Final Report of Internship Program

On

“PREDICT BLOOD DONATIONS”

MEDTOUREASY, NEW DELHI

Date : 22th August,25

By: Tanmoy Saha

ACKNOWLEDGEMENT

The internship opportunity that I had with MedTourEasy was a great change for learning and understanding the intricacies in the field of Data Analytics in Data Science; and also, for personal as well as professional development. I am very obliged for having a chance to interact with so many professionals who guided me throughout the internship project and made it a great learning curve for me.

Firstly, I express my deepest gratitude and special thanks to the Training Head of MedTourEasy, Mr. Ankit Hasija who gave me an opportunity to carry out my internship at their esteemed organization. Also, I express my thanks to him for making me understand the details of the Data Analytics profile and training me in the same so that I can carry out the project properly and with maximum client satisfaction and also for spearing his valuable time in spite of his busy schedule.

I would also like to thank the team of MedTourEasy who made the working environment productive and very conducive.

TABLE OF CONTENTS

INTRODUCTION

1.1 About the Company:

MedTourEasy, an online medical tourism marketplace, provides you the informational resources needed to evaluate your global options. It helps you find the right healthcare solution based on your specific health needs, affordable care, while meeting the quality standards that you expect to have in healthcare.

MedTourEasy, a global healthcare company, provides you the informational resources needed to evaluate your global options. It helps you find the right healthcare solution based on specific health needs, affordable care while meeting the quality standards that you expect to have in healthcare.

MedTourEasy improves access to healthcare for people everywhere. It is an easy to use platform and service that helps patients to get medical second opinions and to schedule affordable, high-quality medical treatment abroad.

MedTourEasy commitment to quality and transparency in healthcare is core to our mission. At MedTourEasy, we integrate the same three factors that physicians themselves agree are most important when selecting or referring a healthcare provider (patient satisfaction, experience match, and the quality of the hospital where a physician provides care.

MedTourEasy aspires to be the leader in making information on physicians and hospitals more accessible and transparent. The purpose is to give people the confidence to make the right healthcare decisions.

MedTourEasy's mission is to provide access to quality healthcare for everyone, regardless of location, time frame, or budget. Patients can connect with internationally-accredited clinics and hospitals.

1.2 About the Project:

The project titled **"Predict Blood Donations"** focuses on addressing one of the most critical challenges in healthcare: **forecasting blood supply**. Blood transfusions save lives—whether by replacing lost blood during surgery or injury, or by treating illnesses and blood disorders. However, ensuring that blood is always available when needed is a serious and recurrent problem faced by blood collection managers.

The **demand for blood fluctuates throughout the year**; for example, blood donations typically slow down during busy holiday seasons. This makes accurate forecasting essential for taking proactive measures to maintain sufficient supply and ultimately save more lives.

To tackle this problem, the project leverages **data-driven modeling** using the Blood Transfusion Dataset. The workflow involves:

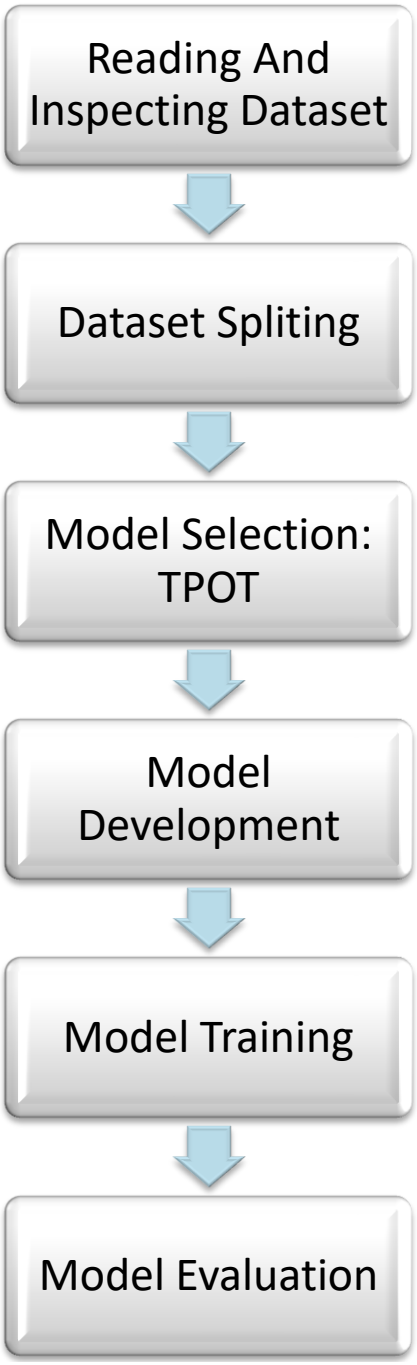
- **Loading and inspecting** the dataset
- **Defining features and target variables**
- **Splitting the dataset** into training and testing sets
- **Automated model selection** using **TPOT** (Tree-based Pipeline Optimization Tool)
- **Building and training** machine learning models
- **Evaluating model performance** using appropriate metrics

By applying machine learning techniques, this project aims to **predict future blood donations**, enabling health professionals to plan collection drives effectively and reduce the risk of shortages.

METHODOLOGY

2.1 Flow of the Project

The flow of the project can be understood via the following diagram:-



2.1 Language and Platform Used

Language : Python

What is Python?

Python is a programming language that is widely used in web applications, software development, data science, and machine learning (ML). Developers use Python because it is efficient and easy to learn and can run on many different platforms. Python software is free to download, integrates well with all types of systems, and increases development speed.

Benefits of Python include:

- Developers can easily read and understand a Python program because it has basic, English-like syntax.
- Python makes developers more productive because they can write a Python program using fewer lines of code compared to many other languages.
- Python has a large standard library that contains reusable codes for almost any task. As a result, developers do not have to write code from scratch.
- Developers can easily use Python with other popular programming languages such as Java, C, and C++.
- The active Python community includes millions of supportive developers around the globe. If you face an issue, you can get quick support from the community.
- Plenty of helpful resources are available on the internet if you want to learn Python. For example, you can easily find videos, tutorials, documentation, and developer guides.
- Python is portable across different computer operating systems such as Windows, macOS, Linux, and Unix.

Data science and machine learning:

Data science is extracting valuable knowledge from data, and machine learning (ML) teaches computers to automatically learn from the data and make accurate predictions. Data scientists use Python for data science tasks such as the following:

- Fixing and removing incorrect data, which is known as data cleaning
- Extracting and selecting various features of data
- Data labeling, which is adding meaningful names for the data
- Finding different statistics from data
- Visualizing data by using charts and graphs such as line charts, bar graphs, histograms, and pie charts

Data scientists use Python ML libraries to train ML models and build classifiers that accurately

classify data. People in different fields use Python-based classifiers to do classification tasks such as image, text, and network traffic classification; speech recognition; and facial recognition. Data scientists also use Python for deep learning, an advanced ML technique

What are Python libraries?

Python libraries are collections of pre-written code that developers use to avoid writing everything from scratch. Python comes with a **Standard Library**, and there are over **137,000 external libraries** for tasks like web development, data science, and machine learning.

Popular Libraries:

- **NumPy** → Arrays, math ops, linear algebra
- **Pandas** → Data handling, cleaning, grouping
- **Matplotlib** → Basic plots (line, bar, scatter)
- **Seaborn** → Advanced statistical visualization
- **Scikit-learn** → ML models (regression, classification, clustering, metrics)
- **Statsmodels** → Statistical tests, regression analysis
- **TensorFlow / Keras** → Deep learning, neural networks
- **OpenCV** → Image/video processing
- **Requests** → HTTP requests, web APIs
- **NLTK / SpaCy** → Natural Language Processing
- **XGBoost / LightGBM** → Gradient boosting ML models
- **Plotly** → Interactive plots & dashboards
- **Streamlit** → Data apps and dashboards
- **PyTorch** → Deep learning research/production

.

Machine Learning

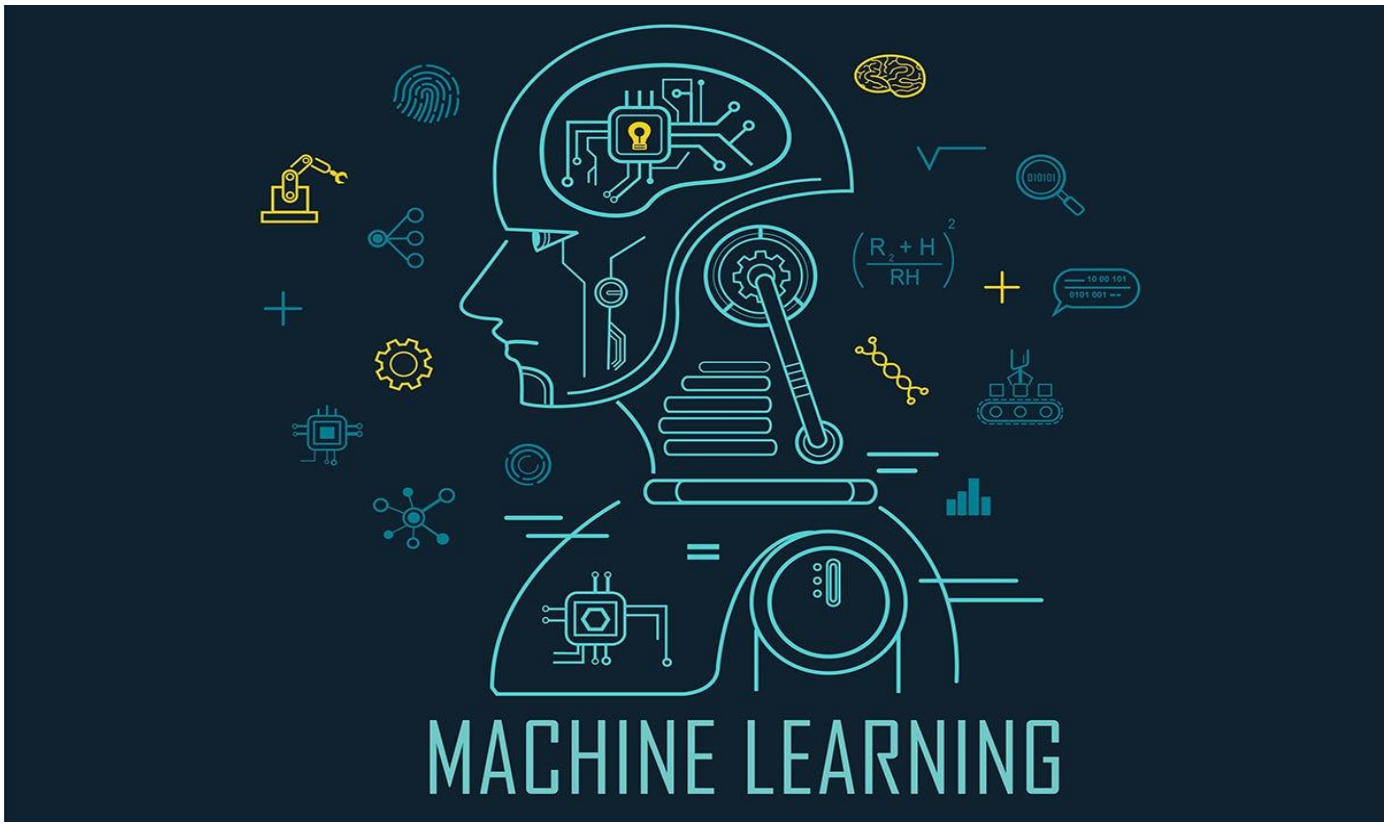
According to Arthur Samuel, *"Machine Learning algorithms enable computers to learn from data, and even improve themselves, without being explicitly programmed."*

Machine Learning (ML) is a subfield of Artificial Intelligence (AI) that focuses on developing algorithms and models that can analyze data, identify patterns, and make decisions with minimal human intervention. Instead of following strict rule-based programming, ML systems learn from experience—meaning, they improve their performance as they are exposed to more data over time.

The core idea behind ML is to create models that can **generalize** from past observations and make accurate predictions on unseen data. For example, ML is widely used in applications such as medical diagnosis, fraud detection, recommendation systems (like Netflix or Amazon), voice recognition, and self-driving cars.

Some key characteristics of ML are:

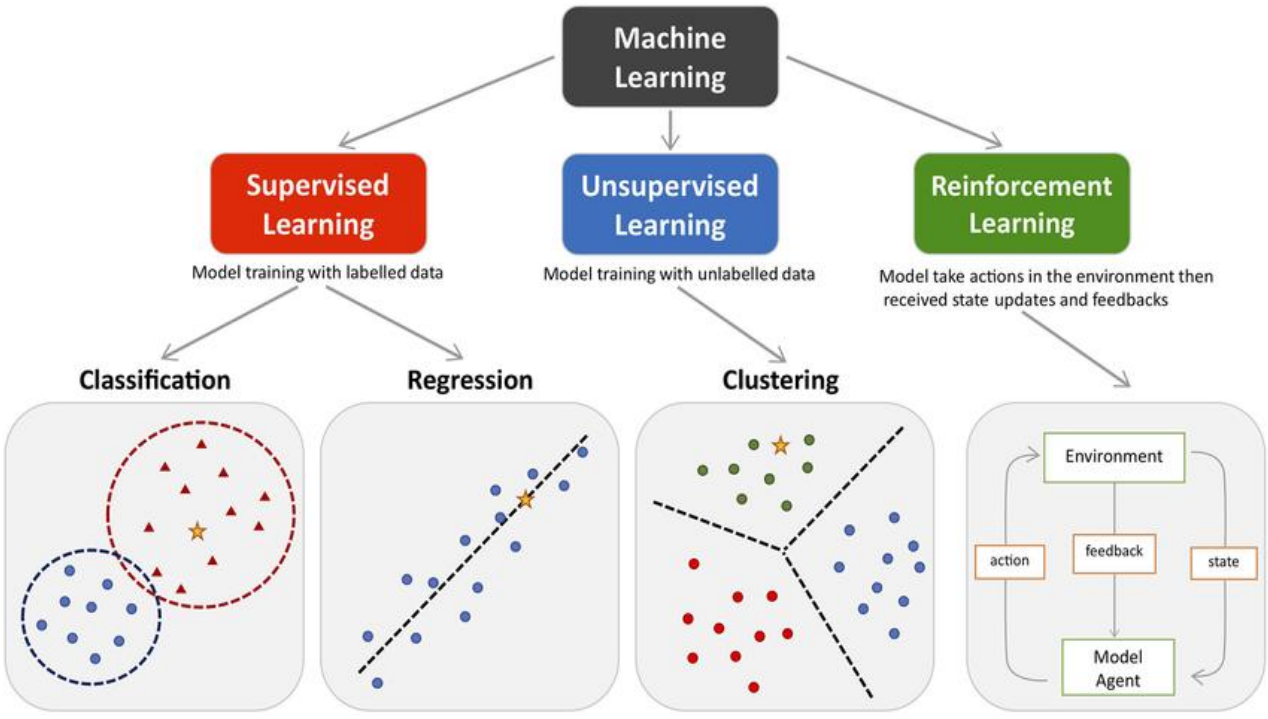
- It enables **automation of decision-making** processes.
- It adapts dynamically as new data becomes available.
- It relies on **mathematical and statistical methods** for pattern recognition.
- It improves **prediction accuracy** over time.



Types of Machine Learning:

Machine Learning can be broadly classified into three categories:

1. **Supervised Learning**
   - In supervised learning, the model is trained using a labeled dataset, meaning the input data is paired with the correct output.
   - The system learns the mapping function from input to output so it can predict results on new unseen data.
   - Examples: Predicting house prices, spam email detection.
2. **Unsupervised Learning**
   - In unsupervised learning, the data is unlabeled, and the model tries to find hidden structures or patterns.
   - It is mainly used for clustering and dimensionality reduction.
   - Examples: Customer segmentation, market basket analysis.
3. **Reinforcement Learning**
   - In reinforcement learning, an agent learns by interacting with an environment and receiving feedback in the form of rewards or penalties.
   - It is used in areas such as robotics, game playing, and resource management.
   - Example: Training an AI to play chess or control a robot.



Platform: Google Colab Notebook

Google Colab (Colaboratory) is a cloud-based Jupyter Notebook environment provided by Google. It allows users to write and execute Python code directly in their web browser without requiring any installation or configuration.

Some major advantages of Google Colab are:

- **Free GPU/TPU Access:** Users can leverage powerful hardware accelerators for machine learning and deep learning tasks without extra cost.
- **Cloud-Based Environment:** Since it runs on the cloud, there's no need to worry about local storage, processing power, or software installation.
- **Easy Collaboration:** Just like Google Docs, multiple users can share, view, and edit a Colab notebook in real-time.
- **Integration with Google Drive:** Notebooks and datasets can be directly saved and loaded from Google Drive.
- **Supports Rich Outputs:** Besides code, you can add text, equations, visualizations, and images, making it ideal for research and data analysis presentations.

Colab is widely used for:

- Data cleaning and preprocessing
- Statistical modeling
- Machine learning and deep learning model training
- Data visualization and storytelling
- Experimentation with Python libraries (NumPy, Pandas, Matplotlib, Scikit-learn, TensorFlow, PyTorch, etc.)

Thus, Google Colab provides a convenient and powerful environment for students, researchers, and professionals to explore and implement machine learning projects.

IMPLEMENTATION

3.1 Dataset Description

The dataset used in this project, **`transfusion.csv`**, has been obtained from the **UCI Machine Learning Repository**. It consists of a random sample of **748 blood donors**, collected by the *Blood Transfusion Service Center* in Taiwan. The data originates from a **mobile blood donation vehicle** that regularly visits different universities to conduct donation drives.

The objective of this dataset is to **predict whether a donor will give blood during the next visit** of the mobile donation vehicle. The dataset follows the **RFMTC model**, which is an extension of the traditional **RFM (Recency, Frequency, Monetary)** model widely used in marketing to identify loyal customers. In this context, the "customers" are blood donors.

The attributes of the dataset are:

- **R (Recency):** Number of months since the last donation.
- **F (Frequency):** Total number of donations made by the donor.
- **M (Monetary):** Total volume of blood donated (in c.c.).
- **T (Time):** Number of months since the donor's first donation.
- **Target Variable:** A binary outcome indicating whether the donor gave blood in **March 2007**.
    - `1` → Donated blood
    - `0` → Did not donate blood

This well-structured dataset provides an excellent foundation for applying **predictive modeling techniques** to solve a real-world healthcare challenge.

3.2 Statistical Insights of Dataset

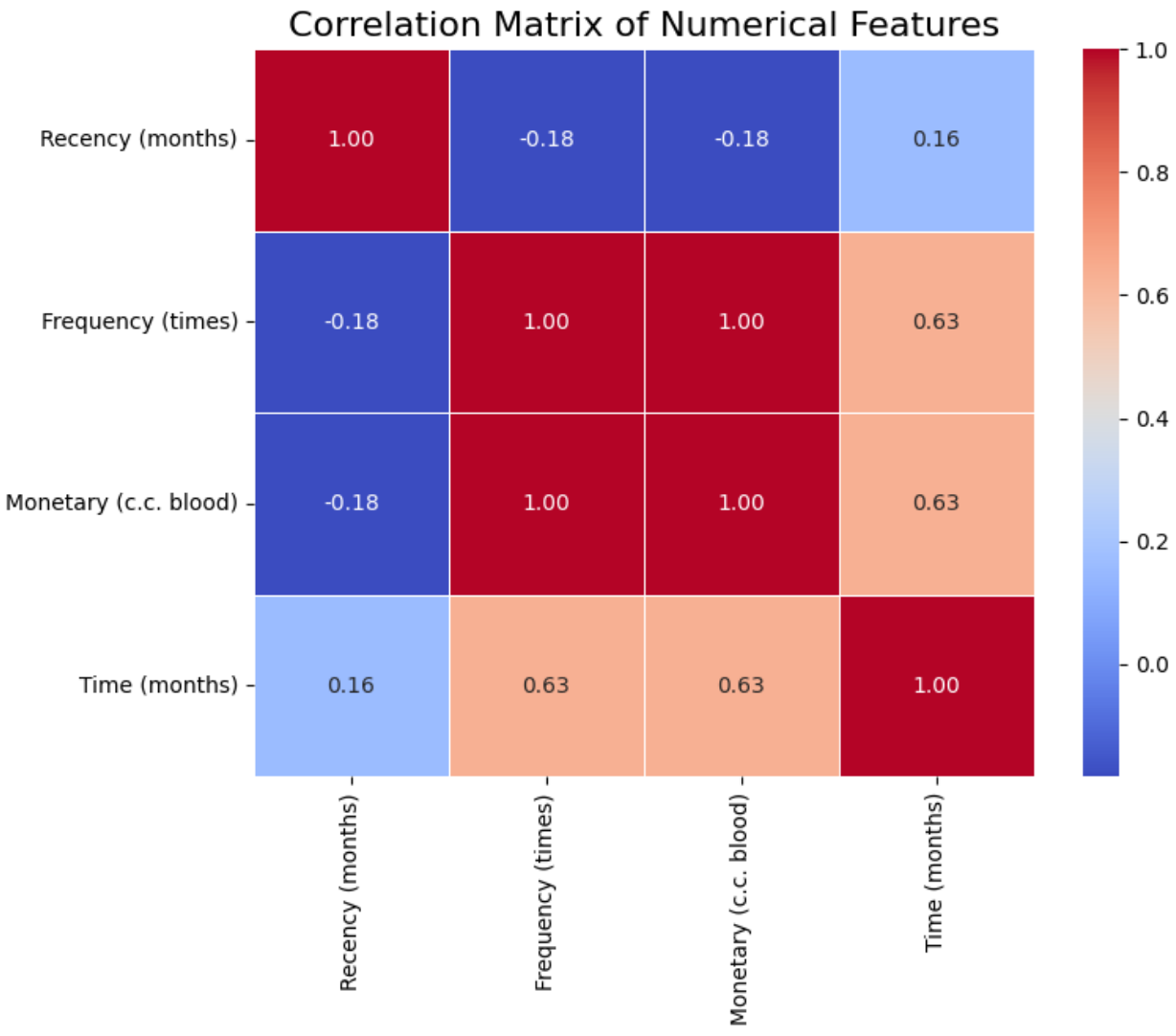
Like most datasets, `transfusion.csv` contains **hidden patterns** that can be uncovered through **statistical analysis and visualization**. Some of the key findings are:

- **Data Type:** All features in the dataset are **numerical**, which makes it suitable for statistical modeling and machine learning without requiring categorical encoding.
- **Target Distribution:** After analyzing the target variable, the class distribution was found to be imbalanced:
    - Class **0 (Did not donate in March 2007): ~76.2%**
    - Class **1 (Donated in March 2007): ~23.8%**

    This imbalance indicates that **most donors did not donate again**, which makes the prediction task more challenging.

- **Data Splitting Strategy:**
  To ensure a balanced evaluation, the dataset was split into training and testing sets using the following parameters:
  - **Test Size:** 25% of the dataset
  - **Random State:** 42 (for reproducibility)
  - **Stratified Sampling:** The target column was used for stratification to preserve the class distribution across train and test sets.
- **Feature Correlation:**
  A **correlation matrix** was computed to study the relationships among features. The heatmap revealed:
  - **High correlation between Frequency and Monetary value**, since the total blood donated directly depends on the number of donations.
  - **Recency and Time** show moderate correlation with other features, suggesting they capture different temporal aspects of donor behavior.
  - No severe multicollinearity was observed, making the dataset well-suited for predictive modeling.

These statistical insights provide an understanding of the dataset's structure, distribution, and relationships, forming the basis for building **robust machine learning models**.



3.3 Model Selection and Development

Selecting the most suitable algorithm and designing the entire machine learning pipeline is often a challenging and time-intensive task. To overcome this, we leverage **Automated Machine Learning (AutoML)**, which systematically evaluates multiple algorithms, preprocessing techniques, and hyperparameters to identify the best-performing pipeline.

For this study, we used **TPOT (Tree-Based Pipeline Optimization Tool)**, an AutoML library in Python. TPOT applies **genetic programming** to explore thousands of potential pipelines and optimizes them based on the dataset's characteristics. It automates repetitive tasks such as model selection, feature preprocessing, and hyperparameter tuning, significantly reducing development time.

TPOT is built on top of **scikit-learn**, ensuring that the generated pipelines are interpretable and follow widely accepted machine learning standards.

After optimization, TPOT selected **Logistic Regression** as the most suitable model for our dataset, requiring no additional preprocessing steps. The model achieved an **AUC score of 0.7850**, making it a strong candidate for our prediction task.

Based on these results, **Logistic Regression** was chosen for final model development and evaluation.

3.4 Model Training & Evaluation

Model training and evaluation are critical steps in the machine learning pipeline, as they determine how well the selected algorithm generalizes to unseen data. In this project, **Logistic Regression** was chosen (as suggested by TPOT) for model development. However, before training the model, it was important to ensure that the dataset features were properly scaled and normalized.

One of the key assumptions of linear models such as Logistic Regression is that the features provided should have relationships that can be expressed in a **linear fashion**, or at least measured using a **linear distance metric**. When features vary drastically in terms of scale or variance, the model may incorrectly assign higher importance to those features with greater variance.

In this dataset, the feature **Monetary (c.c. blood)** displayed a variance significantly higher compared to the other attributes. Without correction, this imbalance could bias the learning process and hinder model performance. To address this, **log normalization** was applied. Log normalization is a common preprocessing technique that reduces skewness and minimizes the impact of large values, thereby ensuring all features contribute more proportionately to the learning process.

Once preprocessing was completed, the **Logistic Regression model** was trained with the following hyperparameters:

- **Solver = liblinear** (a robust solver suitable for small datasets and binary classification)
- **Random State = 42** (for reproducibility of results)

After training, the model was evaluated using the **AUC (Area Under the ROC Curve) Score**, which measures the ability of the model to distinguish between the classes (donors who donate blood again vs. those who do not).

The trained model achieved an **AUC Score of 0.7891**, which indicates a good level of predictive performance and confirms that Logistic Regression is a suitable choice for this dataset.

The code for the project:

https://colab.research.google.com/drive/1HNG8bDrE7z5eO3fMkDi9iUQRbiNNPBKE?usp=sharing

CONCLUSION & FUTURE SCOPE

The demand for blood is highly dynamic and often fluctuates throughout the year. For instance, blood donations tend to decline during busy holiday seasons, creating potential shortages. An accurate forecast for future blood supply enables healthcare organizations to take proactive steps, thereby saving more lives.

In this project, we utilized **Google Colab** as our development environment, benefiting from its cloud-based infrastructure, pre-installed machine learning libraries, and GPU/TPU support. This simplified the model development pipeline, from **data cleaning and normalization** to **training, testing, and evaluation**.

We implemented **automatic model selection** using TPOT, where the best performing model achieved an AUC score of **0.7850**. While this was already better than a naïve model that always predicts '0' (which would yield ~76% success rate), further improvements were made. By applying **log normalization** to features with very high variance (e.g., Monetary variable), we enhanced the performance and achieved an improved AUC score of **0.7891**. In the field of machine learning, even a **0.5% increase in predictive power** is meaningful, especially in critical domains like healthcare.

The choice of **logistic regression** proved effective not only because of its predictive capability but also due to its **interpretability**. Unlike black-box models, logistic regression provides clear insights into how each feature contributes to predicting blood donation likelihood.

Future Scope

- **Feature Engineering:** More advanced feature transformations and interaction terms could be introduced to capture deeper relationships in the data.
- **Alternative Models:** Exploring non-linear models such as Random Forests, Gradient Boosting, or Neural Networks may lead to higher predictive performance.
- **Time-Series Forecasting:** Since blood donation trends vary seasonally, incorporating time-series models like ARIMA, LSTM, or Prophet could provide better long-term forecasts.
- **Donor Profiling & Campaigns:** Insights from the model can help design **personalized donor retention strategies** by identifying individuals more likely to donate again.
- **Integration with Hospital Systems:** Deploying the model in real-time systems can assist blood banks in managing supply and demand efficiently.

Ultimately, this project demonstrates how **machine learning, coupled with tools like Google Colab**, can support life-saving decisions in the healthcare sector.

REFERENCES

The following websites have been referred for input data and statistics:

- WebMD – Blood Transfusion Overview
- KJRH – Red Cross in Blood Donation Crisis
- NCBI – Blood Donation and Transfusion Guide
- TPOT Documentation – Epistasis Lab

The following websites have been referred for coding part:

- Python Official Documentation
- GitHub – Logistic Regression Implementation
- TPOT Documentation – Epistasis Lab