A close up of a logo

Description automatically generated with low confidence ****

**PROJECT REPORT ON**

BLOOD DONATION PREDICTION

**Submitted by:**

Pratham Mohanty 20BCE7176

Aditya Sajith 20BCE7386

Nagulakonda Aditya 20BCE7390

Anandam Vishal Aryan 20BCR7116

**TABLE OF CONTEXT**

|  |  |  |
| --- | --- | --- |
| **S.NO:** | **TOPIC** | **PAGE NO:** |
| **1** | **INTRODUCTION** | **3** |
| **2** | **LITERATURE SURVEY** | **4-7** |
| **3** | **THEORITICAL ANALYSIS** | **7-8** |
| **4** | **EXPERIMENTAL INVESTIGATION** | **9-10** |
| **5** | **FLOWCHART** | **11** |
| **6** | **RESULT** | **12-16** |
| **7** | **ADVANTAGES**  **& DISADVANTAGES** | **17-18** |
| **8** | **APPLICATIONS** | **18** |
| **9** | **CONCLUSION** | **18** |
| **10** | **FUTURE SCOPE** | **19** |
| **11** | **BIBILOGRAPHY**  **&**  **APPENDIX** | **19-20** |

**INTRODUCTION**

**1.1 OVERVIEW:**

Blood donation plays a crucial role in saving lives and maintaining a steady supply of blood for medical treatments and emergencies. However, blood shortages and inconsistent donor participation pose challenges for blood banks and healthcare organizations. To address this issue, predictive analytics and machine learning techniques are being utilized to develop blood donation prediction models. It involves analysing various factors and data points to estimate the likelihood of individuals donating blood in the future. By examining demographic information, past donation history, and behavioural patterns, predictive models can provide insights into the probability of someone becoming a blood donor. These prediction models are typically built using machine learning algorithms that learn from historical data on blood donors. By training on features such as age, gender, location, donation frequency, and other relevant factors, the models can identify significant predictors of blood donation and establish correlations between various variables. Once the prediction model is developed, it can be used to make predictions on new individuals or populations. This information can be invaluable for blood banks, healthcare organizations, and donation campaigns, as it helps them target specific demographics, plan efficient donation drives, and allocate resources effectively. While blood donation prediction models can provide valuable insights, it is important to note that they are not infallible. Individual circumstances and personal choices can vary, and predictions are based on historical data and statistical correlations. However, by leveraging predictive analytics, these models offer a data-driven approach to encourage blood donation and support the needs of patients requiring blood transfusions.

**1.2 PURPOSE:**

The purpose of blood donation prediction is to address the challenges and uncertainties associated with blood supply and donor participation. By utilizing predictive analytics and machine learning techniques, the goals of blood donation prediction. By utilizing blood donation prediction, organizations can make data-driven decisions, improve donor engagement, and optimize blood supply management. These efforts ultimately lead to better patient care, reduced blood shortages, and increased efficiency in the healthcare system.

**LITERATURE SURVEY**

**2.1 EXISTING APPROACHES:**

We inspected the scholarly writing and gathered what we found into a couple unique classifications. To start with, blood donation centres frequently will study benefactor volunteers to attempt to grasp the elements that drove them to give. For instance, Godin, Conner et al. (2007) tracked down that the significant variables that lead to rehashed blood gift among experienced contributors were aim, saw control, expected lament, moral standard, age, and past gift recurrence. Additionally, the elements prompting rehashed blood gift among new benefactors were just aim and age.

Others have planned investigations to comprehend one's thought processes in giving blood. Sojka and Sojka (2008) reviewed more than 500 donators and tracked down that the most regularly revealed inspiration among their members was companion impact (47.2%), trailed by media demands (23.5%).

In conclusion, they tracked down that benevolence (40.3%), social obligation (19.7%), and companion impact (17.9%) were the essential drivers for blood contributors to keep on being blood benefactors later. As expressed already, just around 5% of qualified contributor populace really give (Katsaliaki 2008). The explanations behind this are routinely explored by friendly and conduct researchers to help move along populace investment (Ferguson, France et al. 2007). The examinations just examined are framed underneath in Table 1 are only a negligible portion of the many examinations being performed to more readily comprehend the social and conduct parts of why individuals give blood. We have observed that many investigations are attempting to broaden what is known as a hypothesis of arranged conduct (TPB), which keeps on being created around here. This hypothesis predicts the event of specific conduct given that it is purposeful and under volitional control (Veldhuizen, Fergusonet al. 2011). An efficient survey and meta-examination were acted around here by Bednall, Boveet al. (2013).

|  |  |  |  |
| --- | --- | --- | --- |
| **Authors** | **Methods** | **Data** | **Drivers** |
| (Godin, Conner et al. 2007) | Logistic Regression | Survey (2070 experience donors, 161 new donors) | *Experienced donors*: intention, perceived control, anticipated regret, moral norm, age, and past donation frequency. *New donors*: intention and age |
| (Sojka and Sojka 2008) | Descriptive statistics | Survey (531 participants) | *General motivators*: friend influence (47.2%), media requests (23.5%). *Continued donations*: altruism (40.3%), social responsibility (19.7%), friend influence (17.9%) |
| (Masser, White et al. 2009) | Structural equation modeling | Survey 1 (263 participants); Follow-up survey (182 donors) | Moral norm, donation anxiety, and donor identity indirectly predicted intention through attitude. |
| (Masser, Bednall et al. 2012) | Path analysis | Survey1 (256 participants) | Their extended TPB model showed intention was predicted by attitudes, perceived control, and self-identify |

**2.2 PROPOSED SOLUTION:**

In this capstone project, we aim to develop a predictive model that can help blood banks to identify donors who are likely to come back to donate blood based on a limited number of attributes. The model will leverage machine learning algorithms to analyse the data from previous blood donations and predict the likelihood of donors to return for subsequent donations. The proposed predictive model has the potential to assist blood banks in increasing the number of repeat donors and maintaining a stable supply of blood. Moreover, this model can provide valuable insights into the factors that influence donor behaviour, which can help blood banks to create targeted campaigns to encourage more people to donate blood regularly. Based on the given dataset, the blood donation project can be formulated as a binary classification problem. The goal is to predict whether a donor will donate blood in March 2007 (Made Donation in March 2007) based on their past donation behaviour and demographic attributes. The target variable is binary, and the dataset contains both categorical and continuous features. Therefore, suitable machine learning algorithms for this problem include logistic regression, KNN, decision trees, random forest, support vector machines (SVMs), GradientBoosting and XGBoost.

When evaluating the performance of a binary classification model, several performance metrics can be used, depending on the specific needs of the project. Some of the common performance metrics used for binary classification problems include:

**Accuracy:** The proportion of correct predictions made by the model.

**Precision:** The proportion of true positives (correctly predicted positive class) among all predicted positives.

**Recall:** The proportion of true positives among all actual positive samples.

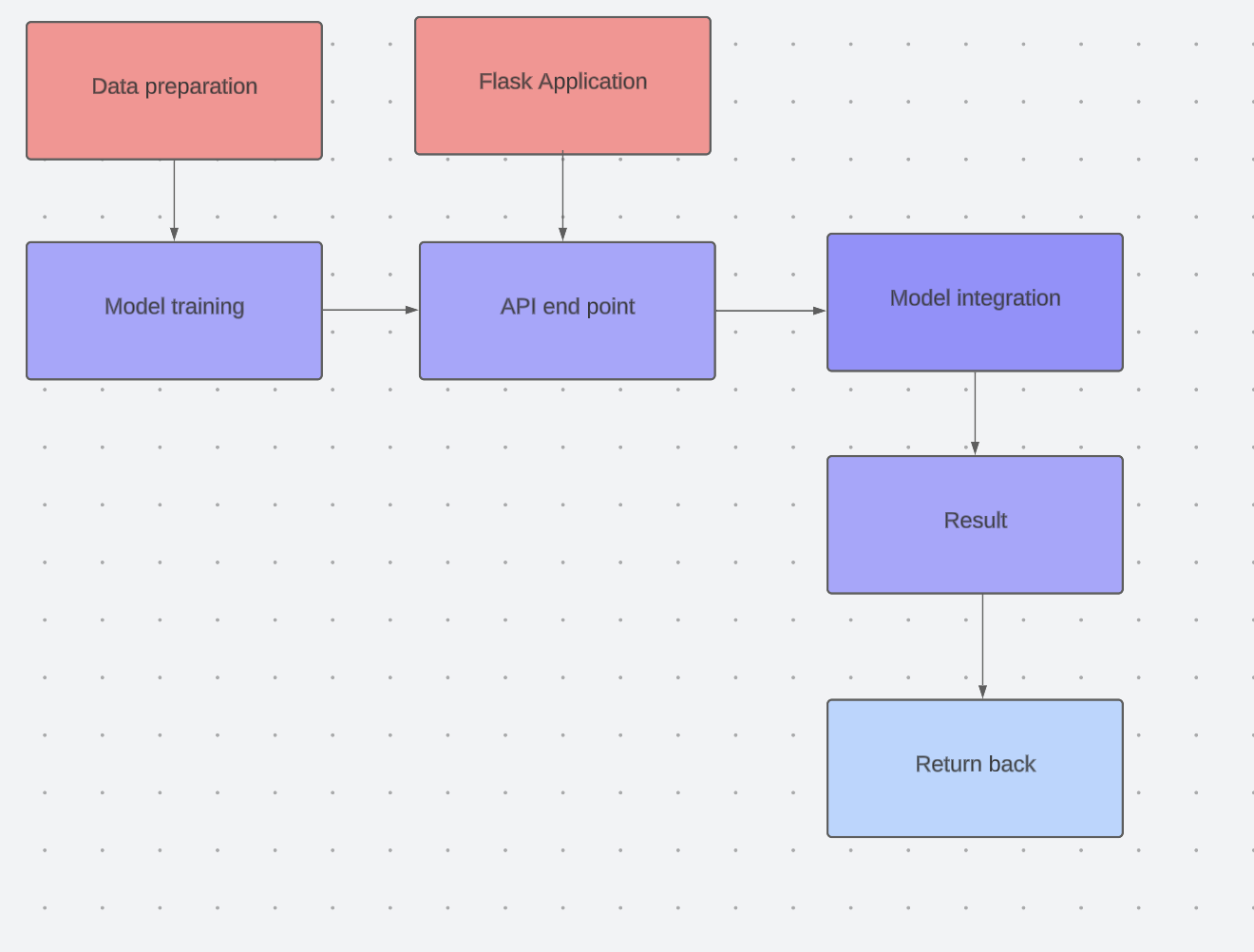
**F1-score:** The harmonic mean of precision and recall, which combines both metrics.

Area Under the ROC Curve (AUC-ROC): The area under the receiver operating characteristic (ROC) curve, which measures the trade-off between true positive rate and false positive rate.

**THEORETICAL ANALYSIS**

**3.1 BLOCK DIAGRAM:**

Diagrammatic overview of the project.



**3.2 HARDWARE/SOFTWARE DESIGNING:**

Hardware and software requirements of the project

**Hardware Specifications:**

The basic components like,

**Processor:** A modern processor with multiple cores for efficient data pre-processing and other capabilities.

**Memory (RAM):** Sufficient RAM to handle data processing and model training and testing.

**WIFI connections:** for supporting the support of the jupyter notebook to run the required packages algorithms and other works regarding flask and the web based UI for the user to input the data to check the predictions of the model

**Software Specifications:**

**Programming:** Knowledge of programming language Python, which is commonly used for data analysis and machine learning tasks, and performing various machine learning algorithms and use of tools like **Anaconda, Spyder, Jupiter notebook** would be essential also the use of **flask**

**Data Analysis Libraries:** Python libraries such as pandas and NumPy are utilized for data cleaning, exploration, and manipulation.

**Machine Learning Libraries:** Libraries like scikit-learn or label encoder train test split the accuracy score as a part of the evaluation metrics used for developing and training machine learning models.

**CSV Data Import:** The project would involve importing the CSV file containing the car performance data. Python libraries like pandas can be used to read and manipulate the data.

**Data Visualization:** Libraries like Matplotlib or Seaborn can be used for visualizing the data and generating insights.

**EXPERIMENTAL INVESTIGATION**

The experimental investigations in a Blood donation prediction project could involve various aspects depending on the specific goals and objectives of the project. Here are some possible experimental investigations that could be conducted:

**Data quality:**

- 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 pre-processing techniques like removing duplicates, transforming data into normal distribution. We have also performed exploratory data analysis (EDA) to identify the patterns in the data and understand the data distribution.

- With given number of attributes, the model's accuracy score, and F1 score are appeared to be stagnant around 0.75 and 0.65 respectively. To improve the model performance, more observations and features are required which are more relevant for prediction. The idea on new features can be gained through domain analysis done at the beginning of the project.

**Limited data:**

- The dataset has only 266 instances which may not be enough for some machine learning algorithms.

- To mitigate this, we have used cross-validation during model selection and evaluated the performance on multiple metrics to ensure that the models are robust and not overfitting.

**Imbalanced dataset:**

- In this project, 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 techniques over-sampling technique, SMOTE.

- We have also used evaluation metrics like precision, recall, F1-score, and ROC-AUC score, which are suitable for imbalanced datasets. We have also analysed the precision and recall trade-off and ROC AUC curve to find the better threshold value for predicting the classes.

**Model selection:**

- Choosing the right machine learning algorithm and tuning its hyperparameters can be challenging and time-consuming.

- To mitigate this, we have tried multiple algorithms such as Logistic Regression, Support Vector Classifier, Random Forest Classifier, and Gradient Boosting Classifier.

- We also used grid search to tune the hyperparameters for each algorithm and chose the best performing one based on evaluation metrics.

- In addition, we have also used techniques like cross-validation to compare different models' performance and choose the best one.

**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.

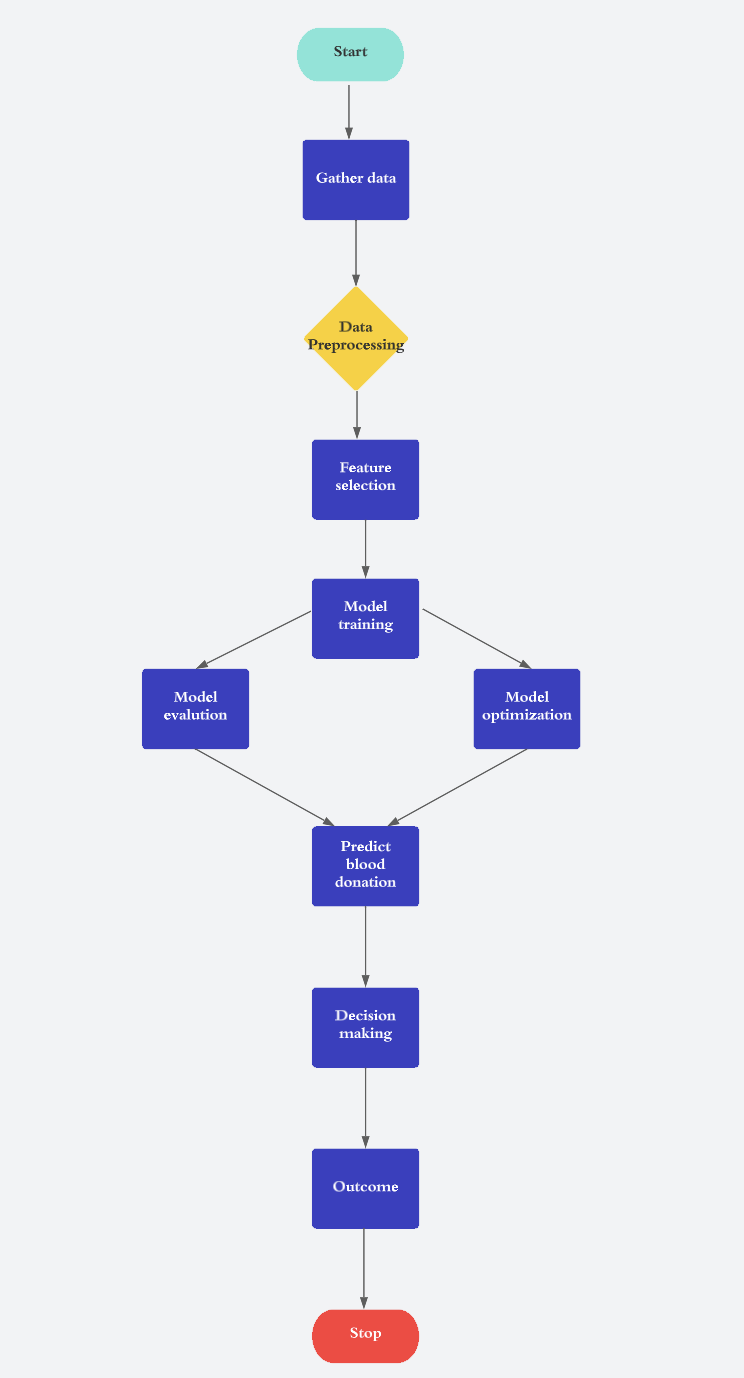
- In this project, we have used GridSearchCV to search for the best hyperparameters. However, we noticed that performance of selected models decreased after tuning. So, we have preferred model without tuning.

**Overfitting and underfitting:**

- 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.

**FLOWCHART**

****

**RESULT**

**CROSS VALIDATION:**

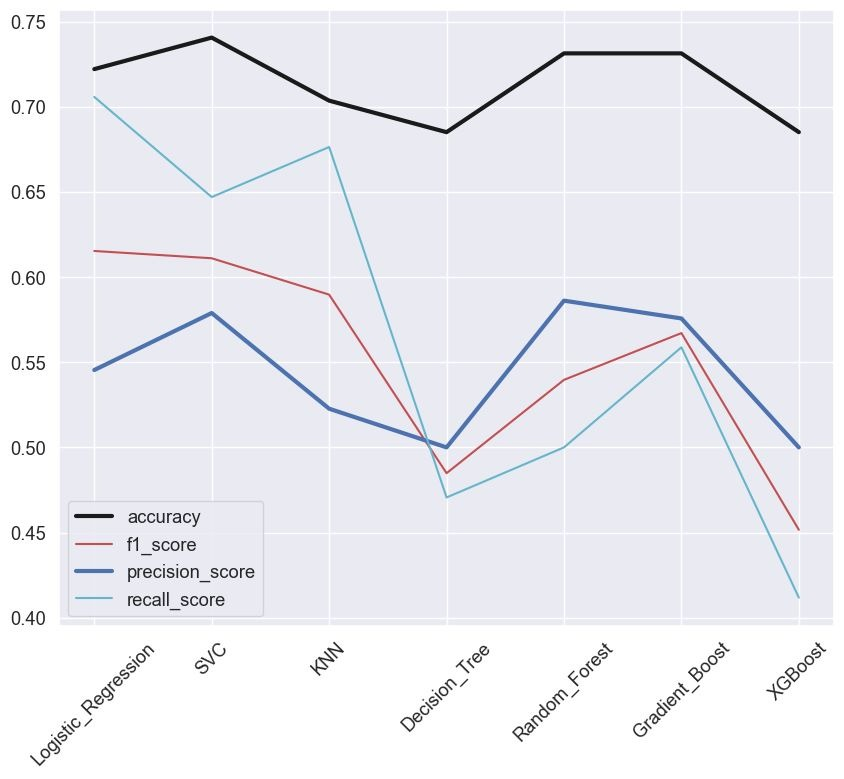
****

**A picture containing text, diagram, screenshot, rectangle

Description automatically generated**

Overall, the results suggest that logistic regression, SVC and Gradient Boost are good candidates for the final model. However, it is important to note that these results are based on the SMOTE-transformed data and may not generalize well to new data. Therefore, we will evaluate the models on a holdout test set before making any final decisions.

**TEST SET PERFORMANCE OF MODELS:**

****

overall, SVC and Logistic Regression are the best models for this dataset based on their performance on the test set.

**SVC HYPER-PARAMETER TUNING:**

**A screenshot of a computer

Description automatically generated with medium confidence**

**LOGISTIC REGRESSION HYPER-PARAMETER TUNING:**

**A screenshot of a computer screen

Description automatically generated with low confidence**

**For SVC model:**

- We trained the SVC model with default hyperparameters and evaluated its performance using various metrics such as accuracy, precision, recall, f1-score, and 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.

**For Logistic Regression model:**

- We then trained the Logistic Regression model with default hyperparameters and evaluated its performance using various metrics such as accuracy, precision, recall, f1-score, and ROC AUC score.

- Again we tried to improve the model's performance by tuning the threshold value.

- We used the grid search technique to optimize the Logistic Regression model's hyperparameters, including C, penalty, and solver, and used the best set of hyperparameters to fit the model and get the predictions.

- We also performed hyperparameter tuning and then analysed results

- However, we found that the tuned Logistic Regression model with a threshold of 0.5 gave unfavourable results compared to the default model in terms of f1-score and ROC AUC score. Therefore, we did not select this model as our final model.

Overall, we can say that the tuned SVC model with a threshold of 0.44 gave the best performance among all the models we developed.

Hence, we will select Support Vector Classifier model with threshold at 0.44.

It gives following results:

**Accuracy 0.76**

**precision 0.60**

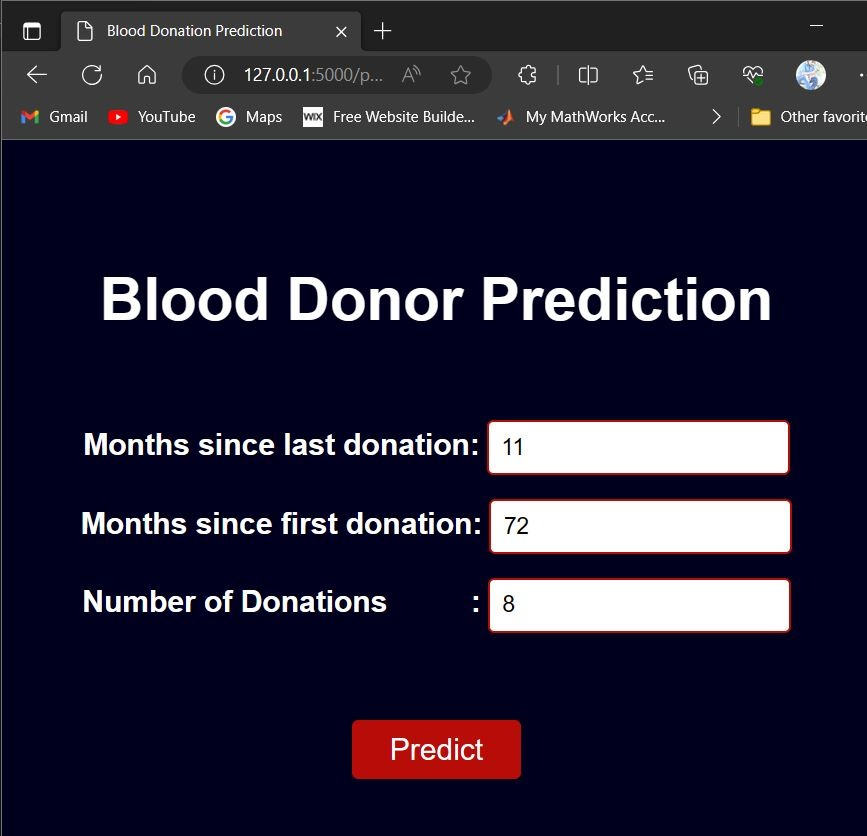
**F1 score 0.64**

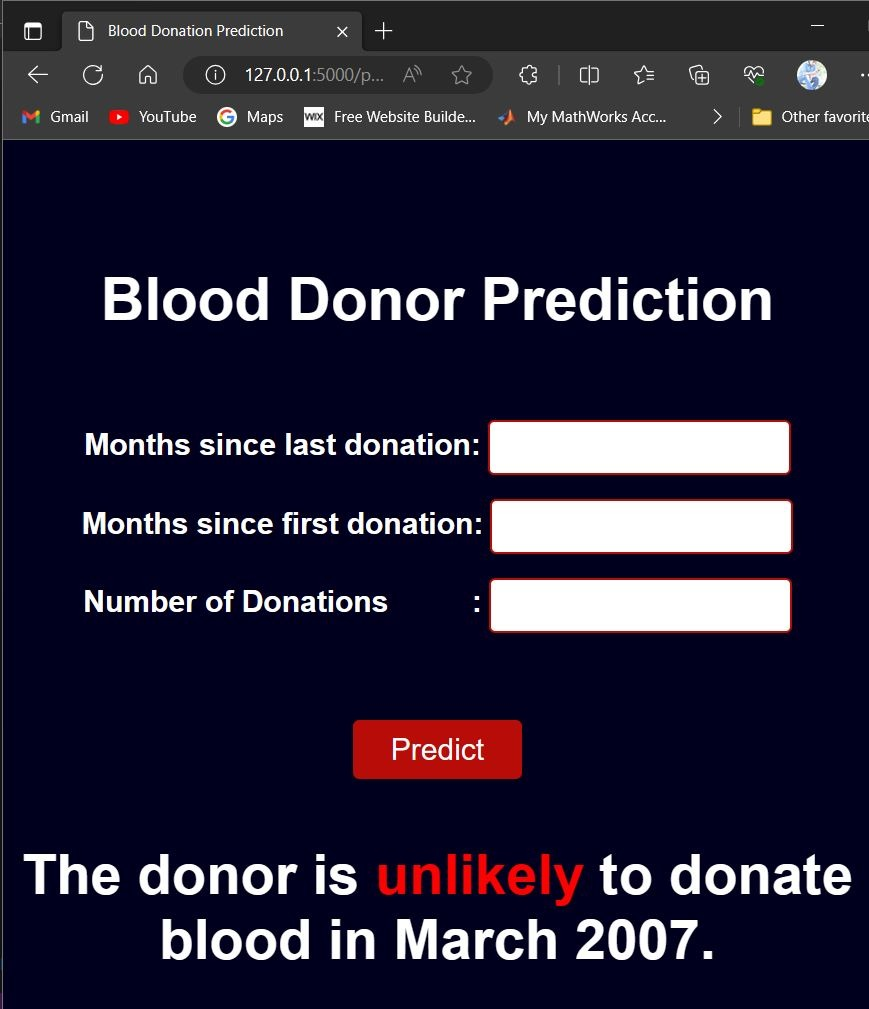
**ROC AUC score 0.74**

A screenshot of a computer

Description automatically generated

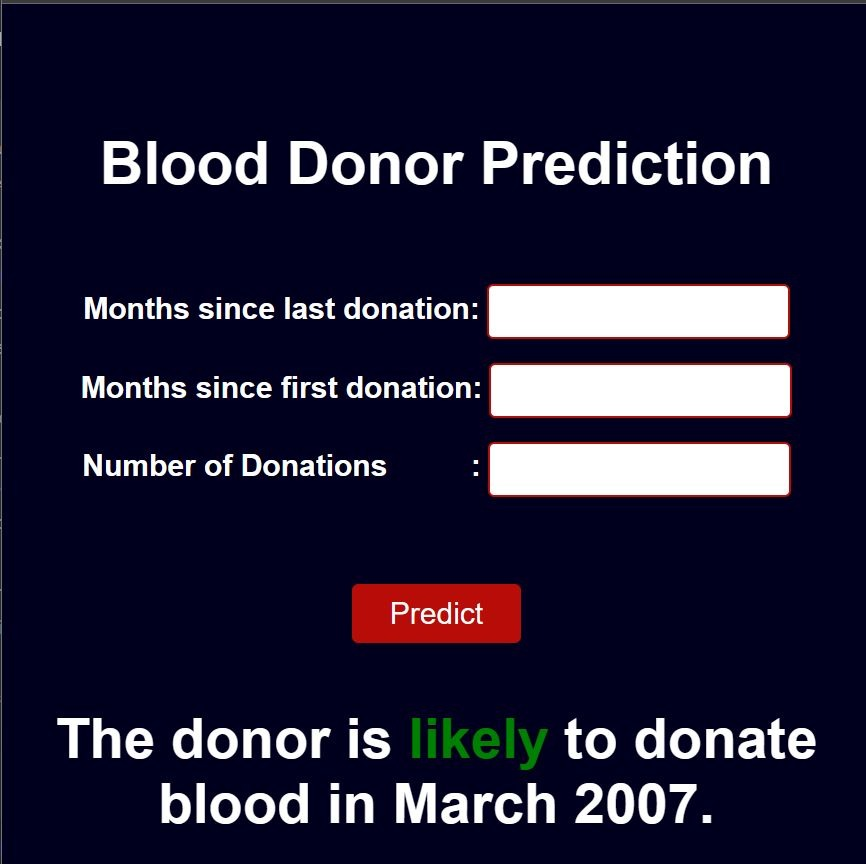
**BASED ON THE GIVEN DATA SET:**





A screenshot of a computer

Description automatically generated



**ADVANTAGES AND DISADVANTAGES**

**ADVANTAGES:**

**Ease of Development:** Flask is a lightweight and user-friendly web framework in Python. It provides a simple and intuitive syntax, making it easy to develop a blood donation prediction system using Flask. The framework's minimalistic design allows for quick prototyping and implementation.

**Scalability:** Flask is scalable and capable of handling multiple simultaneous requests efficiently. It can handle high traffic loads by leveraging additional tools like Gunicorn or uWSGI. This scalability ensures the blood donation prediction system can handle increased user demand without significant performance degradation.

**Deployment Flexibility:** Flask provides flexibility in deployment options. It can be easily deployed on various hosting platforms, cloud services, or containerized environments. Flask's deployment options enable the blood donation prediction system to be deployed in a manner that best suits the specific needs and infrastructure of the organization.

**Active Community and Documentation:** Flask has a large and active community of developers, providing extensive documentation, tutorials, and community support. This can be valuable when encountering issues or seeking guidance during the development of the blood donation prediction application.

**DISADVANTAGES:**

**Limited Functionality:** Flask is a micro-framework, which means it provides only the essential features for web development. Advanced functionalities like user authentication, authorization, and complex data management may require additional packages or custom implementation.

**Manual Configuration:** Flask requires manual configuration for handling more complex deployment scenarios, such as load balancing, containerization, or handling multiple instances of the application. This might require additional effort and technical expertise.

**Performance Concerns:** Although Flask is known for its scalability, its performance might be relatively lower compared to more robust frameworks like Django or asynchronous frameworks like FastAPI. In high-traffic scenarios, Flask might not provide the same level of performance as more optimized frameworks.

**Maintenance and Updates:** Flask is an open-source framework, which means it requires continuous maintenance and updates to address security vulnerabilities or compatibility issues. Staying updated with the latest versions of Flask and its dependencies might require periodic effort.

**APPLICATIONS**

Donor Recruitment and Retention

Blood Supply Management

Resource Allocation and Staffing

Targeted Communication and Education

Donor Health Monitorin

**CONCLUSION**

Hence, we conclude that blood donation prediction using Flask provides a robust and versatile framework for developing a sophisticated blood donation prediction system. Flask's simplicity, flexibility, and web API capabilities make it an ideal choice for integrating machine learning models into web applications. With Flask, organizations can optimize donor recruitment and retention efforts, effectively manage blood supply, allocate resources efficiently, and deliver targeted communication and education campaigns. Flask's scalability ensures the system can handle high traffic loads, while its active community and extensive documentation offer valuable support during development. By leveraging Flask, organizations can enhanse their blood donation initiatives, improve operational efficiency, and ultimately make a positive impact on blood donation efforts.

**FURURE SCOPE**

In terms of medium future scope, blood donation prediction using Flask has several potential areas for further development and improvement:

**Enhanced Feature Engineering:** By incorporating more extensive feature engineering techniques, such as feature selection, feature extraction, and feature transformation, the blood donation prediction model can capture more nuanced relationships and patterns in the data. This can lead to more accurate predictions and better understanding of the factors influencing blood donation behaviour.

I**ntegration of External Data Sources:** Integrating additional external data sources, such as demographic data, socio-economic indicators, and health-related information, can provide a more comprehensive view of potential blood donors. This enriched data can improve the prediction model's accuracy and enable better segmentation and targeting of potential donors.

**Integration of Feedback Mechanisms:** Incorporating feedback mechanisms from donors, such as surveys or rating systems, can provide valuable insights into their experience and satisfaction levels. This feedback can be used to refine the prediction model and improve the overall blood donation process.

**Data Visualization and Reporting**: Developing interactive data visualization tools and reporting functionalities can help organizations analyse and interpret the prediction results effectively. Visual representations of donation patterns, trends, and insights can facilitate decision-making and communication of key findings to stakeholders.

**BIBLIOGRAPHY**

* [**https://doi.org/10.1109/DSAA.2019.00079**](https://doi.org/10.1109/DSAA.2019.00079)**]**
* [**https://doi.org/10.1109/ICOSEC45756.2019.8983454**](https://doi.org/10.1109/ICOSEC45756.2019.8983454)
* [**https://doi.org/10.17306/J.JMIT.0000028604**](https://doi.org/10.17306/J.JMIT.0000028604)
* [**https://doi.org/10.1109/ConTEL.2016.7536350**](https://doi.org/10.1109/ConTEL.2016.7536350)
* [**https://doi.org/10.1109/SMC.2015.449**](https://doi.org/10.1109/SMC.2015.449)
* [**https://doi.org/10.1016/j.eswa.2011.03.081**](https://doi.org/10.1016/j.eswa.2011.03.081)

**APPENDIX**

**Blood donation prediction ipynb file link:**

[**https://github.com/Aryan4032/blood-donation-prediction/blob/main/BLOOD%20DONATION%20PREDICTION/BloodDonationPred.ipynb**](https://github.com/Aryan4032/blood-donation-prediction/blob/main/BLOOD%20DONATION%20PREDICTION/BloodDonationPred.ipynb)

**appy.py:**

[**https://github.com/Aryan4032/blood-donation-prediction/blob/main/BLOOD%20DONATION%20PREDICTION/app.py**](https://github.com/Aryan4032/blood-donation-prediction/blob/main/BLOOD%20DONATION%20PREDICTION/app.py)

**project video:**

[**https://drive.google.com/file/d/1DZ2rKUW5ciorgxIZJn\_m86Lk65Nulwnu/view?usp=drive\_link**](https://drive.google.com/file/d/1DZ2rKUW5ciorgxIZJn_m86Lk65Nulwnu/view?usp=drive_link)