UNIFIED MENTOR

Personalized Healthcare Recommendation

Data Science and Machine Learning Project

Prepared by:
Meghtithi Mitra

UNID: UMID13092558107

Tools & Technologies:
Python, Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, Google Colab

SECTION 1: INTRODUCTION

Personalised healthcare and preventive interventions increasingly rely on data-driven models that can identify high-risk individuals and recommend targeted actions. In this project, we develop a machine learning model that uses blood donation–related features to predict whether an individual is likely to donate blood in the future and to generate personalised recommendations.

The dataset contains information about previous donations, including how recently the person donated (Recency), how often they have donated (Frequency), the total volume of blood donated (Monetary), and how long they have been donating (Time). The target variable, **Class**, indicates whether the individual donated again at a reference point in time.

Although this dataset is donation-focused, the same methodology generalises to healthcare settings where patient history and behaviour are used to inform personalised outreach, reminders, and lifestyle counselling

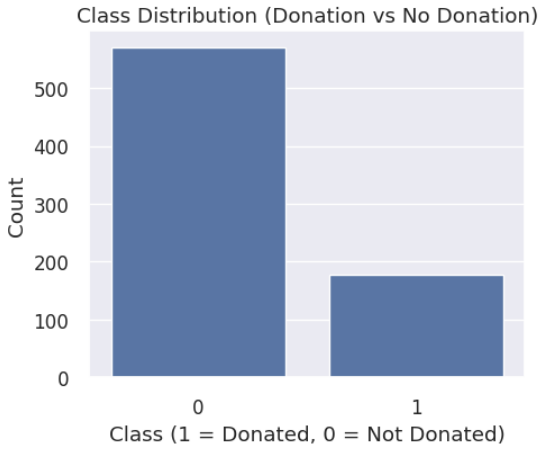
SECTION 2: DATASET DESCRIPTION

The `blood.csv` dataset consists of **748 observations** and **5 variables**:

- **Recency** – Months since last donation.
- **Frequency** – Total number of donations.
- **Monetary** – Total volume of blood donated (in c.c.).
- **Time** – Months since the first donation.
- **Class** – Target variable (1 = donated again, 0 = did not donate).

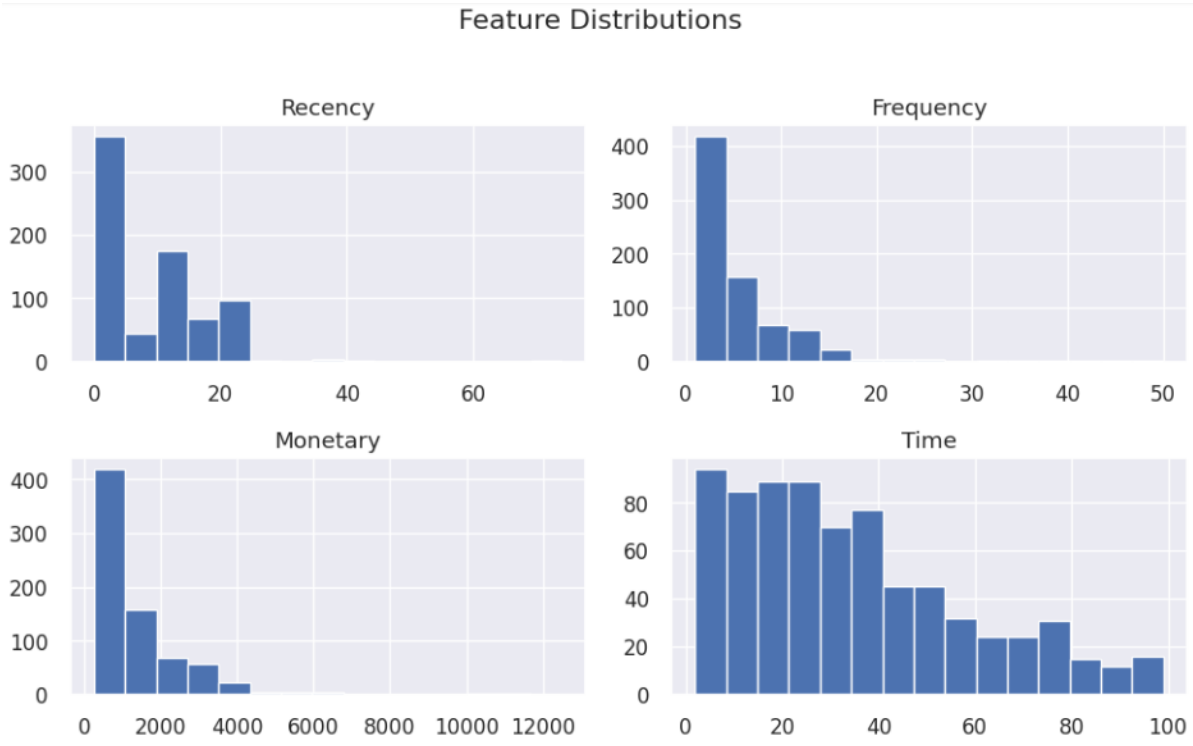
Key Observations

- No missing values were found in the dataset.
- The target variable `Class` is **imbalanced** (approx. 76% Class 0, 24% Class 1).
- Features are already numeric, so preprocessing is straightforward.



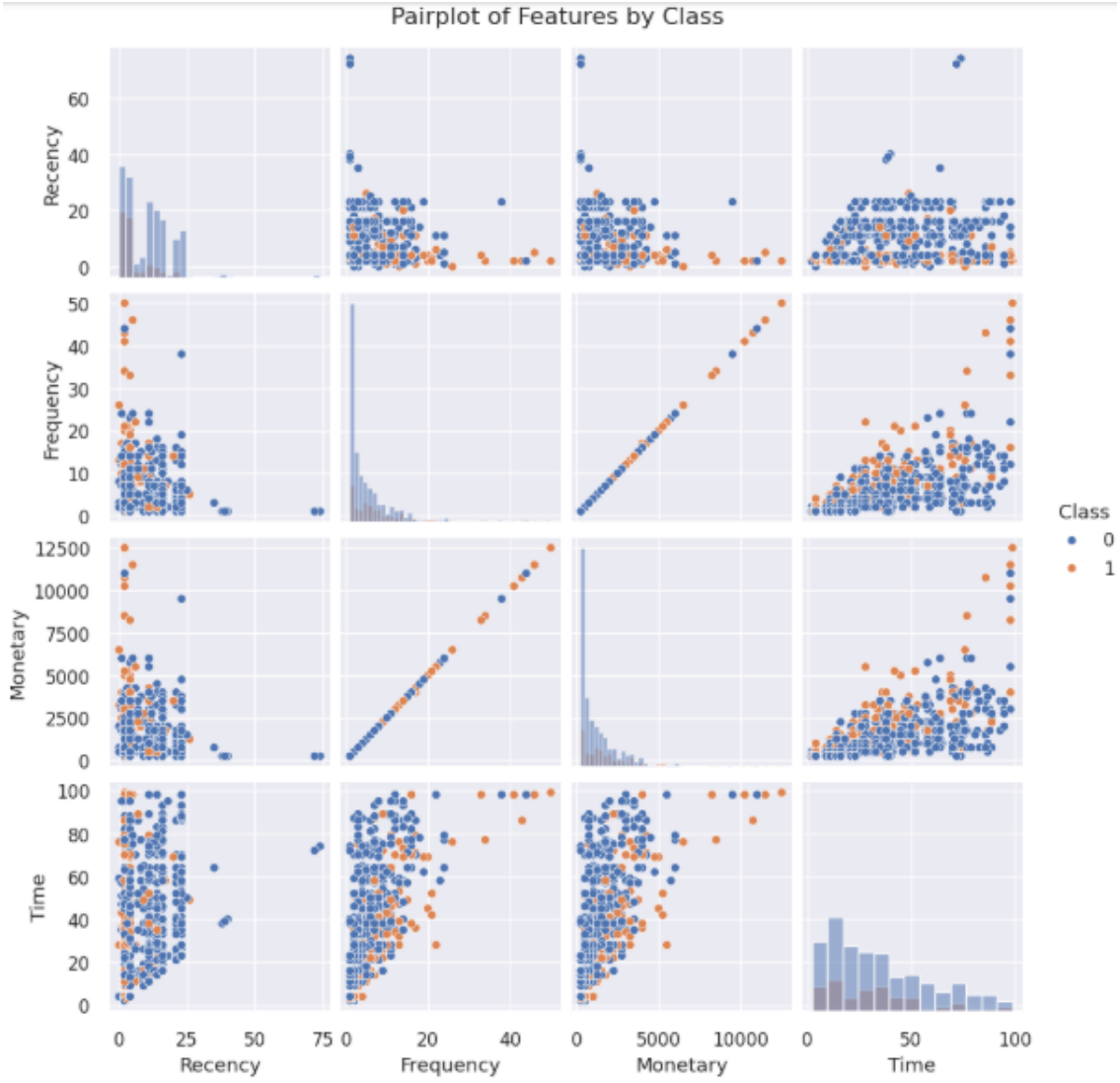
SECTION 3: EDA [EXPLORATORY DATA ANALYSIS] & CORRELATION ANALYSIS

Feature Distribution Analysis



- All variables exhibit **non-normal and skewed distributions**.
- **Recency** is heavily skewed toward lower values, indicating that many donors have donated blood recently.
- **Frequency** shows strong right skewness, with most donors donating only a few times and a small number donating very frequently.
- **Monetary** closely mirrors Frequency, as total blood volume increases with the number of donations.
- **Time** spans a wide range, suggesting the presence of both new donors and long-term donors.

Pairplot Observations



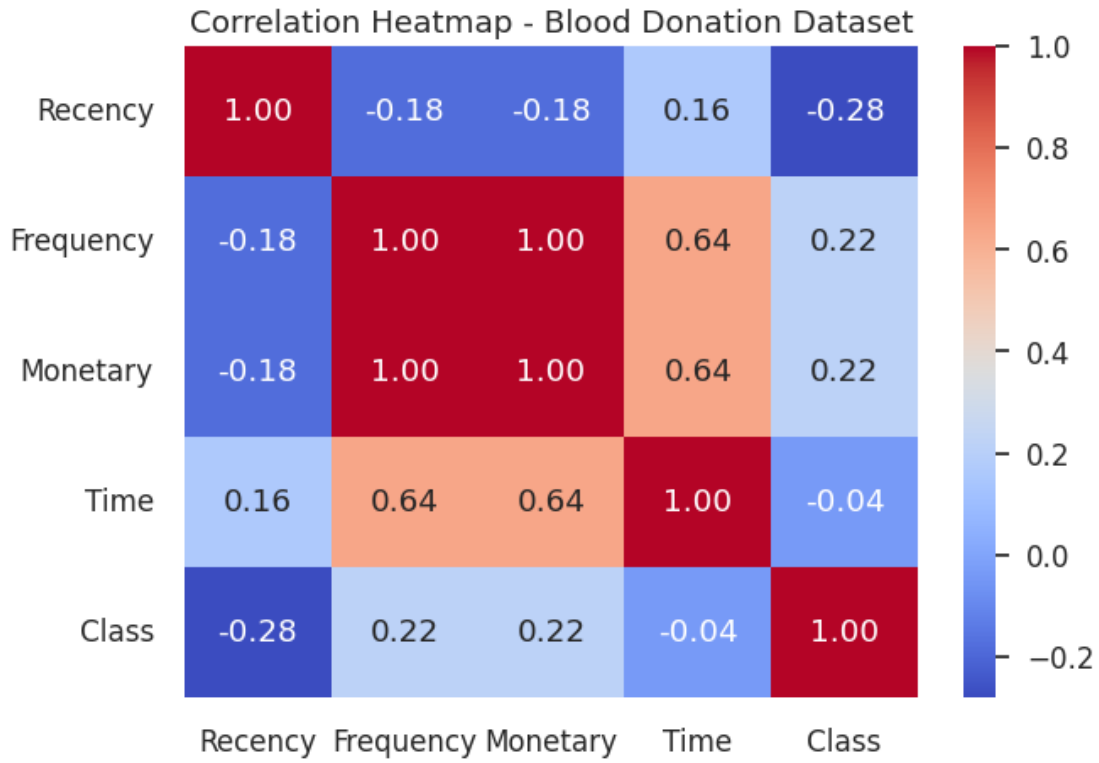
 Donors who belong to **Class 1 (returned donors)** tend to cluster at:

- **Lower Recency values**
- **Higher Frequency values**
- **Higher Monetary values**

 Significant overlap exists between the two classes, indicating that the classification problem is non-trivial.

 Despite the overlap, clear behavioural trends can be observed, supporting the predictive value of donation history.

Correlation Analysis



 Frequency and Monetary exhibit an almost perfect positive correlation, as expected.

 Time shows a moderate positive correlation with Frequency and Monetary, indicating that long-term donors tend to donate more often.

 Recency has a moderate negative correlation with the target variable (Class), suggesting that recent donors are more likely to donate again.

 Frequency and Monetary have positive correlations with Class, indicating that consistent donors have a higher probability of returning.

 Time shows a weak correlation with Class, implying that donation duration alone is less informative than recent and frequent donation behaviour.

SECTION 4: DATA PREPROCESSING

Before building machine learning models, the dataset was preprocessed to ensure it was suitable for training and evaluation.

Handling Missing Values

- The dataset was examined for missing values across all variables.
- No missing values were found in any feature or in the target variable.
- As a result, no imputation or row removal was required.

Feature Scaling

- Although all features were numeric, they existed on different scales (for example, Monetary values were much larger than Recency values).
- To ensure that no feature dominated the learning process due to scale differences, feature standardization was applied.
- **StandardScaler** was used to transform the features to have zero mean and unit variance.

Train–Test Split

- The dataset was split into training and testing sets using an **80:20 ratio**.
- Stratified sampling was applied based on the target variable (Class) to preserve the original class distribution in both sets.
- The training set was used to fit the models, while the test set was used for performance evaluation on unseen data.

This preprocessing step ensured that the dataset was clean, well-scaled, and appropriately structured for fair and reliable model comparison.

SECTION 5: MODEL BUILDING

After preprocessing, supervised machine learning models were developed to predict whether an individual is likely to donate blood again. The objective of this stage was to compare different classification algorithms and identify the model that best captures the underlying patterns in donor behaviour.

Two classification models were implemented and evaluated:

Logistic Regression

- Logistic Regression was used as a baseline classification model.
- It is a widely used algorithm for binary classification problems, particularly in healthcare and behavioural analytics.
- The model estimates the probability of an individual belonging to Class 1 (donated again) based on a linear combination of input features.
- Feature scaling was applied before training to ensure stable and reliable model performance.

Random Forest Classifier

- Random Forest is an ensemble-based learning algorithm that constructs multiple decision trees and combines their predictions.
- This model was selected to capture non-linear relationships and interactions between features that may not be detected by linear models.
- To address the class imbalance present in the dataset, class weighting was applied during training.
- Random Forest also provides feature importance measures, which support interpretability and further analysis.

Both models were trained on the same training dataset using identical preprocessing steps to ensure a fair comparison. The trained models were then evaluated on the test dataset to assess their predictive performance and generalization ability.

SECTION 6: MODEL EVALUATION AND RESULTS

The performance of the trained models was evaluated using the test dataset to assess their ability to generalize to unseen data. Given the imbalanced nature of the target variable, multiple evaluation metrics were considered rather than relying solely on accuracy.

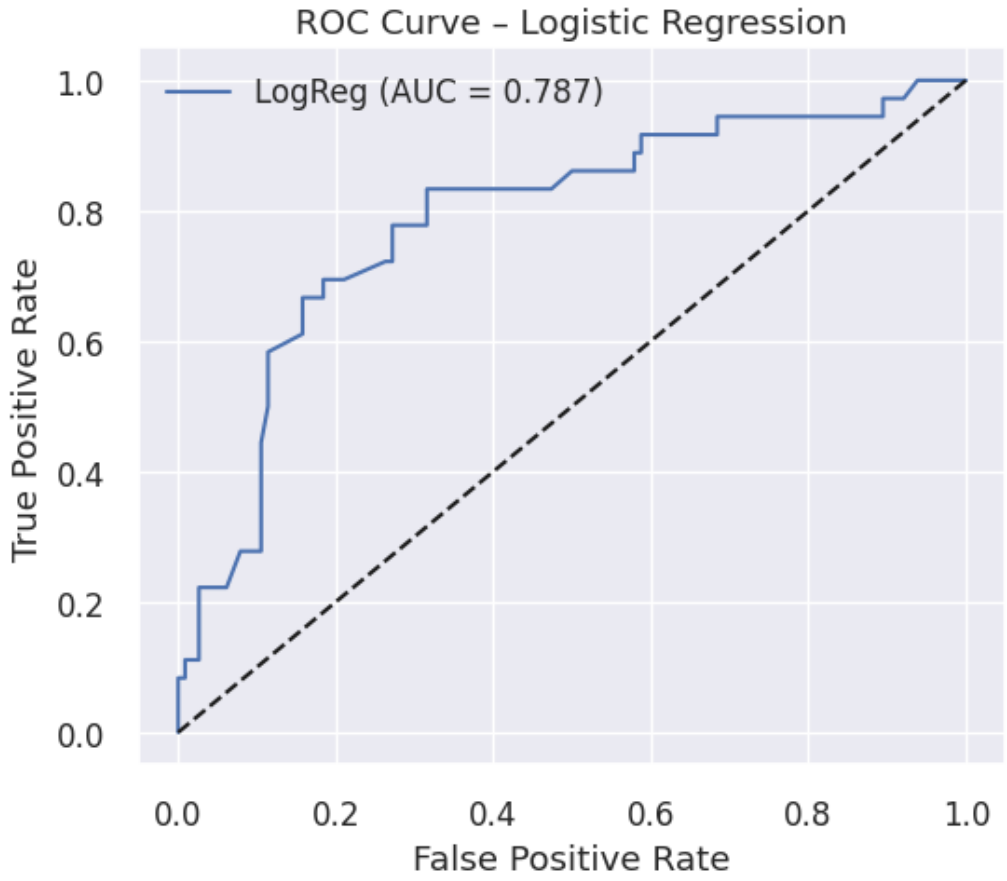
The following metrics were used for evaluation:

- Accuracy
- Precision
- Recall
- F1-score
- Confusion Matrix
- Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC)

6.1 Logistic Regression Performance

The Logistic Regression model demonstrated reasonable overall accuracy; however, its performance on the minority class (Class 1 – returning donors) was limited.

```
Logistic Regression – Classification Report
              precision    recall  f1-score   support

           0       0.78      0.97      0.86       114
           1       0.57      0.11      0.19        36

    accuracy                           0.77       150
   macro avg       0.67      0.54      0.52       150
weighted avg       0.73      0.77      0.70       150

Confusion Matrix (Logistic Regression)
[[111   3]
 [ 32   4]]
{'accuracy': 0.7666666666666667,
 'precision': 0.5714285714285714,
 'recall': 0.1111111111111111,
 'f1': 0.18604651162790697,
 'roc_auc': np.float64(0.7872807017543859)}
```



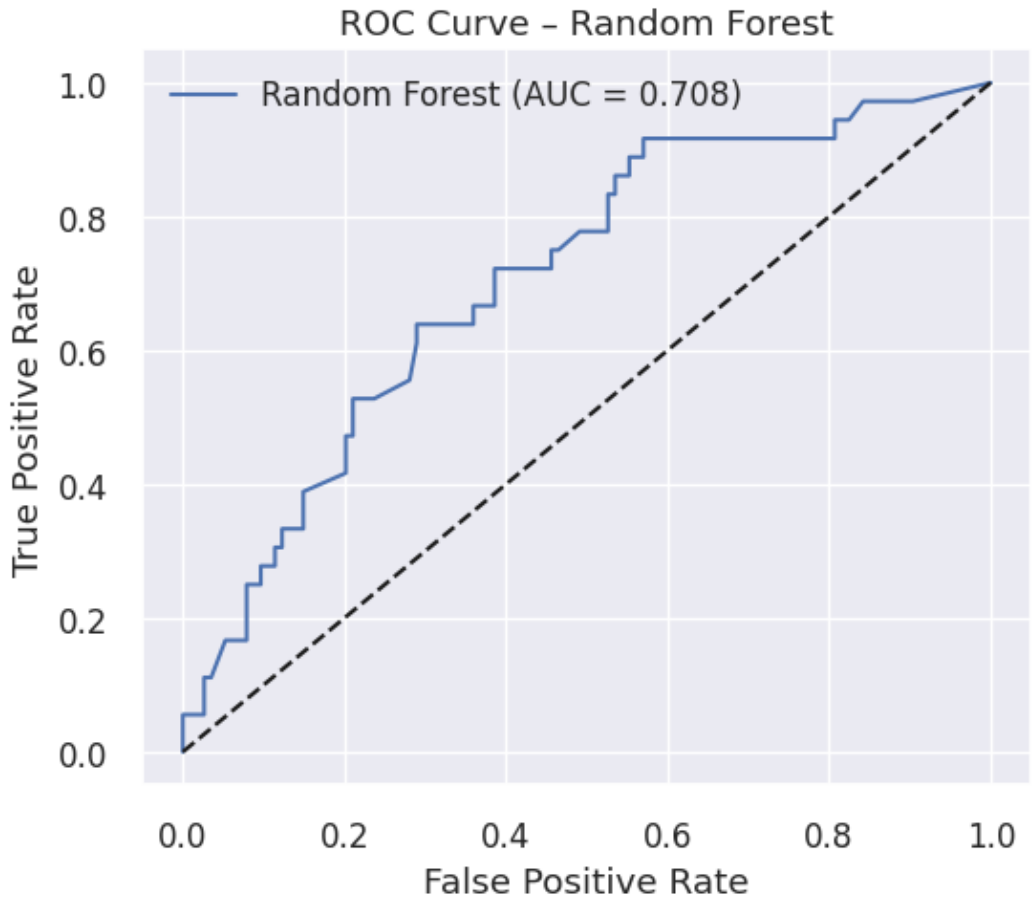
- The confusion matrix indicates that the model correctly identifies most non-returning donors (Class 0).
- However, a large number of returning donors (Class 1) were misclassified as non-returning.
- This resulted in **low recall for Class 1**, meaning the model failed to detect many potential future donors.
- Although the ROC–AUC score was relatively high, indicating good ranking ability, the imbalance in predictions limits its practical usefulness for personalized recommendations.

Overall, Logistic Regression serves as a useful baseline model but is less effective for identifying donors who are likely to donate again.

6.2 Random Forest Performance

The Random Forest classifier showed improved performance in identifying returning donors compared to Logistic Regression.

```
Random Forest – Classification Report
              precision    recall  f1-score   support

           0       0.83      0.79      0.81       114
           1       0.41      0.47      0.44        36

    accuracy                           0.71       150
   macro avg       0.62      0.63      0.62       150
weighted avg       0.73      0.71      0.72       150

Confusion Matrix (Random Forest)
[[90 24]
 [19 17]]
{'accuracy': 0.7133333333333334,
 'precision': 0.4146341463414634,
 'recall': 0.4722222222222222,
 'f1': 0.44155844155844154,
 'roc_auc': np.float64(0.7084551656920077)}
```



- The confusion matrix shows a better balance between correctly classified Class 0 and Class 1 instances.
- Recall and F1-score for Class 1 were notably higher, indicating that the model was more successful in detecting potential future donors.
- Although the overall accuracy was slightly lower than that of Logistic Regression, the improvement in recall and F1-score makes the Random Forest model more suitable for this problem.
- The ROC curve and AUC score further confirm that the model has good discriminatory ability.

This performance suggests that Random Forest is better suited for handling the class imbalance and capturing complex relationships within the dataset.

6.3 Model Comparison and Selection

A comparison of both models highlights the following:

- Logistic Regression achieves higher accuracy but performs poorly in detecting Class 1 instances.
- Random Forest provides a better balance between precision and recall, particularly for the minority class.
- For personalized healthcare and donor engagement applications, **identifying potential donors (Class 1) is more important than maximizing overall accuracy**.

Based on these observations, the **Random Forest classifier was selected as the final model** for subsequent analysis and recommendation generation.

SECTION 7: MODEL ENHANCEMENT USING GRADIENT BOOSTING

 While Random Forest demonstrated strong performance, an additional model was explored to further enhance predictive accuracy and robustness.

 Gradient Boosting was selected as it builds models sequentially, with each new model focusing on correcting the errors made by previous models.

 Gradient Boosting is particularly effective for **structured tabular data** and has been shown to outperform bagging-based models such as Random Forest in many classification tasks.

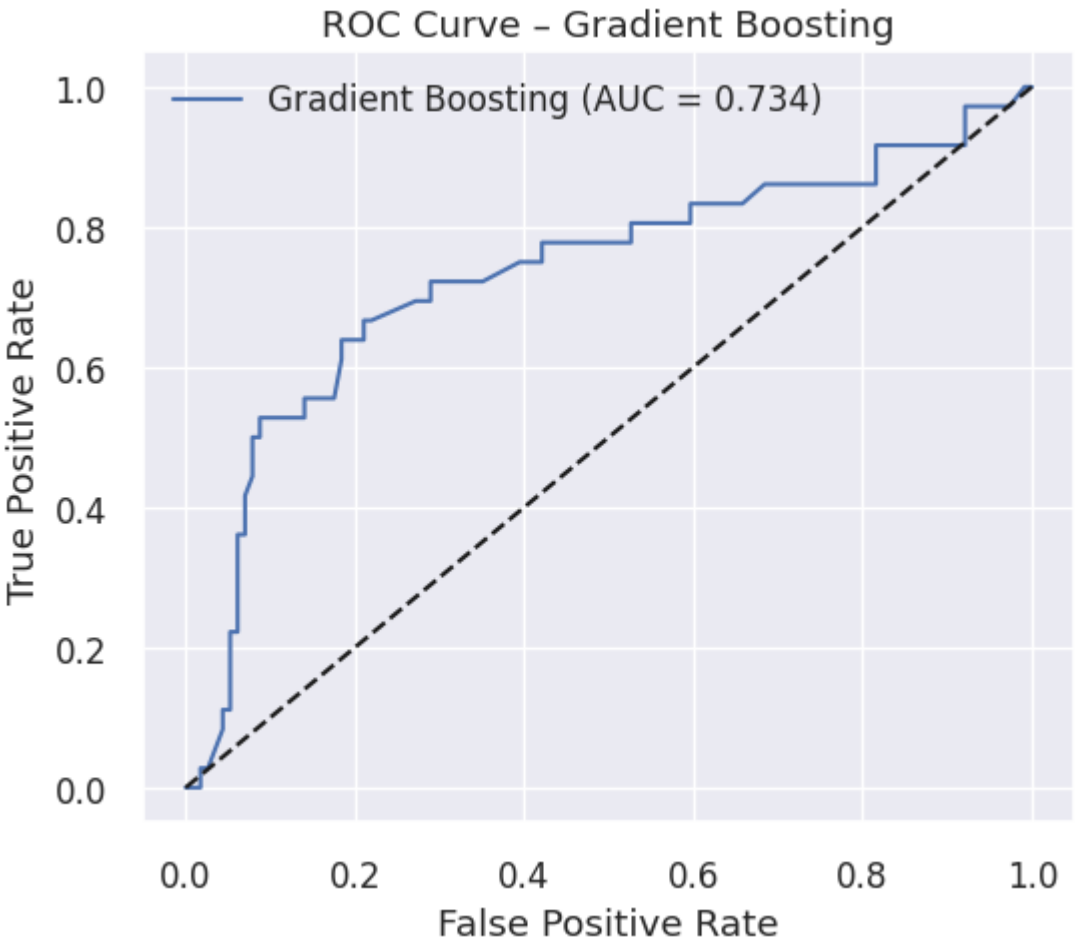
 To ensure a fair comparison, the Gradient Boosting model was trained using the **same training and testing splits** as the baseline models.

 Evaluation results indicate that Gradient Boosting achieved an **improved balance between precision and recall** for the minority class (Class 1).

 The **ROC–AUC score** further suggests enhanced discriminatory ability compared to the baseline models.

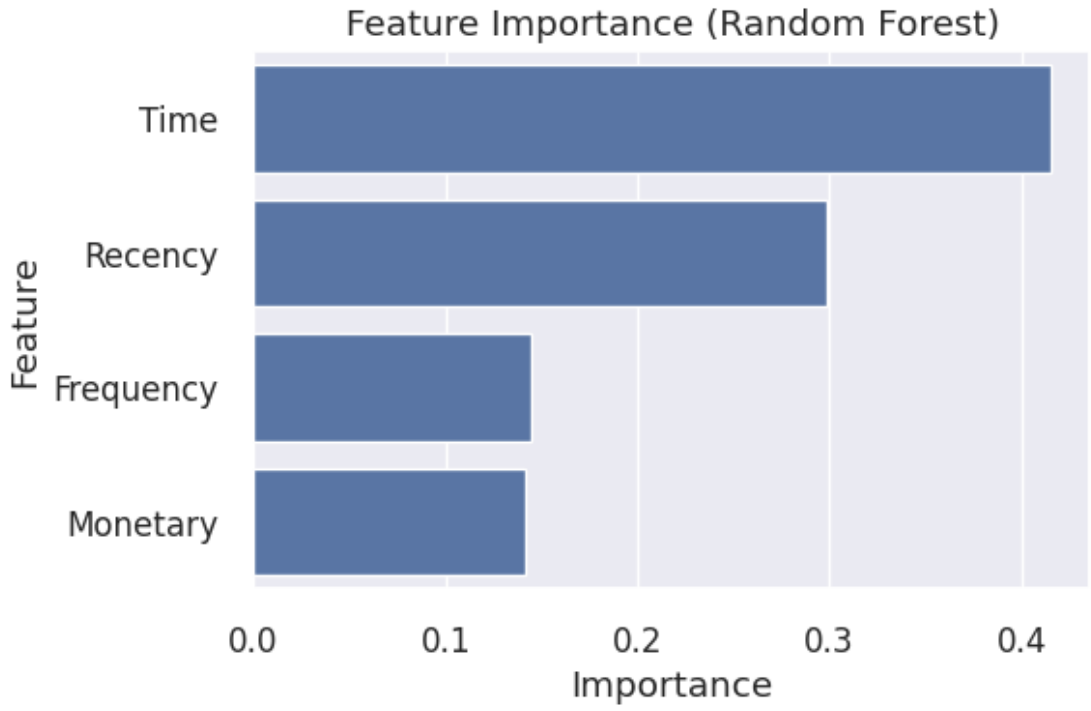
 This experiment demonstrates that model performance can be **incrementally improved beyond Random Forest** while maintaining interpretability and methodological consistency.

```
Gradient Boosting Classification Report
              precision    recall  f1-score   support

           0       0.83      0.93      0.88       114
           1       0.64      0.39      0.48        36

    accuracy                           0.80       150
   macro avg       0.73      0.66      0.68       150
weighted avg       0.78      0.80      0.78       150

Confusion Matrix
[[106   8]
 [ 22  14]]
ROC-AUC: 0.7337962962962964
```


ROC Curve – Gradient Boosting
Gradient Boosting (AUC = 0.734)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

SECTION 8: FEATURE IMPORTANCE ANALYSIS

- Feature importance analysis was conducted to understand which variables contribute most to the model's predictions.
- Since ensemble tree-based models provide inherent measures of feature importance, this analysis was performed using the **Random Forest** model.
- Feature importance scores represent the relative contribution of each feature in reducing impurity across the trees.
- The results indicate that:
    - **Recency** is the most influential feature, highlighting the importance of how recently a donor last donated.
    - **Frequency** also plays a significant role, suggesting that donors with a history of repeated donations are more likely to return.
    - **Monetary**, which reflects total blood donated, closely follows Frequency in importance.
    - **Time** has comparatively lower importance, indicating that donation duration alone is less predictive than recent and frequent donation behaviour.
- These findings align well with insights from the exploratory data analysis and reinforce the behavioural interpretation of donor return patterns.
- Feature importance analysis improves model interpretability and supports the use of machine learning models in healthcare-related decision-making.



SECTION 9: PERSONALIZED RECOMMENDATION SYSTEM

- The primary objective of this stage is to move beyond prediction and convert model outputs into **actionable, personalised recommendations**.
- The recommendation system uses predicted outcomes to support targeted donor engagement strategies.

Model Selection for Recommendations

- Multiple machine learning models were evaluated during this study.
- The **Random Forest classifier** was selected as the primary model for generating personalised recommendations.
- This choice was based on its **higher recall for the minority class (Class 1)** compared to other models.
- In the context of donor engagement and healthcare applications:
    - Missing a potential returning donor (**false negative**) is more costly than incorrectly flagging a non-returning donor.
    - Higher recall ensures that fewer potential donors are overlooked.

Recommendation Logic

- Predicted probabilities from the Random Forest model are used to estimate the likelihood of future donation.
- Based on the predicted class:
    - **Class 1 (likely to donate again):**
        - Recommend proactive engagement such as reminders, follow-up messages, and appreciation campaigns.
    - **Class 0 (unlikely to donate again):**
        - Recommend awareness-building initiatives, educational outreach, or long-term engagement strategies.
- While Gradient Boosting demonstrated strong overall predictive performance, Random Forest was preferred for the recommendation system due to its **greater sensitivity in identifying potential returning donors**.

- This approach illustrates how machine learning models can be effectively integrated into **decision-support systems**, enabling data-driven and personalised interventions in healthcare-related domains.

```
{'predicted_class': 1,
 'probability_of_donation': 0.969413568407454 3,
 'recommendation': 'The model predicts a HIGH probability that this person will donate blood again. Encourage regular donation and provide reminders.'}
```

SECTION 10: CONCLUSION AND FUTURE WORK

Conclusion

- This project demonstrated the application of machine learning techniques to predict future blood donation behaviour and generate personalised recommendations.
- Using donation history features such as Recency, Frequency, Monetary, and Time, predictive models were developed and evaluated.
- Exploratory data analysis revealed strong behavioural patterns, particularly the influence of recent and frequent donations on future donor activity.
- Among the baseline models, **Random Forest** provided the best balance between recall and overall performance, making it suitable for recommendation-based decision making.
- Feature importance analysis further supported the relevance of behavioural donation history in predicting donor return.
- The final recommendation system successfully translated model predictions into actionable engagement strategies, highlighting the practical value of data-driven approaches in healthcare-related domains.

Future Work

- Model performance could be further improved by exploring advanced ensemble techniques such as **Gradient Boosting**, **XGBoost**, or **LightGBM**, combined with careful hyperparameter tuning.
- Techniques for improved handling of class imbalance, such as **threshold tuning** or **synthetic oversampling methods (e.g., SMOTE)**, may enhance recall for returning donors.
- Feature engineering, including derived variables such as donation rate or recent-donor indicators, could provide additional predictive power.
- Incorporating demographic or health-related attributes would allow for more comprehensive and clinically relevant personalised recommendations.
- Deploying the recommendation system as a web-based application or dashboard could support real-time decision making and operational use.