Advanced Web Application for Blood Patron Prediction Using SVC

V Karthick, Associate Professor
Department of CSE
Rajalakshmi Engineering College
Chennai, India
vkarthick86@gmail.com

Sweatha R, UG Student
Department of CSE
Rajalakshmi Engineering College
Chennai, India
210701275@rajalakshmi.edu.in

Thamizh Bharathi M, UG Student
Department of CSE
Rajalakshmi Engineering College
Chennai, India
210701288@rajalakshmi.edu.in

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. 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. Approximately 118.4 million blood donations are collected worldwide every year. 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.

I. INTRODUCTION

Blood transfusion plays a vital role in modern healthcare, with blood transfused worldwide to treat various medical conditions[1]. Maintaining a consistent and adequate blood supply remains a significant challenge for healthcare systems. [2]The stability of blood supply is difficult due to numerous factors, including donor recruitment, retention, and return rates. While one-time donors contribute to the blood pool, regular repeat donors are essential for sustaining the supply over time.

Ensuring the regular return of blood donors is important for blood transfusions. [3]However, understanding and predicting donor behavior is also a complex challenge due to the diverse motivations influencing donation decisions. Traditional approaches include promotional campaigns, but these methods may not effectively focus [4] on individuals most likely to return for frequent donations.

[5]We propose a data-driven solution to predict donor return likelihood based on historical donation data. By using machine learning algorithms and analyzing donor attributes such as the time since the last donation, frequency of donations, volume of donated blood, and overall donation history, we aim to develop a predictive

model that can identify donors willing to return for future donations. [6]Such a model could provide blood banks to prioritize the individuals with the highest likelihood of becoming regular donors thereby contributing to the sustainability of blood transfusion services.

Brian G., Emily A. Wentz, and Jorge L. [5] show how the supply chain of blood donations is affected. This book analyzes all the factors that caused this issue using data-driven approaches to identify similar issues in the future. They also point out how the donors are being affected by this issue. The paper imposes the importance of effective methods to improve donor availability

Coster Chideme and Tedai Makoni [6] used the time series analysis to demonstrate the relationship between blood supply and demand. The research also focuses on the factors accountable for the reduction of blood donors and also suggests some specified solutions to the blood banks. The accuracy of the prediction was hurdled by the pandemic highlighting the need for an efficient strategy

Steven De Bruyne [7] provides the solutions for supervised classification problems. This paper highlights the specific strategies to build an efficient classifier model. It also covers logistic regression, decision trees, and random forest algorithms. It provides an authentic way to understand the basics of classification problems.

Tatsuma Hashizume, and Gaku Kondo [8] demonstrate the risk factors associated with vasovagal reactions (VVR) among the blood donors. This research included a dataset from Japan ranging over 577,000 blood donations. Some of the key factors noted were age, estimated blood volume, and height.

Spencer, Jakub, and Benjamin [9] performed an analysis of the sustainability of the blood system in the US. The research paper highlights the drawbacks of the existing blood donation system and emphasizes the corrective measures to be taken to solve the problems. This paper seems to be unique as it also encounters the changes in payment and health care delivery.

Lipo Wang [10] provides the mathematical knowledge of Support Vector Machines in statistics. This paper includes both theoretical and practical approaches. It shows how bioinformatics is categorized into numbers for pattern recognition.

King Kim Chuan [11] highlights the practical guidance on preparing the data for successful data reduction. It also includes the essential techniques for data reduction, cleaning, and transformation. This paper enhances the use of Python to prepare data for feature engineering.

The paper authored by Strickland gives a detailed explanation of Logistic Regression and its applications.[12] It provides a guide to building and interpreting LR models. This paper also includes the regularisation techniques and multi-class logistic regression. Practical tips are provided for the model evaluation and metrics analysis.

The study[13] demonstrates the enhancement of machine learning models while using a SMOTE(Synthetic Minority Over-Sampling Technique). This paper compares various ML algorithms, including random forests and support vector machines with and without SMOTE. The findings show that SMOTE increased the accuracy and sensitivity of the models.

The paper [15] proposes a Python experimental guide to exploratory data analysis (EDA). The paper covers various EDA techniques such as summarization and visualization of data and provides better insights for better decision-making. It also includes a set of practical examples and code snippets. It also covers the role of EDA in data science workflow, showcasing the patterns and relationships.

II. MATERIALS AND METHODS

The web-based application gets the basic details such as the age of the user, the time since his last donation, the time since his first donation, and the volume of blood he has donated so far. With all this information the application predicts whether the individual will donate blood at a specified time or not. [6]The forecasting idea originated from the paper cited by Makoni T. It helped us to know how time series forecasting helps in predicting blood donation. The prediction is a binary classification and for the binary classification, we have many algorithms such as [7]Logistic Regression, SVC, Decision tree classifier, etc. When we have numerous models we choose the best one based on metrics. The metrics may vary in the range of accuracy score, F1 score, precision score, and recall score. The model with the highest metric value is the well-trained and suited model for prediction.

Hardware Requirements :

- Modern multi-core CPU
- GPU (Optional)
- 16GB RAM

Software Requirements:

- Jupyter Notebook or an equivalent IDE
- Code Editor
- Django

III. EXISTING SYSTEM

The existing projects related to blood donation have contributed to the betterment of society but also have certain limitations. There also exists a critical gap which is the inability to predict individual blood donation behavior at a specified time. There is no effective model to predict the possibility of a donor donating blood at a specified time, and all the existing systems have been compromised by insufficient volumes of high-quality data. There was not an exact existing system for our project idea however some of the closely related projects are, [6] Clustering blood donors based on covariates, [7]blood donor selection based on the guidelines,[4]forecasting of blood donors. Due to the absence of a prediction model, there is a lack of sustained blood supply. [6], [8]This paper validates and predicts the vasovagal reaction upon whole-blood donation. [9]A sustainable alternative to the current system was proposed in the United States for ensuring the proper blood supply but also had the limitation of the dataset.

IV. PROPOSED SYSTEM

Our proposed project introduces a predictive model predicting whether an individual will donate blood at a specified time. [10]This prediction is a binary classification, with the ultimate goal of streamlining the blood donation process and ensuring a sustainable blood supply. After a long research among the numerous classification algorithms, we have found the two algorithms with the highest result: [11]Support Vector Classifier(SVC) and [12]Logistic regression. These algorithms show exceptional performance when enhanced with advanced techniques such as Synthetic Minority Over-sampling Technique (SMOTE) and Best Skewness Transformer. [13]These techniques help us effectively manage the data imbalances and increase the predictive accuracy of our model. The integration of these two algorithms not only provides a way for the strategic allocation of resources but also empowers blood donation centers to make decisions quickly and save lives quickly.

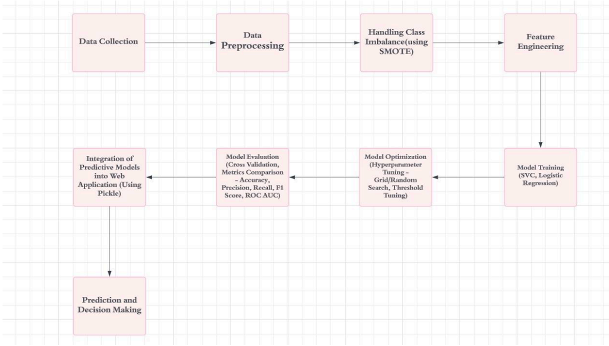

Data Collection
Data Preprocessing
Handling Class Imbalance(using SMOTE)
Feature Engineering
Model Training (SVC, Logistic Regression)
Model Optimization (Hyperparameter Tuning - Grid/Random Search, Threshold Tuning)
Model Evaluation (Cross Validation, Metrics Comparison - Accuracy, Precision, Recall, F1 Score, ROC AUC)
Integration of Predictive Models into Web Application (Using Pickle)
Prediction and Decision Making

Fig.1. Architecture Diagram

Fig.1. Shows the architecture diagram and the flow of our proposed project. Our main motto is to advance the health care delivery and enhance the well-being of individuals worldwide.

V. APPROACH

Our project uses the Support Vector Classifier (SVC) and Logistic Regression algorithms to build a prediction model. These two algorithms have outperformed in all our performance metric tests.

A. 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.

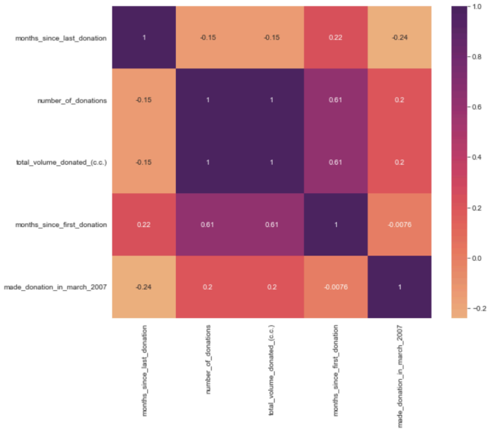


Fig.2. Confusion Matrix

B. Data preprocessing

Fig.2. 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

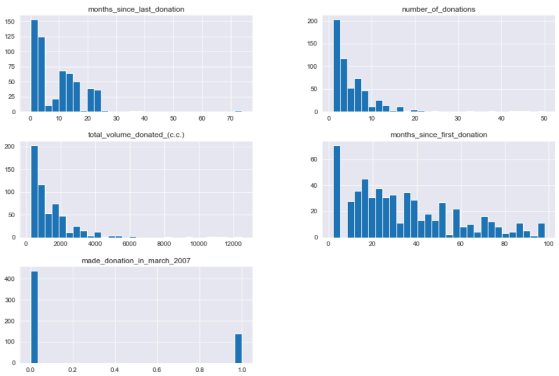


Fig.3. Data preprocessing

In Fig.3. It is evident that 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.

C. Model Selection

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.

D. Hyperparameter Tuning

Hyperparameters are parameters that are not learned from the data but are set before the training process. These hyperparameters significantly affect the model's performance, and choosing the right hyperparameters is a challenge. In this project, we have used GridSearchCV to search for the best hyperparameters. However, we noticed that the performance of selected models decreased after tuning. So we have a preferred model without tuning.

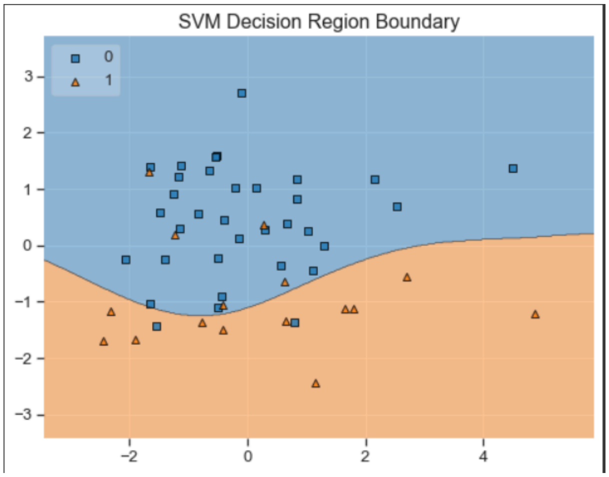


Fig.4. SVM classification

Fig.4. Shows the binary classification of individuals whether they will donate blood or not at a specified time. Overfitting occurs when the model learns the training data too well and performs poorly on unseen data. Underfitting occurs when the model is too simple and cannot capture the underlying patterns in the data. To overcome these challenges, we have used techniques like regularization, early data splitting, and cross-validation to improve the model's performance.

VI. RESULT

We trained the SVC model with default hyperparameters and evaluated its performance using various metrics such as accuracy, precision, recall, f1-score, and [18]ROC AUC score. We also 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. We used the grid search technique to optimize the SVC model's hyperparameters, including C and gamma and used the best set of hyperparameters to fit the model and get the predictions.

We also 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.

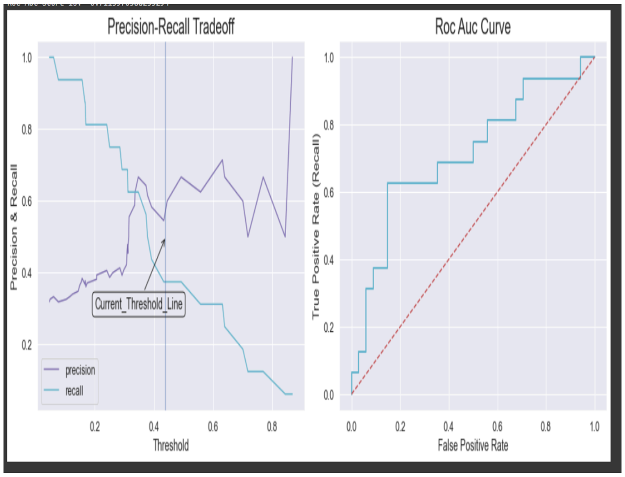
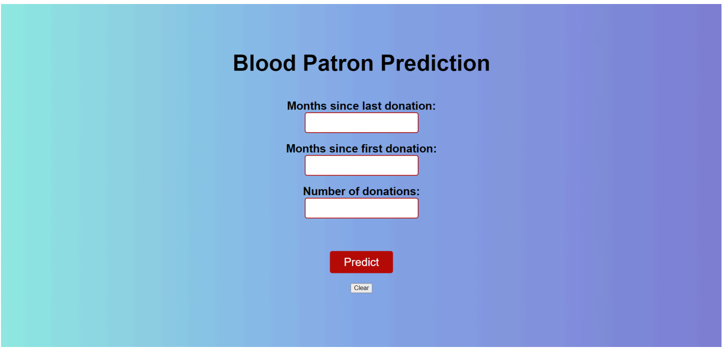


Fig.5. Metric score plot of the SVC model

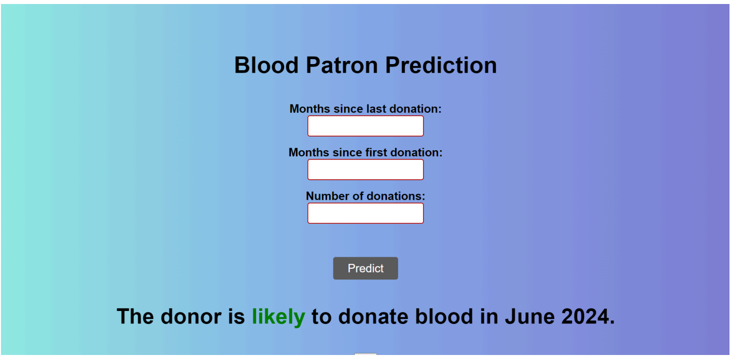
VII. OUTPUT

We have a website 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.



Blood Patron Prediction

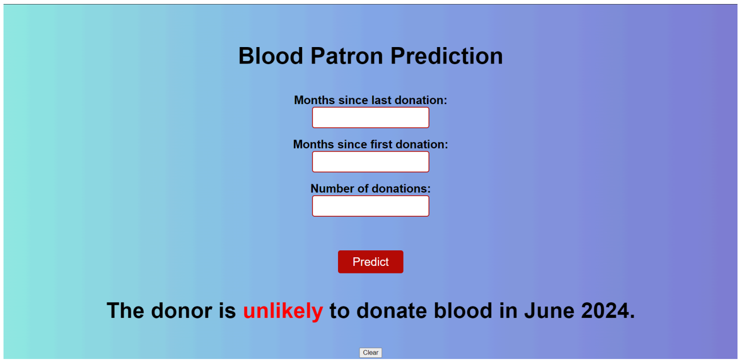
Fig.6. Home Page



Blood Patron Prediction

The donor is likely to donate blood in June 2024.

Fig.7. Positive prediction result



Blood Patron Prediction

The donor is unlikely to donate blood in June 2024.

Fig.8. Negative prediction result

VIII. CONCLUSION & FUTURE WORK

Thus our project highlights the significance of the prediction of blood patron during a specific period by optimizing the Support Vector Classifier (SVC) for prediction. The model's performance has been enhanced by implementing the grid search techniques. The tuned SVC model with a threshold of 0.44, performed better than the regular model in most metrics, including F1 Score and ROC AUC score. Our prediction offers various insights on blood donation management and enhancing healthcare resources. 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.

IX. 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.