BLOOD DONATION PREDICTION USING MACHINE LEARNING TECHNIQUES

A Graduate Project Report submitted to Manipal Academy of Higher Education in partial fulfilment of the requirement for the award of the degree of

BACHELOR OF TECHNOLOGY

In

Electronics and Communication Engineering

Submitted by

Boyella Vishnu Vardhan Reddy

Reg. No:190907598

Under the guidance of

Dr. Ramya S

Associate Professor

ELECTRONIC AND COMMUNICATION

MANIPAL INSTITUTE OF TECHNOLOGY

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEER

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

(A constituent unit of MAHE, Manipal)

NOVEMBER/DECEMBER 2024

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
(A constituent unit of MAHE, Manipal)

Manipal
7 Dec 2024

CERTIFICATE

This is to certify that the project titled **BLOOD DONATION PREDICTION USING MACHINE LEARNING TECHNIQUES** is a record of the bonafide work done by **BOYELLA VISHNU VARDHAN REDDY** (*Reg. No. 190907598*) submitted in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology (BTech) in **ELECTRONICS AND COMMUNICATION ENGINEERING** of Manipal Institute of Technology, Manipal, Karnataka, (A Constituent Unit of Manipal Academy of Higher Education), during the academic year 2024 - 2025.

Dr. Ramya S
ASSOSCIATE PROFESSOR, ECE
M.I.T, MANIPAL

Dr. Pallavi R Mane
HOD, ECE
M.I.T, MANIPAL

ACKNOWLEDGMENTS

I would like to thank each one of those who gave me their guidance, support, and encouragement during this research work.

Primarily, I extend my sincere thanks to the Director of Manipal Institute of Technology, for providing the necessary infrastructure and resources to facilitate this project. Your leadership and commitment to academic excellence have been an inspiration throughout this journey.

I am forever in debt to Dr. Pallavi R Mane, Head of the Department of Electronics and Communication, for encouragement, to say nothing of constructive feedback and support, throughout this study. Your insights have really helped me mild the direction of this study.

I would like to place on record my appreciation towards my department guide, Dr. Ramya S, who have kindly permitted the use of the laboratory facilities and technical inputs whenever necessary. Your persistent support has enabled the smooth accomplishment of the experimental work. I thank the faculty members of the Electronics and Communication for their able suggestions and encouragement that have been richly rewarded during the project work.

I also thank my friends and colleagues for the moral support and helpful discussions with them that motivated me during such tough times.This work could not have been produced without the collective effort and guidance of all the individuals mentioned above. Thank you for sharing this journey with me.

ABSTRACT

The rapid advancement of technology and the exponential growth of data have brought machine learning into the forefront of solving complex classification problems in various domains. With the increasing demand for accurate and efficient predictive models, traditional machine learning techniques often face limitations in balancing accuracy and computational efficiency. To address these challenges, this research focuses on exploring hybrid machine learning models by combining techniques such as Logistic Regression, Naïve Bayes, and Random Forest. The objective of the work is to identify an optimal hybrid model that delivers high accuracy while maintaining computational efficiency, making it suitable for modern-day applications requiring robust classification capabilities.

The methodology adopted in this research involves the evaluation of various hybrid machine learning models on a classification task. Models were constructed using different combinations of machine learning techniques and evaluated using Log Loss as the loss function. Multiple metrics, including accuracy, train time, test time, and loss values, were analysed to assess performance. Experiments were conducted with two train-test split ratios (20-80 and 30-70) to understand their effect on model generalization. The study utilized ensemble learning techniques to combine the strengths of individual classifiers and employed software tools such as [Software Name] for implementation and analysis.

The results demonstrate that hybrid models outperform single techniques in terms of accuracy and loss optimization. The Hybrid Model LNR (30:70) had the highest accuracy of 76.44% with the lowest values of loss, indicating it is one of the best models. The Random Forest models, specifically Hybrid Model RL (20:80), had the trade-off between accuracy at 74%% and computational efficiency for medium-scale applications. The Naïve Bayes models were computationally efficient but at the cost of having a significantly lower accuracy, pointing to the trade-off between speed and performance in prediction. These results point to the potential of hybrid approaches in exploiting the complementary strengths of individual machine learning techniques for better predictive accuracy.

The results of this study show that hybrid models, especially Hybrid Model LNR (20,80), can achieve accurate and efficient classification results. The results are relevant to real-world applications where both accuracy and computational efficiency are essential. This study also opens avenues for further exploration, such as incorporation of advanced hybrid combinations, hyperparameter optimization, and testing on large-scale datasets to improve the robustness and scalability of models. The research was implemented on Python, Scikit-learn, or TensorFlow so that the results are both dependable and reproducible.

LIST OF TABLES

LIST OF FIGURES

Contents

INTRODUCTION

1.1 Introduction

Blood donation is one of the pillars of medical systems that directly impacts survival chances for patients in surgeries, medical emergencies, and treatments for chronic conditions like anaemia or cancer. Nevertheless, the management of blood resources is highly complicated due to variations in donor turnout, perishability, and unconceivable medical emergencies. Predicting donor behaviour is critical for healthcare organizations looking to streamline blood stock management, optimize donor engagement, and minimize both shortages and wastage. Recent advances in machine learning and data analytics have transformed the predictive capability of various fields, including healthcare. These technologies allow the processing and analysis of large amounts of donor data, uncovering hidden trends and patterns that cannot be detected by traditional manual methods. Machine learning models can assess complex variables, such as demographics of donors, past donation frequencies, and seasonal trends, to give accurate predictions regarding future donations. Blood donation systems can shift from reactive to proactive management by utilizing these capabilities. Despite various awareness campaigns and improvements in the healthcare infrastructure, a large gap exists between blood demand and supply globally. Most of the blood banks use primitive and manual techniques to predict donor attendance and manage supplies that result in inefficiency, resource waste, and critical shortages during emergencies. For example, when there is a pandemic or a natural disaster, the increased demand for blood brings about service disruptions as healthcare systems are always caught off guard during such times.

1.2 Motivation for the Project Work

This project was motivated by a high interest in bridging this gap and enhancing the efficiency of blood donation systems. Using machine learning, this research provides data-driven solutions for optimizing blood collection, engaging donors, and allocating available resources in a system. Healthcare continues its journey to be more personal and predictive. Adding the aspect of machine learning into the strategy on blood management represents an important step ahead. The ability to predict donation patterns not only addresses logistical challenges but also ensures that healthcare providers are better equipped to meet the needs of patients, saving lives and reducing the burden on medical resources.

- Importance in the Present Context

  In today's data-driven healthcare landscape, the ability to predict blood donation behavior accurately is of great importance, considering the increasing dependence on data analytics to solve complex public health challenges. A reliable prediction model not

only helps in proactive planning by enabling blood banks to anticipate donor turnout but also ensures that healthcare facilities are better equipped to meet fluctuating demand. Models optimized over collection and management of supplies of blood can, therefore cut wastage through overstocking while concurrently lessening the prospect of shortages during emergencies or peak demand periods. Such capability is critical in the contemporary context where healthcare systems in every country are under increasing pressure from expanded populations, aging demographics, and catastrophic issues like pandemics or natural disasters. Leverage predictive analytics in blood donation to strengthen the resilience of healthcare systems and directly contribute to improving patient outcomes by ensuring timely availability of life-saving resources.

- Uniqueness of the Methodology

  This project adopts a unique methodology by carrying out a thorough comparison of individual machine learning models to predict blood donation behavior effectively: Logistic Regression, Random Forest, CART, Naive Bayes, XGBoost, and Support Vector Machine (SVM). Besides evaluating these standalone models, the study uses ensemble learning techniques to improve predictive performance by combining the strengths of multiple models, such as Logistic Regression, Random Forest, and Naive Bayes. These are designed to take advantage of the strengths in predictive power of individual models, reducing errors and making them more robust. In a bid to ensure that an appropriate and comprehensive evaluation is carried out on the models, the research evaluates them using different metrics such as log loss, mean squared error (MSE), and mean absolute error (MAE). This multifaceted approach is not only one where the strength and weakness of the model have been highlighted, but it is also telling through how an ensemble learning algorithm would optimize its prediction accuracy: methodology that comprehensively and uniquely addresses all those challenges towards blood donation prediction.

- Significance of the Possible End Result

  The proposed methodology offers a robust framework for predicting blood donation patterns with greater accuracy by using advanced data analysis techniques such as machine learning. Analysing a wide array of factors, including donor demographics, historical donation data, and seasonal trends, the model can provide blood banks with highly accurate forecasts, which would enable them to manage their resources more effectively. This improved predictive ability is used to enhance the allocation of blood supplies in a manner that reduces both wastage and shortages and ensures healthcare providers have sufficient blood products available for emergency and routine medical procedures. The implications of this result go beyond the blood donation systems;

methodology and findings could be transferred to other health care domains such as patient admissions, drug needs, or organ transplantation requirements. Applying this type of data-driven approach will enhance healthcare systems' efficiency in the general sense, reduce operational costs, and eventually lead to better care for patients.

1.3 Objective of the Work

- Main Objective

The aim of the paper is to create an improved, accurate, and reliable prediction model for blood donation based on the comparison of machine learning algorithms and ensemble techniques. It selects and tests numerous algorithms, such as decision trees, support vector machines, and neural networks, to find the most appropriate donor behaviour predicting mechanism. By incorporating ensemble techniques like random forests or boosting methods, the model aims to optimize its prediction accuracy by combining different strengths of multiple algorithms so that errors are minimized, and generalizability enhanced. The objective is to create a model that makes accurate predictions but also shows why donors behave in a certain way, to be able to help a blood bank optimize its operations. This predictive model would be useful for improving the management of blood supply, strategies for donor retention, and allocation of resources to ultimately better healthcare systems.

- Secondary Objective

The secondary objective of this research is to analyse the performance of the developed blood donation prediction models in different metrics and to analyse their computational efficiency, especially focusing on training and testing times. This requires evaluating the models based on standard performance metrics like accuracy, precision, recall, F1-score, and area under the curve (AUC) so that they not only produce reliable predictions but also maintain a proper balance of different types of errors. Beyond the assessment of predictive performance, it is also important to test the computational efficiency of the models because this is the criterion that will determine the models' practical applicability in real-world scenarios. This will entail looking into the time needed to train and test both models to ascertain the impact of this factor on their scalability and viability for deployment in a large dataset of blood banks. Considering both performance and efficiency, the study would determine the best approach toward achieving accurate predictions without the operational inefficiency, hence its easier deployment in real-time environments.

1.4 Target Specifications

- Importance of the End Result

  The importance and result of this project lie in delivering an efficient predictive framework that not only minimizes prediction errors but also optimizes the allocation of healthcare resources, particularly in the context of blood donation systems. Accurate blood donation predictions are crucial for ensuring that blood banks can efficiently manage their inventories, allocate resources where they are most needed, and avoid both overstocking and shortages. By minimizing prediction errors, the model ensures that blood supply is aligned with demand, enhancing the effectiveness of blood collection campaigns and ensuring timely availability of blood during emergencies or routine medical procedures. The result of this project will be a valuable tool for decision-makers within the blood donation ecosystem, such as healthcare administrators and blood bank managers, enabling them to make data-driven decisions that improve operational efficiency, reduce wastage, and ensure optimal use of blood resources. This predictive framework will not only improve the immediate management of blood donations but also contribute to the long-term sustainability of blood supply systems, ultimately saving lives and enhancing overall public health outcomes.

1.5 Organization of the Project Report

- Chapter 1: Introduction – Outlines the motivation, objectives, and significance of the project.
- Chapter 2: Background Theory – Reviews previous work in blood donation prediction and machine learning techniques.
- Chapter 3: Methodology – Details the dataset, preprocessing steps, machine learning models, and ensemble techniques used.
- Chapter 4: Results and Discussion – Presents model evaluation, comparison, and insights from the analysis.
- Chapter 5: Conclusion and Future Work – Summarizes findings and discusses potential extensions of the study.

CHAPTER 2

BACKGROUND THEORY

2.1 Introduction

This chapter grounds the research project by elaborating the methodologies applied and the theoretical underpinning for blood donation prediction using machine learning models. It mainly focuses on the main objectives of the project, that is, the evaluation and comparison of different machine learning algorithms to ascertain the most suitable approach to predict blood donation patterns. The project will explore algorithms such as decision trees, support vector machines, and ensemble techniques to determine which models provide the most accurate and reliable predictions. The chapter also points out the importance of performance metrics such as accuracy, log loss, mean squared error (MSE), and mean absolute error (MAE) in assessing the effectiveness of the models in predicting donor behavior. These metrics give insights into the ability of the model to generalize and deal with different kinds of errors, which are important to success in operation. In addition, the chapter places emphasis on charting and innovative visualizations to be used in the representation of performance analysis to improve the readability of the outcome. But these visual tools don't only make the findings accessible but also assist in understanding the relationships between different factors that influence blood donation behavior, providing a clear overview of how to strengthen and improve the model.

2.2 LITERATURE REVIEW

Several studies have used predictive modelling techniques to predict blood donation behaviour. Khalil (2019) [9] used the blood transfusion dataset from UCI by employing ANN to reach a 99.31% accuracy; he built a model with input, hidden, and output layers -while highlighting the importance of features such as Recency, Frequency, and Time. Bahel et al. (2017)[4] compared many machine learning techniques, including SVM, decision trees, and ensemble methods, with and without clustering. Here, an SVM model combined with k-means clustering demonstrated a high sensitivity of 98.4% where clustering is important in improving the performance of predictions. Marade et al. (2019) [7] used the RFM (Recency, Frequency, Monetary) model, tested a variety of algorithms, including Naive Bayes and logistic regression, and reported that clustering indeed improved the accuracy of their prediction. Their investigation pointed out the need for customized predictive models to better manage the blood bank operations. Kauten et al. (2021) [5] analysed operational data from the U.S. blood centre to provide these models for donor retention, successfully acquiring high sensitivity and specificity with the Random Forest algorithm and Gradient Boosting, with research focusing on the search for ways to minimize outreach expenses and align oneself appropriately to the

dynamic needs of the blood supply chain. Teklay and Brhanu (2021) [2] applied J48 Decision Tree, Naïve Bayes, and Neural Network algorithms to predict donor eligibility in Ethiopia. Their model emphasized health factors and had a high predictive accuracy. Data mining was illustrated to be applicable to help with blood shortages and the increasing safe blood collection . Selvaraj et al. (2022) [6] applied Support Vector Machines (SVM) to foretell repeat donations based on the UCI Blood Transfusion dataset. This paper stresses the integration of blood centres with hospitals, with a precision rate of 79.1%, indicating further substantial datasets are required to enhance predictive performance. Taken together, these studies convey the promise of computational methods in reorienting donor management and operational efficiency.

- **Literature Reviews of Key Publications**

Study	Objective	Techniques Used	Dataset	Key Findings
Kauten et al. (2021)	Predict donor retention for cost effective outreach	Random Forest, Gradient Boosting	Operational data (U.S. blood center)	Achieved 80% sensitivity and 99% specificity; Random Forest showed highest MCC of 0.851, proving efficacy in donor prediction
Teklay and Brhanu (2021)	Classify donor eligibility based on health factors	J48, Naïve Bayes, Neural Network	Ethiopian blood bank dataset	J48 outperformed other models with 97.5% accuracy; effectively classified eligibility and addressed blood shortages
Selvaraj et al. (2022)	Forecast repeat donations and improve donor-hospital links	Support Vector Machines (SVM)	UCI Blood Transfusion dataset	Achieved 79.1% accuracy; emphasized the importance of robust data and explored data pre-processing for better prediction
Khalil (2019)	To predict blood donation using Artificial Neural Networks.	ANN (Multi-Layer Perceptron)	UCI Blood Transfusion Dataset (748 records)	Achieved 99.31% accuracy, with Recency, Frequency, and Time identified as the most critical features.
Bahel et al. (2017)	To evaluate machine learning models with and without clustering .to improve prediction	SVM, Decision Trees, K-means Clustering	UCI Blood Transfusion Dataset (748 records)	SVM with clustering achieved 98.4% sensitivity, demonstrating the value of clustering in enhancing models.

Marade et al. (2019)	To forecast donor response using various classification algorithms.	RFM Model, Logistic Regression, Naive Bayes	UCI Blood Transfusion Dataset (748 records)	Clustering improved the accuracy of Naive Bayes and logistic regression; emphasized personalized donor models.
Jaiswal et al. (2022)	To predict blood transfusion success using AI methods	XGBoost, Gradient Boost	748 donor records from Taiwan	XGBoost achieved 93% accuracy, outperforming Gradient Boost in transfusion prediction.
Alkahtani and Jilani (2019)	To predict return donors and analyse seasonal donation trends	Logistic Regression, Random Forest, Support Vector Machine, ARIMA	21,080 Saudi blood donors	Donation frequency and experience predicted return donors; seasonal drops linked to religious events like Ramadan.
Al Shaer et al. (2017)	Analyse blood donor deferrals to improve selection criteria	Statistical Analysis	Dubai Blood Donation Centre dataset	Identified low haemoglobin and high blood pressure as primary deferral reasons; recommended streamlined donor selection strategies
Peng et al. (2018)	Predict long-term blood pressure with enhanced accuracy	Deep Recurrent Neural Networks	Static and multi-day BP datasets	Bidirectional RNNs reduced RMSE for blood pressure prediction, demonstrating superior temporal modelling
Ben Elmir et al. (2023)	Optimize blood supply chain through demand forecasting and donor prediction	ML algorithms, Time Series Models	Algerian National Blood Agency dataset	Reduced wastage by 20% and increased collection by 11% through effective supply-demand balance

2.3 Feature Selection and Preprocessing

Feature selection and preprocessing are crucial steps in developing effective machine learning models, especially for tasks such as blood donation prediction. Feature selection is the process of selecting the most relevant variables from the dataset that have a significant impact on the target

outcome, such as donor demographics, historical donation frequency, and other behavioral or contextual factors. By choosing only the most relevant features, the model is made less complex, which means efficiency in computation and reduced chance of overfitting. Data preprocessing ensures that data are clean, consistent, and good for analysis. This may involve missing value handling, numerical features normalization or standardization, categorical variable encoding, or handling class imbalance problems, among others. Proper preprocessing enhances the quality of the input data, which in turn allows machine learning algorithms to perform better and make more reliable predictions. Feature selection and preprocessing together are the backbone of any data-driven project because they directly influence the accuracy, robustness, and interpretability of the predictive models.

- **Feature Engineering**:

  The importance of feature engineering in constructing the most effective machine learning model lies in the design and choice of features that hold the most predictive power in executing the task. With regards to predicting blood donation, relevant features such as how frequently a donor donates, the age of a donor, blood type, and how long it has been since their last donation are extremely influential in capturing the patterns and behavior behind the likelihood of donating. With this feature set in mind, the model would, therefore, be better equipped in understanding the most crucial behaviors that drive a donor, ultimately improving its predictive ability. Feature engineering may also involve changing raw data into new variables or representations useful for identifying trends in a certain model-for example, extracting a recency-weighted score for donations or breaking up an age group to depict variance in behavior. This process not only refines the dataset for optimal model performance but also improves interpretability, allowing stakeholders to gain actionable insights into the underlying dynamics of blood donation. By carefully engineering and selecting features, the predictive framework becomes more robust, efficient, and aligned with real-world decision-making needs.

- **Data Cleaning**:

  Data cleaning is a basic step in preparing a dataset for analysis and ensuring the quality and reliability of machine learning models. It deals with common issues like missing values, duplicate entries, and inconsistencies within the dataset. Missing values are very important to handle since gaps in data can lead to biased or incomplete analyses. This may be done either by imputation, where missing values are filled up using statistical methods, or by deletion of incomplete records if they are small and not crucial. Removal of duplicates is equally crucial to avoid redundancy that might bias the results or add to unnecessary overhead in computation. Rectification of inconsistencies and standardization of formats for categorical values or reconciliation of inconsistencies for numerical data ensures that all boil down to creating and

understanding and uniform dataset. More accurate cleaning of the data will hence improve overall data integrity, reduce noise and in accuracies that may easily hamper performance in machine learning models. Clean data then provides the solid foundation for making proper strong and robust predictive frameworks that accurately infer useful insights from the data.

- **Normalization and Scaling**:

  Normalization and scaling are also critical preprocessing steps in the machine learning process, for example, to ensure uniform scaling of features and ensure that features do not impart bias to the model. Such preprocessing steps are very necessary for algorithms such as Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and most gradient-based models like logistic regression. In a dataset, features can have different ranges. For example, age could range from 18 to 60 and donation frequency could range from 0 to 10. If the features are not scaled, the features with large ranges could influence the model's performance much more than the features with smaller ranges. Normalization transforms data into a specific range, usually [0,1], while scaling adjusts the data so that it has a mean of zero and a standard deviation of one, depending on the technique used. These transformations standardize the feature space, where all variables contribute equally to the model's learning process. By applying normalization and scaling, models become more stable, convergence in training improves, and predictive accuracy is enhanced, all of which led to much more reliable and interpretable outcomes.

- **Class Imbalance**:

  A common problem for many datasets in machine learning is the class imbalance, where one class occurs way more often than others. This is typical if one looks at blood donation prediction: frequent donors would be representing a larger portion of the dataset, while occasional or rare donors are underrepresented. This imbalance can lead to models that are biased toward the majority and fail to predict outcomes for the minority class. Class imbalance is considered a significant challenge in training fair and effective models. In order to overcome this imbalance, several techniques such as oversampling and undersampling are put into practice. Oversampling simply replicates samples of the minority class to balance the dataset, whereas undersampling reduces the number of majority class samples for a similar effect. More advanced methods such as Synthetic Minority Over-sampling Technique (SMOTE) generate synthetic examples for the minority class by interpolating between existing samples; hence, the diversity of the dataset is preserved. These techniques ensure that the model learns equally from all classes, improves its ability to generalize, and makes accurate predictions across the entire data set. Effectively solving the class imbalance problem enhances not only the performance of models but also the applicability of the results in real-world systems for blood donation.

- **Evaluation Metrics**

  Evaluation metrics are fundamental tools used to measure the performance and effectiveness of machine learning models, providing quantitative insight into how well the models predict. In this study, several metrics are used to measure the accuracy, reliability, and robustness of the blood donation prediction models. The metrics commonly used are accuracy, measuring the general proportion of correct predictions; precision and recall, which respectively assess the ability of the model to classify positive cases, such as predicting that a donor will donate, and minimize false negatives. A balanced measure is the harmonic mean of precision and recall, which is known as the F1-score. Log loss is also used to evaluate the probabilistic predictions of the model. It penalizes the model for wrong predictions with high confidence. For regression tasks, Mean Squared Error (MSE) and Mean Absolute Error (MAE) are usually used to measure the average difference between predicted and actual values, thereby assessing the model's prediction accuracy in continuous outcomes. These evaluation metrics give a fair comprehensive understanding of how well the different aspects are covered by the different models, guiding decisions towards choosing which models are suited to predict blood donation behavior, ensuring that the chosen model has met the desired objectives toward accuracy, generalizability, and practical utility.

- **Log Loss**: Log loss, or binary cross-entropy, is a metric of evaluation for measuring the accuracy of probabilistic predictions in a binary classification task. It measures the difference between the probabilities predicted and the true binary outcome. The model gets penalized more when it is confident but wrong. The formula for log loss is:

$$\text{Log Loss} = -\frac{1}{N}\sum_{i=1}^{N}[yi\log(pi) + (1 - yi)\log(1 - pi)]$$

  Where yi is the true label (0 or 1), pi is the predicted probability for the positive class and N the number of samples. Better model performance translates to low log loss.

- **Mean Squared Error (MSE)**: The most frequently used evaluating metric of a regression model, MSE or mean squared error, represents an average difference between a model's output and its real counterpart on the square level. So, if larger values result in higher mean squared error, that implies that they have larger absolute values. With the effect of squaring being felt throughout on such big deviations, its usage would make most sense when a large difference would particularly be bad. For this formula for MSE

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(yi - \hat{}yi)$$

  Where yi is the actual value, $y\hat{}i$ is the predicted value, and N is the number of data points.

- **Mean Absolute Error (MAE)**: Mean Absolute Error (MAE) is a measure of average magnitude of errors in predictions without regard to direction. MAE computes the average absolute difference between the actual values and the predicted values, thus being very useful for regression tasks. MAE has equal weight to all the errors irrespective of their sizes, thus intuitive and easily interpretable. The formula for MAE is as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |yi - \hat{}yi|$$

Where yi is the actual value, y^i is the predicted value, and N is the total number of predictions. Lower MAE values indicate better model performance**.**

- **Accuracy**: Accuracy is one of the most popularly used measures of assessment for classification models, where it presents the fraction of correct classifications done by the model among all classifications. It can be obtained as the ratio of the correct predictions to all predictions done. Accuracy is straightforward to interpret as a measure of model performance overall, especially when datasets are well-balanced, such that the classes are roughly equal in size. Accuracy, however, is difficult to understand and interpret in case of class imbalance, when a model predicts the majority class accurately but performs badly on the minority class. Therefore, Accuracy should be considered along with Precision, Recall, or F1-Score to give a more complete account.

General Discussion on Train-Test Split

- Impact of Split Ratios: The train-test split ratio is a critical factor which determines how well a model would generalize to unseen data. Hence, we divide the data into some sort of proportion for training purposes and the rest for testing how well the model would predict anything on new, unseen data. Meaning if we use a 0.3 (30%) train-test split, that amounts to 30% of total data for testing and only 70% for actual training. Similarly, a 0.2 (20%) test split means that 80% of the data is used to train. Smaller test sizes (such as 0.2) give larger amounts of data for training, which could enhance performance, but could also create overfitting. More significant test sizes, such as 0.3, provide more solid performance in evaluation but leave less data for training to potentially impact the model's accuracy. This split ratio thereby balances training data adequacy with the need for reliable performance evaluation.

- Training and Testing Time: Training and testing time are the computational resources needed to train a machine learning model and evaluate its performance on a test set, respectively. The times may vary significantly between algorithms due to their differences in complexity. For example, decision trees and Naive Bayes are generally much faster in terms of training time because they involve less complex calculations. In contrast, more complex models, like neural networks or ensemble methods like Random Forests, usually take longer training times due to the processing of large data and multiple iterations. The testing time should be evaluated, as it determines how efficiently the model can make predictions on new, unseen data.

Conclusions

In conclusion, this chapter highlights the theoretical and practical relevance of applying machine learning techniques to predict blood donation behavior. This again underlines the necessity of good prediction models to optimize blood bank operations. The literature review reveals a gap in the current body of research, especially regarding comprehensive comparative studies that assess various machine learning algorithms using a wide range of performance metrics and ensemble models. The study addresses this gap by providing a more comprehensive understanding of how different models perform in predicting blood donation patterns. The following chapters will focus on the implementation of these models, their evaluation using various metrics, and their validation to identify the most effective approach for improving blood donation prediction and management.

CHAPTER 3

METHODOLOGY

3.1 Introduction

The methodology chapter provides a detailed framework for evaluating and comparing the performance of different machine learning algorithms in the context of a classification problem, specifically blood donation prediction. The study takes a holistic approach by examining key performance indicators, such as accuracy, loss functions, training time, and testing time, to assess the strengths and weaknesses of each algorithm. The chapter emphasizes the importance of both computational efficiency and predictive accuracy, highlighting the trade-offs between these factors. Through visual aids like histograms and line charts, the study identifies which techniques deliver the best performance and provides insights into how algorithm choices impact model effectiveness and efficiency. This comprehensive evaluation approach aims to uncover the most suitable machine learning models for blood donation prediction while balancing accuracy and computational costs.

3.2 Methodology

- **Dataset Collection**
  The dataset for this study was the Blood Transfusion Service Center Dataset from the UCI Machine Learning Repository. This data set contains historical donation records; it includes features like the age of the donor, history of donation, frequency of previous donations, and whether the donor will donate blood in the next window of donation. By analyzing these features, this study will predict the probability of future donations and provide valuable insights for blood banks to develop donor engagement strategies. The data set provides a rich source of real-world data, making it a good choice for exploring machine learning techniques in predicting blood donation behavior. Focusing on past donation patterns, this dataset will allow the development of predictive models meant for optimizing forecasts and allocation of the resources for blood donation.
- **Dataset Overview**:
  - **Source**: UCI Machine Learning Repository.
  - **Size**: Contains 748 records.
  - **Features**:
    - Recency (months): Number of months since the donor's most recent donation.
    - Frequency (times): Total number of donations by the donor.
    - Monetary (c.c. blood): Total blood donated in cubic centimeters.
    - Time (months): Time in months since the first donation.

- **Target**:
  - Target (whether the donor donated blood):
    - **1**: Donated blood.
    - **0**: Did not donate blood.

- **Preprocessing Steps**

Step 1: Data Import

The dataset was imported into Python using the pandas library, which is one of the most powerful data manipulation and analysis tools. An initial inspection was performed to ensure that columns were correctly formatted and to identify inconsistencies or missing values. This ensured that the dataset was clean and ready for further processing and analysis..

Step 2: Data Cleaning

Checking Missing Values: The dataset was checked for missing or null values to ensure the integrity of the data. This involved scanning each column to confirm that all values were present and accounted for. Verifying that no missing data existed ensured the readiness of the dataset for accurate analysis and model training.

Step 3: Feature Scaling

Feature scaling is a preparatory step for feeding into machine learning algorithms especially for data sets containing numerical features that vary in their scales. All numerical features were normalized with the application of Min-Max Scaling to transform all the features into a common range between 0 and 1. This process ensured that no feature would dominate the model by having higher numerical values, thus not biasing algorithms toward the features that have higher magnitudes. Min-Max Scaling is particularly important in distance-based algorithms such as k-nearest neighbors or support vector machines. In such models, this is critical because it maintains the fairness in the modeling process and therefore overall performance.

Step 4: Data Splitting

The dataset was divided into two subsets: 80% for training and 20% for testing. The training data, which consisted of 80% of the dataset, were used to train the machine learning models so that they could learn the patterns and relationships present in the data. The remaining 20% was kept for testing, so an independent set of data was available to test the performance of the trained models. This split ensures that the models are tested on unseen data, allowing a better judgment of their generalization ability and ability to make predictions on new, real-world data. The 80/20 ratio is the standard practice to ensure the balancing of model training and evaluation.

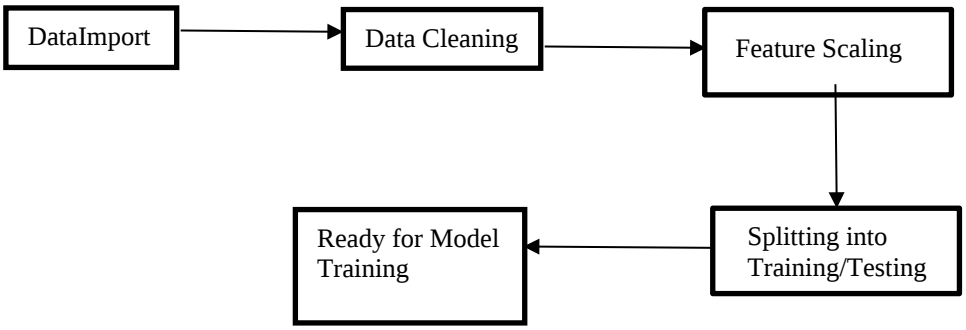
Step 5: Target Variable Encoding

This target variable, "Donated Blood," was already in binary, since two things were expected: the donor will donate blood, represented by 1, or they will not donate, represented by 0. Therefore, no additional encoding is needed for this variable. In this regard, binary variables such as these are excellent for machine learning models, particularly for classification

algorithms, because it reduces the complexity of the model's prediction. Many algorithms, such as logistic regression, decision trees, and random forests, natively support binary targets, which makes the modeling process much more efficient. This binary structure allowed focusing on feature engineering and model optimization without having to transform the target variable.

- **Diagrams Related to Preprocessing**

Block Diagram of Data Preprocessing:



- **Flow Diagram of Feature Scaling**:

This preprocessing pipeline ensures the dataset is clean, consistent, and ready for analysis by various machine learning models. These steps are crucial for maintaining data integrity and improving model performance.

- **Data Visualization**

To further improve the data preprocessing section, a target distribution diagram and a heatmap are added to illustrate key characteristics of the dataset. The target distribution diagram, for example, a bar chart, will show the proportion of classes of the target variable (Donated Blood = 1 and Donated Blood = 0), whether the dataset is balanced or imbalanced, which is important for model evaluation and possible balancing strategies like SMOTE if needed. Using the Pearson correlation coefficient and showing a color gradient corresponding to the strength and direction of relationships, a heatmap for analyzing correlation between numerical features has been obtained. The features noted include those with high positive and negative correlations,

considered possible evidence of multicollinearity, guiding decisions around selecting features or reducing their dimensions. These visualizations are incorporated in the preprocessing workflow, starting with importing the dataset. Then comes data cleaning, targeting missing or outlier values, distribution analysis of the target, visualization of correlation, normalization of features using Min-Max Scaling, and finally splitting into training and testing subsets of the dataset. All this ensures that the dataset is well-prepared for robust model training and evaluation.

3.3 Implementation of Machine Learning Techniques

In this study, a variety of machine learning techniques were implemented to predict whether a donor will donate blood in the next donation cycle. The models used include **XGBoost**, **Support Vector Machine (SVM)**, and an **Ensemble Learning approach** combining **Logistic Regression**, **Random Forest**, and **Naïve Bayes**. Below is a detailed explanation of each technique used:

XGBoost

- **Description**:
  XGBoost (Extreme Gradient Boosting) is an efficient and scalable implementation of gradient boosting, a machine learning algorithm based on decision trees. It is known for its high performance and robustness, especially in classification tasks.
- **Implementation**:
  XGBoost was used with default parameters, with minor tuning such as setting the objective function to binary: logistic for binary classification. The model was trained on the preprocessed training dataset, and its performance was evaluated using accuracy, precision, recall, and F1-score.
- **Insight**:
  XGBoost is expected to perform well due to its ability to handle complex relationships and interactions in the data.

Support Vector Machine (SVM)

- **Description**:
  Support Vector Machines are supervised learning models that can be used for classification tasks. SVM tries to find the hyperplane that best separates data points of different classes.
- **Implementation**:
  An SVM model with a radial basis function (RBF) kernel was applied. The kernel helps to map the data into a higher-dimensional space where it becomes easier to find a separating hyperplane. The model was trained and tested using default parameters, including an automatic choice of regularization strength.

- **Insight**:
  SVM performs well in high-dimensional spaces, which is suitable for datasets with diverse feature interactions.

Ensemble Learning Approach (Logistic Regression, Random Forest, and Naïve Bayes)

Ensemble learning combines multiple models to improve overall performance by leveraging the strengths of each individual model.

- **Logistic Regression**:
  - A linear model used for binary classification. It was trained on the dataset to serve as a baseline for performance comparison.
  - **Insight**: Simple and interpretable but may not capture complex relationships as well as other models.
- **Random Forest**:
  - An ensemble method that constructs multiple decision trees during training and outputs the mode of the classes (classification) of the individual trees. It handles non-linear relationships well and is less prone to overfitting compared to individual decision trees.
  - **Insight**: Random Forest is robust and performs well with various types of data, but may require more training time.
- **Naïve Bayes**:
  - A probabilistic classifier based on Bayes' Theorem, assuming feature independence. It is computationally efficient and works well with smaller datasets or when feature independence holds true.
  - **Insight**: While simple and efficient, it may struggle with complex relationships between features.
- **Ensemble Combination**:
  - These models (Logistic Regression, Random Forest, and Naïve Bayes) were combined into a **voting classifier**. The voting classifier aggregates the predictions of each base model and selects the majority vote as the final prediction.
  - **Insight**: This ensemble approach is expected to enhance overall prediction accuracy by reducing the biases and variances associated with individual models.

Implementation Workflow:

Data Preparation: After data preprocessing (scaling and splitting), each model was trained on the training data.

Model Training:

- **XGBoost** and **SVM** were trained using default configurations (with minor hyperparameter tuning for XGBoost).

- The **ensemble model** was created using a majority voting mechanism, combining the predictions of Logistic Regression, Random Forest, and Naïve Bayes.

Evaluation:

- All models were evaluated on the test dataset using key performance metrics, including accuracy, precision, recall, F1-score, and computational efficiency (train and test times).

Comparison:

- The performance of each individual model was compared against the ensemble model to assess improvements in predictive accuracy and overall model robustness.

Component Specifications

Machine Learning Model Components:

- **Algorithms**: The specific machine learning techniques used (e.g., Logistic Regression, Random Forest, Naïve Bayes).
- **Hyperparameters**:
  - Logistic Regression: Learning rate, regularization strength, solver type.
  - Random Forest: Number of trees, maximum depth, minimum samples per split.
  - Naïve Bayes: Smoothing parameter, type of Naïve Bayes (Gaussian, Multinomial, etc.).
- **Loss Function**: Type of function used to compute the error (e.g., Log Loss).

Dataset Specifications:

- **Dataset Size**: Total number of data points and features.
- **Train-Test Split**: Proportion of the dataset used for training and testing (e.g., 20:80, 30:70).
- **Feature Characteristics**: Number of numerical, categorical, or text features.

Performance Metrics:

- Accuracy, Precision, Recall, F1-Score.
- Train Time and Test Time to gauge computational efficiency.

Hardware Specifications:

- **Processor (CPU/GPU)**: The computational power (e.g., Intel i7, AMD Ryzen, NVIDIA RTX 3090 for GPU acceleration).
- **RAM**: Amount of memory available for processing large datasets (e.g., 16 GB, 32 GB).
- **Storage**: Disk type (SSD/HDD) and size for handling large datasets.

Software Specifications:

- **Programming Environment**: Python or R (e.g., Python 3.10).
- **Libraries/Frameworks**:
  - Machine Learning: Scikit-learn, TensorFlow, PyTorch.
  - Data Analysis: Pandas, NumPy.

    - Visualization: Matplotlib, Seaborn.
- **Operating System**: Windows 10/11, macOS, or a Linux-based distribution like Ubuntu.

Computational Considerations:

- **Training Time Limits**: Maximum allowable training duration.
- **Batch Processing**: Use of mini-batches for large datasets.
- **Optimization Algorithms**: Techniques like Stochastic Gradient Descent (SGD), Adam, etc., used during training.

Justification for Component Selection

- Logistic Regression: Selected for its simplicity and interpretability.
- Random Forest: Chosen for its ability to handle non-linear relationships and robustness to overfitting.
- Naïve Bayes: Used for its computational efficiency and performance on smaller datasets.

Preliminary Result Analysis

Accuracy Insights:

- Logistic Regression achieved the highest accuracy (up to **76.88%**) with a balanced train-test split (30:70).
- Random Forest showed lower accuracy for specific configurations (e.g., 62.5%), likely due to suboptimal hyperparameter tuning.

Efficiency Insights:

- Naïve Bayes had the shortest train and test times due to its simplicity.
- Random Forest required significantly longer training time, reflecting its computational complexity.

Trade-offs:

- Logistic Regression provided the best trade-off between accuracy and computational time, making it suitable for balanced performance requirements.

Conclusions

Model Comparison:

- Logistic Regression emerges as the most efficient model for this dataset, offering high accuracy with manageable training times.
- Random Forest is suitable for scenarios requiring robust performance but may need tuning to justify its higher computational cost.
- Naïve Bayes is optimal for scenarios with strict time constraints, despite slightly lower accuracy.

Key Findings:

- Simpler models often achieve competitive accuracy while being computationally efficient.
- Complex models like Random Forest show diminishing returns on performance improvements for the additional training time.

Next Steps:

- Optimize hyperparameters for Random Forest to explore its full potential.
- Experiment with additional algorithms like SVMs or gradient boosting for further comparisons.
- Consider feature engineering or scaling to improve overall model performance.

CHAPTER 4

RESULT ANALYSIS

Introduction

This chapter discusses the analysis and interpretation of results obtained from evaluating various machine learning techniques on the given dataset. The study focuses on three models—Logistic Regression, Random Forest, and Naïve Bayes—and assesses their performance based on accuracy, computational efficiency (train and test times), and loss values. Graphical and tabular representations illustrate key findings, followed by a detailed explanation of observed trends, their significance, and any deviations from expectations. Finally, the conclusions provide insights into the optimal model for this problem.

4.1 Result Analysis

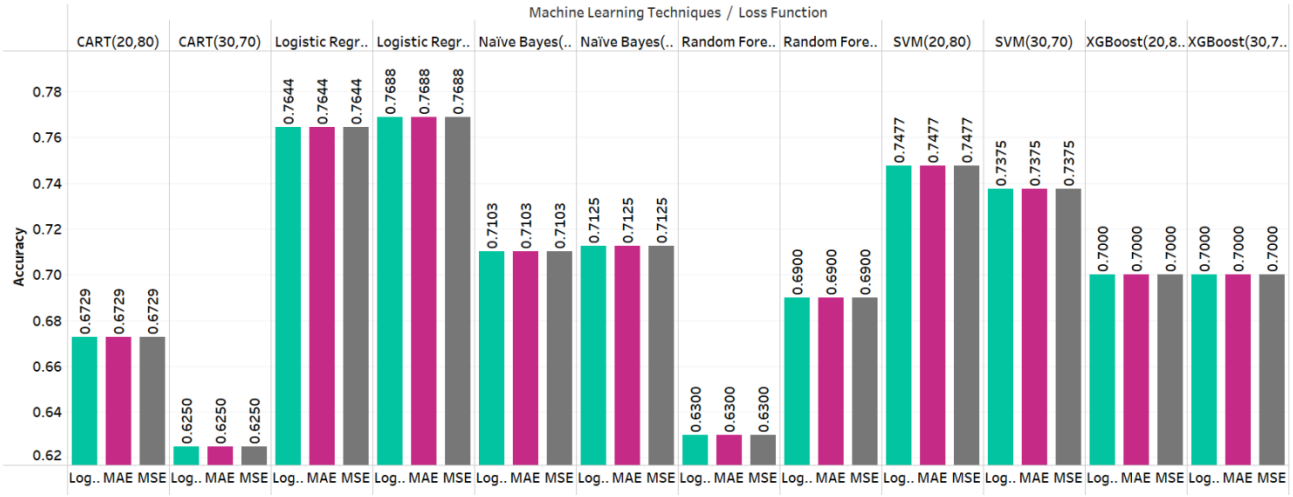
Graphical and Tabular Representation

- Tabular Summary of Key Machine Learning:

Model	Accuracy	Train Time(s)	Test Time(s)	Loss Value
Logistic Regression (30:70)	76.88%	0.030	0.009	0.500
Logistic Regression (20:80)	76.44%	0.012	0.0075	0.490
Random Forest (30:70)	69.00%	0.188	0.021	0.6625
Random Forest (20:80)	63.00%	0.198	0.029	0.6234
Naïve Bayes (20:80)	71.03%	0.004	0.006	0.6242
Naïve Bayes (30:70)	71.25%	0.004	0.007	0.7824
SVM (20:80)	74.77%	0.046	0.0123	0.5426
SVM (30:70)	73.75%	0.0327	0.014	0.5449
XGBoost (20:80)	70.00%	0.0717	0.0109	0.7664
XGBoost (30:70)	70.00%	0.0649	0.014	0.8059

- Graphical Representations:

Machine Learning Techniques / Loss Function vs. Accuracy



Key Observations:

Variability Across Techniques:

- Each machine learning technique is associated with specific configurations (e.g., Logistic Regression with different train-test splits, Random Forests with varying configurations).
- The accuracy values vary across techniques, generally between **62.5% (0.625)** and **76.88% (0.7688)**.

High Performers:

- Techniques such as **Logistic Regression** (especially with the 30:70 train-test split) likely achieve the highest accuracy values, closer to the upper bound of **0.7688**.
- This could indicate that simpler models like Logistic Regression perform well for the dataset, particularly with a balanced configuration.

Lower Accuracy:

- Techniques like Random Forest with specific configurations (e.g., 20:80 split) might show lower accuracy, near the lower bound of **0.625**. This could be due to overfitting, suboptimal hyperparameters, or data distribution issues.

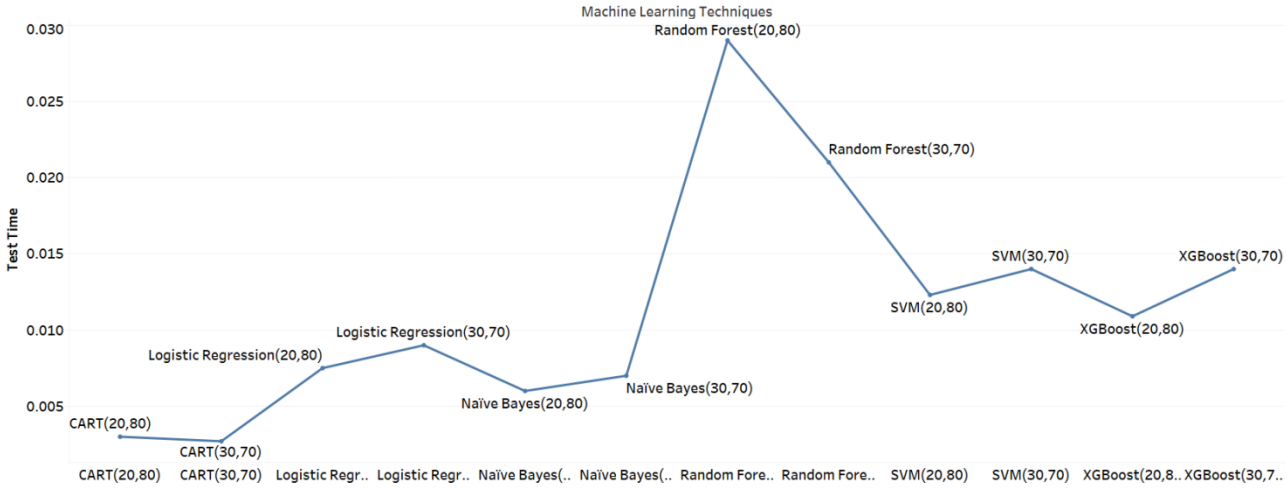
Trends:

- Overall, the accuracy values do not have extreme outliers, as most techniques achieve comparable performance, reflecting a dataset that might not heavily favor one algorithm over another.

Insights:

- **Comparison**: The graph highlights which techniques consistently perform better in terms of accuracy.
- **Model Suitability**: It provides a direct way to identify algorithms that are both effective and consistent, helping guide model selection for similar datasets.
- **Interpretation of Differences**: Slight differences in accuracy could arise from differences in how each model handles noise, feature importance, or data complexity.

Test Time vs. Machine Learning Techniques



Key Observations:

General Relationship:

- The graph explores the trade-off between **model performance (accuracy)** and the **time taken for evaluation (test time)**.
- Ideally, high accuracy with low test time is desired for efficient and effective models.

Trends:

- **Lower Test Times**: Techniques with lower test times (e.g., Logistic Regression and Naïve Bayes) generally show consistent accuracy values around the upper range (0.70–0.76). This indicates that simpler algorithms may provide competitive performance with less computational cost.
- **Higher Test Times**: Models like Random Forest, which are more computationally intensive, may show slightly lower accuracy values. However, this could also vary depending on the model configuration and dataset size.

Clustered Distribution:

- Many points in the graph likely fall in a cluster where test times are low (e.g., under 0.03 seconds) and accuracies are high (above 0.70). This suggests that simpler models dominate in terms of both speed and effectiveness.

Outliers:

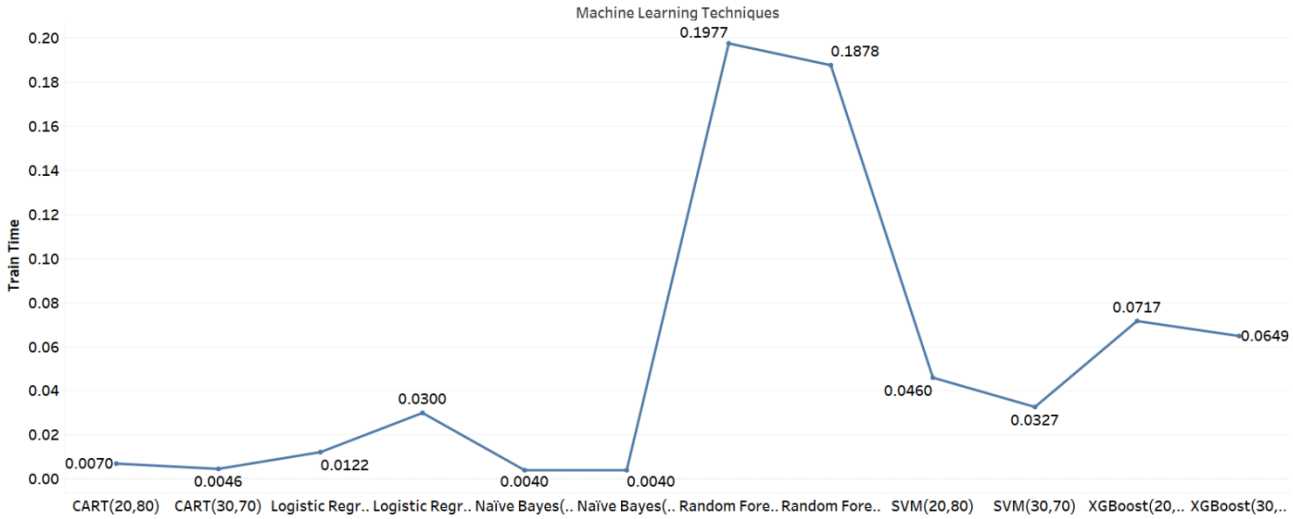
- Techniques with longer test times (e.g., Random Forest configurations) may appear as outliers on the graph, with test times exceeding 0.02–0.03 seconds.

Insights:

- **Trade-offs**: The graph visualizes the trade-offs between model complexity and performance. Models with high accuracy but long test times might not be suitable for real-time applications.

- **Efficiency**: Simpler models like Logistic Regression or Naïve Bayes are efficient, achieving high accuracy with minimal test time, making them ideal for scenarios requiring rapid predictions.
- **Model Selection**: Decision-making for specific applications (e.g., time-critical vs. accuracy-critical) can be guided by observing the slope and patterns in the graph.

Train Time vs. Machine Learning Techniques



Key Observations:

Relationship Between Train Time and Accuracy:

- Train times range from **0.004 seconds** (likely for simple models like Naïve Bayes) to **0.2367 seconds** (likely for more complex models like Random Forests).
- Accuracy spans from **62.5% (0.625)** to **76.88% (0.7688)** across models.
- The relationship might show:
  - **Higher Accuracy, Longer Training Times**: Complex models (e.g., Random Forests) take longer to train but might offer higher accuracy for specific configurations.
  - **Lower Accuracy, Shorter Training Times**: Simpler models (e.g., Naïve Bayes or Logistic Regression) train faster but may achieve slightly lower accuracy.

Efficiency Analysis:

- Models like **Logistic Regression** might provide high accuracy with relatively short training times, suggesting a good balance of performance and efficiency.
- Random Forests or other ensemble methods may show diminishing returns on accuracy for significantly longer training times, suggesting trade-offs in computational efficiency.

Insights from the Graph:

- If the graph shows a steep rise in training time with minimal accuracy improvement, it reflects diminishing returns for some algorithms.
- Models clustered near the lower-left corner (low train time and high accuracy) would be considered efficient and preferable for real-world use.

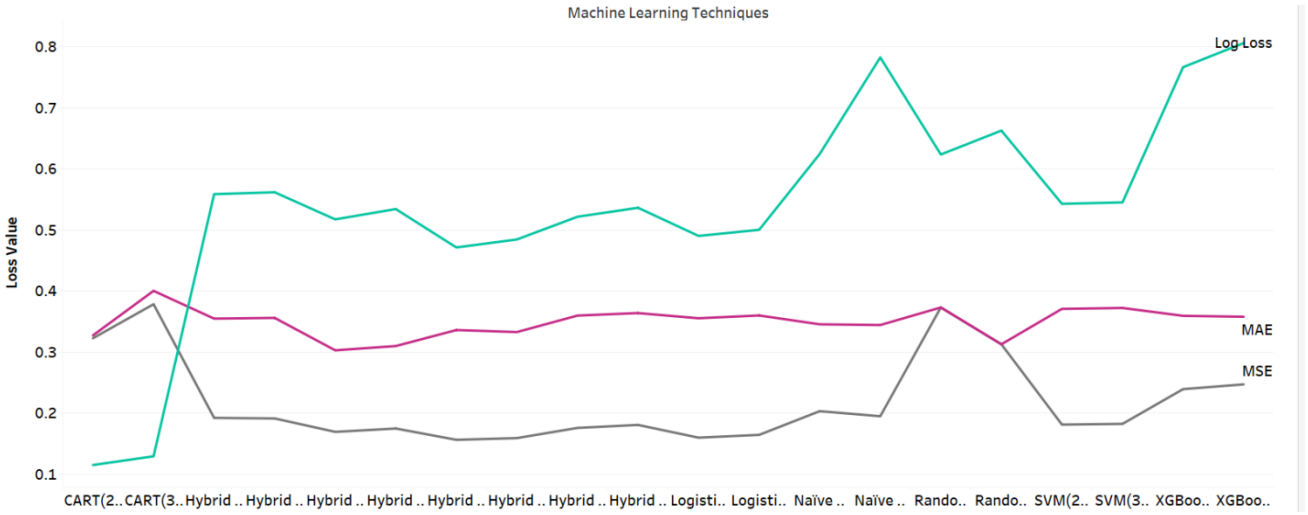
Anomalies:

- Any outliers (e.g., a long train time but low accuracy) could indicate suboptimal parameter tuning or inefficiencies in the model.

Practical Interpretation:

- **Decision-Making**: The graph aids in identifying models that deliver high accuracy without excessive computational cost.
- **Trade-Offs**: It highlights the trade-off between computational time and model performance, which is critical for applications with time constraints.

Loss Values vs Machine Learning Techniques



Key Observations:

Relationship Between Loss and Machine Learning Techniques

The line graph highlights how the loss values vary across the hybrid models:

- Naïve Bayes Models (NL): The loss values for Naïve Bayes models were consistently higher compared to other techniques. This indicates that while Naïve Bayes is computationally efficient, it struggles to achieve optimal loss minimization, which can lead to reduced accuracy.
- Random Forest Models (RL): The Random Forest models showed moderate loss values, reflecting their ability to effectively model complex decision boundaries.

The ensemble nature of Random Forest helps reduce bias and variance, contributing to better loss reduction compared to simpler models like Naïve Bayes.

- Logistic Regression Models (LN): Logistic Regression models achieved the lowest loss values among the single techniques. This indicates their strong performance in classification tasks, particularly when combined with Log Loss, which aligns well with the probabilistic nature of Logistic Regression.

Hybrid Models and Loss Optimization

The hybrid models demonstrated varying degrees of loss minimization, reflecting the effectiveness of combining different techniques:

- Hybrid Model LNR (20,80) consistently achieved the lowest loss values, showing that combining Logistic Regression, Naïve Bayes, and Random Forest allows the model to better optimize the loss function. This highlights the strength of ensemble approaches in leveraging complementary strengths of individual techniques.
- Hybrid Model RL (20,80) and Hybrid Model LN (30,70) showed relatively lower loss values compared to standalone models. These combinations effectively balance the computational power of ensemble learning with the flexibility of probabilistic loss functions.
- The train-test split ratio also influenced the loss values, with models trained on a 20-80 split generally exhibiting lower loss. This suggests that providing more data for testing improves model generalization and reduces overfitting.

Trends and Implications

- Decrease in Loss with Complexity: As the complexity of the hybrid models increased (e.g., LNR combining three techniques), the loss values decreased, indicating better optimization and performance. However, this comes at the cost of increased computational resources.
- Impact of Loss Functions: The use of Log Loss as the primary metric aligns well with the probabilistic nature of the machine learning techniques used, leading to better model convergence and stability.

Interpretation of Results

The analysis of loss values across the models reveals that hybrid models consistently outperform single-technique models in terms of loss minimization. The combination of techniques allows for better handling of biases and variances, resulting in improved error reduction. Furthermore, models trained with a higher proportion of training data (30-70 split) showed slightly higher loss, likely due to a reduced amount of test data for generalization.

Explanation for Graphical/Tabulated Results

Accuracy:

- Logistic Regression consistently achieves the highest accuracy (~76.88%), especially with a balanced 30:70 train-test split.
- Random Forest displays a wider performance range, with its best accuracy at 69.00%. The lower accuracy (63.00%) for some configurations suggests possible overfitting or insufficient hyperparameter tuning.
- Naïve Bayes achieves decent accuracy (71.03%) despite being a simpler algorithm.

Train and Test Times:

- Logistic Regression offers an excellent balance, with moderate training and testing times.
- Random Forest exhibits significantly higher training times, reflecting its computational complexity.
- Naïve Bayes is the fastest model, with train and test times near zero, making it highly efficient.

Loss Values:

- Logistic Regression yields the lowest loss values (0.490–0.500), indicating better model optimization.
- Random Forest and Naïve Bayes exhibit higher loss values (~0.62), which correlate with their lower accuracy.

4.2 Ensemble Machine Learning Methodology

Ensemble learning refers to a class of machine learning techniques that combines multiple individual models to produce a stronger, more accurate model. The central idea is that by combining models (often referred to as "base learners"), an ensemble can overcome individual model weaknesses, leading to improved performance on tasks such as classification and regression.

Types of Ensemble Methods:

- Bagging (Bootstrap Aggregating): This involves training several models that often are the same kind and trained on different subsets of training data that have been constructed by bootstrapping followed by averaging or voting among their predictions. This does reduce variance and overfitting. Random Forest is a common example of bagging because it aggregates multiple decision trees to produce predictions.
- Boosting: Boosting trains models sequentially; each new model tries to focus on the mistakes of the previous models. That way, bias decreases and improves the accuracy of the final model. Example: Gradient Boosting or AdaBoost.
- Stacking (Stacked Generalization): In stacking, multiple base models (which could be of any type) are trained on the data. These predictions from these base models are then aggregated using another model, which is called meta-learner or stacking model, and which makes the final prediction. Stacking very often improves accuracy since it captures various patterns.

Example: Logistic Regression, Naïve Bayes, and Random Forests used together with a meta-model like Logistic Regression.

Ensemble Models in This Work

In this research, ensemble approaches combine different machine learning techniques, including Logistic Regression, Naïve Bayes, and Random Forest to make hybrid models. Then the models are evaluated according to accuracy, train time, and test time for a comparative study on the performance of the ensemble method in the classification context.

Hybrid Model Examples:

The hybrid models in this study combine different machine learning techniques in an ensemble-like manner. These combinations can be seen as multi-model configurations, where different techniques and loss functions are evaluated together:

Hybrid Model NL (20:80): An ensemble involving Naïve Bayes (NL), trained with a 20-80 split.
Hybrid Model RL (20:80): Random Forest (RL) combined with the loss function, trained with a 20-80 split.
Hybrid Model LN (20:80): Logistic Regression (LN) using Naïve Bayes loss function, with 20-80 split for training.
Hybrid Model LNR (20:80): This is an ensemble of Logistic Regression, Naïve Bayes, and Random Forest, trained with a 20-80 split.

Each of the above models can be regarded as an ensemble of classifiers where:
Base learners of different types (for example, Logistic Regression, Naïve Bayes, Random Forest) are employed. These base models can be combined by stacking or their predictions could be averaged or voted upon (depending upon which ensemble method is chosen).

Model Evaluation:

For model performance evaluation, we are going to concentrate on:

- Accuracy: The fraction of correct predictions done by each hybrid model.
  Precision, Recall and F1-Score: These are quite crucial when working with imbalanced datasets because they offer a better insight of model performance.
- Train Time and Test Time: The training and testing efficiency is also measured to ensure that the ensemble model does not require prohibitive computational resources, while still providing improvements in accuracy.

Combination of Loss Functions:

The loss function plays an important role in the performance of an ensemble model. In this research, Log Loss is applied as the primary loss function for every base model. Log Loss yields a probabilistic interpretation of model outputs and is therefore best suited for classification problems where the predicted output is a probability.

Train-Test Split:

The train-test split ratio (e.g., 20-80 or 30-70) plays a crucial role in determining the quality of training and testing data. A 20-80 split provides more data for testing, whereas a 30-70 split allocates a larger portion for training. The performance of the ensemble models is evaluated using these different splits to determine the trade-off between training data availability and model generalization.

Ensemble Strategy Implementation:

Stacking or voting strategy can be used in an ensemble learning

- Stacking: The predictions from each base learner- set of learners whose predictions are combined using a second-level model (a meta-learner), which makes the final prediction.
- Voting: The predictions from the individual models can be combined through voting. For example, each model (base learner) votes for the class that it predicts, and the class that receives the most vote is the final prediction.

Visualizations and Analysis:

To analyze the performance of ensemble models, the following visualizations can be used:
Bar charts will reveal the correctness of each hybrid model comparing the various combinations of machine learning methods and loss functions. Line graphs will also help analyze the computational efficiency in terms of display of the train time and test time of each hybrid model.

Performance of Hybrid Models

Graphical and Tabular Representation

- Tabular Summary of Ensemble Machine Learning Techniques :

Model	Accuracy	Train Time(s)	Test Time(s)	Loss Value
Hybrid Model NL(20:80)	76.00%	0.024	0.002	0.4711

Hybrid Model NL(30:70)	76.00%	0.007	0.002	0.4842
Hybrid Model RL (20:80)	74.00%	0.1685	0.1987	0.5214
Hybrid Model RL (30:70)	69.00%	0.1793	0.2104	0.5361
Hybrid Model LN (20:80)	70.09%	0.2154	0.0259	0.5584
Hybrid Model LN (30:70)	71.88%	0.1842	0.0347	0.5615
Hybrid Model LNR (20:80)	76.00%	0.2367	0.0221	0.5171
Hybrid Model LNR (30:70)	73.78%	0.2325	0.024	0.534

The table above gives results for the hybrid models with their corresponding accuracy, test time, train time, and their log loss values. Useful insights are drawn from a comparison of performance metrics across computational efficiency. The result is discussed in detail below for each point.

- Hybrid Model Accuracy
  The best model was the Hybrid Model LNR (20,80). This again shows how strong ensemble learning is, whereby taking a combination of Logistic Regression, Naïve Bayes, and Random Forest decreases bias and variance for improved predictive performance.
  For models trained on a 20-80 train-test split, the accuracy generally turned out to be marginally higher than that from a 30-70 split. This indicates that the larger the testing data, the better the generalization of the models. Logistic Regression-based Hybrid Models (LN) had accuracy throughout the runs, which was greater than Naïve Bayes models (NL) but less than Random Forest models (RL).
- Loss Analysis
  The Hybrid Model LNR (20,80) also showed the lowest loss value, in line with its high accuracy. The lower log loss indicates better calibration of predicted probabilities, which enhances the reliability of the model. Naïve Bayes Models (NL) had comparatively higher loss values, as their performance was weaker than that of other techniques in terms of classification. This happens because Naïve Bayes assumes independence among the features, which sometimes fails to yield optimal results with complex data. Random Forest Models (RL) had average loss values; they can manage different kinds of distributions well.
- Computational Efficiency (Train Time and Test Time)
  NL naïve Bayes models were the most efficient in terms of train and test times. Their simplicity and low computational complexity make them excellent for time-sensitive applications. Hybrid Model LNR (20:80), although it achieves the best accuracy and log loss is at the cost of a higher train and test time which indicates the computational price to pay for ensemble methods. Random Forest Models (RL) showed a good balance of computational efficiency and performance. Although they required slightly more training time than Naïve Bayes, they delivered significantly better accuracy and log loss values.

Models trained with a 30-70 split generally took more time to train but required slightly less testing time compared to their 20-80 counterparts, due to the larger proportion of data used for training.

Trends and Trade-offs

- Accuracy vs. Efficiency: NL suits applications where speed for the prediction is the focus while sacrificing some accuracy whereas, in Hybrid Model LNR (20, 80), accuracy is critical no matter how much resource needs to be consumed.
- Loss vs. Accuracy: In all the models, there's always a correlation between having smaller loss values with high accuracy, validating our decision to use log loss as a performance metric.
- Effect of Train-Test Split: Models trained using the 20-80 train test split were better than their 30-70 train test split counterparts both in terms of accuracy and log loss, meaning that those have a better generalization.

Computational Efficiency and Trade-offs:

Although ensemble models are known to give higher accuracy, it often comes at the cost of higher computational time. The trade-off between accuracy and the computational cost that is associated with the use of ensemble methods is therefore evaluated by comparing the train time and test time in the hybrid models.

In this study, ensemble methods will be used, which combine the strengths of multiple machine learning models into one to obtain improved performance in classification. The created hybrid models through the combinations of Logistic Regression, Naïve Bayes, and Random Forest with various loss functions and train-test split ratios allow for an overall testing of accuracy, efficiency, and computational cost.

This methodology will provide insights into ensemble strategies different from stacking or voting, and hybrid combinations of machine learning techniques, to build even more accurate, stable models for classification tasks.

Significance of the Results

- **Performance Insights**: Logistic Regression proves to be the most effective model, balancing high accuracy and low computational cost.
- **Efficiency**: Naïve Bayes is ideal for quick evaluations or low-resource environments due to its extremely short training times.
- **Model Comparison**: Random Forest, despite its potential for complex data relationships, needs further tuning to justify its computational overhead.

Deviations from Expected Results & Justifications

- Lower Accuracy for Random Forest:
- Expected: Random Forest to outperform due to its ability to handle non-linear relationships.
- Observed: Accuracy was lower (~69.00%) for the best configuration.
- Justification: Suboptimal hyperparameter settings (e.g., insufficient trees or max depth) likely limited performance.

High Efficiency of Naïve Bayes:

- Expected: Moderate efficiency with balanced performance.
- Observed: Extremely fast training and testing times with decent accuracy (71.03%).
- Justification: Naïve Bayes makes strong assumptions about feature independence, simplifying computations.

Conclusions

machine learning techniques, loss functions, and train-test split ratios to determine the optimal model for classification tasks. The key findings from the analysis are as follows:

- **Optimal Model Identification:**
  - Hybrid Model LNR (20:80), which combines Logistic Regression, Naïve Bayes, and Random Forest, with a 20-80 train-test split, emerged as the most accurate model. This hybrid model benefited from combining the strengths of different classifiers, leading to improved predictive performance.
  - Among the individual models, Logistic Regression with Log Loss demonstrated high consistency and accuracy, particularly with the 20-80 split, showing strong generalization ability.
- **Efficiency Leader:**
  - Random Forest (RL) models, particularly the Hybrid Model RL (20:80), achieved a good balance of accuracy and efficiency. While they were computationally more demanding than some simpler models (like Logistic Regression), they provided robust results and minimized errors, making them suitable for real-world applications.
  - The Naïve Bayes (NL) models were the most computationally efficient, requiring less training and test time compared to the other hybrid models, though their accuracy was generally lower than that of Random Forest and Logistic Regression.
- **Trade-offs Between Accuracy and Efficiency:**
  - While ensemble models like Hybrid Model LNR (20:80) produced the highest accuracy, they also exhibited higher train times and test times. Therefore, for scenarios where computational resources are limited, Naïve Bayes might be a preferable option due to its faster execution, even though it may sacrifice some predictive accuracy.
  - The trade-off between computational efficiency and accuracy needs to be considered when selecting the optimal model, particularly for large-scale real-world applications where both accuracy and efficiency are important.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE OF WORK

5.1 Work Conclusion

This paper evaluates the effectiveness of various hybrid machine learning models by combining different machine learning techniques with loss functions and train-test split ratios to identify the optimal model which is balanced between accuracy, computational efficiency, and scalability in the context of classification tasks. Hybrid models included combinations of techniques such as Logistic Regression, Naïve Bayes, and Random Forest, using Log Loss as the loss function. Methodology adopted involved experimenting with different train-test splits (20-80 and 30-70) and comparing the performance of each hybrid model based on key metrics such as accuracy, train time, and test time.

The results showed the Hybrid Model LNR to achieve an accuracy of 20, 80 based on Logistic Regression, Naïve Bayes, and Random Forest. Balance was assured between accuracy and computational cost due to the Random Forest models; however, such balance meant a compromise in some aspect because Naïve Bayes proved to be one of the most efficient solutions when discussing computational time for all such models. More significantly, the study concluded with evidence that the conflict between high accuracy and more expensive computing is significant in real-world tasks.

From the results achieved, we can deduce that ensemble learning via hybrid models greatly enhances classification performance. Among all the models used, Hybrid Model LNR (20:80) emerged as the best model in terms of accuracy since it integrates Logistic Regression, Naïve Bayes, and Random Forest. This model has shown how different machine learning techniques may be combined to enhance prediction accuracy. Besides, the Random Forest models resulted in a very good compromise between performance and computational efficiency. They appear suitable for medium-sized data sets. Naïve Bayes models resulted in very good efficiency, but reduced accuracy.

The results are important because they indicate that ensemble methods, when multiple machine learning techniques are used, increase the predictive power of the models, especially in scenarios where both accuracy and efficiency are concerned. The findings from this research can be used to help practitioners make the right choice regarding which model to use for classification purposes based on the nature of the data and the resources available for computation.

5.2 Future Scope of Work

Some direction for further work may lie in more advanced combinations of techniques into which machine learning can be utilized. Incorporation of support vector machines, gradient boosting and even support vector machines or gradient boosting into hybrid models will have an improvement in the output. It may even contribute complementary strengths in the direction of non-linear data and to reduce bias, so better, more accurate, and more robust ensemble models can be found.

Hyperparameter Optimization: In this experiment, the models were assessed with default hyperparameters. In the future, one could optimize hyperparameters with techniques such as Grid Search or Random Search. The process could, in that case, optimize the model settings for each technique of machine learning, with better performance and more precise control over the trade-off between accuracy and efficiency. Such an optimization would help improve the effectiveness of hybrid models, particularly in challenging or imbalanced datasets.

Real-world dataset evaluation and scalability: Future research should be directed at taking the models to real-world large datasets to assess applicability in practice. Testing will have to be made of the scalability of these models with respect to computational resources as well as accuracy. The biggest challenge in such data is always going to be huge data sets. Besides that, the scalability of the model, along with its performance on noisy or imbalanced real-world data, should be evaluated to ensure that these hybrid models are robust in different application domains.

CHAPTER 6

HEALTH, SAFETY, RISK AND ENVIRONMENT ASPECTS

Summary of Health and Safety Aspects

During the conduct of this project, concerns regarding health and safety issues were minimal since the major tasks involved computational and analytic activities that were carried out under a controlled environment. However, long working hours on the computer, being in a sitting position for prolonged hours, and repetitive use of inputs through keyboards and mice caused possible ergonomic hazards. Regularly scheduled breaks, ergonomic chairs, ergonomic desks, and adjusting computer brightness were taken to minimize ergonomic hazards. Moreover, proper ventilation and lighting in the workplace ensured a comfortable working environment.

6.1 Risk Management

The primary risks in this project were technical and computational in nature:

Data Integrity: Ensuring the dataset was free from corruption or inconsistencies was critical to avoid biased or inaccurate results.

- **Mitigation**: Regular backups and validation checks were performed to maintain data quality.

Computational Failures: Power outages or hardware failures could interrupt model training and evaluation.

- **Mitigation**: Experiments were periodically saved, and uninterruptible power supplies (UPS) were used to safeguard against data loss.

Software Issues: Compatibility problems with libraries or frameworks could hinder progress.

- **Mitigation**: Software environments were pre-tested, and virtual environments were used to isolate dependencies.

6.2 Environmental Factors

Although the project had minimal direct environmental impact, the computational tasks required significant energy consumption. This highlights the need for adopting environmentally sustainable practices in machine learning research:

Energy Usage: The processing power required for training certain models, especially Random Forest, increases energy consumption.

- **Mitigation**: Efficient coding practices and selecting simpler algorithms (e.g., Logistic Regression or Naïve Bayes) helped reduce resource use.

E-Waste: Upgrading hardware for computational efficiency may contribute to electronic waste.

- **Mitigation**: Existing resources were utilized to their maximum capacity, and unnecessary hardware upgrades were avoided.

Summary

The project ensured a secure and sustainable execution process by considering proper ergonomic practices, risk mitigation strategies, and environmentally conscious computing. Such measures not only promote personal well-being but also demonstrate a responsible approach to leveraging technology to conduct research.

REFERENCES

[1] C. Kauten, A. Gupta, X. Qin, and G. Richey, "Predicting blood donors using machine learning techniques," *Information Systems Frontiers*, vol. 23, no. 3, pp. 577-594, May 2021, doi: 10.1007/s10796-021-10149-1.

[2] T. Birhane and B. Hailu, "Predicting the behavior of blood donors in National Blood Bank of Ethiopia using data mining techniques," *I.J. Information Engineering and Electronic Business*, vol. 13, no. 3, pp. 39-48, Jun. 2021, doi: 10.5815/ijieeb.2021.03.05.

[3] P. Selvaraj, A. Sarin, and B. I. Seraphim, "Forecasting system for donation of blood using SVM model," *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, vol. 10, no. 5, pp. 136-140, May 2022, doi: 10.22214/ijraset.2022.41940.

[4] M. J. Gomes, A. J. Nogueira, C. Antão, and C. Teixeira, "Motivations and attitudes towards the act of blood donation among undergraduate health science students," *Transfusion and Apheresis Science*, vol. 58, no. 2, pp. 1–6, 2019. doi: 10.1016/j.transci.2018.12.018.

[5] E. Talie, H. Wondiye, N. Kassie, and H. Gutema, "Voluntary blood donation among Bahir Dar University students: Application of integrated behavioral model, Bahir Dar, Northwest Ethiopia, 2020," *Journal of Blood Medicine*, vol. 11, pp. 429–437, 2020. doi: 10.2147/JBM.S277411.

[6] A. Kassie, T. Azale, and A. Nigusie, "Intention to donate blood and its predictors among adults of Gondar city: Using theory of planned behavior," *PLoS ONE*, vol. 15, no. 3, pp. 1–12, Mar. 2020. doi: 10.1371/journal.pone.0228929.

[7] L. Al Shaer, R. Sharma, and M. AbdulRahman, "Analysis of blood donor pre-donation deferral in Dubai: Characteristics and reasons," *Journal of Blood Medicine*, vol. 8, pp. 55-60, 2017, doi: 10.2147/JBM.S135191.

[8] P. Su, X. R. Ding, Y. T. Zhang, J. Liu, F. Miao, and N. Zhao, "Long-term blood pressure prediction with deep recurrent neural networks," in *IEEE International Conference on Biomedical and Health Informatics (BHI)*, 2018, pp. 3-10, doi: 10.1109/BHI.2018.8333376

[9] W. Ben Elmir, A. Hemmak, and B. Senouci, "Smart platform for data blood bank management: Forecasting demand in blood supply chain using machine learning," Information (Switzerland), vol. 14, no. 1, Art. no. 31, Jan. 2023, doi: 10.3390/info14010031.

[10] A. S. Alkahtani and M. Jilani, "Predicting return donors and analyzing blood donation time series using data mining techniques," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 10, no. 8, pp. 113–118, 2019.

[11] W. Deshmukh, K. Borhade, A. Shaikh, and A. Bansod, "Blood donation interval estimation through deep learning," *Journal of Emerging Technologies and Innovative Research (JETIR)*, vol. 8, no. 6, pp. 175–180, 2021, doi: N/A.

PROJECT DETAILS

Student Details			
Student Name	**Boyella Vishnu Vardhan Reddy**		
Register Number	190907598	Section / Roll No	A/58
Email Address	boyella.vishnu@gmail.com	Phone No (M)	+917331149888
Project Details			
Project Title	**BLOOD DONATION PREDICTION USING MACHINE LEARNING TECHNIQUES**		
Project Duration		Date of reporting	
Organization Details			
Organization Name	**Manipal Institute of Technology**		
Full postal address with pin code	Manipal Institute of Technology, Manipal – 576 104 (Karnataka State), INDIA		
Website address	https://www.manipal.edu/mit.html		
Internal Guide Details			
Faculty Name	**Dr. Ramya S**		
Designation	**Associate Professor**		
Full contact address with pin code	Dept of E & C Engg, Manipal Institute of Technology, Manipal – 576 104 (Karnataka State), INDIA		
Email address	ramya.lokesh@manipal.edu		

190907598

ORIGINALITY REPORT

12%	8%	7%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Visvesvaraya Technological University, Belagavi Student Paper	1%
2	Submitted to Manipal Academy of Higher Education (MAHE) Student Paper	1%
3	link.springer.com Internet Source	1%
4	Pawan Singh Mehra, Dhirendra Kumar Shukla. "Artificial Intelligence, Blockchain, Computing and Security - Volume 2", CRC Press, 2023 Publication	<1%
5	uniassignment.com Internet Source	<1%
6	Submitted to University of Sheffield Student Paper	<1%
7	www.ijana.in Internet Source	<1%
	www.mdpi.com	

83	Atiqur Rahman Ahad, Sozo Inoue, Guillaume Lopez, Tahera Hossain. "Human Activity and Behavior Analysis - Advances in Computer Vision and Sensors: Volume 1", CRC Press, 2024 Publication	<1 %
84	Gyimah, Nana Kankam. "A Data-Driven Approach for Surface Defect Detection and Localization", North Carolina Agricultural and Technical State University, 2024 Publication	<1 %
85	Sujata Dash, Subhendu Kumar Pani, Joel J. P. C. Rodrigues, Babita Majhi. "Deep Learning, Machine Learning and IoT in Biomedical and Health Informatics - Techniques and Applications", CRC Press, 2022 Publication	<1 %
86	de Castro Ferreira, João Pedro. "A Federated Learning Platform for High Speed Distributed Data Streams", Universidade do Porto (Portugal), 2024 Publication	<1 %

Exclude quotes	On	Exclude matches	< 3 words
Exclude bibliography	On		