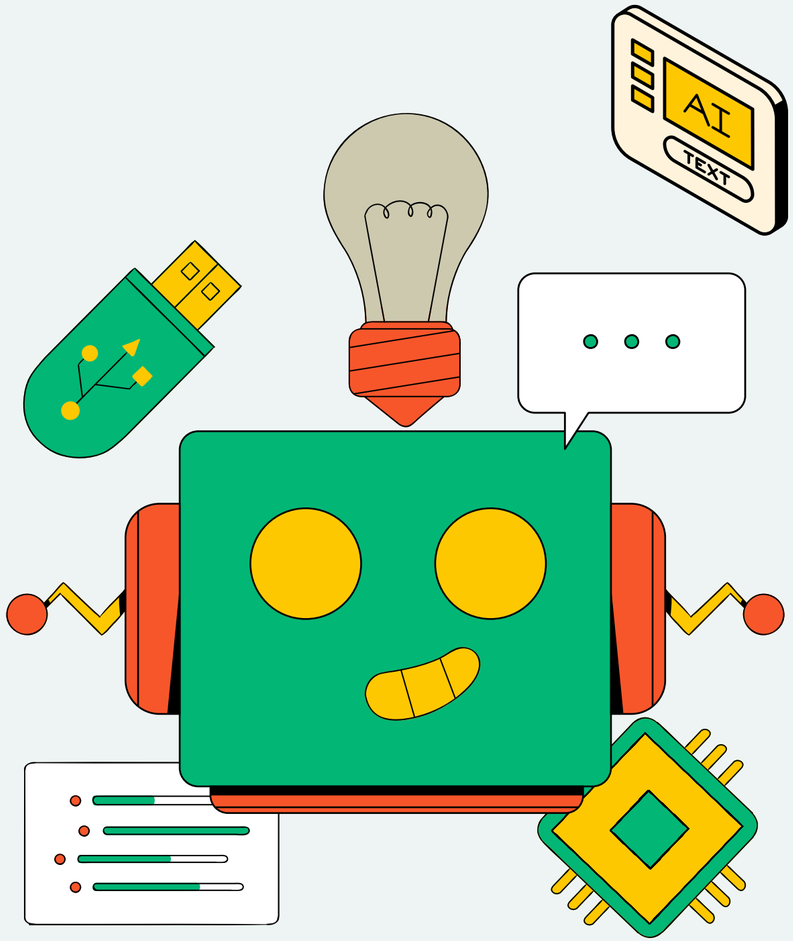
WE LEARN FOR THE FUTURE



MULTIPLE DISEASE PREVEDNTION USING MACHINE LEARNING

PRESENTATION

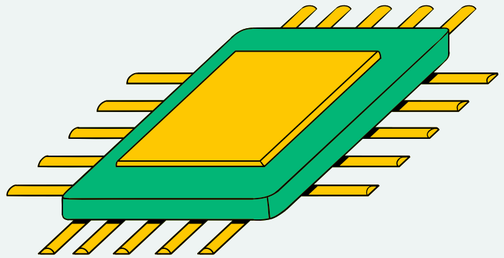
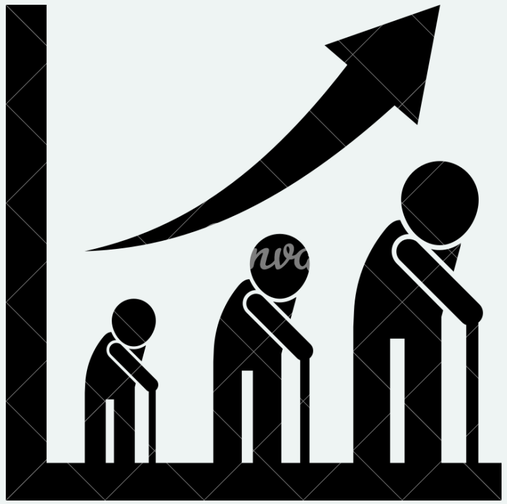


TABLE OF CONTENTS

INTRODUCTION

With an increase of 1% every year since 2011. Singapore is one of the most rapidly aging populations in the world. With 16.6% of its citizens being 65 years and older.

Naturally, the older we get, our immune system gets weaker, making us more prone to get ill. This is why yearly medical check-ups are essential for elderlies.



PROBLEM FORMULATION

Some medical tests can be invasive which is not ideal for elderlies. However, most elderlies would go for a yearly medical check-up where they will get their blood drawn which makes their blood sample quite accessible.

Can we use this blood sample to test if they have any underlying illnesses?

```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")

# Load the dataset
file_path = '/work/Blood_samples_dataset_balanced_2(f).csv'
pre_data = pd.read_csv(file_path)

# Display the first few rows and some general information about the dataset
data_info = pre_data.info()
data_head = pre_data.head()
data_description = pre_data.describe()

data_info, data_head, data_description
```

```
 Glucose  Cholesterol  Hemoglobin  Platelets  White Blood Cells  \
0.739597     0.650198    0.713631   0.868491           0.687433
0.121786     0.023058    0.944893   0.905372           0.507711
0.452539     0.116135    0.544560   0.400640           0.294538
0.136609     0.015605    0.419957   0.191487           0.081168
0.176737     0.752220    0.971779   0.785286           0.443880

Red Blood Cells  Hematocrit  Mean Corpuscular Volume  \
       0.529895    0.290006                 0.631045
       0.403033    0.164216                 0.307553
       0.382021    0.625267                 0.295122
       0.166214    0.073293                 0.668719
       0.439851    0.894991                 0.442159

Mean Corpuscular Hemoglobin  Mean Corpuscular Hemoglobin Concentration
                   0.001328                                   0.795829
                   0.207938                                   0.505562
                   0.868369                                   0.026808
                   0.125447                                   0.501051
                   0.257288                                   0.805987

...     HbA1c  LDL Cholesterol  HDL Cholesterol       ALT       AST  \
...  0.502665         0.215560         0.512941  0.064187  0.610827
...  0.856810         0.652465         0.106961  0.942549  0.344261
...  0.466795         0.387332         0.421763  0.007186  0.506918
...  0.016256         0.040137         0.826721  0.265415  0.594148
...  0.429431         0.146294         0.221574  0.015280  0.567115
```

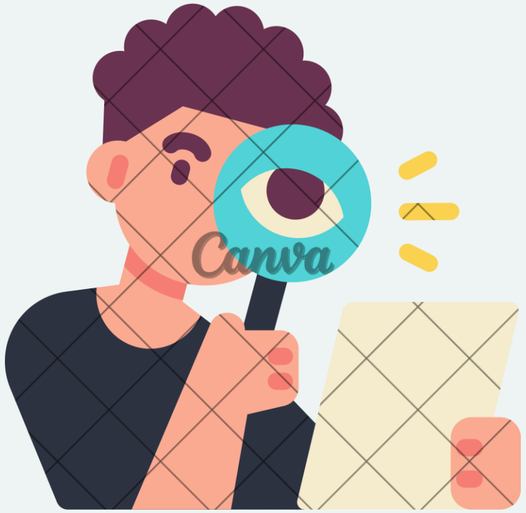
DATASET

The dataset displays different blood parameters(eg glucose level, insulin, etc.)

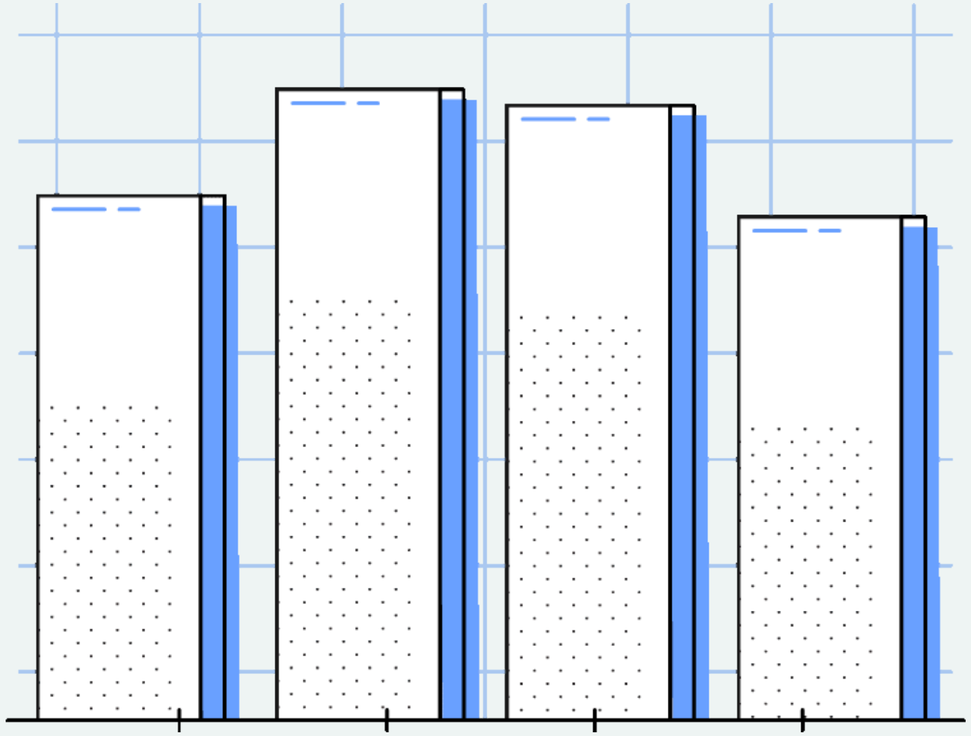
This dataset was already pre-cleaned. Hence, there was little to no preparation and cleaning needed to analyze our data better.

Clean Data ready for analysing

EXPLORATORY ANALYSIS



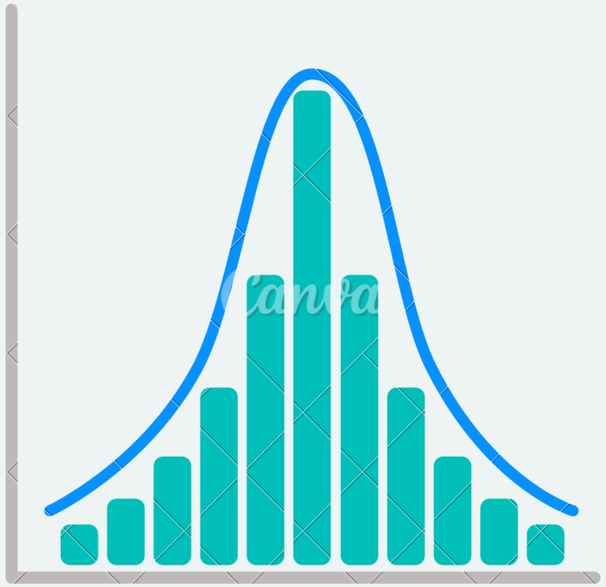
- Data Visualisation
- Observe trends and impact of variables
- Deeper Cleaning of Data



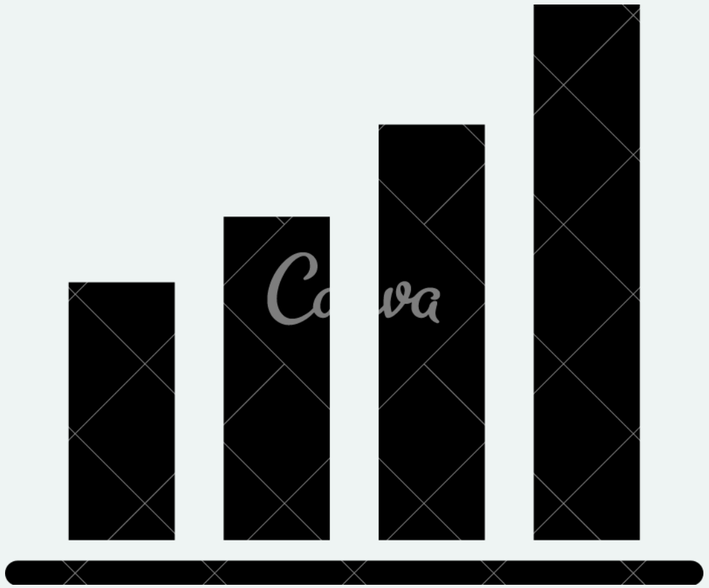
DATA VISUALIZATION

Mainly used to see the trends of each variable
3 Visualisation Methods Used:

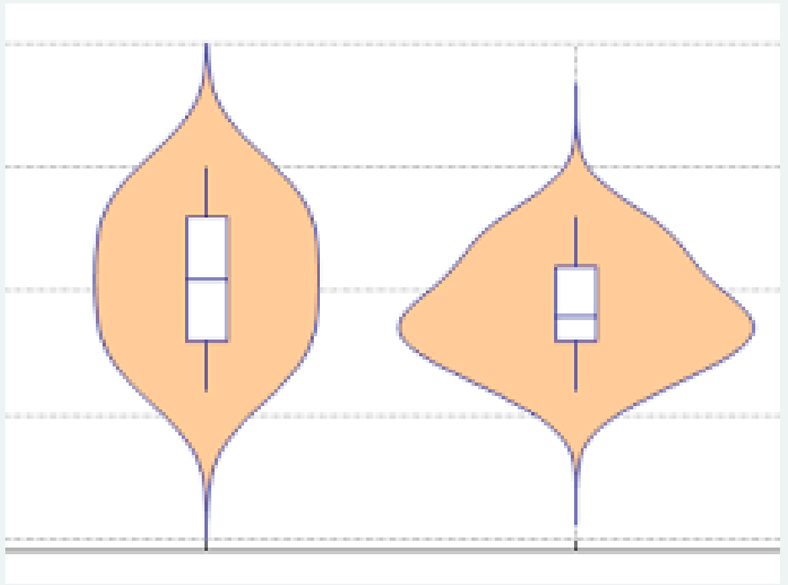
Histogram



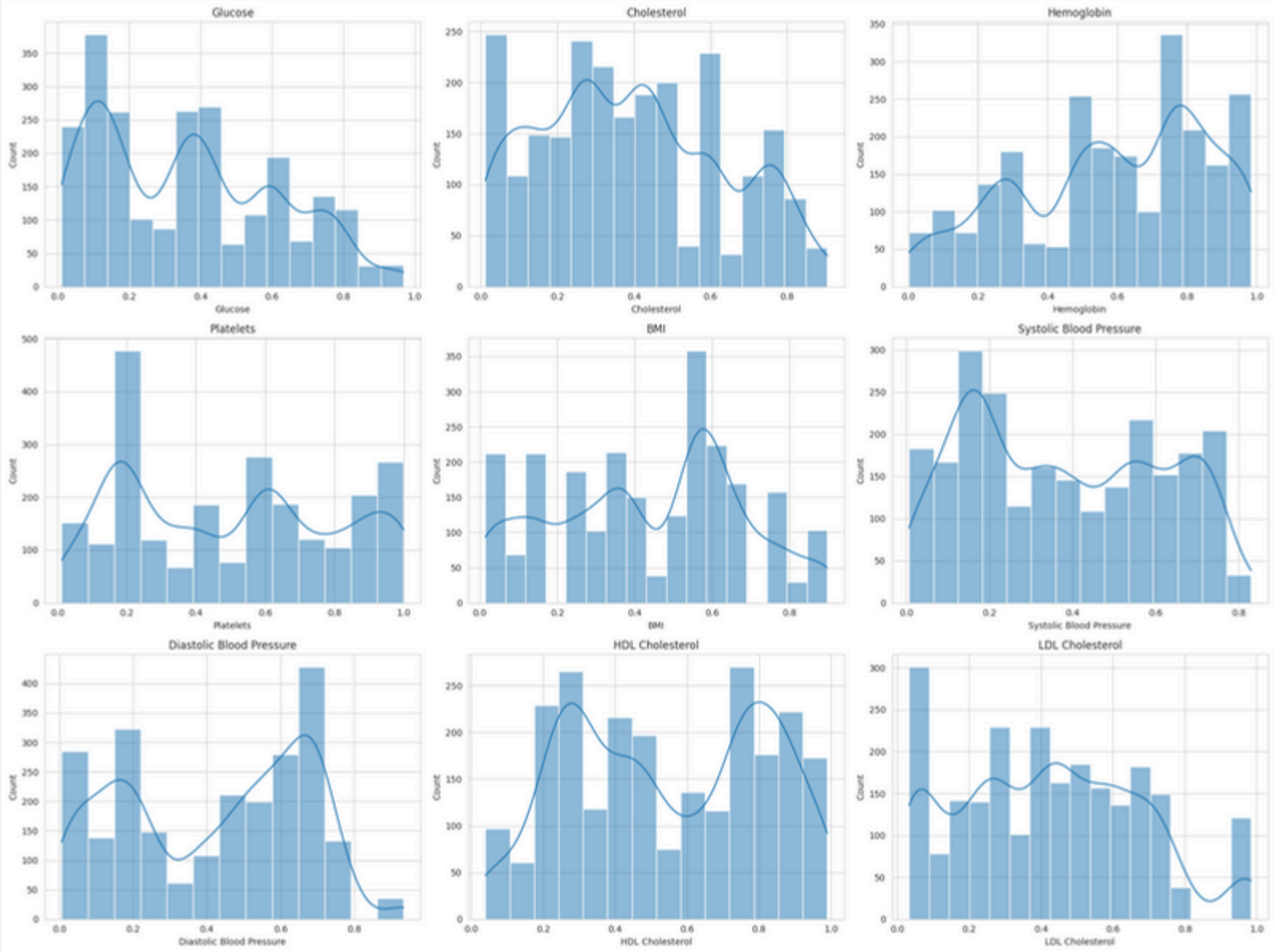
Bar Chart



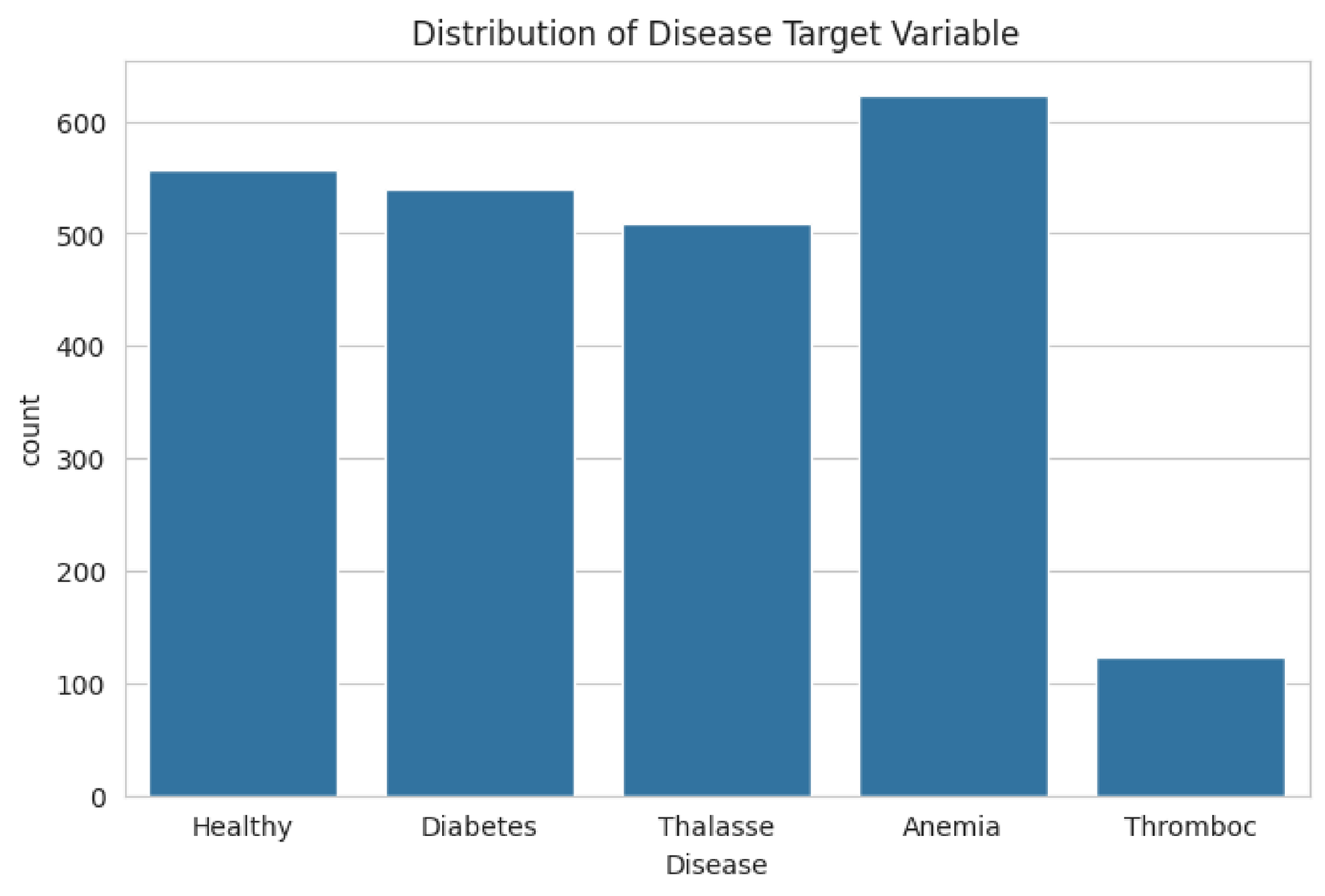
Violin Plot



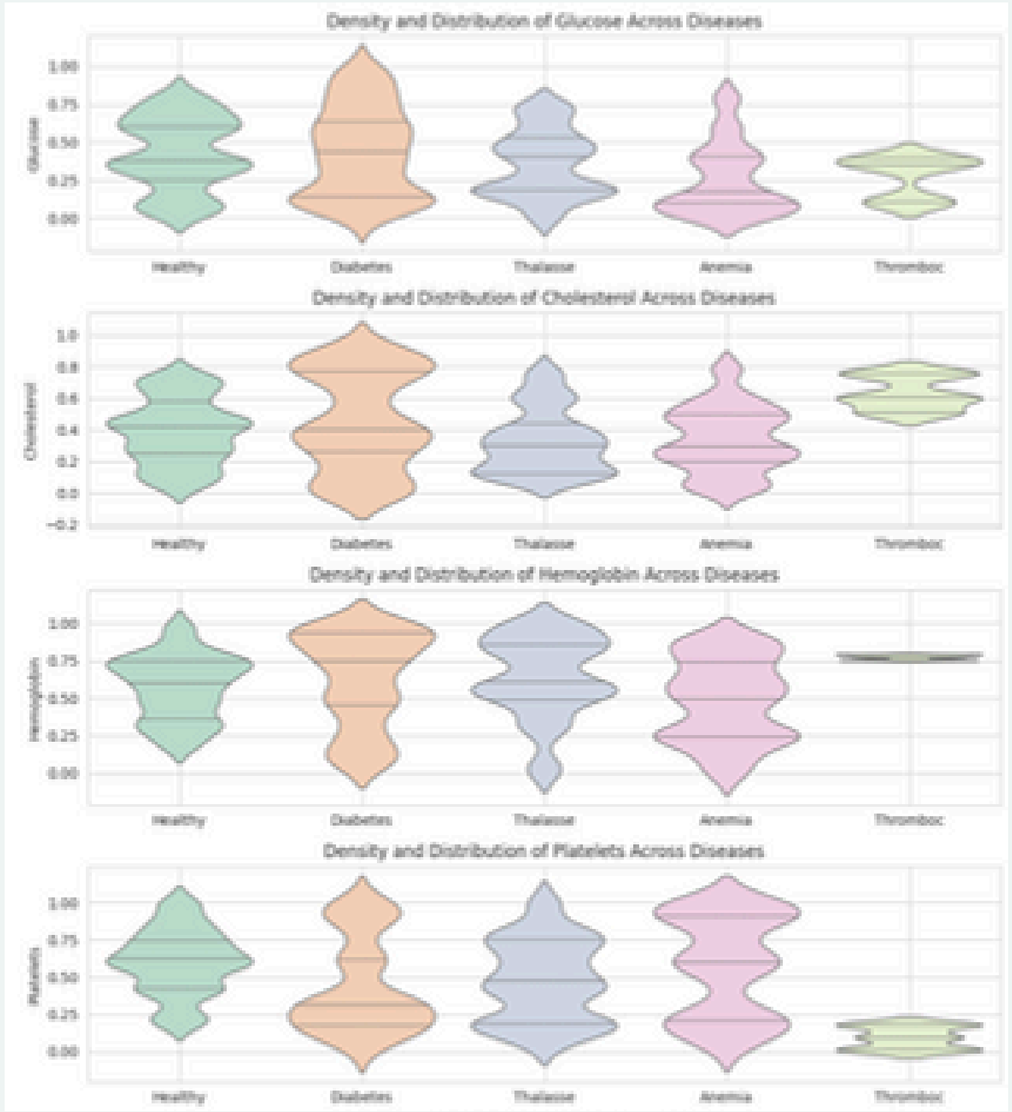
HISTOGRAM

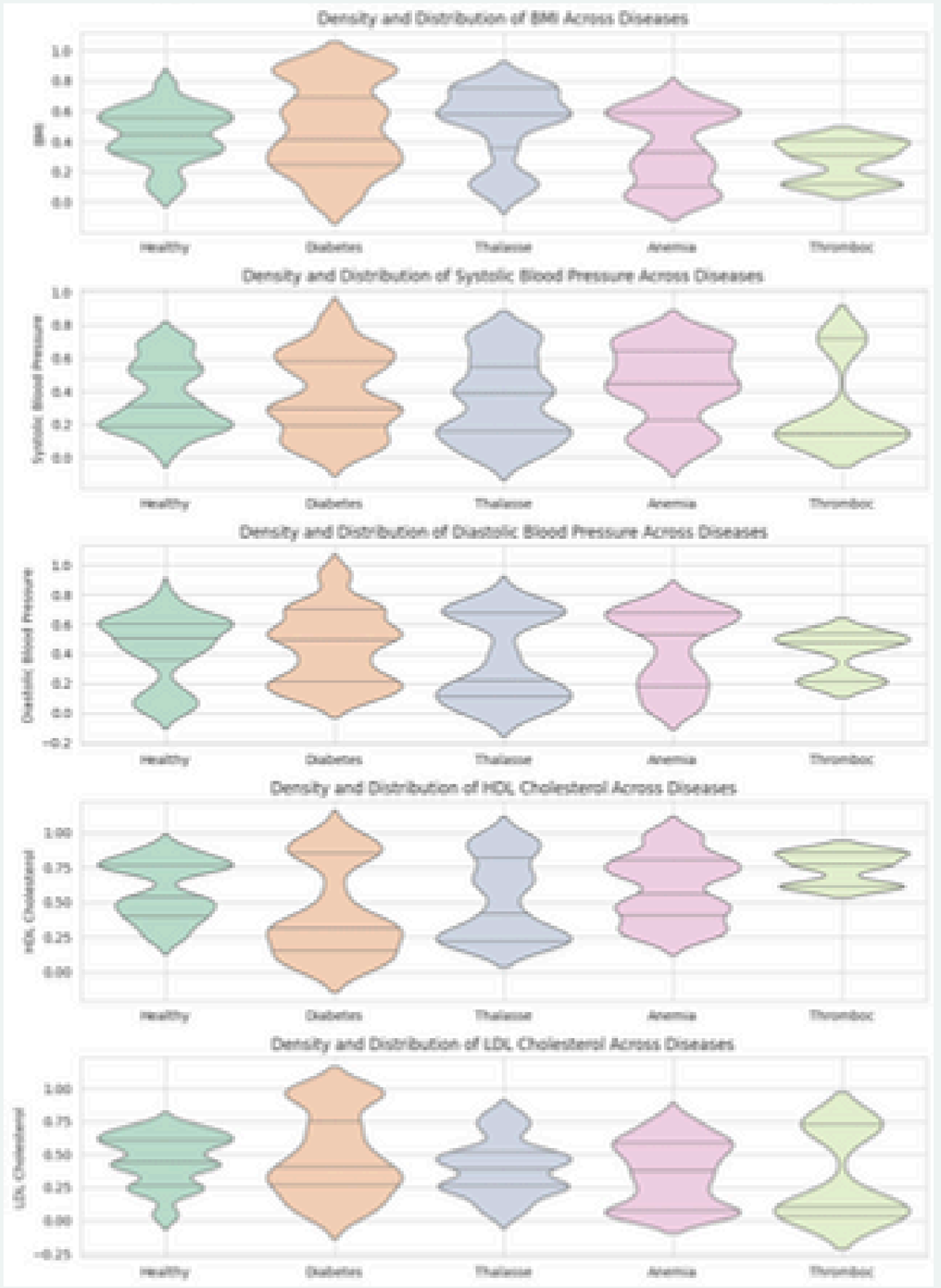

Glucose
Count
Glucose
Cholesterol
Count
Cholesterol
Hemoglobin
Count
Hemoglobin
Platelets
Count
Platelets
BMI
Count
BMI
Systolic Blood Pressure
Count
Systolic Blood Pressure
Diastolic Blood Pressure
Count
Diastolic Blood Pressure
HDL Cholesterol
Count
HDL Cholesterol
LDL Cholesterol
Count
LDL Cholesterol

BAR CHART

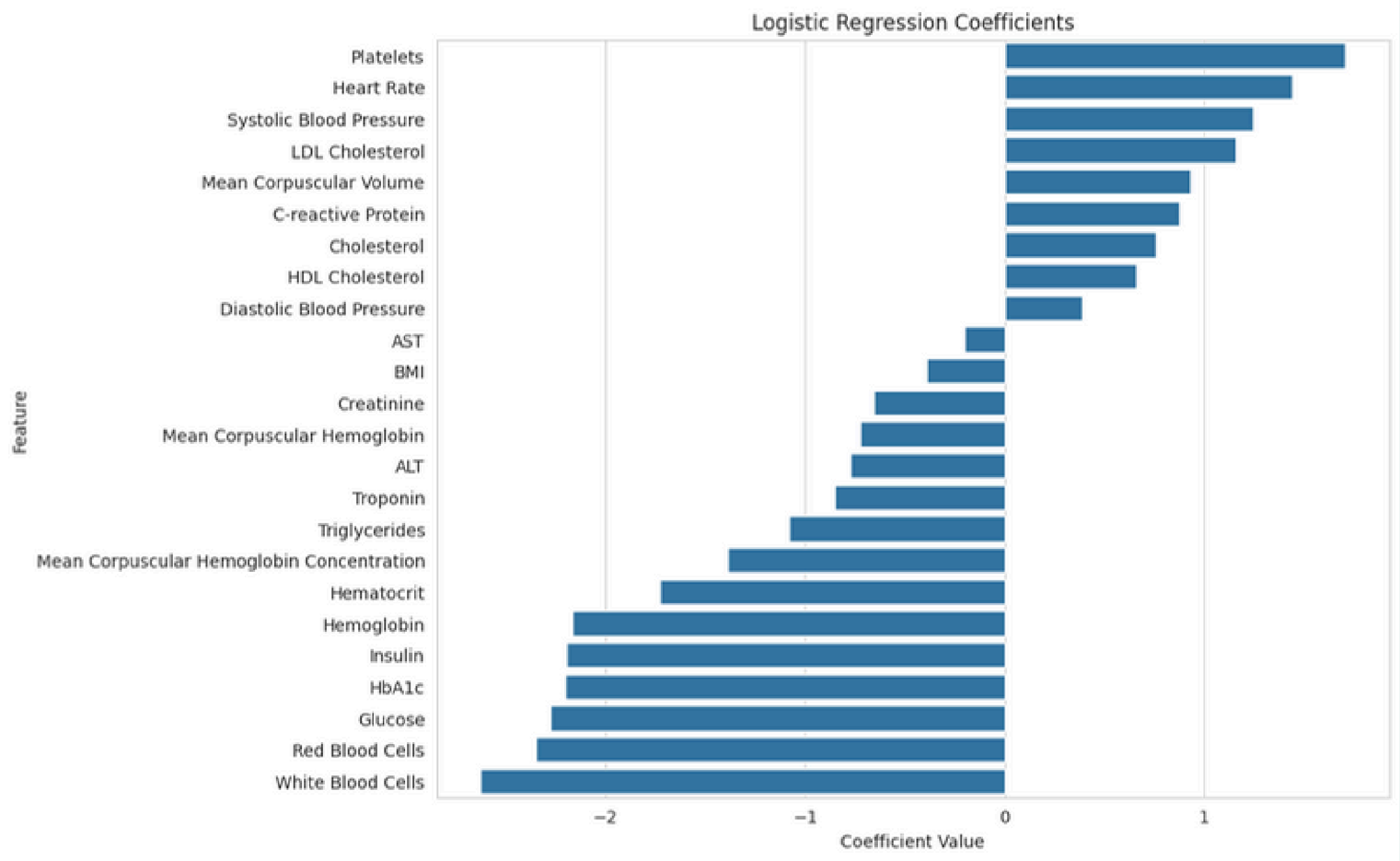

Distribution of Disease Target Variable
count
600
500
400
300
200
100
0
Healthy
Diabetes
Thalasse
Anemia
Thromboc
Disease

VIOLIN PLOT


Density and Distribution of Glucose Across Diseases
Glucose
Density and Distribution of Cholesterol Across Diseases
Cholesterol
Density and Distribution of Hemoglobin Across Diseases
Hemoglobin
Density and Distribution of Platelets Across Diseases
Platelets
Healthy
Diabetes
Thalasse
Anemia
Thromboc


Density and Distribution of BMI Across Diseases
BMI
Density and Distribution of Systolic Blood Pressure Across Diseases
Systolic Blood Pressure
Density and Distribution of Diastolic Blood Pressure Across Diseases
Diastolic Blood Pressure
Density and Distribution of HDL Cholesterol Across Diseases
HDL Cholesterol
Density and Distribution of LDL Cholesterol Across Diseases
LDL Cholesterol
Healthy
Diabetes
Thalasse
Anemia
Thromboc

NOTICEABLE TREND



With Logistic Regression supporting:
Diastolic Blood Pressure, AST (Aspartate Aminotransferase) and BMI (Body Mass Index) does not impact the likelihood/unlikelihood of diseases as much as the other variables due to the low coefficent value.

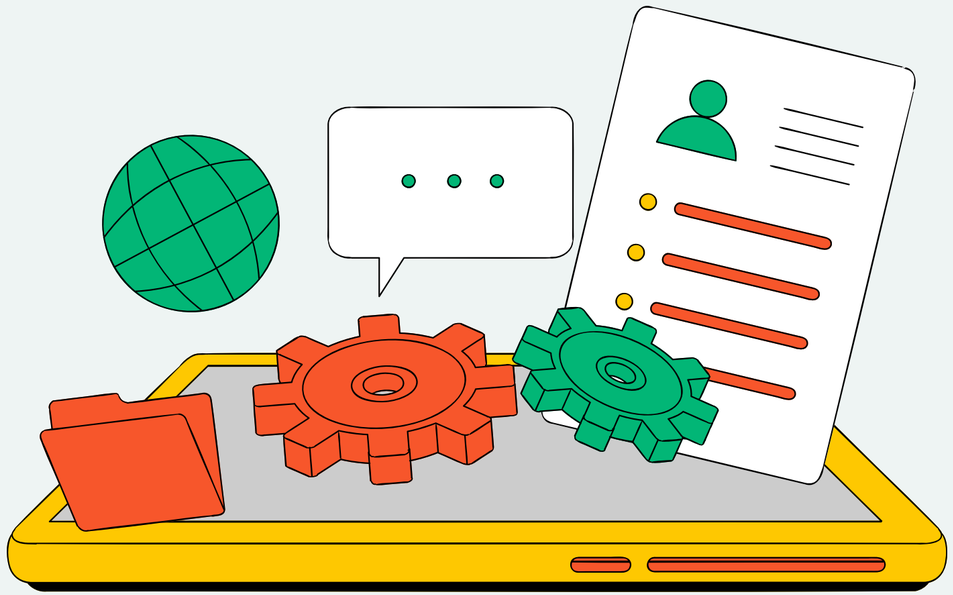
This can also be noticed on the violin plot.

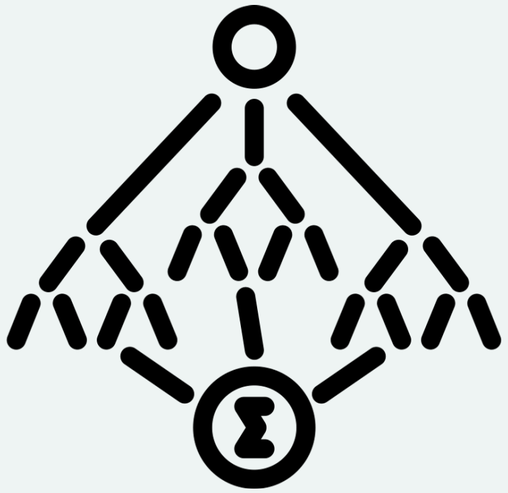
MODELS

Random forest

Gradient Boosting

Support vector machine





RANDOM FOREST

Random forest is a supervised machine learning algorithm used to solve classification as well as regression problems. It is a type of ensemble learning technique in which multiple decision trees are created from the training dataset and the majority output from them is considered as the final output.

After implementing it , it was found that it produced an accuracy score of 100% on testing against the test dataset

Model: Random Forest

	precision	recall	f1-score	support
Anemia	1.00	1.00	1.00	134
Diabetes	1.00	1.00	1.00	112
Healthy	1.00	1.00	1.00	102
Thalasse	1.00	1.00	1.00	103
Thromboc	1.00	1.00	1.00	20
accuracy			1.00	471
macro avg	1.00	1.00	1.00	471
weighted avg	1.00	1.00	1.00	471

Accuracy: 1.0

GRADIENT BOOSTING

Model: Gradient Boosting

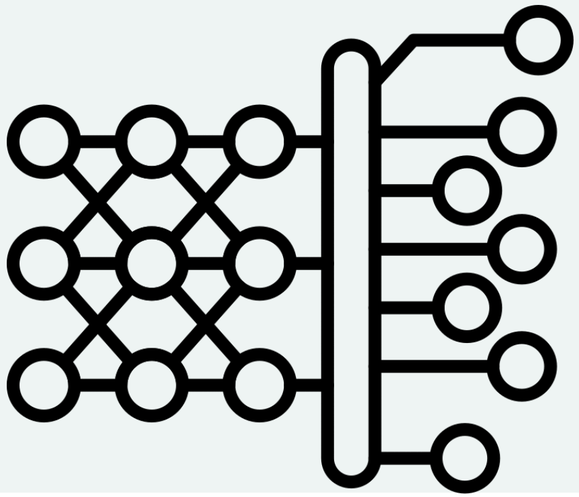
	precision	recall	f1-score	support
Anemia	1.00	1.00	1.00	134
Diabetes	1.00	1.00	1.00	112
Healthy	1.00	1.00	1.00	102
Thalasse	1.00	1.00	1.00	103
Thromboc	1.00	1.00	1.00	20
accuracy			1.00	471
macro avg	1.00	1.00	1.00	471
weighted avg	1.00	1.00	1.00	471

Accuracy: 1.0

Gradient Boosting is an machine learning technique that combines the predictions from several models to improve the overall predictive accuracy. It is particularly useful for regression and classification problems like detecting diseases.

After its implementation, It was found that the model had an accuracy score of 100% against the test dataset.

SUPPORT VECTOR MACHINE



Model: Support Vector Machine

	precision	recall	f1-score	support
Anemia	1.00	1.00	1.00	134
Diabetes	1.00	1.00	1.00	112
Healthy	1.00	1.00	1.00	102
Thalasse	1.00	1.00	1.00	103
Thromboc	1.00	1.00	1.00	20
accuracy			1.00	471
macro avg	1.00	1.00	1.00	471
weighted avg	1.00	1.00	1.00	471

Accuracy: 1.0

Support Vector Machine, is a powerful supervised learning algorithm used for classification and regression tasks. It works by finding the optimal hyperplane that best separates different classes in the input data while maximizing the margin between them. SVM is effective in high-dimensional spaces and is particularly useful when dealing with complex data that is not linearly separable.

After running the model , it was observed to have an accuracy score of 100% against the test dataset.

Testing against a new dataset

A test dataset provided by the same kaggle source was used to test the models:

- **Support Vector Machine** performed the best in terms of overall accuracy (46.91%), but the performance is still quite modest.
- **Gradient Boosting** showed the next best accuracy at (44.65%), which is similar but slightly lower than the SVM.
- **Random Forest** struggled with an accuracy score of (39.17%) , the lowest out of the three models.

Evaluation for Random Forest:

	precision	recall	f1-score	support
Anemia	0.34	0.61	0.44	84
Diabetes	0.63	0.39	0.48	294
Healthy	0.06	1.00	0.11	5
Heart Di	0.00	0.00	0.00	39
Thalasse	0.33	0.46	0.39	48
Thromboc	0.33	0.06	0.11	16
accuracy			0.40	486
macro avg	0.28	0.42	0.25	486
weighted avg	0.48	0.40	0.41	486

Accuracy: 0.3991769547325103

Evaluation for Gradient Boosting:

	precision	recall	f1-score	support
Anemia	0.36	0.58	0.44	84
Diabetes	0.68	0.48	0.56	294
Healthy	0.05	0.60	0.10	5
Heart Di	0.00	0.00	0.00	39
Thalasse	0.33	0.44	0.38	48
Thromboc	0.17	0.25	0.20	16
accuracy			0.45	486
macro avg	0.27	0.39	0.28	486
weighted avg	0.51	0.45	0.46	486

Accuracy: 0.44650205761316875

Evaluation for Support Vector Machine:

	precision	recall	f1-score	support
Anemia	0.38	0.37	0.38	84
Diabetes	0.68	0.62	0.65	294
Healthy	0.04	0.60	0.07	5
Heart Di	0.00	0.00	0.00	39
Thalasse	0.19	0.25	0.22	48
Thromboc	0.00	0.00	0.00	16
accuracy			0.47	486
macro avg	0.22	0.31	0.22	486
weighted avg	0.50	0.47	0.48	486

Accuracy: 0.4691358024691358

Improving the Model

```
Classification Report:
              precision    recall  f1-score   support

      Anemia       0.40      0.63      0.49        84
    Diabetes       0.70      0.53      0.60       294
     Healthy       0.07      1.00      0.13         5
    Heart Di       0.00      0.00      0.00        39
    Thalasse       0.35      0.40      0.37        48
    Thromboc       0.43      0.19      0.26        16

    accuracy                           0.48       486
   macro avg       0.32      0.46      0.31       486
weighted avg       0.54      0.48      0.50       486

Accuracy: 0.4835390946502058
```

- We noticed that each model predicts each variable differently with varying accuracy. Applying advanced ensemble techniques such as stacking, we can combine the predictions from multiple models and make a final prediction, possibly improving overall performance.

CONCLUSION

Although the accuracy of the model may not be high overall, with the usage of the stacked model, the models could be used as a sidekick for medical professionals to highlight possible diseases (Diabetes (Moderate/High Confidence), Anemia/Thalasse/Thoromboc (Low/Moderate Confidence)).

This would prove to be extremely useful to accurately (100%) predict every disease in the future