Advanced Web Application for Blood Patron Prediction Using SVC

MINI PROJECT REPORT

Submitted by

Sweatha R	**210701275**
Thamizh Bharathi M	**210701288**

in partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE
An AUTONOMOUS Institution
Affiliated to ANNA UNIVERSITY, Chennai

ANNA UNIVERSITY
PROGRESS THROUGH KNOWLEDGE

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

ANNA UNIVERSITY:: CHENNAI 600 025

APRIL 2024

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

BONAFIDE CERTIFICATE

Certified that this Report titled **Advanced Web Application for Blood Patron Prediction Using SVC** is the bonafide work of **Sweatha R (210701275) and Thamizh Bharathi M (210701288)** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Karthick V
Associate Professor,
Department of Computer Science and Engineering,
Rajalakshmi Engineering College,
Chennai – 602015

Submitted to Mini Project Viva-Voce Examination held on ___________________

Internal Examiner **External Examiner**

ABSTRACT

Among all donations, blood donation is the most important one, due to its nature of saving lives. It's considered to be the most important donation. Blood donation is not a bed of roses! Blood transfusion is a critical aspect of modern healthcare, saving millions of lives worldwide every year. Approximately 118.4 million blood donations are collected worldwide every year. When someone wants blood, they have to request the blood bank and if there’s a match they’ll get it. If not, it will be chaos to search for an individual. Even if we got the correct individual, we don’t know whether the individual will be eligible to donate blood or not. We need individuals who are interested in donating blood but mostly everyone is dependent on the blood bank. Blood banks are dependent on the donors who have been donating subsequently. So we are concentrating on the latter part such as whether an individual will donate blood again or not in a particular period. Our model analyzes factors such as the time since the last donation, number of donations, volume of donated blood, and total months since the first donation to predict donor behavior. Through various evaluations using cross-validation and performance metrics, we identify the Support Vector Classifier (SVC) model as the most effective in predicting donor return likelihood. Our approach aims to provide blood banks with a data-driven tool to increase the number of repeat donors, thereby maintaining a stable blood supply and saving more lives. Additionally, insights gained from our model can inform targeted campaigns to encourage regular blood donation, contributing to a sustainable blood transfusion ecosystem. We are using machine learning algorithms to analyze the dataset and predict whether a single person will donate or not in a particular period. We also consider the volume of blood donated and the last time the individual donated blood.

ACKNOWLEDGEMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavour to put forth this report. Our sincere thanks to our Chairman **Mr. S.MEGANATHAN, B.E, F.I.E.**, our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN**, **Ph.D.,** for providing us with the requisite infrastructure and sincere endeavoring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N. MURUGESAN, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. P. KUMAR, Ph.D.**, Professor and Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We convey our sincere and deepest gratitude to our internal guide, **Kathick V** Associate Professor, Department of Computer Science and Engineering. Rajalakshmi Engineering College for his valuable guidance throughout the course of the project.

Sweatha R-210701275
Thamizh Bharathi M-210701288

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

SVM	Support Vector Machines
KNN	K Nearest Neighbors
SVC	Support Vector Classifier
SMOTE	Synthetic Minority Over-sampling Technique

CHAPTER 1
INTRODUCTION

1.1 GENERAL

Blood donation is a critical component of healthcare systems, essential for saving lives and supporting medical treatments. However, predicting whether individuals will donate blood at specific times poses a significant challenge. Accurate predictions are vital for ensuring a steady and reliable blood supply to meet ongoing medical needs.

1.2 OBJECTIVE

The objective of this project is to develop an advanced predictive model to accurately forecast individual blood donation behaviors at specific times. Utilizing machine learning techniques such as Support Vector Classifier (SVC) and Logistic Regression, and addressing data imbalances with SMOTE and Best Skewness Transformer, we aim to enhance prediction accuracy.

1.3 EXISTING SYSTEM

Current systems in blood donor management primarily focus on clustering donors based on covariates, adhering to selection guidelines, and forecasting overall donor trends. These systems lack the capability to predict individual donation behaviors accurately at specific times. Often constrained by limited datasets, they result in suboptimal predictive models. Consequently, there is a significant gap in addressing the precise prediction of individual blood donations.

1.4 PROPOSED SYSTEM

Our proposed system involves developing a precise predictive model to anticipate individual blood donation behaviors at specified times. Through advanced machine learning techniques like Support Vector Classifier (SVC) and Logistic Regression, and data enhancement methods such as SMOTE and Best Skewness Transformer, we aim to achieve higher prediction accuracy.

CHAPTER 2
LITERATURE SURVEY

This book provides a comprehensive overview of blood transfusion[1] science, covering the principles, procedures, and management of blood donation and transfusion. It emphasizes the critical role of transfusions in saving lives and discusses the latest advancements in the field. Additionally, it addresses the challenges in blood collection, storage, and compatibility testing. The authors highlight the importance of donor recruitment and retention strategies[1] to ensure a stable blood supply. It also includes case studies and real-world examples to illustrate best practices in blood transfusion services.

Dangeti's work focuses on the application of statistical methods in machine learning. It covers essential topics like probability distributions, hypothesis testing, and regression analysis, providing practical examples and case studies relevant to machine learning practitioners[2]. The book explains how to apply these statistical techniques to real-world datasets, making it accessible for both beginners and experienced data scientists. It also includes sections on advanced topics such as Bayesian statistics and time series analysis. The author emphasizes the importance of statistical thinking in building robust machine learning models.

This review explores the sociological aspects of voluntary blood donation, examining motivations, barriers, and the impact of social factors on donor behavior. It highlights the importance of understanding these elements to improve donor retention and recruitment strategies[3]. The study synthesizes findings from various sociological studies to provide a comprehensive view of the factors influencing donor behavior. It discusses the role of cultural, economic, and psychological factors in shaping donation practices. The authors suggest policy interventions to address the barriers to voluntary blood donation.

Prasad's book offers insights into various facets of blood donation, including donor eligibility, the donation process, and the significance of regular donations. It aims to educate and encourage more individuals to participate in blood donation programs[4]. The book covers the history and evolution of blood donation practices, highlighting key milestones and advancements. It also discusses the ethical considerations and regulations governing blood donation. The author includes personal anecdotes and experiences to make the content relatable and engaging for readers.

This study investigates data-driven strategies to predict blood donor deferral, using machine learning models to analyze patterns and factors leading to deferrals. The findings aim to enhance the efficiency of donor selection and reduce deferral rates[5]. The researchers employed various algorithms such as logistic regression, decision trees, and neural networks to build predictive models. They also conducted feature selection to identify the most significant factors influencing donor deferral. The study provides recommendations for optimizing donor recruitment and screening processes.

This paper presents a hierarchical time series forecasting model to project blood donation trends in Zimbabwe. The model aims to assist blood banks in planning and managing their blood supply more effectively by predicting future donation patterns[6]. The authors used historical donation data and applied advanced time series analysis techniques to develop the model. They validated the model using real-world data and found it to be highly accurate in predicting donation trends. The study also discusses the potential impact of external factors such as economic conditions and public health campaigns on donation rates.

De Bruyne discusses the development and application of supervised classification models in various domains. The book emphasizes the importance of data preprocessing, feature selection, and model evaluation to achieve accurate and reliable classification results[7]. It includes practical guidelines for implementing and fine-tuning classifiers using popular machine learning frameworks. The author also explores the challenges and limitations of different classification algorithms. Case studies from diverse fields such as healthcare, finance, and marketing are presented to illustrate the application of classification models.

This study develops a scoring system to predict the likelihood of vasovagal reactions during whole-blood donation. The system uses donor characteristics and previous donation history to provide a risk assessment, aiming to improve donor safety and experience[8]. The researchers conducted extensive data analysis to identify the key predictors of vasovagal reactions. They validated the scoring system through a series of clinical trials, demonstrating its effectiveness in reducing adverse reactions. The study also discusses the implementation of the scoring system in blood donation centers.

Mulcahy's analysis examines the sustainability of the U.S. blood supply system, identifying challenges and potential solutions. The report discusses policy recommendations and alternative approaches to ensure a stable and sufficient blood supply[9]. It provides an in-depth analysis of the current blood supply chain, highlighting inefficiencies and areas for improvement. The author also explores the role of emerging technologies and innovations in enhancing blood collection and distribution. The report includes case studies from other countries to draw lessons and best practices for the U.S. context.

Mudd explores Briggs' information processing model applied to binary classification tasks. The book provides insights into cognitive processes involved in classification and offers practical applications for improving model accuracy and efficiency[10]. It discusses the psychological underpinnings of decision-making in classification tasks. The author integrates concepts from cognitive psychology and machine learning to provide a holistic understanding of classification processes. Real-world applications in fields such as medical diagnosis and financial forecasting are examined.

Wang's book delves into the theoretical foundations and practical applications of support vector machines (SVMs). It covers various SVM algorithms, kernel functions, and their use in classification and regression tasks[11]. The book provides detailed mathematical formulations and proofs for the key concepts in SVM. It also includes practical guidelines for implementing SVMs using popular programming languages and software tools. Case studies from different domains such as bioinformatics and image processing are presented to demonstrate the versatility of SVMs.

Strickland provides a detailed explanation of logistic regression, covering its theoretical underpinnings and practical applications. The book includes examples and case studies to illustrate the use of logistic regression in different scenarios[12]. It offers a step-by-step guide to building and interpreting logistic regression models. The author also discusses advanced topics such as regularization techniques and multi-class logistic regression. Practical tips for model evaluation and validation are provided to ensure robust and reliable results.

This paper discusses the enhancement of machine learning models using the Synthetic Minority Over-sampling Technique (SMOTE) to predict the recurrence of postoperative chronic subdural hematoma. The study demonstrates improved model performance and predictive accuracy[13]. The researchers compared various machine learning algorithms, including random forests and support vector machines, with and without SMOTE. The results showed that SMOTE significantly increased the accuracy and sensitivity of the models. The study provides recommendations for applying SMOTE in other medical prediction tasks.

Jafari's book focuses on data preprocessing techniques in Python, essential for preparing data for analytics and machine learning. It covers cleaning, transforming, and normalizing data, providing practical examples and code snippets[14]. The book emphasizes the importance of data quality and consistency for successful data analysis. It also includes advanced topics such as handling missing data, feature engineering, and data augmentation. The author provides best practices and tips for efficient data preprocessing in real-world projects.

Mukhiya and Ahmed provide a comprehensive guide to exploratory data analysis (EDA) using Python. The book covers various EDA techniques to understand, summarize, and visualize data, facilitating better insights and decision-making[15]. It includes practical examples and code snippets to demonstrate the application of EDA techniques. The authors also discuss the importance of EDA in the data science workflow, highlighting its role in uncovering patterns and relationships. Advanced visualization tools and techniques are presented to enhance data interpretation.

This book offers practical guidance on implementing gradient boosting algorithms using XGBoost and scikit-learn in Python. It includes step-by-step instructions, examples, and case studies to demonstrate the effectiveness of these techniques in machine learning[16]. The authors explain the theoretical foundations of gradient boosting and its advantages over other algorithms. They also provide practical tips for hyperparameter tuning and model optimization.

Jung's book provides an introduction to the fundamentals of machine learning, covering key concepts, algorithms, and applications[17]. It is designed for beginners and includes practical examples to illustrate machine learning principles. The book covers a wide range of topics, from ba sic supervised and unsupervised learning to advanced techniques such as deep learning. It also discusses the ethical considerations and challenges in machine learning[17]. Practical exercises and case studies are provided to reinforce the learning experience.

Nakas' work focuses on ROC (Receiver Operating Characteristic) analysis for evaluating and comparing classification models. The book provides theoretical background and practical applications, highlighting the importance of ROC analysis in model assessment[18]. It includes detailed explanations of key concepts such as sensitivity, specificity, and the area under the ROC curve. The author also discusses advanced topics such as multi-class ROC analysis and the interpretation of ROC curves[18]. Practical examples and case studies are provided to illustrate the application of ROC analysis in various fields.

CHAPTER 3
SYSTEM DESIGN

3.1 DEVELOPMENT ENVIRONMENT

3.1.1 HARDWARE SPECIFICATIONS

This project uses minimal hardware but in order to run the project efficiently without any lack of user experience, the following specifications are recommended

Table 3.1.1 Hardware Specifications

PROCESSOR	Intel Core i5
RAM	4GB or above (DDR4 RAM)
GPU	Intel Integrated Graphics
HARD DISK	6GB
PROCESSOR FREQUENCY	1.5 GHz or above

3.1.2 SOFTWARE SPECIFICATIONS

The software specifications in order to execute the project has been listed down in the below table. The requirements in terms of the software that needs to be pre-installed and the languages needed to develop the project has been listed out below.

Table 3.1.2 Software Specifications

FRONT END	HTML, CSS
BACK END	Python, Django
FRAMEWORKS	Scikit-learn, XGBoost
SOFTWARES USED	Visual Studio Code, Jupyter Notebook

3.2 SYSTEM DESIGN

3.2.1 ARCHITECTURE DIAGRAM

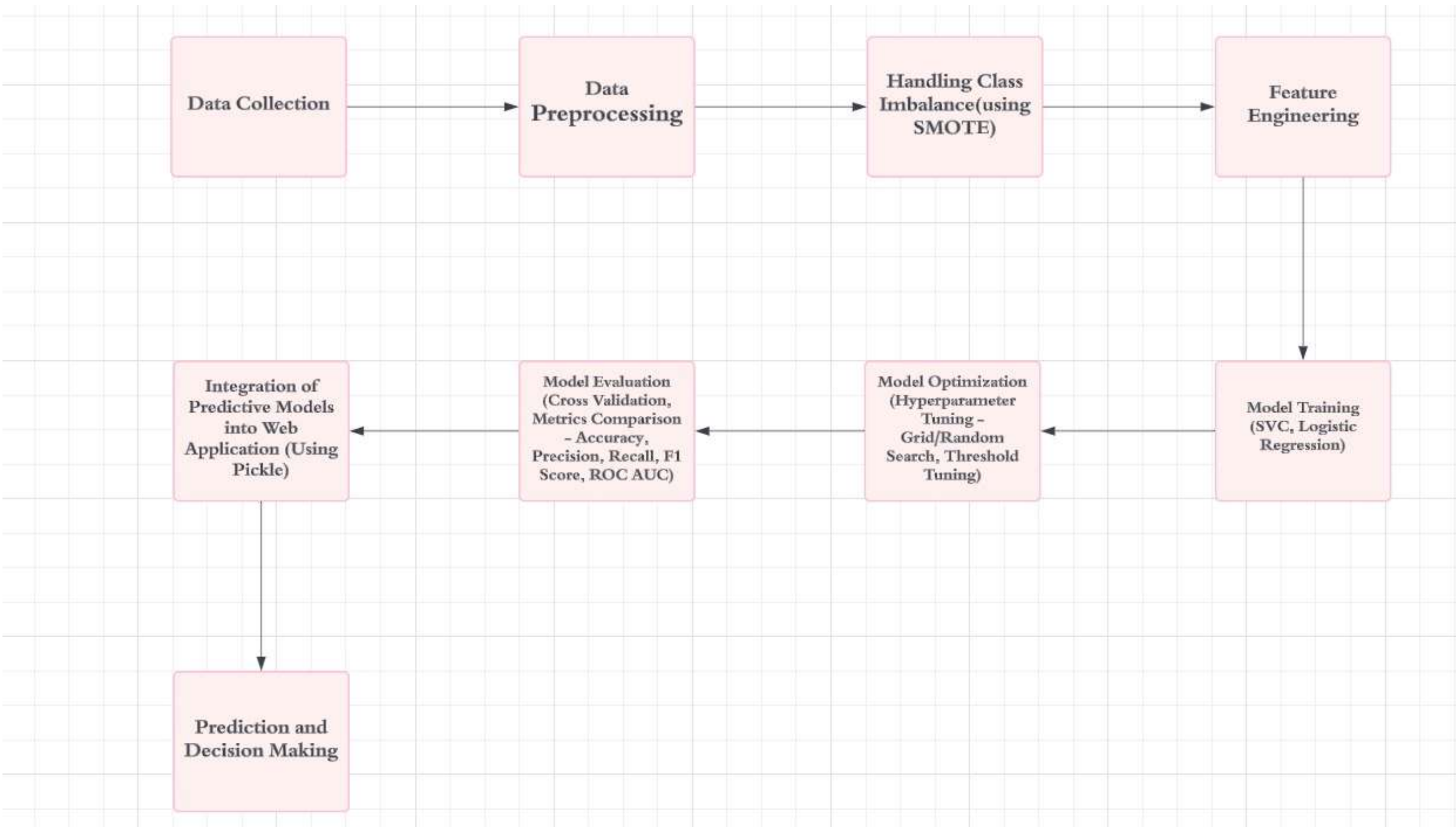

Fig 3.2.1 Architecture Diagram

PRE-PROCESSING:

The confusion matrix gives us a clear idea of which factors contribute more to the prediction. [14]The quality of data is a crucial factor that affects the model's performance. In this project, if the data is not pre-processed and cleaned properly, it may lead to incorrect results. To mitigate this risk, we have performed data cleaning and preprocessing techniques like removing duplicates and transforming data into normal distribution. We have also performed [15]exploratory data analysis (EDA) to identify the patterns in the data and understand the data distribution. The dataset is imbalanced, i.e., the number of instances in one class may be much higher than the other. This led to bias in the model's predictions towards the majority class. To overcome this challenge, we have used an over-sampling technique, SMOTE.

CHAPTER 4
PROJECT DESCRIPTION

4.1 MODULE DESCRIPTION

4.1.1 DATA COLLECTION:

The model will be developed using data obtained from a mobile blood donation vehicle, including information such as the last donation, number of donations, volume of donated blood, and total months past since the first donation of donors. The data includes details of the person who donated blood to various universities. The accuracy of the proposed model in forecasting future donor behavior will be assessed through the application of multiple metrics. Additionally, this approach can offer insightful information about the variables impacting donor behavior, which can assist blood banks in developing focused campaigns that motivate more people to donate blood regularly.

4.1.2 DATA PREPROCESSING:

The confusion matrix gives us a clear idea of which factors contribute more to the prediction. [14]The quality of data is a crucial factor that affects the model's performance. In this project, if the data is not preprocessed and cleaned properly, it may lead to incorrect results. To mitigate this risk, we have performed data cleaning and preprocessing techniques like removing duplicates and transforming data into normal distribution. We have also performed [15]exploratory data analysis (EDA) to identify the patterns in the data and understand the data distribution. The dataset is imbalanced, i.e., the number of instances in one class may be much higher than the other. This led to bias in the model's predictions towards the majority class. To overcome this challenge, we have used an over-sampling technique, SMOTE.

4.1.3 TRAINING MODEL:

Selecting the appropriate machine-learning algorithm and optimizing hyperparameters can present challenges related to time and effort. To address this, we evaluated multiple algorithms including [12]Logistic regression, support vector classification, random forest classification, and [16]gradient boosting classifier. Grid search was used to tune hyperparameters for each algorithm, and model performance was assessed using evaluation metrics to select the top-performing model. [17]Additionally, cross-validation techniques were applied to compare performance across models and identify the best-performing one. Initial analysis of algorithm metrics including accuracy, F1 score, and recall revealed superior results from support vector classification, logistic regression, and random forest classification before integrating synthetic minority over-sampling technique (SMOTE). [18]The dataset's asymmetric properties were also addressed using the best skewness transformer technique. Following the integration of SMOTE and the transformer, support vector classification, and logistic regression demonstrated the highest accuracy compared to other algorithms.

CHAPTER 5
IMPLEMENTATION AND RESULTS

5.1 IMPLEMENTATION

For the implementation of the blood donation prediction system, various methods and algorithms were utilized to ensure robustness and efficiency in predicting donor behavior. Firstly, extensive data was collected from multiple sources, including blood banks, donation drives, and healthcare facilities. This data encompassed donor demographics, donation history, medical records, and other pertinent information.

The collected data underwent meticulous preprocessing, employing techniques such as data cleaning, formatting, and encoding. Missing values were handled, and categorical variables were encoded using advanced methods like one-hot encoding. Numerical features underwent normalization and standardization to ensure uniformity and enhance model performance.

For model selection and training, two primary machine learning algorithms were evaluated: the Support Vector Classifier (SVC) and the Random Forest Classifier. These algorithms were chosen for their ability to handle complex datasets and provide accurate predictions. The data was divided into training and testing sets, and extensive hyperparameter tuning was performed using techniques such as grid search and cross-validation to optimize model performance.

Model evaluation was conducted rigorously, focusing on metrics like accuracy, precision, recall, F1-score, and ROC AUC score. Confusion matrices and ROC curves were generated to visualize the models' performance and assess classification accuracy thoroughly. Upon selecting the most suitable model, it was deployed into a production environment, where it could make predictions on new, unseen data. The deployment process involved developing APIs or web applications to seamlessly integrate the model into existing blood donation systems or healthcare platforms.

Continuous monitoring and maintenance of the deployed model were crucial to ensure its continued accuracy and reliability. Regular updates and retraining were conducted to adapt to changing data patterns or requirements, ensuring the system remained effective in predicting blood donation behaviors. Overall, the implementation prioritized the utilization of advanced methods and algorithms to achieve high efficiency and accuracy in predicting donor behavior and enhancing the prediction rate of blood donor behavior.

5.2 OUTPUT SCREENSHOTS

The visualized the confusion matrix to understand the misclassifications made by the model. We found that the default threshold of 0.5 gave suboptimal results and we tried to improve the model's performance by tuning the threshold value. The evaluated the performance of the tuned model using the same set of metrics and compared it with the default model. Finally, we selected the tuned SVC model with a threshold of 0.44 as our final model, which gave the best results in terms of f1-score and ROC AUC score. Thus, the model predicts whether an individual will donate blood or not at a specific period by using SVC.

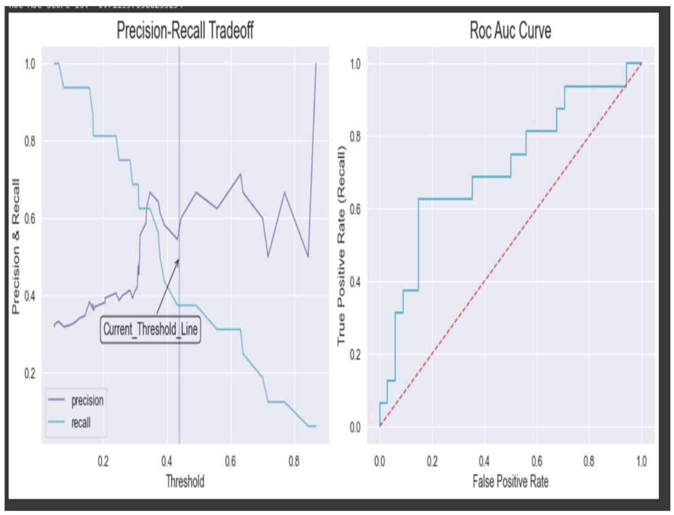

Fig 5.2.1 Accuracy Curve

The website is built using Django as a backend where we get the necessary information from the user and using our prediction model, we will predict whether the individual will donate blood or not in that period. This serves to help blood banks to filter and find the donor sooner and save the lives of the people.

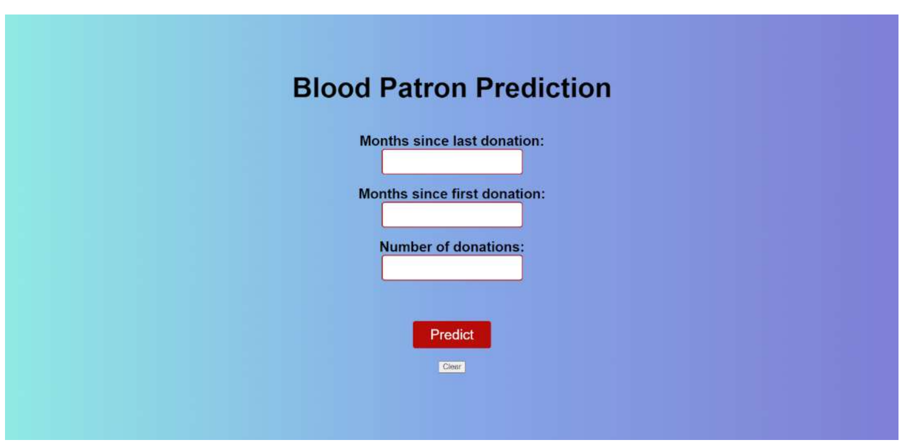

Fig 5.2.2 Home page

CHAPTER 6
CONCLUSION AND FUTURE ENHANCEMENTS

6.1 CONCLUSION

In conclusion, our project on blood donation prediction has successfully addressed the critical need for forecasting donor behaviours. By employing advanced machine learning techniques, we developed robust predictive models capable of accurately determining whether individuals are likely to donate blood at specific times. Through extensive data collection, preprocessing, and model training, we have created a valuable tool for optimizing blood donation campaigns and ensuring a sustainable blood supply.

The implementation of Support Vector Classifier (SVC) and Random Forest Classifier algorithms, combined with thorough evaluation and deployment processes, has enabled us to achieve high predictive accuracy. This system holds significant potential for blood banks, healthcare facilities, and donation drives to streamline their operations, allocate resources effectively, and maximize donation efforts.

Looking ahead, ongoing monitoring and refinement of the predictive models will be essential to adapt to changing donor behaviours and external factors influencing blood donation patterns. Collaborative efforts with stakeholders and further research can enhance the system's performance and contribute to the advancement of blood donation initiatives globally. Overall, our project represents a significant step forward in leveraging data-driven approaches to improve blood donation practices and support public health efforts.

6.2 FUTURE ENHANCEMENTS

As a future work, this project can be scaled to be used for the entire world by collecting data from all the countries and also by collaborating with the government this model will streamline the process of finding a blood donor. We can also use more advanced machine learning techniques in case of future scaling to tune the model to predict with more accuracy. Thus, our project contributes to the advancement of the prediction of blood patrons facilitating informed decision-making and better outcomes.

.

REFERENCES

[1] Erhabor and Adias, *Essentials of Blood Transfusion Science*. Author House, 2013.

[2] P. Dangeti, *Statistics for Machine Learning*. Packt Publishing Ltd, 2017.

[3] V. Raivola and R. Thorpe, "A scoping review of sociology of voluntary blood donation," *Vox Sang.*, Apr. 2024, doi: 10.1111/vox.13638.

[4] S. V. N. Prasad, *SNIPPETS ON BLOOD DONATION*. SVN Prasad, 2024.

[5] *Data-driven Donation Strategies: Understanding and Predicting Blood Donor Deferral*. 2024.

[6] C. Chideme, D. Chikobvu, and T. Makoni, "Blood donation projections using hierarchical time series forecasting: the case of Zimbabwe's national blood bank," *BMC Public Health*, vol. 24, no. 1, p. 928, Apr. 2024.

[7] S. De Bruyne, *Process, Data and Classifier Models for Accessible Supervised Classification Problem Solving*. ASP / VUBPRESS / UPA, 2010.

[8] T. Hashizume *et al.*, "Development and validation of a scoring system to predict vasovagal reaction upon whole-blood donation," *Vox Sang.*, vol. 119, no. 4, pp. 300–307, Apr. 2024.

[9] A. W. Mulcahy, *Toward a Sustainable Blood Supply in the United States: An Analysis of the Current System and Alternatives for the Future*. RAND Corporation, 2016.

[10] S. Mudd, *Briggs' Information Processing Model of the Binary Classification Task*. Psychology Press, 2019.

[11] L. Wang, *Support Vector Machines: Theory and Applications*. Springer Science & Business Media, 2005.

[12] J. Strickland, *Logistic Regression Inside and Out*. Lulu.com, 2017.

[13] Z. Ni *et al.*, "Synthetic minority over-sampling technique-enhanced machine learning models for predicting recurrence of postoperative chronic subdural hematoma," *Front. Neurol.*, vol. 15, p. 1305543, Apr. 2024.

[14] R. Jafari, *Hands-On Data Preprocessing in Python: Learn how to effectively prepare data for successful data analytics*. Packt Publishing Ltd, 2022.

[15] S. K. Mukhiya and U. Ahmed, *Hands-On Exploratory Data Analysis with Python: Perform EDA techniques to understand, summarize, and investigate your data*. Packt Publishing Ltd, 2020.

[16] C. Wade and K. Glynn, *Hands-On Gradient Boosting with XGBoost and scikit-learn: Perform accessible machine learning and extreme gradient boosting with Python*. Packt Publishing Ltd, 2020.

[17] A. Jung, *Machine Learning: The Basics*. Springer, 2022.

[18] C. Nakas, *ROC Analysis for Classification and Prediction in Practice*. CRC Press, 2023.