Model Comparison Report

1. Introduction

Blood transfusions are critical in saving lives, addressing blood loss during surgery, injuries, and managing various illnesses. However, maintaining a sufficient blood supply is a consistent challenge. Predictive models can assist in identifying potential repeat donors, aiding blood transfusion services in proactive planning and improved outreach.

This report presents an analysis of data from a mobile blood donation vehicle in Taiwan, focusing on predicting donor behaviour during a blood drive in March 2007. The ultimate goal is to identify the best-performing model for practical implementation.

2. Dataset Description

The dataset originates from a mobile blood donation service operating across universities in Taiwan. Key details include:

- **Features:**

1. Recency: Months since the last donation.
2. Frequency: Total number of donations.
3. Monetary: Total blood donated in c.c.
4. Time: Months since the first donation.

- **Label:**
  Binary variable indicating whether the individual donated blood in March 2007 (Yes = 1, No = 0).

- **Preprocessing Steps:**

- Normalization of values for consistent scaling.

- Differentiation between features and labels.
- Addressing class imbalance through oversampling techniques to improve model performance for the minority class (donors).

3. Model Evaluation

Six machine learning models were trained, tested, and evaluated using accuracy and F1-scores for both classes (Class 0: Non-donors, Class 1: Donors).

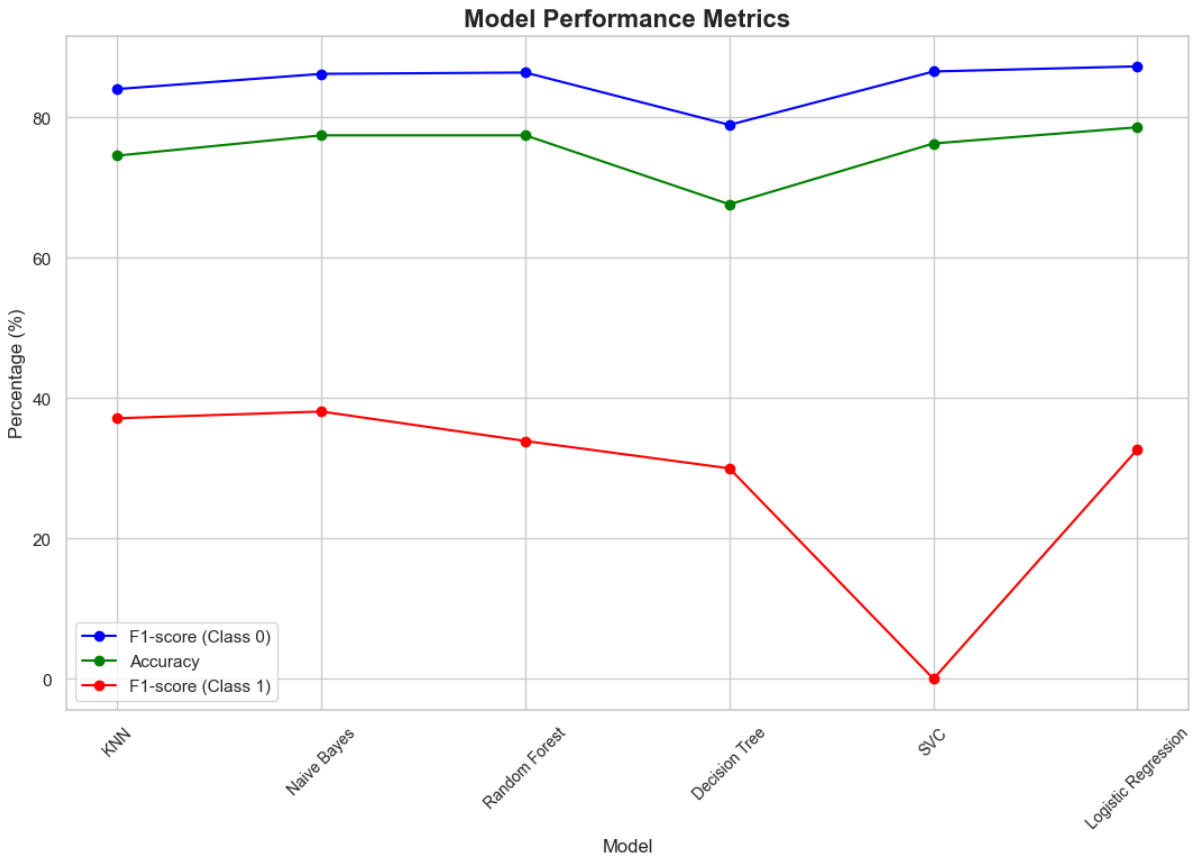
Initial Results:

Model	Accuracy (%)	F1-Score (Class 0)	F1-Score (Class 1)
K-Nearest Neighbours	74.57	84.06	37.14
Naive Bayes	77.46	86.22	38.10
Random Forest	77.46	86.41	33.90
Decision Tree	67.63	78.95	30.00
Support Vector Classifier	76.30	86.56	0.00
Logistic Regression	78.61	87.29	32.73

Analysis:

- Logistic Regression achieved the highest accuracy (78.61%) in the initial evaluation, followed by Random Forest and Naive Bayes (77.46% each).
- F1-scores for Class 1 (donors) were relatively low across models, indicating class imbalance challenges.

The plot shown below is a line plot, where performance metrics (F1-score for class 0, F1-score for class 1, and accuracy) are compared across different machine learning models (e.g., LR, KNN, RF, SVC, NB). Each line represents the trend of a specific metric across the models.



Insights:

- The F1-score for Class 1 (donors) is consistently lower across all models, with SVC scoring 0.0, highlighting significant challenges in predicting the minority class despite achieving reasonable accuracy.

- Logistic Regression achieves the highest accuracy and balanced F1-scores for Class 0 and Class 1, making it the most reliable model for predicting blood donation behaviour.

- While SVC achieves a high F1-score for Class 0 (non-donors), its F1-score for Class 1 drops to 0.0, indicating it completely fails to predict donors and is unsuitable for this imbalanced dataset.

4. Model Optimization

Top 3 models based on accuracy Random Forest, Logistic Regression, and Naive Bayes were optimized using Grid Search for hyperparameter tuning.

Optimized Results:

Model	Accuracy (%)	Remarks
Random Forest	79.15	Improved after Grid Search
Logistic Regression	79.19	Best performance after optimization
Naive Bayes	77.46	No improvement despite optimization

Hyperparameters:

1. Random Forest: Number of estimators, maximum depth, and minimum samples split were tuned.
2. Logistic Regression: Regularization strength (C) and solver parameters were adjusted.
3. Naive Bayes: No significant tuneable parameters resulted in improvements.

5. Final Model Selection

Logistic Regression is identified as the best-performing model due to:

- Highest accuracy (79.19%) post-optimization.
- Consistency in handling class imbalance after preprocessing.
- Simplicity and ease of interpretation for practical use in blood donation prediction.

Additionally, its lightweight implementation and explainability make it suitable for deployment in real-world scenarios.

6. Implementation

The Logistic Regression model was saved as a Pickle file for easy reuse. It can help blood donation services:

- Predict potential repeat donors during future drives.
- Plan targeted outreach campaigns.
- Enhance donor retention strategies with limited attributes.

7. Conclusion

This analysis highlights the importance of leveraging machine learning to improve blood donation planning. Logistic Regression outperformed other models after tuning and is recommended for production use. Future improvements could involve incorporating more features or exploring advanced algorithms like ensemble methods.