should be done in a seamless manner so as to ensure minimal disruptions to current operations.

The blood banks can utilize this model to increase their efficiency. By identifying donors who are at risk of not donating again, they can concentrate their efforts on retaining them. This can contribute to a substantial increase in the number of available blood donors who can donate blood and help to save lives additionally the model will assist blood centers in developing focused retention strategies that are more effective in maintaining donors over time by determining factors that influence donor behavior. This can eliminate the need for continued donor recruitment, which is both costly and time-consuming.

By identifying blood donors who may not likely return for donations blood banks can come up with tailored communication strategies, offering targeted incentives as well as personalized follow up schedules to these donors.

The model can also be used to quickly identify donors who are most likely to respond to emergency appeals and donate promptly based on their past donation patterns. Blood banks can then prioritize outreach efforts towards those individuals who exhibit a high likelihood of timely responses. This targeted approach will enhance the efficiency of donor engagement strategies as well as optimize the allocation of the scarce resources during critical times.

5.7. Publications

N. Kiarie, A. C. Kirongo, and M. Mwadulo, "A systematic review of predictive blood donor retention models.," *African Journal of Science, Technology and Social Sciences*, vol. 2, no. 2, pp. 35–41, 2023, doi: https://doi.org/10.58506/ajstss.v2i2.206.

N. Kiarie, "Challenges and Prospects of Implementation of ISO 9001:2015 in TVET Institutions: The Case of Nkabune Technical Training Institute," *Africa Journal of Technical and Vocational Education and Training*, vol. 5, no. 1, pp. 84–95, Apr. 2020, Available: http://www.afritvetjournal.org/index.php/Afritvet/article/view/106

N. Kiarie, "Smartphones for Smart Learning in TVET: The Case of Nkabune Technical Training Institute," *Africa Journal of Technical and Vocational Education and Training*, vol. 1, no. 1, pp. 66–74, May 2016, doi: https://doi.org/10.69641/afritvet.2016.1113.

H. Muturi and N. Kiarie, "Effects of online tax system on tax compliance among small taxpayers in meru county, kenya," *International Journal of Economics, Commerce and Management United Kingdom*, vol. III, 2015, Available: https://ijecm.co.uk/wp-content/uploads/2015/12/31219.pdf

References

Adepoju, P. (2019). Blood transfusion in Kenya faces an uncertain future. *The Lancet*, *394*(10203), 997–998. https://doi.org/10.1016/S0140-6736(19)32140-3

Ajzen, I. (1985). From Intentions to Actions: A Theory of Planned Behavior. In *Action Control* (pp. 11–39). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-69746-3_2

Anand, A., Pugalenthi, G., Fogel, G. B., & Suganthan, P. N. (2010). An approach for classification of highly imbalanced data using weighting and undersampling. *Amino Acids*, *39*(5), 1385–1391. https://doi.org/10.1007/s00726-010-0595-2

Asamoah-Akuoko, L., Hassall, O. W., Bates, I., & Ullum, H. (2017). Blood donors' perceptions, motivators and deterrents in Sub-Saharan Africa - a scoping review of evidence. *British Journal of Haematology*, *177*(6), 864–877. https://doi.org/10.1111/bjh.14588

Awwalu, J., Ogwueleka, F., & Nonyelum, O. F. (2019). On Holdout and Cross Validation: A Comparison between Neural Network and Support Vector Machine. *International Journal of Trend in Research and Development*, *6*(2), 2394–9333. www.ijtrd.com

Baron, D. M., Franchini, M., Goobie, S. M., Javidroozi, M., Klein, A. A., Lasocki, S., Liumbruno, G. M., Muñoz, M., Shander, A., Spahn, D. R., Zacharowski, K., & Meybohm, P. (2020). Patient blood management during the COVID–19 pandemic: a narrative review. *Anaesthesia*, *75*(8), 1105–1113. https://doi.org/10.1111/anae.15095

Bates, S., Hastie, T., & Tibshirani, R. (2023). Cross-Validation: What Does It Estimate and How Well Does It Do It? *Journal of the American Statistical Association*, 1–12. https://doi.org/10.1080/01621459.2023.2197686

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021a). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021b). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Bi, Q., Goodman, K. E., Kaminsky, J., & Lessler, J. (2019). What is Machine Learning? A Primer for the Epidemiologist. *American Journal of Epidemiology*. https://doi.org/10.1093/aje/kwz189

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A., Deng, D., & Lindauer, M. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *WIREs Data Mining and Knowledge Discovery*, *13*(2). https://doi.org/10.1002/widm.1484

BM, K., BN, B., NN, M., WO, K., JC, K., TC, L., JK, K., F, T., & A, O. (2022). Challenges facing blood transfusion services at a regional blood transfusion center in Western Kenya. *International Journal of Blood Transfusion and Immunohematology*, *12*(2), 12–20. https://doi.org/10.5348/100075z02km2022ra

Brodeur, Z. P., Herman, J. D., & Steinschneider, S. (2020). Bootstrap Aggregation and Cross-Validation Methods to Reduce Overfitting in Reservoir Control Policy Search. *Water Resources Research*, *56*(8). https://doi.org/10.1029/2020WR027184

Bui, Q.-T., Chou, T.-Y., Hoang, T.-V., Fang, Y.-M., Mu, C.-Y., Huang, P.-H., Pham, V.-D., Nguyen, Q.-H., Anh, D. T. N., Pham, V.-M., & Meadows, M. E. (2021). Gradient Boosting Machine and Object-Based CNN for Land Cover Classification. *Remote Sensing*, *13*(14), 2709. https://doi.org/10.3390/rs13142709

Chang, Y.-C., Chang, K.-H., & Wu, G.-J. (2018). Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Applied Soft Computing*, *73*, 914–920. https://doi.org/10.1016/j.asoc.2018.09.029

Chengsheng, T., Huacheng, L., & Bing, X. (2017). AdaBoost typical Algorithm and its application research. *MATEC Web of Conferences*, *139*, 00222. https://doi.org/10.1051/matecconf/201713900222

Cloutier, M., Grégoire, Y., Choucha, K., Amja, A., & Lewin, A. (2021). Prediction of donation return rate in young donors using machine-learning models. *ISBT Science Series*, *16*(1), 119–126. https://doi.org/10.1111/voxs.12618

Dei-Adomakoh, Y., Asamoah-Akuoko, L., Appiah, B., Yawson, A., & Olayemi, E. (2021). Safe blood supply in sub-Saharan Africa: challenges and opportunities. *The Lancet Haematology*, *8*(10), e770–e776. https://doi.org/10.1016/S2352-3026(21)00209-X

Delaney, M., Telke, S., Zou, S., Williams, M. J., Aridi, J. O., Rudd, K. E., Puyana, J. C., Kumar, P., Appiah, B., Dei-Adomakoh, Y., Asamoah-Akuoko, L., Olayemi, E., Singogo, E., Hosseinipour, M. C., M'baya, B., Chipeta, E., Reilly, C., Kamu, R., Aridi, J., … Joshua, M. (2022). The Bloodsafe program: Building the future of access to safe blood in Sub-Saharan Africa. *Transfusion*, *62*(11), 2282–2290. https://doi.org/10.1111/trf.17091

Dong, G., Tang, M., Wang, Z., Gao, J., Guo, S., Cai, L., Gutierrez, R., Campbel, B., Barnes, L. E., & Boukhechba, M. (2023). Graph Neural Networks in IoT: A Survey. *ACM Transactions on Sensor Networks*, *19*(2), 1–50. https://doi.org/10.1145/3565973

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). *CatBoost: gradient boosting with categorical features support*. http://arxiv.org/abs/1810.11363

Eertink, J. J., Heymans, M. W., Zwezerijnen, G. J. C., Zijlstra, J. M., de Vet, H. C. W., & Boellaard, R. (2022). External validation: a simulation study to compare cross-validation versus holdout or external testing to assess the performance of clinical

prediction models using PET data from DLBCL patients. *EJNMMI Research*, *12*(1), 58. https://doi.org/10.1186/s13550-022-00931-w

Fei, Z., Liang, S., Cai, Y., & Shen, Y. (2023). Ensemble Machine-Learning-Based Prediction Models for the Compressive Strength of Recycled Powder Mortar. *Materials*, *16*(2), 583. https://doi.org/10.3390/ma16020583

Ferguson, E. (2021). Strategies and theories to attract and retain blood donors: fairness, reciprocity, equity and warm-glow. *ISBT Science Series*, *16*(3), 219–225. https://doi.org/10.1111/voxs.12640

Gandin, I., Scagnetto, A., Romani, S., & Barbati, G. (2021). Interpretability of time-series deep learning models: A study in cardiovascular patients admitted to Intensive care unit. *Journal of Biomedical Informatics*, *121*, 103876. https://doi.org/10.1016/j.jbi.2021.103876

Giles, M. (2004). An application of the Theory of Planned Behaviour to blood donation: the importance of self-efficacy. *Health Education Research*, *19*(4), 380–391. https://doi.org/10.1093/her/cyg063

Golas, S. B., Shibahara, T., Agboola, S., Otaki, H., Sato, J., Nakae, T., Hisamitsu, T., Kojima, G., Felsted, J., Kakarmath, S., Kvedar, J., & Jethwani, K. (2018). A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. *BMC Medical Informatics and Decision Making*, *18*(1), 44. https://doi.org/10.1186/s12911-018-0620-z

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, & Tie-Yan Liu. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, 3147–3155.

Guo, R., Fu, D., & Sollazzo, G. (2022). An ensemble learning model for asphalt pavement performance prediction based on gradient boosting decision tree. *International Journal of Pavement Engineering*, *23*(10), 3633–3646. https://doi.org/10.1080/10298436.2021.1910825

Hamid, N. Z. A., Basiruddin, R., & Hassan, N. (2013). The Intention to Donate Blood: An Analysis of Socio-Demographic Determinants. *International Journal of Social Science and Humanity*, 503–507. https://doi.org/10.7763/IJSSH.2013.V3.292

Hanieza, W., Sarkan, H. M., Sjarif, N. N. A., & Yahya, Y. (2019). A Prediction Model for Blood Donation Using Multiple Logistic Regression. *Open International Journal of Informatics (OIJI)*, *7*(2), 147–157.

Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised Machine Learning: A Brief Primer. *Behavior Therapy*, *51*(5), 675–687. https://doi.org/10.1016/j.beth.2020.05.002

Kalcheva, N., Todorova, M., & Marinova, G. (2020). *Naive Bayes Classifier, Decision Tree and Adaboost Ensemble Algorithm – Advantages and Disadvantages*. 153–157. https://doi.org/10.31410/ERAZ.2020.153

Kanagasabai, U., Selenic, D., Chevalier, M. S., Drammeh, B., Qualls, M., Shiraishi, R. W., Bock, N., Benech, I., & Mili, F. D. (2021). Evaluation of the WHO global database on blood safety. *Vox Sanguinis*, *116*(2), 197–206. https://doi.org/10.1111/vox.13001

Kauten, C., Gupta, A., Qin, X., & Richey, G. (2022). Predicting Blood Donors Using Machine Learning Techniques. *Information Systems Frontiers*, *24*(5), 1547–1562. https://doi.org/10.1007/s10796-021-10149-1

Kenya Blood Transfusion and Transplant Service. (2023). *Blood Donation Process*. Kenya Tissue and Transplant Authority Website. https://www.ktta.go.ke/

Kervanci, I. S., Akay, M. F., & Özceylan, E. (2024). Bitcoin price prediction using LSTM, GRU and hybrid LSTM-GRU with bayesian optimization, random search, and grid search for the next days. *Journal of Industrial and Management Optimization*, *20*(2), 570–588. https://doi.org/10.3934/jimo.2023091

Kewat, A., & Sharma, A. K. (2018). Evaluating The Performance of Naive Bayes Classification Algorithm for Blood Donors Problem. *Journal of Emerging Technologies and Innovative Research*, *5*(6), 298–304.

Leipnitz, S., de Vries, M., Clement, M., & Mazar, N. (2018). Providing health checks as incentives to retain blood donors — Evidence from two field experiments. *International Journal of Research in Marketing*, *35*(4), 628–640. https://doi.org/10.1016/j.ijresmar.2018.08.004

Lestandy, M., Syafa'ah, L., & Faruq, A. (2020). Classification of potential blood donors using machine learning algorithms approach. *Jurnal Teknologi Dan Sistem Komputer*, *8*(3), 217–221. https://doi.org/10.14710/jtsiskom.2020.13619

Liang, Y., Wu, J., Wang, W., Cao, Y., Zhong, B., Chen, Z., & Li, Z. (2019). Product marketing prediction based on XGboost and LightGBM algorithm. *Proceedings of the 2nd International Conference on Artificial Intelligence and Pattern Recognition*, 150–153. https://doi.org/10.1145/3357254.3357290

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2018). Feature Selection. *ACM Computing Surveys*, *50*(6), 1–45. https://doi.org/10.1145/3136625

Liu, H. (2021). Single-point wind forecasting methods based on ensemble modeling. In *Wind Forecasting in Railway Engineering* (pp. 215–250). Elsevier. https://doi.org/10.1016/B978-0-12-823706-9.00006-5

Liu, S., Zhou, R., Xia, X.-Q., Ren, H., Wang, L.-Y., Sang, R.-R., Jiang, M., Yang, C.-C., Liu, H., Wei, L., & Rong, R.-M. (2021). Machine learning models to predict red blood

cell transfusion in patients undergoing mitral valve surgery. *Annals of Translational Medicine*, *9*(7), 530–530. https://doi.org/10.21037/atm-20-7375

Li, X., Yi, S., Cundy, A. B., & Chen, W. (2022). Sustainable decision-making for contaminated site risk management: A decision tree model using machine learning algorithms. *Journal of Cleaner Production*, *371*, 133612. https://doi.org/10.1016/j.jclepro.2022.133612

Loua, A., Kasilo, O. M. J., Nikiema, J. B., Sougou, A. S., Kniazkov, S., & Annan, E. A. (2021). Impact of the COVID-19 pandemic on blood supply and demand in the WHO African Region. *Vox Sanguinis*, *116*(7), 774–784. https://doi.org/10.1111/vox.13071

Lourençon, A. de F., Almeida, R. G. dos S., Ferreira, O., & Martinez, E. Z. (2011). Evaluation of the return rate of volunteer blood donors. *Revista Brasileira de Hematologia e Hemoterapia*, *33*(3), 190–194. https://doi.org/10.5581/1516-8484.20110052

Lukmanto, R. B., Suharjito, Nugroho, A., & Akbar, H. (2019). Early Detection of Diabetes Mellitus using Feature Selection and Fuzzy Support Vector Machine. *Procedia Computer Science*, *157*, 46–54. https://doi.org/10.1016/j.procs.2019.08.140

Mahesh, T. R., Dhilip Kumar, V., Vinoth Kumar, V., Asghar, J., Geman, O., Arulkumaran, G., & Arun, N. (2022). AdaBoost Ensemble Methods Using K-Fold Cross Validation for Survivability with the Early Detection of Heart Disease. *Computational Intelligence and Neuroscience*, *2022*, 1–11. https://doi.org/10.1155/2022/9005278

Malek, N. H. A., Yaacob, W. F. W., Wah, Y. B., Md Nasir, S. A., Shaadan, N., & Indratno, S. W. (2022). Comparison of ensemble hybrid sampling with bagging and boosting machine learning approach for imbalanced data. *Indonesian Journal of Electrical Engineering and Computer Science*, *29*(1), 598. https://doi.org/10.11591/ijeecs.v29.i1.pp598-608

Marade, C., Pradeep, A., Mohanty, D., & Patil, C. (2019a). Forecasting Blood Donor Response Using Predictive Modelling Approach. *International Journal of Computer Science and Mobile Computing*, *8*(4), 73–77.

Marade, C., Pradeep, A., Mohanty, D., & Patil, C. (2019b). Forecasting Blood Donor Response Using Predictive Modelling Approach. *International Journal of Computer Science and Mobile Computing*, *8*(4), 73–77.

Mbuthia, A. N., Mwangi, E. M., & Ong'Ombe, M. O. (2019). Organisational management of hospital blood transfusion services in Nairobi County, Kenya: Evidence of implementation. *African Journal of Laboratory Medicine*, *8*(1). https://doi.org/10.4102/AJLM.V8I1.676

McElfresh, D. C., Kroer, C., Pupyrev, S., Sodomka, E., Sankararaman, K., Chauvin, Z., Dexter, N., & Dickerson, J. P. (2021). *Matching Algorithms for Blood Donation*. https://doi.org/https://doi.org/10.48550/arXiv.2108.04862

Medvedev, I. N. (2021). Physiological response of the morphological characteristics of mammalian blood to the intake of selenium preparations into the body. *IOP Conference Series: Earth and Environmental Science*, *677*(4). https://doi.org/10.1088/1755-1315/677/4/042060

Ministry of health Kenya. (2023). *Blood Banking Management System (DamuKE)*. Blood Banking Management System (DamuKE). https://www.health.go.ke/

Moore, M. B., Gitau, T., & Kerochi, A. (2020). Factors influencing blood donation practices among students of private universities in Thika Town, Kiambu County, Kenya. *International Journal Of Community Medicine And Public Health*, *7*(6), 2090. https://doi.org/10.18203/2394-6040.ijcmph20202457

Nadali, A., Kakhky, E. N., & Nosratabadi, H. E. (2011). Evaluating the success level of data mining projects based on CRISP-DM methodology by a Fuzzy expert system. *2011 3rd International Conference on Electronics Computer Technology*, 161–165. https://doi.org/10.1109/ICECTECH.2011.5942073

Nagassou, M., Mwangi, R. W., & Nyarige, E. (2023). A Hybrid Ensemble Learning Approach Utilizing Light Gradient Boosting Machine and Category Boosting Model for Lifestyle-Based Prediction of Type-II Diabetes Mellitus. *Journal of Data Analysis and Information Processing*, *11*(04), 480–511. https://doi.org/10.4236/jdaip.2023.114025

Nanni, L., Interlenghi, M., Brahnam, S., Salvatore, C., Papa, S., Nemni, R., & Castiglioni, I. (2020). Comparison of Transfer Learning and Conventional Machine Learning Applied to Structural Brain MRI for the Early Diagnosis and Prognosis of Alzheimer's Disease. *Frontiers in Neurology*, *11*. https://doi.org/10.3389/fneur.2020.576194

Naveen, Sharma, R. K., & Ramachandran Nair, A. (2019). Efficient Breast Cancer Prediction Using Ensemble Machine Learning Models. *2019 4th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT)*, 100–104. https://doi.org/10.1109/RTEICT46194.2019.9016968

Okuthe, J. O., Muitta, E. W., & Odongo, A. O. (2022). Determinants of blood donation among selected tertiary college students in Homa Bay County Kenya. *International Journal Of Community Medicine And Public Health*, *9*(3), 1250. https://doi.org/10.18203/2394-6040.ijcmph20220682

Ou-Yang, J., Bei, C.-H., He, B., & Rong, X. (2017). Factors influencing blood donation: a cross-sectional survey in Guangzhou, China. *Transfusion Medicine*, *27*(4), 256–267. https://doi.org/10.1111/tme.12410

Pabreja, K., & Bhasin, A. (2021). A Predictive Analytics Framework for Blood Donor Classification. *International Journal of Big Data and Analytics in Healthcare*, *6*(2), 1–14. https://doi.org/10.4018/ijbdah.20210701.oa1

Panesar, A. (2019). *Machine Learning and AI for Healthcare*. Apress. https://doi.org/10.1007/978-1-4842-3799-1

Rajeh, M. T. (2022). Modeling the theory of planned behavior to predict adults' intentions to improve oral health behaviors. *BMC Public Health*, *22*(1), 1391. https://doi.org/10.1186/s12889-022-13796-4

Rivera-Lopez, R., Canul-Reich, J., Mezura-Montes, E., & Cruz-Chávez, M. A. (2022). Induction of decision trees as classification models through metaheuristics. *Swarm and Evolutionary Computation*, *69*, 101006. https://doi.org/10.1016/j.swevo.2021.101006

Rodríguez-Tomàs, E., Arenas, M., Baiges-Gaya, G., Acosta, J., Araguas, P., Malave, B., Castañé, H., Jiménez-Franco, A., Benavides-Villarreal, R., Sabater, S., Solà-Alberich, R., Camps, J., & Joven, J. (2022). Gradient Boosting Machine Identified Predictive Variables for Breast Cancer Patients Pre- and Post-Radiotherapy: Preliminary Results of an 8-Year Follow-Up Study. *Antioxidants*, *11*(12), 2394. https://doi.org/10.3390/antiox11122394

Saad Alkahtani, A., & Jilani, M. (2019). Predicting Return Donor and Analyzing Blood Donation Time Series using Data Mining Techniques. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 10, Issue 8). www.ijacsa.thesai.org

Salazar-Concha, C., & Ramírez-Correa, P. (2021). Predicting the Intention to Donate Blood among Blood Donors Using a Decision Tree Algorithm. *Symmetry*, *13*(8), 1460. https://doi.org/10.3390/sym13081460

Selvaraj, P., Archit Raj, & Anurag Gupta. (2018). Predicting Donors Likelihood of Donating Blood Given Various Factors. *International Journal of Pure and Applied Mathematics*, *118*(22), 491–495.

Selvaraj, P., Sarin, A., & Seraphim, B. I. (2022). Forecasting System for Donation of Blood Using SVM Model. *International Journal for Research in Applied Science and Engineering Technology*, *10*(5), 136–140. https://doi.org/10.22214/ijraset.2022.41940

Shama, A. T., Teka, G., Yohannes, S., Tesfaye, B., Ebisa, H., Gebre, D. S., & Terefa, D. R. (2022). Assessment of Blood Donation Practice and Its Associated Factors Among Wollega University Undergraduate Students, Ethiopia. *Journal of Blood Medicine*, *Volume 13*, 711–724. https://doi.org/10.2147/JBM.S385348

Shashikala, B. M., Pushpalatha, M. P., & Vijaya, B. (2019). Machine Learning Approaches for Potential Blood Donors Prediction. *Lecture Notes in Electrical Engineering*, *545*(January 2019), 483–491. https://doi.org/10.1007/978-981-13-5802-9_44

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of Data Warehousing 5*, *5*(4), 13–22.

Shehu, E., Veseli, B., Clement, M., & Winterich, K. P. (2024a). Improving Blood Donor Retention and Donor Relationships with Past Donation Use Appeals. *Journal of Service Research*, *27*(3), 346–363. https://doi.org/10.1177/10946705231202244

Shehu, E., Veseli, B., Clement, M., & Winterich, K. P. (2024b). Improving Blood Donor Retention and Donor Relationships with Past Donation Use Appeals. *Journal of Service Research*, *27*(3), 346–363. https://doi.org/10.1177/10946705231202244

Suessner, S., Niklas, N., Bodenhofer, U., & Meier, J. (2022). Machine learning-based prediction of fainting during blood donations using donor properties and weather data as features. *BMC Medical Informatics and Decision Making*, *22*(1), 222. https://doi.org/10.1186/s12911-022-01971-x

Van Dongen, A. (2015). Easy come, easy go. Retention of blood donors. *Transfusion Medicine*, *25*(4), 227–233. https://doi.org/10.1111/tme.12249

Vujovic, Ž. Đ. (2021). Classification Model Evaluation Metrics. *International Journal of Advanced Computer Science and Applications*, *12*(6). https://doi.org/10.14569/IJACSA.2021.0120670

Weidmann, C., Derstroff, M., Klüter, H., Oesterer, M., & Müller-Steinhardt, M. (2022). Motivation, blood donor satisfaction and intention to return during the <scp>COVID</scp> -19 pandemic. *Vox Sanguinis*, *117*(4), 488–494. https://doi.org/10.1111/vox.13212

WHO. (2018). *Who can give blood*.

WHO. (2023). *Blood Safety Key Facts*. https://www.afro.who.int/health-topics/blood-safety

Wongvorachan, T., He, S., & Bulut, O. (2023). A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. *Information*, *14*(1), 54. https://doi.org/10.3390/info14010054

World Bank. (2022, May 6). *Ensuring Access to Safe Blood in Kenya Amid COVID-19 Pandemic*. https://www.worldbank.org/en/news/feature/2022/05/06/ensuring-access-to-safe-blood-in-kenya-enhanced-amid-covid-19-pandemic

World Health Organization. (2017). *Global status report on blood safety and availability*.

Wu, H., Li, Z., Sun, X., Bai, W., Wang, A., Ma, Y., Diao, R., Fan, E., Zhao, F., Liu, Y., Hong, Y., Guo, M., Xue, H., & Liang, W. (2022a). Predicting willingness to donate blood based on machine learning: two blood donor recruitments during COVID-19 outbreaks. *Scientific Reports*, *12*(1), 19165. https://doi.org/10.1038/s41598-022-21215-2

Wu, H., Li, Z., Sun, X., Bai, W., Wang, A., Ma, Y., Diao, R., Fan, E., Zhao, F., Liu, Y., Hong, Y., Guo, M., Xue, H., & Liang, W. (2022b). Predicting willingness to donate blood based on machine learning: two blood donor recruitments during COVID-19

outbreaks. *Scientific Reports*, *12*(1), 19165. https://doi.org/10.1038/s41598-022-21215-2

Xiao, C., Choi, E., & Sun, J. (2018a). Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, *25*(10), 1419–1428. https://doi.org/10.1093/jamia/ocy068

Xiao, C., Choi, E., & Sun, J. (2018b). Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, *25*(10), 1419–1428. https://doi.org/10.1093/jamia/ocy068

Yang, K., Yu, Z., Chen, C. L. P., Cao, W., You, J., & Wong, H.-S. (2022). Incremental Weighted Ensemble Broad Learning System for Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *34*(12), 5809–5824. https://doi.org/10.1109/TKDE.2021.3061428

Zhang, B., Ren, J., Cheng, Y., Wang, B., & Wei, Z. (2019). Health Data Driven on Continuous Blood Pressure Prediction Based on Gradient Boosting Decision Tree Algorithm. *IEEE Access*, *7*, 32423–32433. https://doi.org/10.1109/ACCESS.2019.2902217

Zhang, Z., Zhao, Y., Canes, A., Steinberg, D., Lyashevska, O., & Group, written on behalf of A. B.-D. C. T. C. (2019). Predictive analytics with gradient boosting in clinical medicine. *Annals of Translational Medicine*, *7*(7), 152–152. https://doi.org/10.21037/ATM.2019.03.29

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2021). A Comprehensive Survey on Transfer Learning. *Proceedings of the IEEE*, *109*(1), 43–76. https://doi.org/10.1109/JPROC.2020.3004555

Zulfikar, W. B., Gerhana, Y. A., & Rahmania, A. F. (2018). An Approach to Classify Eligibility Blood Donors Using Decision Tree and Naive Bayes Classifier. *2018 6th International Conference on Cyber and IT Service Management (CITSM)*, 1–5. https://doi.org/10.1109/CITSM.2018.8674353

Appendices

Appendix A. MIRERC Ethics clearance

MERU UNIVERSITY INSTITUTIONAL RESEARCH & ETHICS REVIEW COMMITTEE
(MIRERC)
Email: mirerc@must.ac.ke ***Website:*** *https://research.must.ac.ke/research-ethics/*

REF: MU/1/39/28 Vol.3 (025) Date: 15th April, 2024

TO: **Nahashon Kieria** (MSc. Information Technology -MUST)
Dr. Amos Chege Kirongo, Dr. Mary Mwadulo

Dear Sir/madam

RE: An Ensemble Gradient Boosting Model for blood Dona Retention

This is to inform you that ***MIRERC*** has reviewed and approved your above research proposal. Your application approval number is ***MIRERC001/2024.*** The approval period is 15th ***April, 2024– 14th April, 2025.***

This approval is subject to compliance with the following requirements;

i. Only approved documents including (informed consents, study instruments, MTA) will be used
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by ***MIRERC.***
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to ***MIRERC*** within 72 hours of notification
iv. Any changes, anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to ***MIRERC*** within 72 hours
v. Clearance for export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days upon completion of the study to ***MIRERC.***

You may also be required to obtain a research license from National Commission for Science, Technology and Innovation (NACOSTI), visit: https://research-portal.nacosti.go.ke and also obtain any other clearances needed for your study.

Yours sincerely

Prof. Peter Masinde, Ph.D.
Chair, MIRERC

MUST IS ISO 9001:2015 and ISO/IEC 27001:2013 CERTIFIED

Appendix B. NACOSTI Research License

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Ref No: **296169**

Date of Issue: **13/May/2024**

RESEARCH LICENSE

This is to Certify that Mr.. Nahashon Kiarie of Meru University of Science and Technology, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev.2014) in Meru on the topic: An Ensemble Gradient Boosting Model for blood Donor Retention for the period ending : 13/May/2025.

License No: **NACOSTI/P/24/35188**

296169

Applicant Identification Number

Director General
NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.

See overleaf for conditions

Appendix C. County Government of Meru Authorization

COUNTY GOVERNMENT OF MERU
DEPARTMENT OF HEALTH

Telephone: 0772207572
Website: www.metrh.or.ke
Email: ceo@metrh.or.ke
info@metrh.or.ke
When replying should be to:
Chief Executive Officer

MERU TEACHING & REFERRAL HOSPITAL

Meru Teaching and Referral Hospital
P.O. Box 8 - 60200
Meru

DATE... 8/5/2024

STUDY DEMOGRAPHIC

1. TITTLE OF THE STUDY: ENSEMBLE GRADIENT BOOSTING MODEL FOR BLOOD DONOR RETENTION
2. REF NO:
3. NAME OF THE PRINCIPAL INVESTIGATOR: NATHASTON KIARIE
4. AFFILIATED INSTITUTION: MERU UNIVERSITY OF SCIENCE & TECHNOLOGY
5. DEPARTMENT NAME: DEPARTMENT OF IT
6. MAILING ADDRESS:
7. PHONE NUMBER: 0725661051
8. E-MAILADDRESS: Kevinghachora12@gmail.com
9. CATEGORY OF PROPOSAL (TICK as appropriate)

UNDERGRADUATE: ☐ CERTIFICATE ☐ DIPLOMA ☐ HIGHER DIP ☐ DEGREE

POSTGRADUATE: ☐ POSTGRADUATE DIPLOMA ☑ MASTERS ☐ PhD POST DOCTORAL

OPERATIONAL RESEARCH: ☐ CLINICAL ☐ SOCIAL SCIENCE ☐ EPIDEMIOLOGY ☐ INTERVENTIONAL CLINICAL TRIAL

10. SOURCE OF FUNDING: Self
11. PROPOSED DATA COLLECTION SITE: Meru Satellite blood bank
12. STUDY DURATION: 1 year
13. FOR ACADEMIC PROPOSALS
    a. NAME OF PRIMARY SUPERVISOR: DR AMOS CHEGE
    b. AFFILIATED INSTITUTION: MERU UNIVERSITY
    c. PHONE NUMBER: 0720984636 EMAILADDRESS: akinyi@must.ac.ke
    d. NAME OF OTHER SUPERVISORS: DR MR MWANDILO
    e. AFFILIATE DINSTITUTION: MERU UNIVERITY
    f. PHONE NUMBER: 0725.789102 E-MAILADDRESS: mmwandilo@must.ac.ke

12 .HOW WOULD YOU WISH TO GET YOUR FEEDBACK: ☐ PICK FROM OFFICE ☐ EMAIL ☑ TEL.

SIGNATURE... Date: 8/5/2024

Appendix D. Plagiarism Report

DrillBit

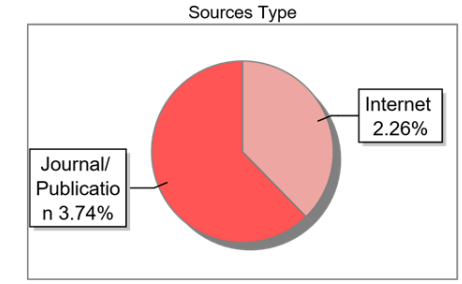
The Report is Generated by DrillBit Plagiarism Detection Software

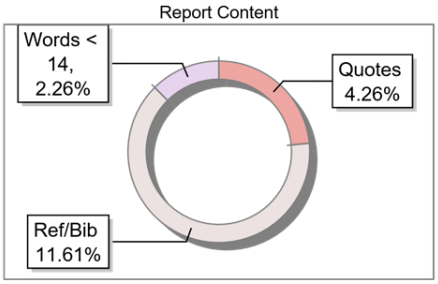
Submission Information

Author Name	NAHASHON KIARIE
Title	A HYBRID ENSEMBLE BOOSTING MODEL FOR ENHANCED BLOOD DONOR RETENTION
Paper/Submission ID	2125051
Submitted by	mmusungu@must.ac.ke
Submission Date	2024-07-17 08:56:36
Total Pages, Total Words	156, 30007
Document type	Thesis

Result Information

Similarity **6 %**





Exclude Information

Quotes	Not Excluded	Language	English
References/Bibliography	Excluded	Student Papers	Yes
Source: Excluded < 14 Words	Not Excluded	Journals & publishers	Yes
Excluded Source	**0 %**	Internet or Web	Yes
Excluded Phrases	Not Excluded	Institution Repository	Yes

Database Selection

A Unique QR Code use to View/Download/Share Pdf File

Appendix E. Publication

AFRICAN JOURNAL OF SCIENCE, TECHNOLOGY AND SOCIAL SCIENCES. 2(2) 2023. TE 35–41

OPEN ACCESS

AFRICAN JOURNAL OF SCIENCE, TECHNOLOGY AND SOCIAL SCIENCES

Journal website: https://journals.must.ac.ke

A Publication of Meru University of Science and Technology

A systematic review of predictive blood donor retention models

Nahashon Kiarie[1*], Amos Chege Kirongo[1], Mary Mwadulo[1]

[1] Meru University of Science and Technology, Meru, Kenya

ARTICLE INFO

KEYWORDS

blood donor retention

machine learning

predictive model

healthcare

blood donation

ABSTRACT

Demand for blood and blood products is increasing due to population growth, medical advances, and increased disease. Availability of a stable blood supply is critical for healthcare organizations and requires effective donor recruitment and retention strategies. This systematic review paper examines the development and implementation of predictive models using machine learning techniques to classify and predict blood donor retention rates. The aim is to analyze the existing literature and provide insights into the design, performance and potential of such models through a systematic search of relevant databases. The reviewed studies include a variety of machine learning approaches and algorithms used to predict blood donor retention rates. These models use various demographic, behavioral, and historical donation data to predict the likelihood of a donor returning to donate blood. The utilization of machine learning techniques, such as decision trees, logistic regression, support vector machines, and neural networks, enables accurate predictions and enable healthcare organizations to implement targeted donor retention interventions to increase blood supply. The models' predictive performance reveals their capacity to recognize donors who are not likely to return and donate blood and target retention strategies appropriately, improving donor engagement and fostering long-term commitment. Several challenges and limitations face the identified existing models. They include the need for comprehensive and high-quality data, interpretability of complex models as well as the requirement for regular model updates to accommodate changing donor behaviors. There is need for development of versatile and comprehensive models with improved accuracy that can reduce the need for constant recruitment of new donors, which is costly and time-consuming enabling blood agencies to accurately predict donor retention rates, inform donor retention strategies, and prioritize resources appropriately and ultimately saving lives

Introduction

Background and Motivation

The demand for blood and blood products is constantly increasing due to population growth, advancements in medical procedures, and rising incidence of diseases such as cancer and chronic conditions that require regular transfusions. However, this increasing demand is not being met adequately, resulting in blood shortages and their subsequent impact on healthcare systems worldwide[1]. One of the primary reasons for blood donation scarcity is the insufficient recruitment and

*Corresponding author: Kiarie Nahashon Email: kiarienahashon12@gmail.com

https://doi.org/10.58506/ajstss.v2i2.206

AFRICAN JOURNAL OF SCIENCE, TECHNOLOGY AND SOCIAL SCIENCES ISSN:2958-0560