**1. Introduction:**

The main aim of this project is to predict whether a donor will give blood the next time the blood donation campaign is held again at the university campus using machine learning models with various features of independent variables. This report will give you more information about the factors influencing the blood donation campaign.

**2. Dataset Overview:**

1. The dataset includes 6 explanatory variables describing different aspects of blood donors.
2. A binary variable ‘Made Donation in March 2007’ is the target variable, which represents whether he/she donated blood in March 2007.

**Attribute Information:**

* Unnamed: 0- You can ignore this column.
* Months since Last Donation- This is the number of months since this donor’s most

recent donation.

* Number of Donations- This is the total number of donations that the donor has made.
* Total Volume Donated- This is the total amount of blood that the donor has donated in cubuc centimeters.
* Months since First Donation- This is the number of months since the donor’s first donation.
* Made Donation in March 2007- A binary variable representing whether he/she donated blood in March 2007 (1 stand for donating blood; 0 stands for not donating blood).

**3. Data** **Summary**:

Number of records: 576

Number of features: 6

**Data Types:**

* Numerical Features – 6

**4. Null Values:**

|  |  |  |
| --- | --- | --- |
| Column Name | Null Values | Type |
| Unnamed:0 | 0 | Numerical |
| Months since last Donation | 0 | Numerical |
| Number of Donations | 0 | Numerical |
| Total Volume Donated | 0 | Numerical |
| Months since First Donation | 0 | Numerical |
| Made Donation in March 2007 | 0 | Numerical |

**5. Data Cleaning and Preprocessing:**

* 1. Removed outliers using statistical methods.

**6. Exploratory Data Analysis:**

**Correlation Analysis**

* Identified strong correlations between Number of Donations and features like Donation Consistency, Relative Recency and between Total Volume Donated (c.c.) with Months since First Donation, Donation frequency, Donation Consistency and Relative Recency.

**Data Visualization**

* Univariate and Bivariate Analysis (Countplot, Histogram)
* Feature relationships (Scatter plots, Heatmaps)
* Numerical feature multivariate analysis using pair plot.

**7. Feature Engineering**

* Created new features such as Recency of Donations, Donation frequency, Donation History Interaction, Donation Consistency, Relative Recency, Log Total Volume Donated.
* Selected important features using feature selection techniques.

**8. Modeling & Prediction**

**Model Selection**

* Implemented various models:
* Logistic Regression
* Random Forest
* Gradient Boosting Classifier

**Model Evaluation**

* Performance measured using:
  + Accuracy Score
  + Precision Score
  + Recall Score
  + F1-Score
* Best performing model: Gradient Boosting model shows the highest accuracy point.

**9. Model Comparison Report**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| Logistic Regression | 0.76 | 0.77 | 0.98 | 0.86 |
| Random Forest | 0.74 | 0.83 | 0.83 | 0.83 |
| **Gradient Boosting** | **0.77** | **0.83** | **0.87** | **0.85** |

**Best Model Recommendation**: Gradient Boosting

**10. Challenges Faced and Solutions**

**Data Quality Issues**

* Outliers were removed using statistical methods.

**Model Performance**

* Hyperparameter tuning was performed to improve accuracy.
* Feature selection was used to avoid overfitting.

**11. Deployment**

Link - <http://127.0.0.1:5000>

**12. Conclusion**

* This project aimed to predict whether a donor would donate blood in the next campaign at the university campus. After conducting thorough data cleaning, feature engineering, and exploratory analysis, various machine learning models were implemented to make predictions.
* The Gradient Boosting model outperformed others, achieving the highest accuracy and overall performance, with strong precision, recall, and F1-scores. Key factors influencing the donation behaviour included donation frequency, recency, and total volume donated.
* Despite challenges such as data quality issues and model performance, effective solutions like outlier removal, hyperparameter tuning, and feature selection were applied to enhance the model’s reliability. Based on the results, the Gradient Boosting model is recommended for predicting future blood donations.