FINAL REPORT OF
TRAINEESHIP PROGRAM 2023

On

PREDICT BLOOD DONATIONS

MEDTOUREASY

26th July 2023

ACKNOWLEDGMENTS

The traineeship opportunity that I had with MedTourEasy was a great change for learning and understanding the intricacies of the subject of Data Visualizations in Data Analytics; 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 traineeship project and made it a great learning curve for me.

Firstly, I express my deepest gratitude and special thanks to the Training & Development Team of MedTourEasy who gave me an opportunity to carry out my traineeship at their esteemed organization. Also, I thank the team 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 despite his busy schedule.

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

TABLE OF CONTENTS

ABSTRACT

Human blood is undeniably crucial for sustaining life, and its importance cannot be overstated. In light of this significance, the present study aims to develop and evaluate machine learning algorithms to predict whether a person will donate blood in the future.

Blood transfusion is a life-saving medical procedure utilized in diverse scenarios, ranging from replenishing lost blood during major surgeries or severe injuries to treating various illnesses and blood-related disorders. The process of blood donation plays a pivotal role in maintaining an adequate blood supply, ensuring that it is readily available whenever required by health professionals.

The need for a sufficient and reliable blood supply poses a significant challenge for healthcare systems worldwide. According to WebMD, an estimated 5 million Americans require blood transfusions annually, highlighting the immense demand for blood products in just one country. This statistic underscores the critical nature of predicting blood donation behavior to bolster the availability of this precious resource.

With its ability to analyze vast amounts of data, machine learning offers promising prospects in identifying patterns and factors that influence individuals' decisions to donate blood. By developing accurate prediction models, healthcare providers and blood banks can effectively tailor their strategies to encourage and motivate potential donors. Such predictive systems could help optimize blood supply management, ensuring that enough blood is readily accessible whenever emergencies or medical treatments necessitate it.

INTRODUCTION

1.1 ABOUT THE COMPANY

MedTourEasy, a global healthcare company, provides you with the informational resources needed to evaluate your global options. MedTourEasy provides analytical solutions to our partner healthcare providers globally.

MedTourEasy is an online medical tourism marketplace that provides informational resources needed to evaluate global options for healthcare solutions based on specific health needs, and affordable care while meeting the quality standards that you expect to have in healthcare. They help patients find the right healthcare solution based on their specific health needs and budget while meeting the quality standards that they expect to have in healthcare. They connect patients with trusted healthcare providers and partner with internationally accredited institutions.

1.2 ABOUT THE PROJECT

“In 1628 British physician William Harvey discovers the circulation of blood. The first known blood transfusion is attempted soon afterward. 1658 In Microscopist Jan Swammerdam observes and describes red blood cells. In 1665 The first recorded successful blood transfusion occurs in England: Physician Richard Lower keeps a dog alive by transfusing blood from other dogs. In 1667 Jean-Baptiste Denis in France and Richard Lower and Edmund King in England separately report successful transfusions from sheep to humans.

In 1818 British obstetrician James Blundell performs the first successful transfusion of human blood to a patient for the treatment of postpartum hemorrhage. From 1873-1880 U.S. physicians attempt to transfuse milk from cows, goats, and humans. In 1884 Saline infusion replaces milk as a "blood substitute" due to the increased frequency of adverse reactions to milk. In 1901 Karl Landsteiner, an Austrian physician, discovers the first three human blood groups. In 1907 Ludvig Hektoen suggests that the safety of transfusion might be improved by cross-matching blood between donors and patients to exclude incompatible mixtures. Reuben Ottenberg performs the first blood transfusion using blood typing and cross-matching. In 1914 Long-term anticoagulants, among them sodium citrate, are developed, allowing longer preservation of blood. In 1939-1940 The Rh blood group system is discovered by Karl Landsteiner, Alexander Wiener, Philip Levine, and R.E. Stetson. In 1940 The U.S. government establishes a national blood collection program. Edwin Cohn develops cold ethanol fractionation, the process of breaking down plasma into components and products. Albumin, gamma globulin, and fibrinogen are isolated and become available for clinical use. John Elliott develops the first blood container, a vacuum bottle extensively used by the Red Cross. An early blood processing program for the relief of English war victims, called Plasma for Britain, begins under the direction of Charles R. Drew, MD. 1941 The Red Cross begins National Blood Donor Service to collect blood for the U.S. military with Dr. Charles R. Drew, formerly of the Plasma for Britain program, as medical director. Soldiers injured during the Pearl Harbor attack are treated with albumin for shock. 1944 Dried plasma becomes a vital element in the treatment of wounded soldiers during World War II. In 1945

The Red Cross ends its World War II blood program for the military after collecting more than 13 million pints. Robin Coombs, Arthur Mourant, and Rob Race describe the use of anti-human globulin to identify incomplete antibodies. The process became known as the Coombs test, also known as the antiglobulin test. In 1947 ABO blood-typing and syphilis testing is performed on each unit of blood. In 1948 The Red Cross begins the first nationwide blood program for civilians by opening its first collection center in Rochester, N.Y. In 1949 The U.S. blood system is comprised of 1,500 hospital blood banks, 46 community blood centers, and 31 American Red Cross regional blood centers. In 1950 Audrey Smith reports the use of glycerol cryoprotectant for red blood cells. The U.S. enters the Korean War. Red Cross becomes a blood collection agency for the military during Korean War.

In 1956 Establishment of National Blood Clearinghouse. In 1957 The American Association of Blood Banks forms its Committee on Inspection and Accreditation to monitor the implementation of standards for blood banking. In 1961 Platelet concentrates are recognized for reducing the mortality from hemorrhage in cancer patients. In 1964 Plasmapheresis is introduced as a means of collecting plasma for fractionation. In 1967 American National Red Cross Board of Governors receives a report that national headquarters will host a national Rare Blood Donor Registry for blood types occurring less than once in 200 people. In 1969 S. Murphy and F. Gardner demonstrate the feasibility of storing platelets at room temperature, revolutionizing platelet transfusion therapy. In 1970 U.S. blood banks moved toward an all-volunteer blood donor system. 1992 Testing of donor blood for HIV-1 and HIV-2 antibodies is implemented.

First National Testing Laboratory, applying standardized tests to ensure the safety of Red Cross blood products, opens in Dedham, Mass.2002 Nucleic acid amplification test (NAT) for HIV and hepatitis C virus.

1.3 PROBLEM STATEMENT

Humans can predict or analyze the probability of blood donations, but it is difficult. So, for the easiness of people, a model is required to predict future donations. The purpose of "Predict Blood Donations" is to predict the probability of blood donation.

OBJECTIVES

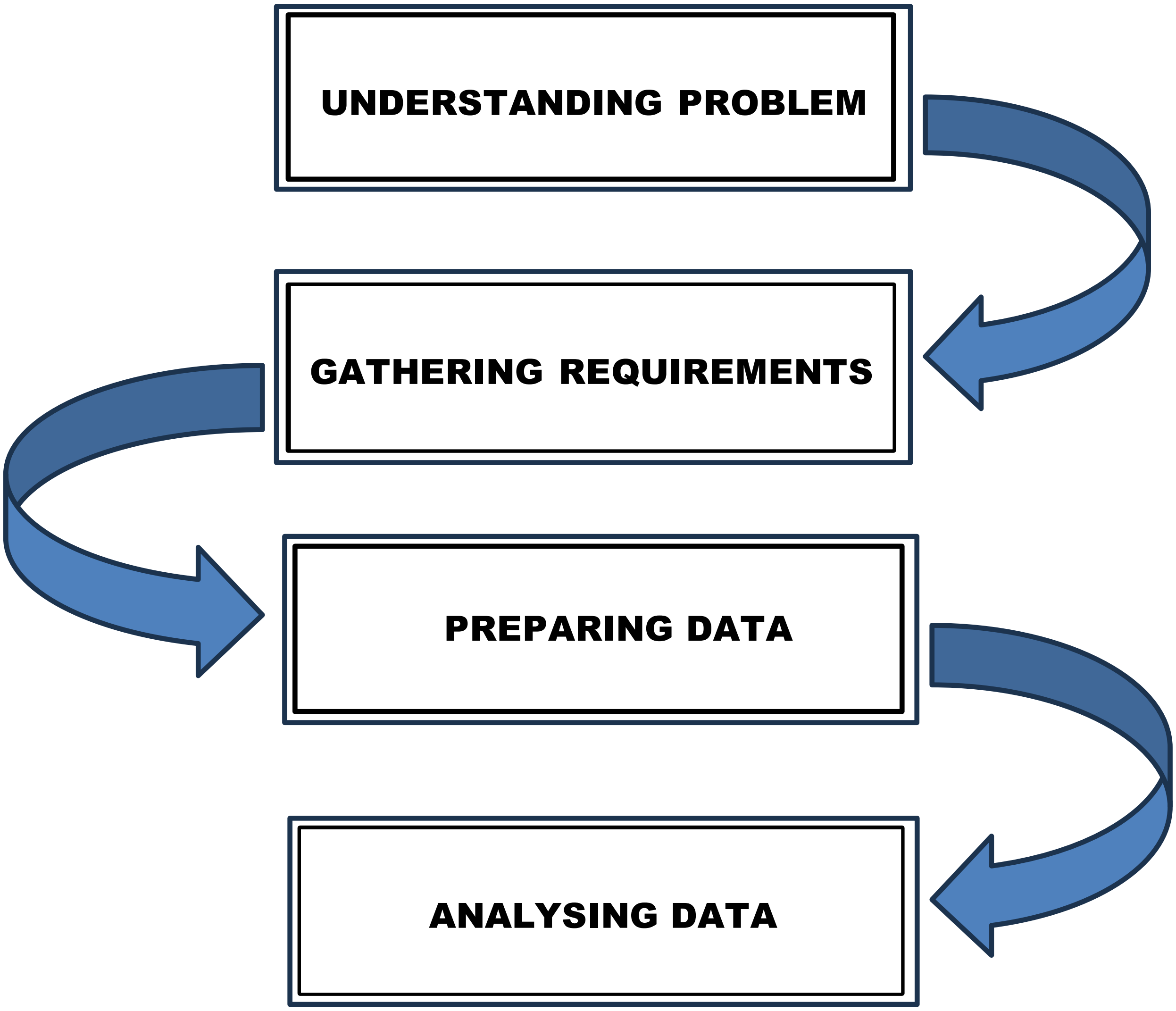
- To develop a model that is going to evaluate blood donations
- To avoid manual blood donation predictions
- To utilize the model in every blood donation camp and hospitals

METHODOLOGY

2.1 FLOW OF THE PROJECT

The flow model is used to work on the project Predict Blood Donations. The output of one stage in our model will serve as an input for the next step.

The phases of the flow model according to my project are as follows:



UNDERSTANDING PROBLEM

Understanding a problem refers to the ability to grasp the nature, intricacies, and implications of a particular situation, challenge, or question. It involves gaining insight into the key components, underlying factors, and possible solutions related to the problem at hand. It refers to the challenge of accurately comprehending the factors and patterns that influence an individual's decision to donate blood. This problem is often encountered in predictive analytics and data science projects aiming to forecast donation behavior or identify potential blood donors.

GATHERING REQUIREMENTS

Gathering requirements is a critical initial phase in the software development and project management process. It involves systematically collecting, analyzing, and documenting the needs and expectations of stakeholders for a particular software system, application, or project. The primary goal of gathering requirements is to understand what the end product should achieve and how it should function to meet the stakeholders' objectives.

PREPARING DATA

Preparing data is a crucial step in the data analysis process. It involves cleaning, transforming, and organizing the raw data into a structured format that can be easily analyzed. Proper data preparation ensures that the data is accurate, complete, and relevant for the analysis, leading to more reliable and meaningful insights.

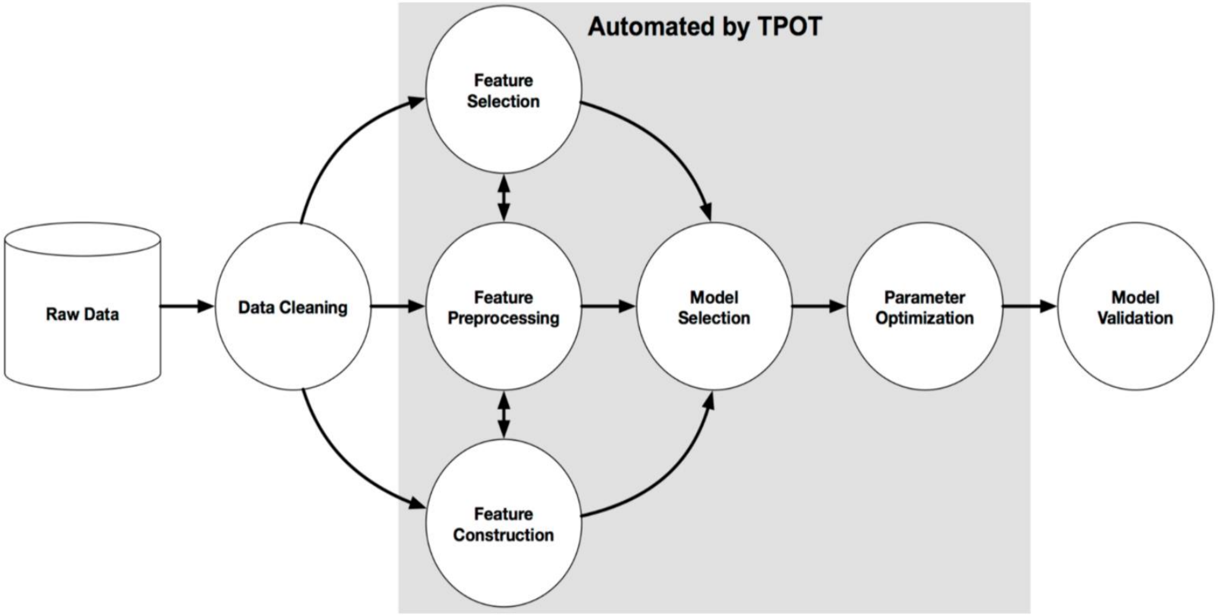
ANALYSING DATA

Analyzing data is a core component of data science projects. It involves applying various statistical, computational, and machine learning techniques to extract insights, patterns, and knowledge from the prepared data. Data analysis is an iterative process that helps data scientists understand the underlying trends, relationships, and dependencies within the data and draw meaningful conclusions

2.2 CORE METHODOLOGY

The project sets up a pipeline to dynamically configure and assess models to find the most accurate fit to predict if a previous donor would donate at a given instance.

The overall workflow of the project is structured as follows:



Data Source

The project uses data sourced from the Blood Transfusion Service Center in Taiwan. The dataset contains 847 blood donors, including the following information about each individual: - Months since last transfusion - Total number of donations - Total blood donated - Timespan since first donation Along with the above-mentioned information, the dataset also included weather each individual donated for a specific month (March 2007) - our target variable.

TPOT Pipeline

This project used the TPOTCLASSIFIER from the TPOT package, an automated pipeline exploring different algorithms and models; the ROC_AUC_SCORE from SKlearn was specified as the validation metric to compare models and choose the best fit based on.

2.3 LANGUAGE AND TOOLS USED

2.3.1 Language: **PYTHON**

Python is one of the most popular and widely used programming languages for data analysis and data science tasks. It offers a rich ecosystem of libraries and tools that make it an excellent choice for processing, manipulating, and visualizing data

Python has a simple and intuitive syntax, making it accessible to beginners and experienced programmers alike. The readability of Python code makes it easier to write and maintain data analysis scripts

Python has a vast array of open-source libraries dedicated to data analysis, such as NumPy (numerical computing), Pandas (data manipulation), Matplotlib (data visualization), SciPy (scientific computing), and Scikit-learn (machine learning). These libraries provide powerful and efficient tools for handling various aspects of data analysis.

LIBRARIES

NumPy

NumPy (Numerical Python) is a fundamental Python library for numerical computing. It provides support for multi-dimensional arrays and a collection of functions for performing efficient operations on these arrays. NumPy is a foundational library in the data science ecosystem and serves as the basis for many other data analysis and scientific computing libraries in Python.

PANDAS

Pandas is a powerful open-source Python library widely used for data manipulation, analysis, and preparation. It provides data structures and functions to efficiently work with structured data, such as tabular and time-series data. Pandas is built on top of NumPy and is a fundamental tool in the data science ecosystem, complementing NumPy by adding higher-level data manipulation capabilities

SCIKIT-LEARN

Scikit-learn, also known as sklearn, is a popular and widely used open-source Python machine-learning library. It is built on top of NumPy, SciPy, and Matplotlib and provides a user-friendly and efficient framework for various machine-learning tasks. Scikit-learn is designed to be simple and accessible while maintaining the flexibility to handle complex machine-learning problems.

TPOT

TPOT (Tree-based Pipeline Optimization Tool) is an open-source automated machine learning (AutoML) library in Python. It is designed to automatically search for and build the best machine learning pipelines for a given dataset, saving time and effort in the model selection and hyperparameter tuning process.

2.3.2 PLATFORM: **JUPYTER NOTEBOOK**

Jupyter Notebook is an interactive web-based computing environment that allows users to create and share documents containing live code, equations, visualizations, and narrative text. It is a popular tool among data scientists, researchers, and educators for data exploration, analysis, and communication of findings. It supports various languages, but it is primarily used with Python.

IMPLEMENTATION

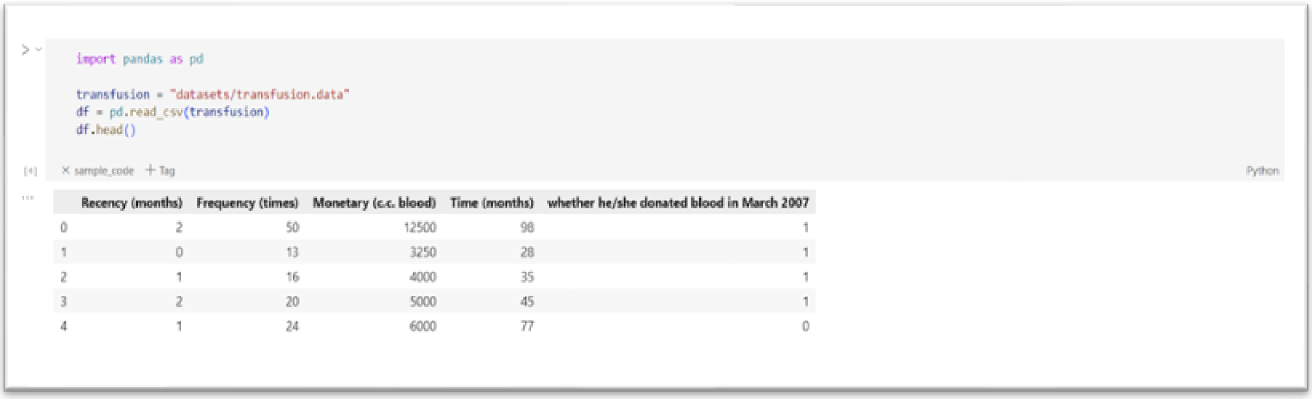
3.1 DATA PROCESSING AND BASIC EXPLORATION

This is the first step wherein the requirements are collected from the clients to understand the deliverables and goals to be achieved after which a problem statement is defined which has to be adhered to while developing the project

3.2 DATA COLLECTION AND IMPORTING

Data collection is a systematic approach for gathering and measuring information from a variety of sources in order to obtain a complete and accurate picture of an interesting area. It helps an individual or organization to address specific questions, determine outcomes and forecast future probabilities and patterns.

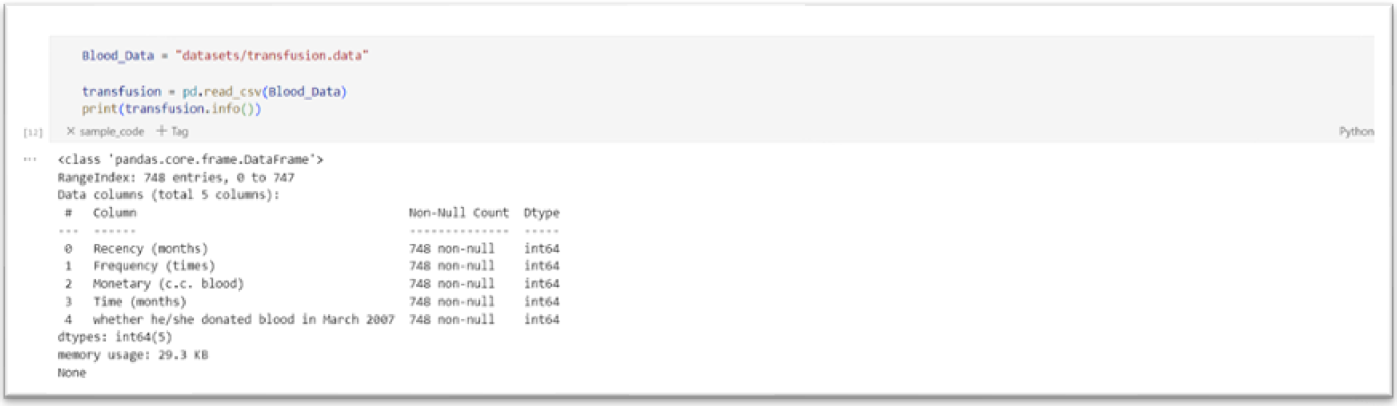
3.3 LOADING THE BLOOD DONATIONS DATA



```python
import pandas as pd

transfusion = "datasets/transfusion.data"
df = pd.read_csv(transfusion)
df.head()
```

```
   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
3                 2                 20                   5000             45                                           1
4                 1                 24                   6000             77                                           0
```

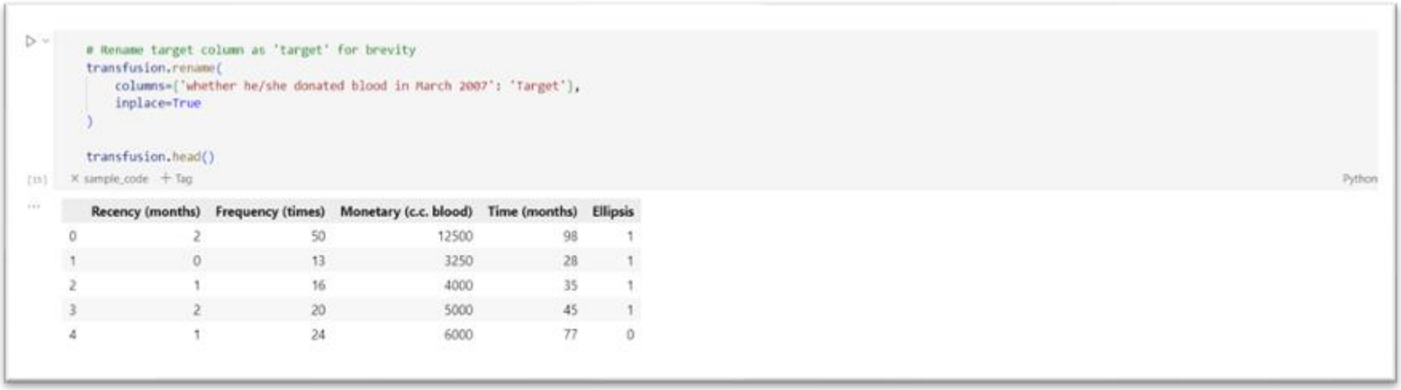
3.4 INSPECTING TRANSFUSION DATA FRAME



```
Blood_Data = "datasets/transfusion.data"

transfusion = pd.read_csv(Blood_Data)
print(transfusion.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 748 entries, 0 to 747
Data columns (total 5 columns):
 #   Column                                      Non-Null Count  Dtype
---  ------                                      --------------  -----
 0   Recency (months)                            748 non-null    int64
 1   Frequency (times)                           748 non-null    int64
 2   Monetary (c.c. blood)                       748 non-null    int64
 3   Time (months)                               748 non-null    int64
 4   whether he/she donated blood in March 2007  748 non-null    int64
dtypes: int64(5)
memory usage: 29.3 KB
None
```

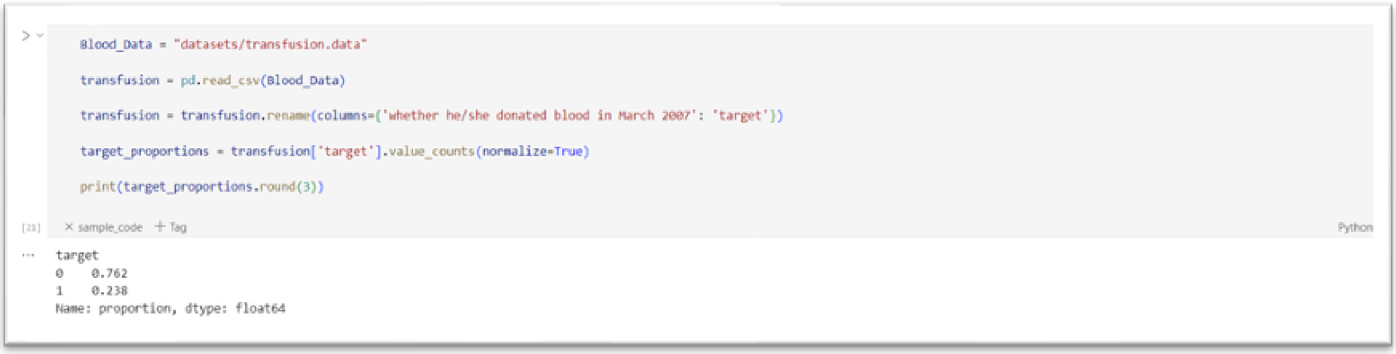
3.5 CREATING TARGET COLUMN



```
# Rename target column as 'target' for brevity
transfusion.rename(
    columns={'whether he/she donated blood in March 2007': 'Target'},
    inplace=True
)

transfusion.head()
```

	Recency (months)	Frequency (times)	Monetary (c.c. blood)	Time (months)	Ellipsis
0	2	50	12500	98	1
1	0	13	3250	28	1
2	1	16	4000	35	1
3	2	20	5000	45	1
4	1	24	6000	77	0

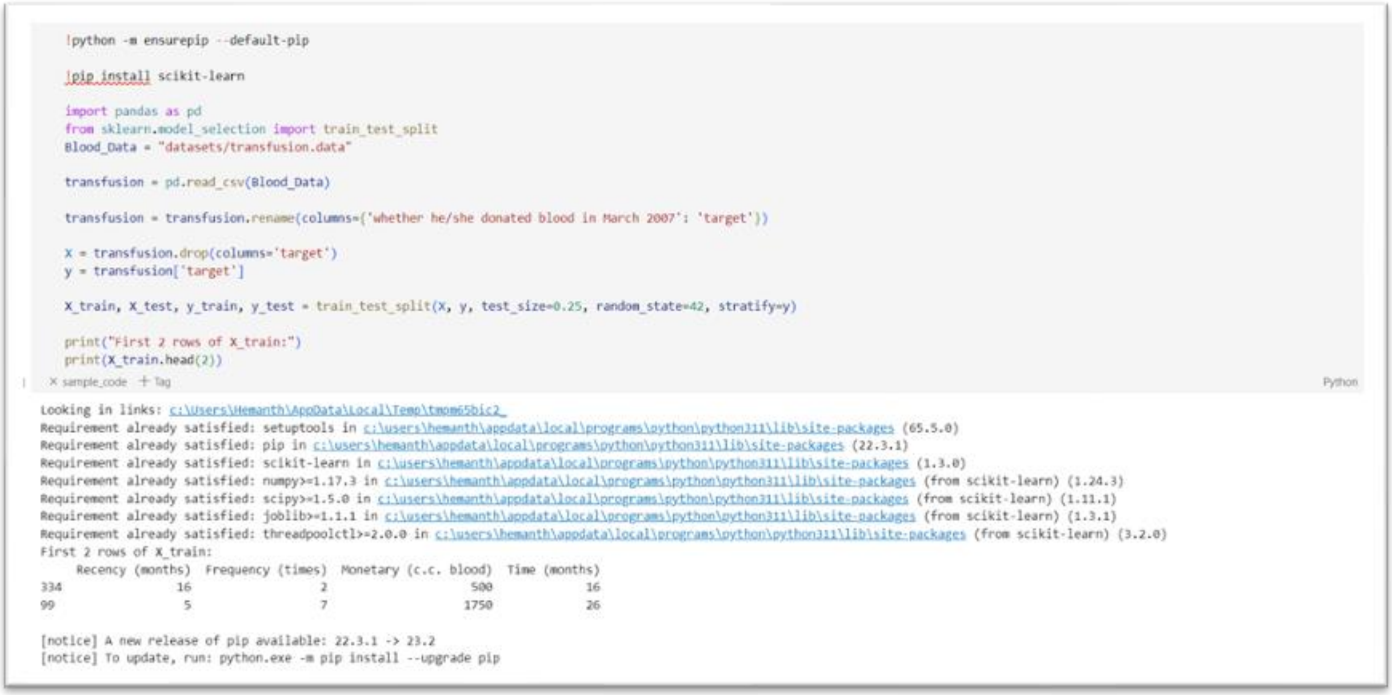
3.6 CHECKING TARGET INCIDENCE



```
Blood_Data = "datasets/transfusion.data"

transfusion = pd.read_csv(Blood_Data)

transfusion = transfusion.rename(columns={'whether he/she donated blood in March 2007': 'target'})

target_proportions = transfusion['target'].value_counts(normalize=True)

print(target_proportions.round(3))
```

```
target
0    0.762
1    0.238
Name: proportion, dtype: float64
```

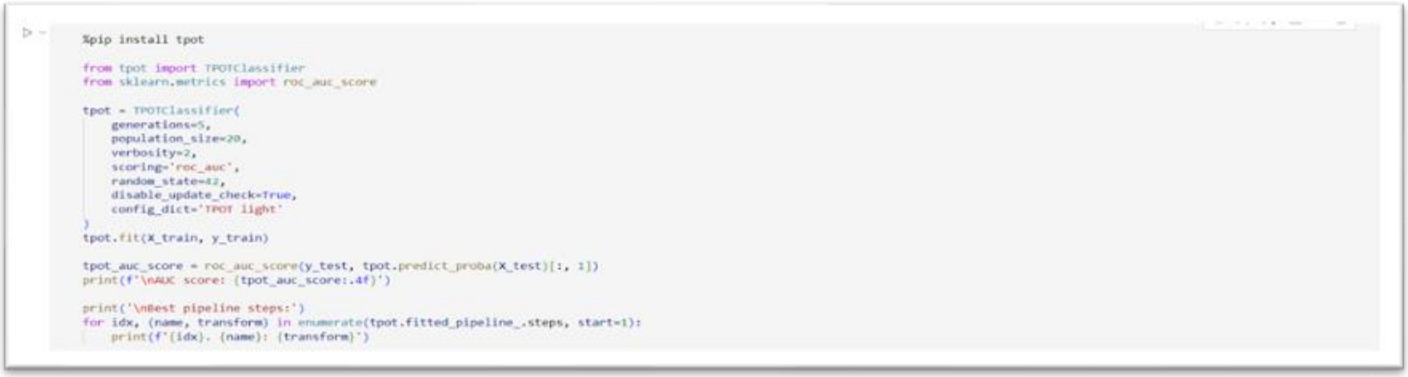
3.7 SPLITTING TRANSFUSION INTO TRAIN AND TEST DATASET



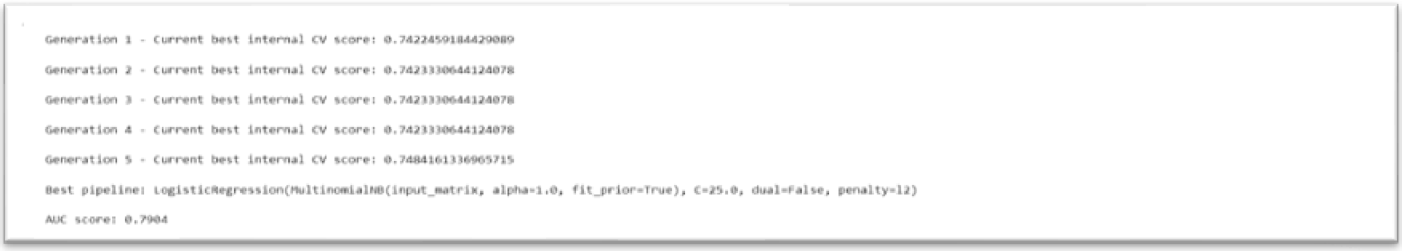
```
!python -m ensurepip --default-pip

!pip install scikit-learn

import pandas as pd
from sklearn.model_selection import train_test_split
Blood_Data = "datasets/transfusion.data"

transfusion = pd.read_csv(Blood_Data)

transfusion = transfusion.rename(columns={'whether he/she donated blood in March 2007': 'target'})

X = transfusion.drop(columns='target')
y = transfusion['target']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=42, stratify=y)

print("First 2 rows of X_train:")
print(X_train.head(2))
```

```
Looking in links: c:\Users\Hemanth\AppData\Local\Temp\tmom65bic2_
Requirement already satisfied: setuptools in c:\users\hemanth\appdata\local\programs\python\python311\lib\site-packages (65.5.0)
Requirement already satisfied: pip in c:\users\hemanth\appdata\local\programs\python\python311\lib\site-packages (22.3.1)
Requirement already satisfied: scikit-learn in c:\users\hemanth\appdata\local\programs\python\python311\lib\site-packages (1.3.0)
Requirement already satisfied: numpy>=1.17.3 in c:\users\hemanth\appdata\local\programs\python\python311\lib\site-packages (from scikit-learn) (1.24.3)
Requirement already satisfied: scipy>=1.5.0 in c:\users\hemanth\appdata\local\programs\python\python311\lib\site-packages (from scikit-learn) (1.11.1)
Requirement already satisfied: joblib>=1.1.1 in c:\users\hemanth\appdata\local\programs\python\python311\lib\site-packages (from scikit-learn) (1.3.1)
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\users\hemanth\appdata\local\programs\python\python311\lib\site-packages (from scikit-learn) (3.2.0)
First 2 rows of X_train:
     Recency (months)  Frequency (times)  Monetary (c.c. blood)  Time (months)
334                16                  2                    500             16
99                  5                  7                   1750             26

[notice] A new release of pip available: 22.3.1 -> 23.2
[notice] To update, run: python.exe -m pip install --upgrade pip
```

3.8 SELECT MODEL USING TPOT

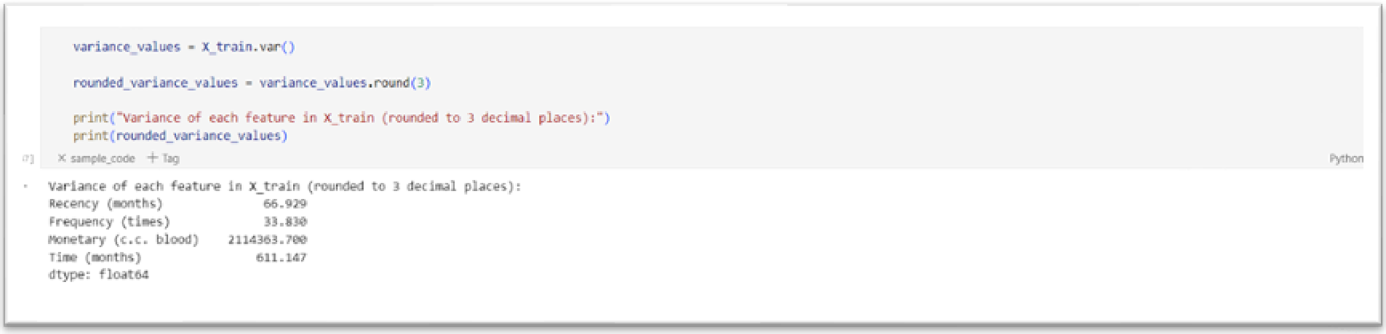


```
%pip install tpot

from tpot import TPOTClassifier
from sklearn.metrics import roc_auc_score

tpot = TPOTClassifier(
    generations=5,
    population_size=20,
    verbosity=2,
    scoring='roc_auc',
    random_state=42,
    disable_update_check=True,
    config_dict='TPOT light'
)
tpot.fit(X_train, y_train)

tpot_auc_score = roc_auc_score(y_test, tpot.predict_proba(X_test)[:, 1])
print(f'\nAUC score: {tpot_auc_score:.4f}')

print('\nBest pipeline steps:')
for idx, (name, transform) in enumerate(tpot.fitted_pipeline_.steps, start=1):
    print(f'{idx}. {name}: {transform}')
```



```
Generation 1 - Current best internal CV score: 0.7422459184429089

Generation 2 - Current best internal CV score: 0.7423330644124078

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: 0.7484161336965715

Best pipeline: LogisticRegression(MultinomialNB(input_matrix, alpha=1.0, fit_prior=True), C=25.0, dual=False, penalty=l2)

AUC score: 0.7904
```

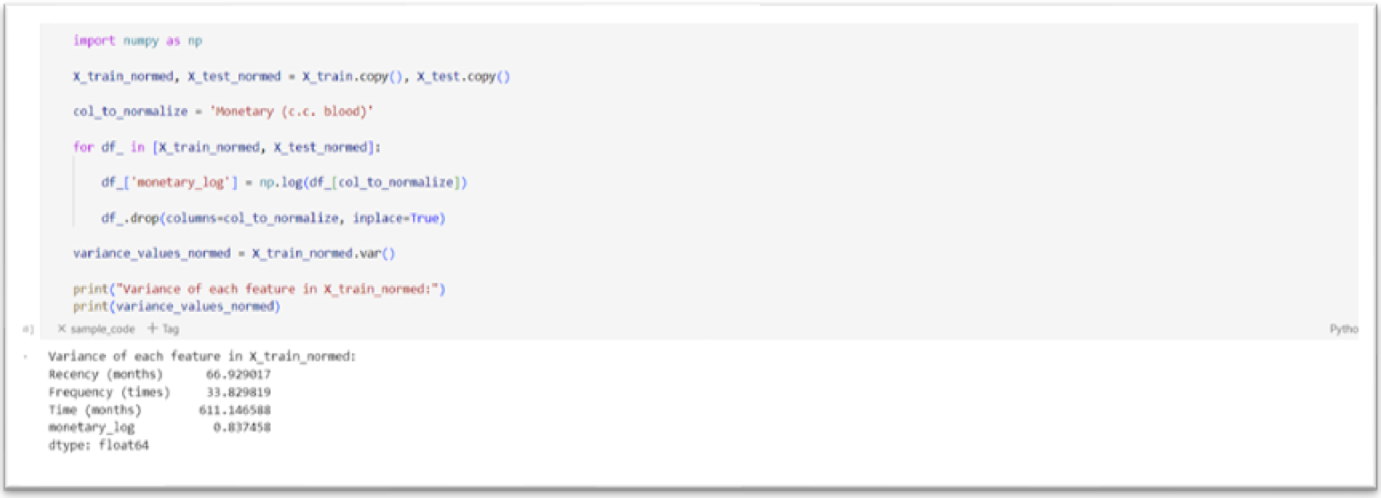
3.9 CHECKING THE VARIANCE



```python
variance_values = X_train.var()

rounded_variance_values = variance_values.round(3)

print("Variance of each feature in X_train (rounded to 3 decimal places):")
print(rounded_variance_values)
```

```
Variance of each feature in X_train (rounded to 3 decimal places):
Recency (months)               66.929
Frequency (times)              33.830
Monetary (c.c. blood)     2114363.700
Time (months)                 611.147
dtype: float64
```

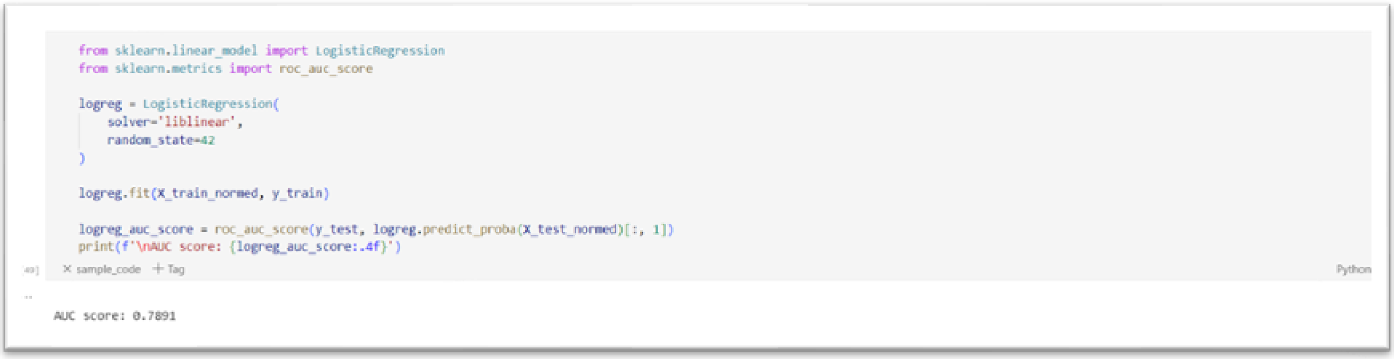
3.10 LOG NORMALIZATION



```python
import numpy as np

X_train_normed, X_test_normed = X_train.copy(), X_test.copy()

col_to_normalize = 'Monetary (c.c. blood)'

for df_ in [X_train_normed, X_test_normed]:

    df_['monetary_log'] = np.log(df_[col_to_normalize])

    df_.drop(columns=col_to_normalize, inplace=True)

variance_values_normed = X_train_normed.var()

print("Variance of each feature in X_train_normed:")
print(variance_values_normed)
```

```
Variance of each feature in X_train_normed:
Recency (months)      66.929017
Frequency (times)     33.829819
Time (months)        611.146588
monetary_log           0.837458
dtype: float64
```

3.11 TRAINING THE LINEAR REGRESSION MODEL



```python
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import roc_auc_score

logreg = LogisticRegression(
    solver='liblinear',
    random_state=42
)

logreg.fit(X_train_normed, y_train)

logreg_auc_score = roc_auc_score(y_test, logreg.predict_proba(X_test_normed)[:, 1])
print(f'\nAUC score: {logreg_auc_score:.4f}')
```

```
AUC score: 0.7891
```

CONCLUSION

Given an individual's history of the four given variables – time since last donation, number of times donated, overall donated amount, and timespan since the first donation – a model was developed to predict if the individual would donate or not with an accuracy of 0.76. The model produced the following regression coefficients for the variables;

Recency: - 0.09
Frequency: 0.96
Amount: - 0.03
Timespan: 0.29

Generally associating more established donors to have a higher probability to donate again, with frequency and timespan having the most impact.

Medical facilities and healthcare providers having a better ability to estimate blood donations ensures a steady supply of blood and the treatment that is contingent, periods of relative scarcity can be avoided if foreseen early either by taking direct measures or by coordination within a health network and its supply chains.

REFERENCES

Below are links to sources that were referenced in this report:

DATA SOURCE

https://drive.google.com/file/d/1S2o3wEAfEPha06ECh6kirwUijcCq54nY/view

SCIKITLEARN PACKAGE: LOGISTIC REGRESSION

sklearn.linear_model.LogisticRegression — scikit-learn 1.3.0 documentation

TPOT PACKAGE: TPOTCLASSIFIER

Using TPOT - TPOT (epistasislab.github.io)

RED CROSS BLOOD DONATION ESTIMATES

50 Quick Blood Facts from the American Red Cross (umms.org)