FINAL REPORT ON TRAINEESHIP-2025 ON

PREDICT BLOOD DONATION PROJECT - REPORT

28th May 2025

INTRODUCTION:

Blood is an essential and often limited resource in healthcare. Because blood donations fluctuate especially during holidays and busy seasons accurately predicting future donations can greatly improve planning and potentially save lives.

In this project, our goal is to predict whether a donor will give blood in the future based on historical donation data. To achieve this, we will:

- Explore and understand the structure and distribution of the data.
- Use automated machine learning through the TPOT library to find the best-performing model pipeline.
- Train and compare a traditional Logistic Regression model against TPOT's output.

- Analyze which model performs best and how data preprocessing impacts performance.

By the end of this project, we aim to deliver a predictive model that is both accurate and interpretable valuable qualities in a healthcare forecasting context.

PROJECT DESCRIPTION:

The dataset used in this project is sourced from the Blood Transfusion Service Center and contains 748 donor records. Each record includes:

- Recency (months since last donation)
- Frequency (total number of donations)
- Monetary (total blood donated in c.c.)
- Time (months since first donation)
- Whether the donor donated in March 2007 (target variable)

We explore the entire machine learning pipeline — from initial data inspection to automated model selection using TPOT and evaluation using the AUC score. The project also includes feature normalization and traditional logistic regression modeling.

OBJECTIVES:

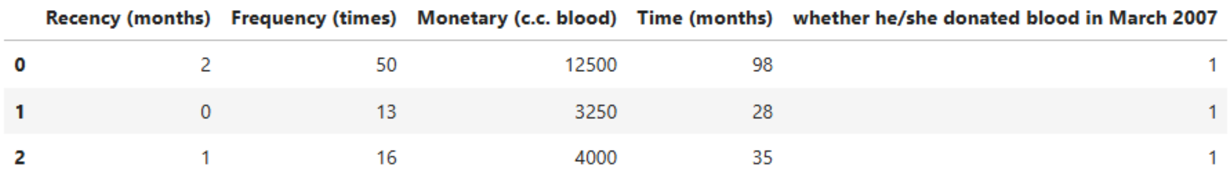
- Predict if a donor will donate again within a time window.
- Evaluate models using the AUC (Area Under Curve) metric.
- Compare traditional and AutoML (TPOT) approaches.
- Improve model performance via normalization.

DATA EXPLORATION:

The first step in building an effective machine learning model is to thoroughly understand and preprocess the data. In this project, we began by loading the dataset using pandas.read_csv(). The dataset, sourced from the UCI Machine Learning Repository, contains 748 entries with five features related to donor behavior. Upon initial inspection, we confirmed that there were no missing values in the dataset, which is crucial for ensuring the integrity of our machine learning pipeline.

To improve readability and streamline future operations, the target column named *"whether he/she donated blood in March 2007"* was renamed simply to target. We then analyzed the class distribution of the target variable using the value_counts() method with normalization. The results showed a class imbalance, with 76% of donors classified as Class 0 (did not donate) and 24% as Class 1 (did donate). This imbalance justified the use of evaluation metrics such as the AUC score instead of simple accuracy, as the latter could be misleading in skewed datasets.

Following this, we performed a train-test split using scikit-learn’s train_test_split() function. Stratification was applied based on the target variable to ensure that both training and testing datasets preserved the original class distribution. This careful attention to stratification helps prevent biased training and ensures fair model evaluation. These preprocessing steps laid the groundwork for reliable and meaningful modeling in the subsequent stages of the project.



	Recency (months)	Frequency (times)	Monetary (c.c. blood)	Time (months)	whether he/she donated blood in March 2007
0	2	50	12500	98	1
1	0	13	3250	28	1
2	1	16	4000	35	1

MODEL EVALUATION:

Model evaluation helps determine how effectively our machine learning model predicts outcomes and guides improvements. In this project, we evaluated two models: an automated pipeline using TPOTClassifier and a manually implemented Logistic Regression model. To measure performance, we used the AUC score, which is especially useful for imbalanced datasets. In our case, only 24% of the donors had donated blood within the target timeframe.

The TPOTClassifier explored multiple pipelines and selected a Logistic Regression model as the best performer. This model achieved an AUC score of 0.7850, establishing our initial benchmark. On further analysis, we found that the *Monetary (c.c. blood)* feature had disproportionately high variance compared to other features. To correct this, we applied log normalization, transforming the feature to reduce variance skew and improve model learning.

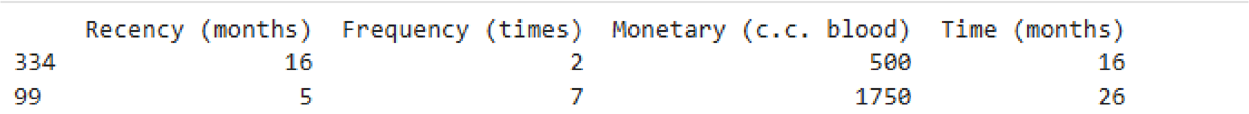
After normalization, we retrained the Logistic Regression model, achieving an improved AUC score of 0.7899. Though the gain may seem small, even minor improvements can be significant in healthcare applications. This evaluation confirms that logistic regression is both effective and interpretable for this problem, and that proper preprocessing such as variance correction can enhance model performance.

MODEL COMPARISON:

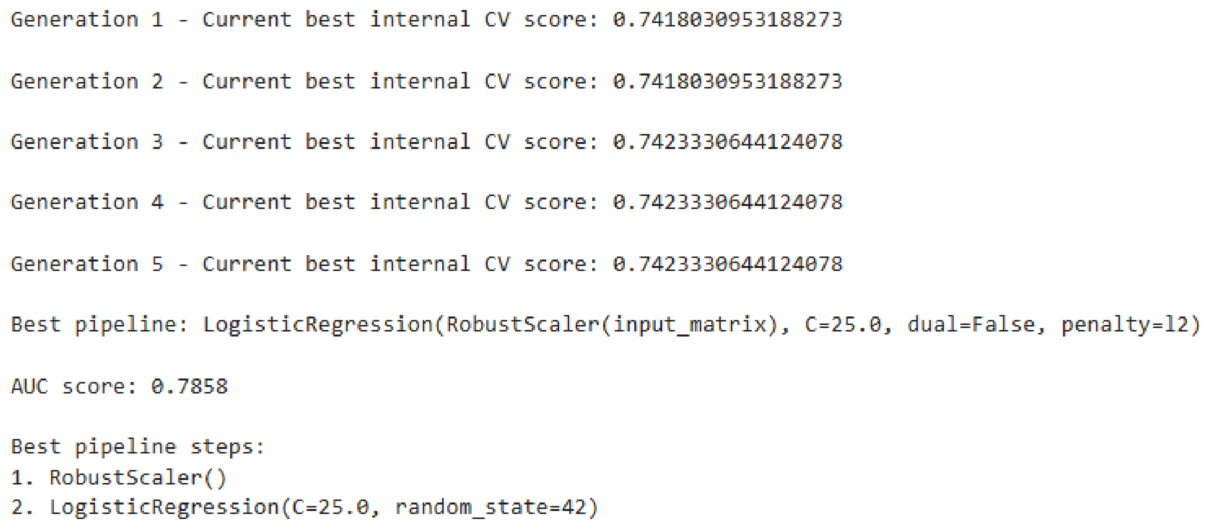
In this project, we compared two machine learning approaches for predicting whether a blood donor would donate within a specific time frame: an automated pipeline via TPOTClassifier and a manually implemented Logistic Regression model. TPOT is an AutoML tool that uses genetic programming to explore and optimize machine learning pipelines. It automatically tested various models and preprocessing combinations and selected Logistic Regression as the best-performing pipeline with an AUC score of 0.7850.

To further refine performance, we manually analyzed the dataset's variance and discovered that the Monetary (c.c. blood) feature had significantly higher variance than the others. This could bias the model. After applying log normalization to this feature and retraining the Logistic Regression model, we achieved a slightly improved AUC score of 0.7899.

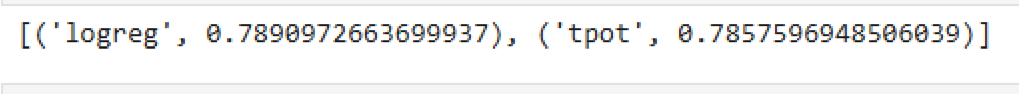
While the TPOTClassifier was useful in identifying a strong baseline model, manual intervention—particularly data normalization—helped marginally boost performance. This comparison highlights that while AutoML tools like TPOT accelerate the modeling process and provide strong baselines, combining them with domain knowledge and manual data preprocessing can further improve results. Moreover, logistic regression proves to be both performant and interpretable for this medical prediction task.



```
     Recency (months)  Frequency (times)  Monetary (c.c. blood)  Time (months)
334                16                  2                    500             16
99                  5                  7                   1750             26
```



```
Generation 1 - Current best internal CV score: 0.7418030953188273

Generation 2 - Current best internal CV score: 0.7418030953188273

Generation 3 - Current best internal CV score: 0.7423330644124078

Generation 4 - Current best internal CV score: 0.7423330644124078

Generation 5 - Current best internal CV score: 0.7423330644124078

Best pipeline: LogisticRegression(RobustScaler(input_matrix), C=25.0, dual=False, penalty=l2)

AUC score: 0.7858

Best pipeline steps:
1. RobustScaler()
2. LogisticRegression(C=25.0, random_state=42)
```



```
[('logreg', 0.7890972663699937), ('tpot', 0.7857596948506039)]
```

KEY INSIGHTS:

This project demonstrated the value of combining AutoML tools with manual preprocessing to enhance model performance. TPOT identified logistic regression as the optimal model with a solid AUC score of 0.7850. A deeper analysis revealed that log normalization of a high-variance feature further improved the model's AUC to 0.7899. Even a small improvement is meaningful in healthcare, where predictive accuracy can impact lives. Additionally, the simplicity and interpretability of logistic regression make it well-suited for medical applications, offering not just predictive power but also clarity on how features influence the prediction.

CONCLUSION:

This project successfully demonstrated how machine learning can be applied to predict future blood donations using a real-world healthcare dataset. Through data exploration, normalization, and model comparison, we built and evaluated predictive models to assist blood banks in planning ahead. TPOT provided a strong baseline using logistic regression, while manual log normalization further improved the model’s performance. The final AUC score of 0.7899 indicates a reliable model that balances accuracy and interpretability. These insights can help healthcare providers optimize donor outreach and inventory planning, ultimately contributing to more efficient and life-saving blood donation strategies.

FUTURE SCOPE:

The current blood donation prediction model lays a strong foundation for practical applications in healthcare logistics and donor management. However, there are several avenues to enhance its performance and broaden its impact.

Firstly, integrating more diverse features can significantly improve prediction accuracy. The current dataset includes limited historical donation behavior, but adding variables such as donor age, gender, blood type, geographic location, and lifestyle indicators (e.g., health history or occupation) could enrich the model's understanding. Time-series data capturing donation patterns over longer periods would also allow for trend-based forecasting, rather than relying solely on recent behaviors.

Secondly, advanced machine learning models can be explored. While logistic regression provides interpretability, models such as Gradient Boosting Machines (GBM), XGBoost, or neural networks could potentially capture more complex, non-linear relationships in the data. Ensemble models that combine the strengths of multiple algorithms may further improve predictive power.

Additionally, integrating this model into a real-time system could help blood banks forecast supply and demand dynamically. Coupling it with a dashboard can assist decision-makers with live updates on donor probability scores, making outreach campaigns more targeted and timely.

From a deployment standpoint, the model could be integrated into web or mobile applications used by healthcare centers. This would help them automate scheduling reminders for likely donors and efficiently allocate resources during high-demand periods like holidays or natural disasters.

Finally, ethical considerations such as data privacy and model fairness must be continuously evaluated, especially when dealing with sensitive health data. Ensuring compliance with healthcare regulations (e.g., HIPAA or GDPR) is essential for responsible deployment.

In summary, expanding the dataset, exploring advanced models, operationalizing the solution, and ensuring ethical compliance can significantly enhance the model’s future potential, making it a valuable asset in public health infrastructure.