DISEASE PREDICTION

By Mallela Preethi

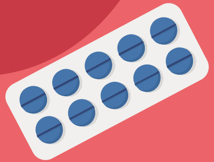
Problem Statement

Background:

- Healthcare professionals rely on various diagnostic tests and biomarkers for health assessment and disease diagnosis.
- Accurate diagnosis is critical for effective treatment and disease management.

Dataset:

- Contains multiple health-related attributes:
  - Cholesterol levels
  - Blood cell counts
  - Hormone levels
  - Other physiological measurements
- Includes the corresponding disease diagnosed for each individual.
- Labels include Healthy, Anemia, Diabetes, Heart Di, Thalasse, Thrombac



Task:

- Create a reliable tool(predictive model) using machine learning algorithms to assist healthcare providers in disease diagnosis and prognosis.
- Enhance the accuracy of disease diagnosis.
- Evaluate the model using accuracy, precision, recall, and F1-score to ensure its reliability and effectiveness in diagnosis.

METHODOLOGY

1. Preprocessing and Exploratory Data Analysis
2. Model Building
3. Evaluation
4. Model Tuning
5. Creation of final prediction model

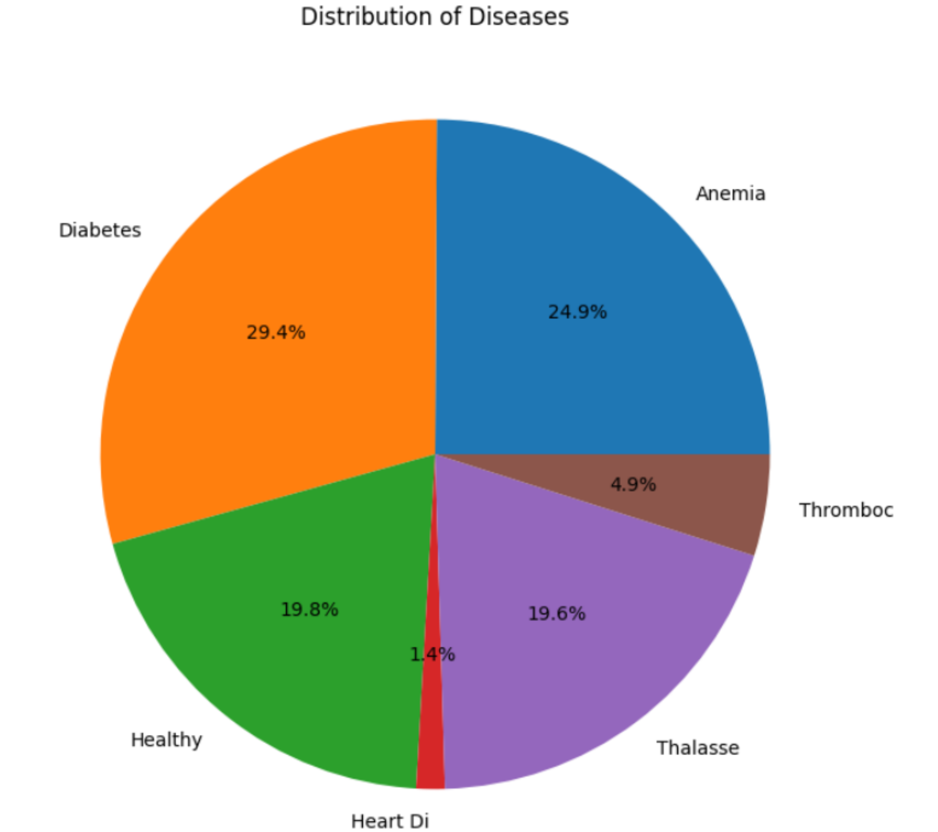
1. Exploratory Data Analysis

Imbalanced Dataset:

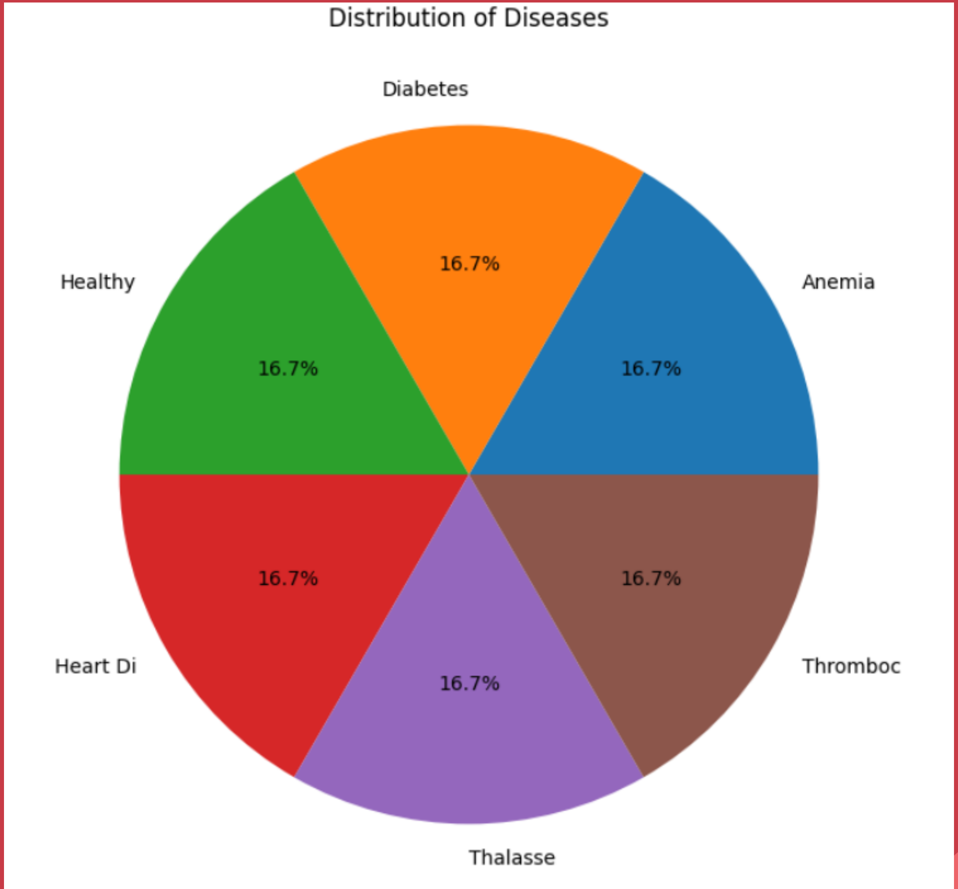
On observing the right side Pie chart we can see Thromboc, Heart disease examples are very less this leads to bias in the model. We can understand that given dataset is a imbalanced dataset. So we have to resample this.

Resampling with SMOTE:

Synthetic Minority Over-sampling Technique (SMOTE) is an effective method for handling class imbalance in datasets by generating synthetic samples for the minority class.



After Application of StratifiedKFold and SMOTE



Train and test data shapes:

```
X_train shape: (4002, 24)
y_train shape: (4002,)
X_test shape: (567, 24)
y_test shape: (567,)
```

2. MODEL BUILDING

Models built and compared:

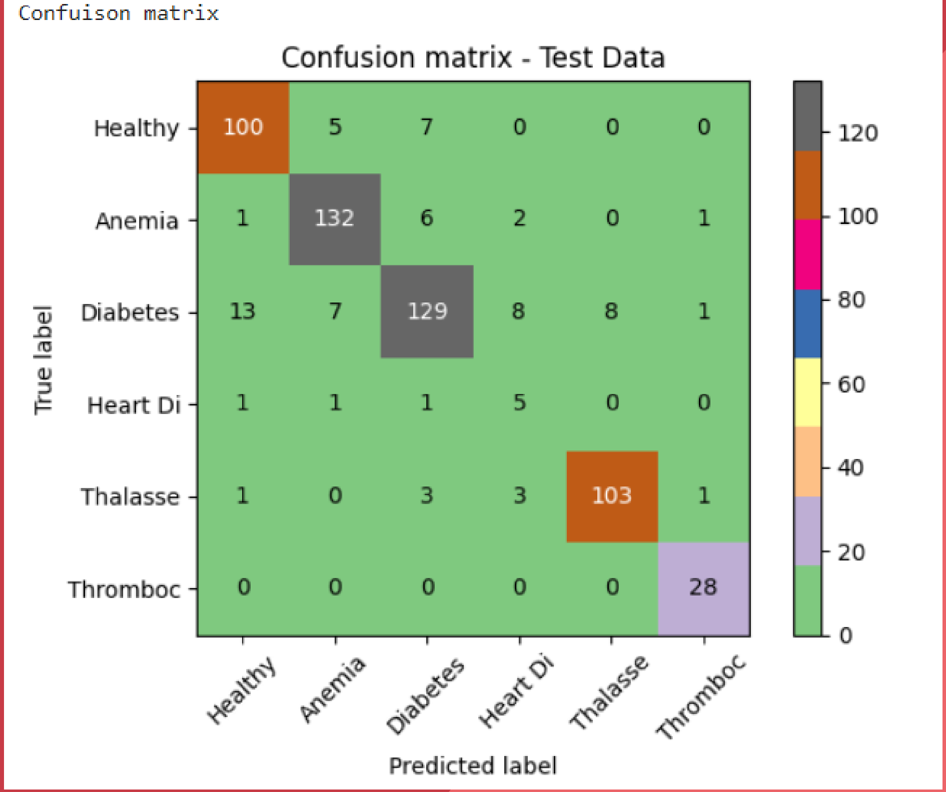
- Logistic Regression
- KNN
- Decision Tree with gini criteria
- Decision Tree with entropy criteria
- Random Forest Classifier
- XGBoost Classifier

These models are built and trained on this balanced dataset and their performances were observed.

3. (i) Evaluation – Logistic Regression

Logistic Regression:

- Accuracy: 0.8765432098765432
- Precision score: 0.887133099355824
- Recall score: 0.8765432098765432
- F1 score: 0.8796970866435904



```
Confuison matrix
```

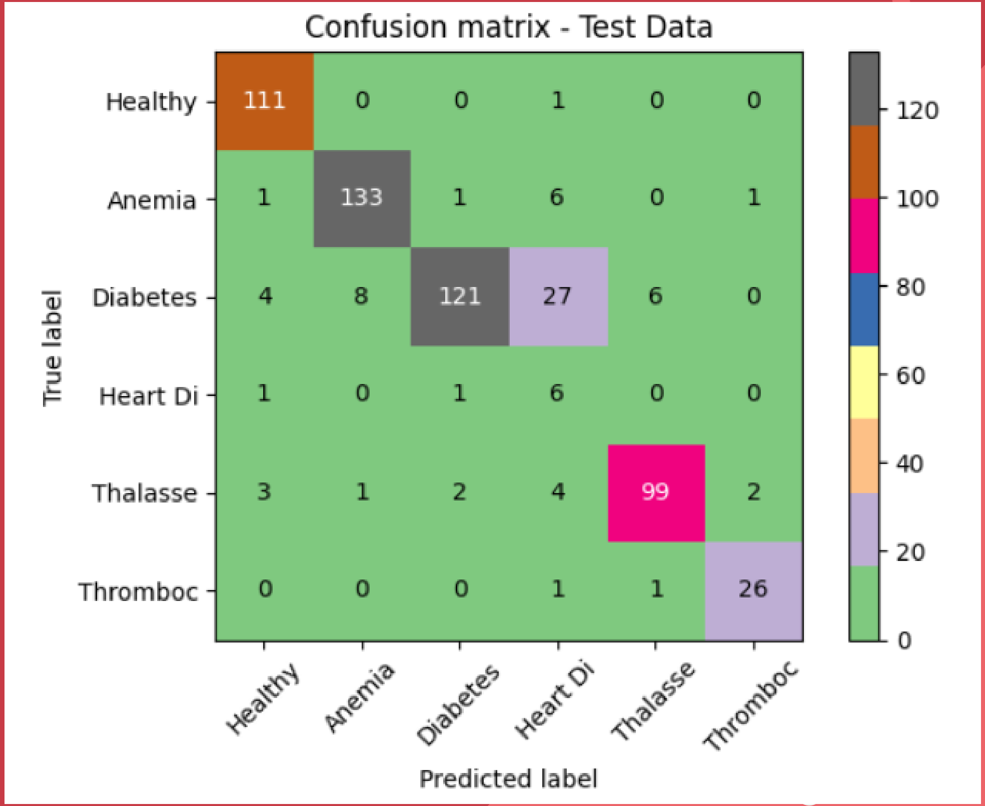
Confusion matrix - Test Data

True label \ Predicted label	Healthy	Anemia	Diabetes	Heart Di	Thalasse	Thromboc
Healthy	100	5	7	0	0	0
Anemia	1	132	6	2	0	1
Diabetes	13	7	129	8	8	1
Heart Di	1	1	1	5	0	0
Thalasse	1	0	3	3	103	1
Thromboc	0	0	0	0	0	28

(ii) Evaluation – KNN

KNN model

- Accuracy: 0.8747795414462081
- Precision score: 0.9296788822985599
- Recall score: 0.8747795414462081
- F1 score: 0.8939273547678349



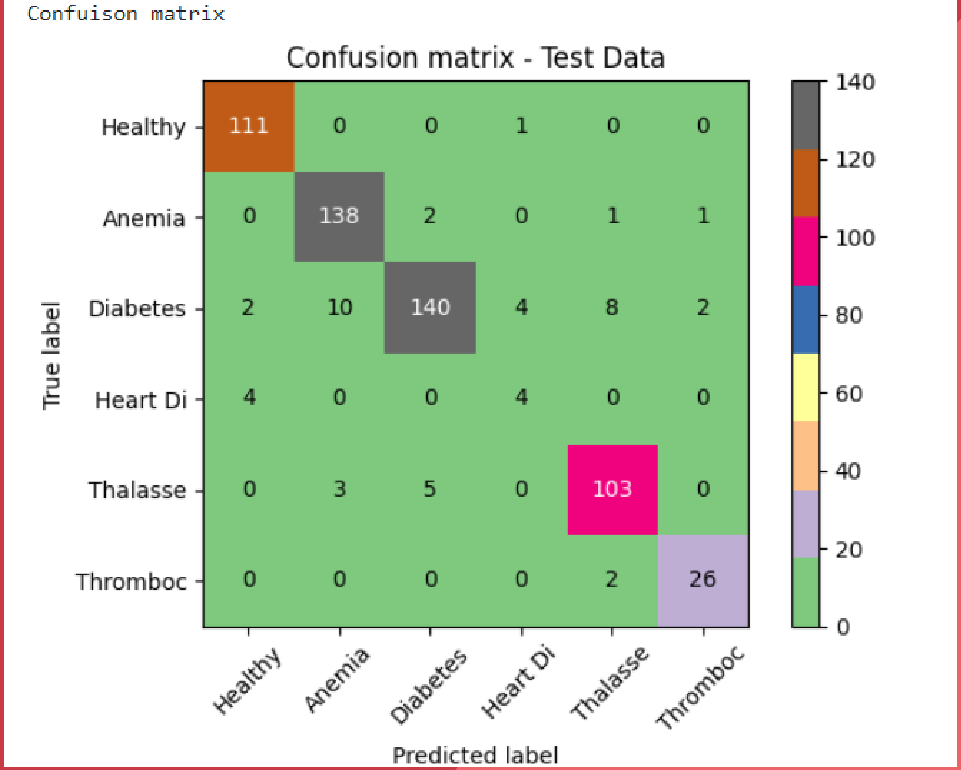
Confusion matrix - Test Data

True label \ Predicted label	Healthy	Anemia	Diabetes	Heart Di	Thalasse	Thromboc
Healthy	111	0	0	1	0	0
Anemia	1	133	1	6	0	1
Diabetes	4	8	121	27	6	0
Heart Di	1	0	1	6	0	0
Thalasse	3	1	2	4	99	2
Thromboc	0	0	0	1	1	26

(iii) Evaluation – Decision Tree

Decision Tree with gini criteria

- Accuracy:0.9206349206349206
- Precision score: 0.9225307941875883
- Recall score: 0.9206349206349206
- F1 score: 0.920231298268425 8



Confuison matrix

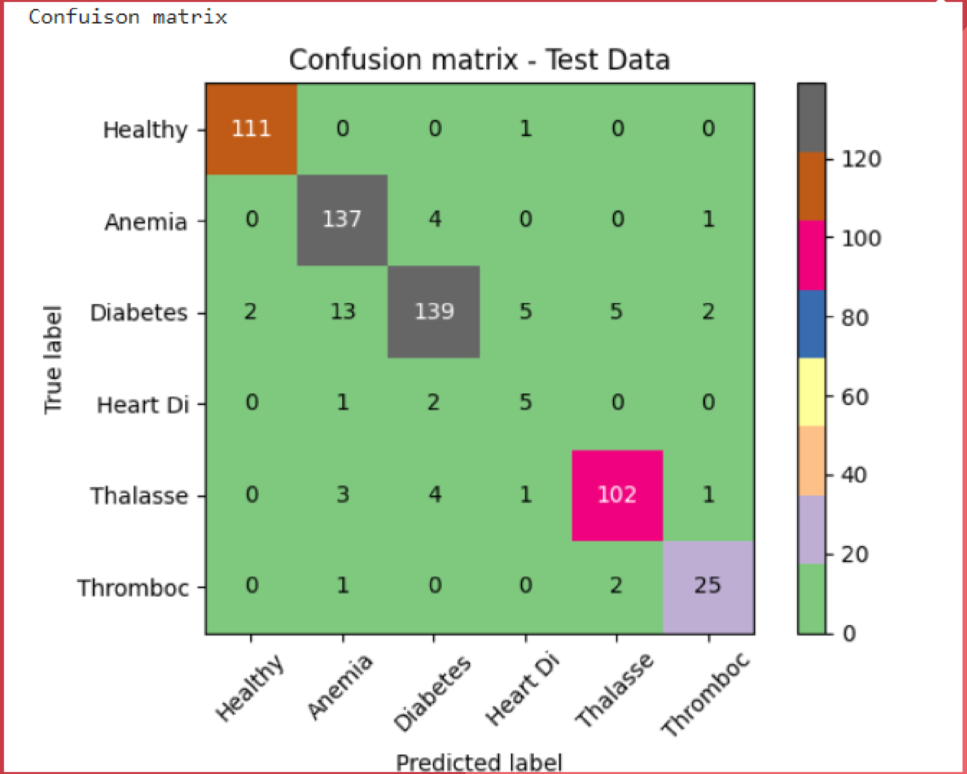
Confusion matrix - Test Data

True label \ Predicted label	Healthy	Anemia	Diabetes	Heart Di	Thalasse	Thromboc
Healthy	111	0	0	1	0	0
Anemia	0	138	2	0	1	1
Diabetes	2	10	140	4	8	2
Heart Di	4	0	0	4	0	0
Thalasse	0	3	5	0	103	0
Thromboc	0	0	0	0	2	26

(iv) Evaluation – Decision Tree

Decision Tree with entropy criteria

- Accuracy: 0.9153439153439153
- Precision score: 0.9201574266982845
- Recall score: 0.9153439153439153
- F1 score: 0.9162262737348601



Confuison matrix

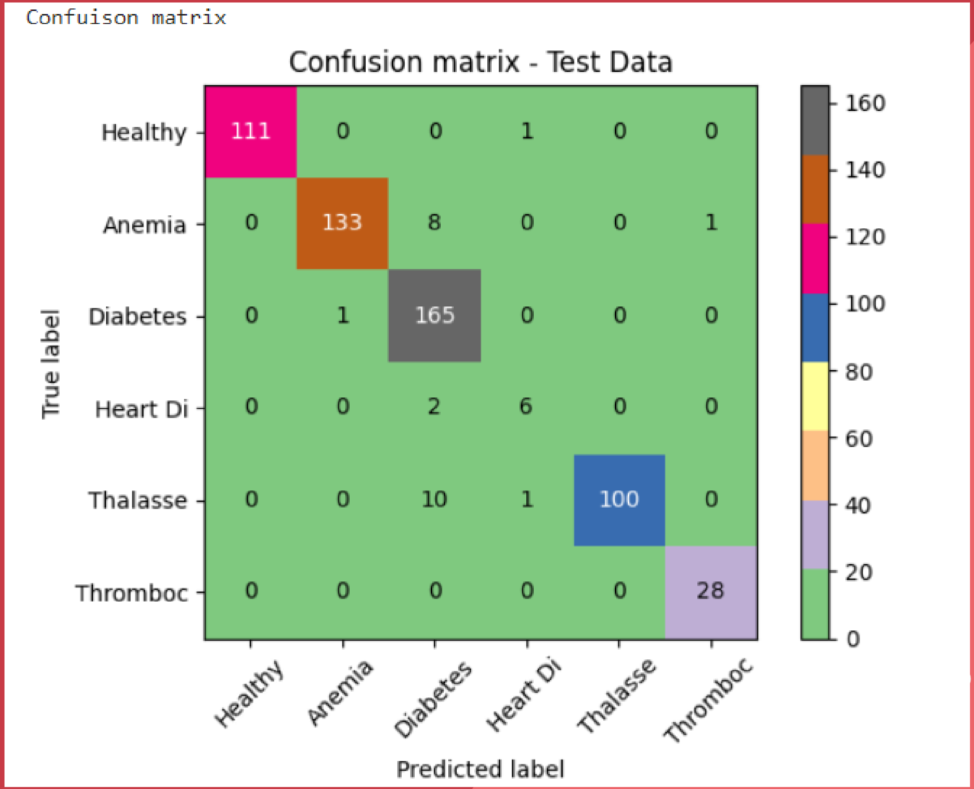
Confusion matrix - Test Data

True label \ Predicted label	Healthy	Anemia	Diabetes	Heart Di	Thalasse	Thromboc
Healthy	111	0	0	1	0	0
Anemia	0	137	4	0	0	1
Diabetes	2	13	139	5	5	2
Heart Di	0	1	2	5	0	0
Thalasse	0	3	4	1	102	1
Thromboc	0	1	0	0	2	25

(v) Evaluation – Random Forest

Random Forest:

- Accuracy: 0.9576719576719577
- Precision score: 0.9612501504764328
- Recall score: 0.9576719576719577
- F1 score: 0.9579246004438842



```
Confuison matrix
```

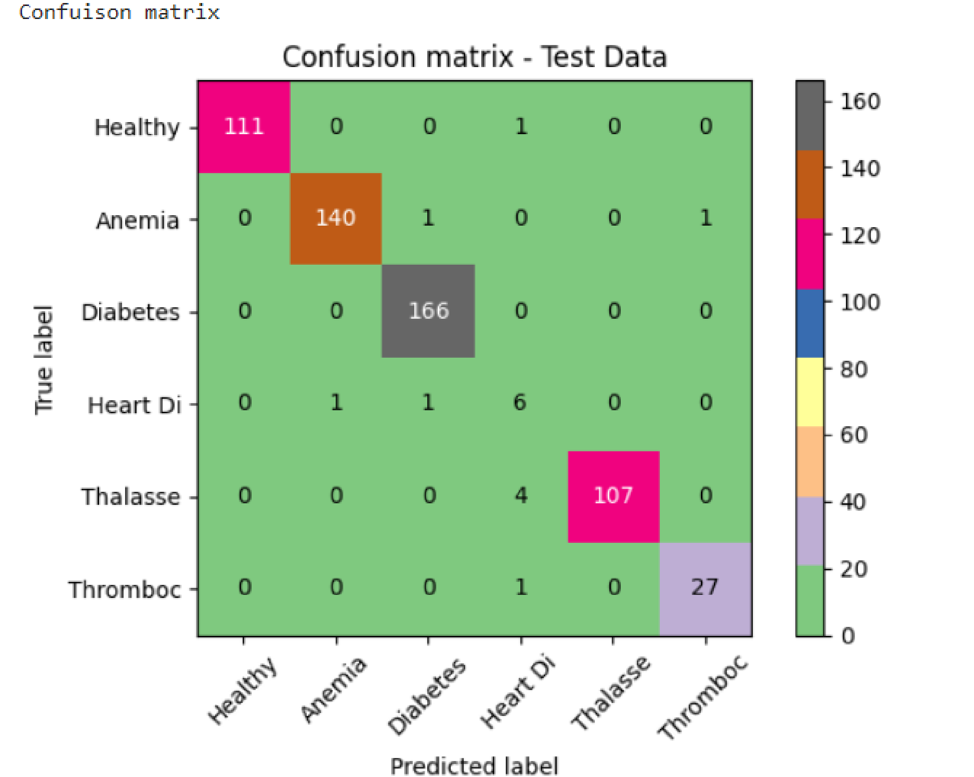
Confusion matrix - Test Data

True label \ Predicted label	Healthy	Anemia	Diabetes	Heart Di	Thalasse	Thromboc
Healthy	111	0	0	1	0	0
Anemia	0	133	8	0	0	1
Diabetes	0	1	165	0	0	0
Heart Di	0	0	2	6	0	0
Thalasse	0	0	10	1	100	0
Thromboc	0	0	0	0	0	28

(vi) Evaluation – XGBoost Classifier

XGBoost:

- Accuracy: 0.982363315696649
- Precision score: 0.9859201363760635
- Recall score: 0.982363315696649
- F1 score: 0.9837067870730555

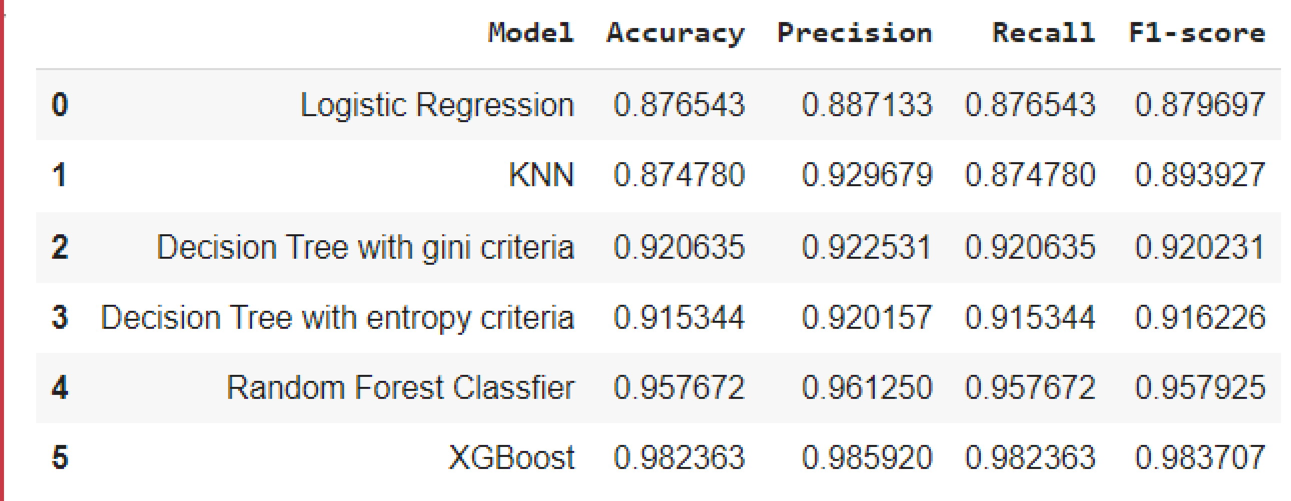


```
Confuison matrix
```

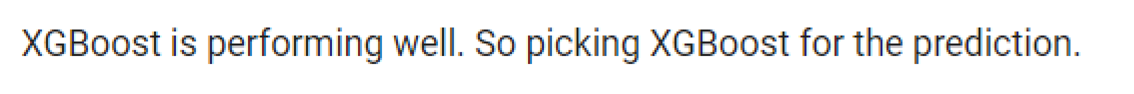
Confusion matrix - Test Data

	Healthy	Anemia	Diabetes	Heart Di	Thalasse	Thromboc
Healthy	111	0	0	1	0	0
Anemia	0	140	1	0	0	1
Diabetes	0	0	166	0	0	0
Heart Di	0	1	1	6	0	0
Thalasse	0	0	0	4	107	0
Thromboc	0	0	0	1	0	27

Comaprison of models and Choosing the best



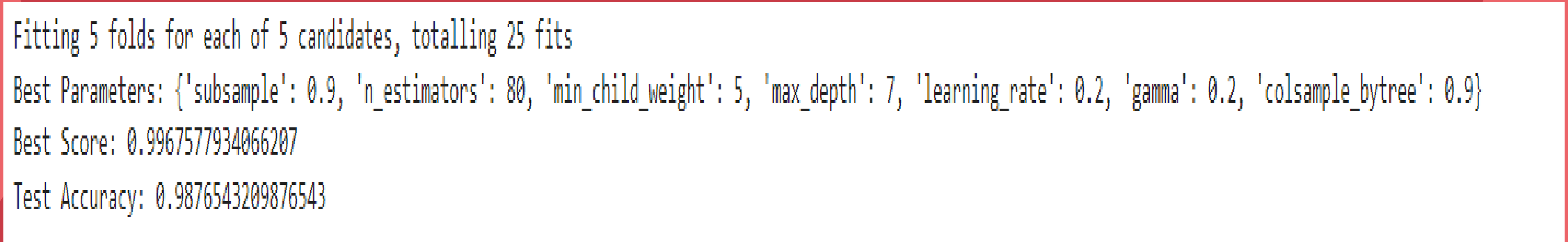
	Model	Accuracy	Precision	Recall	F1-score
0	Logistic Regression	0.876543	0.887133	0.876543	0.879697
1	KNN	0.874780	0.929679	0.874780	0.893927
2	Decision Tree with gini criteria	0.920635	0.922531	0.920635	0.920231
3	Decision Tree with entropy criteria	0.915344	0.920157	0.915344	0.916226
4	Random Forest Classfier	0.957672	0.961250	0.957672	0.957925
5	XGBoost	0.982363	0.985920	0.982363	0.983707



XGBoost is performing well. So picking XGBoost for the prediction.

4. Model Tuning

Best parameters after model tuning :

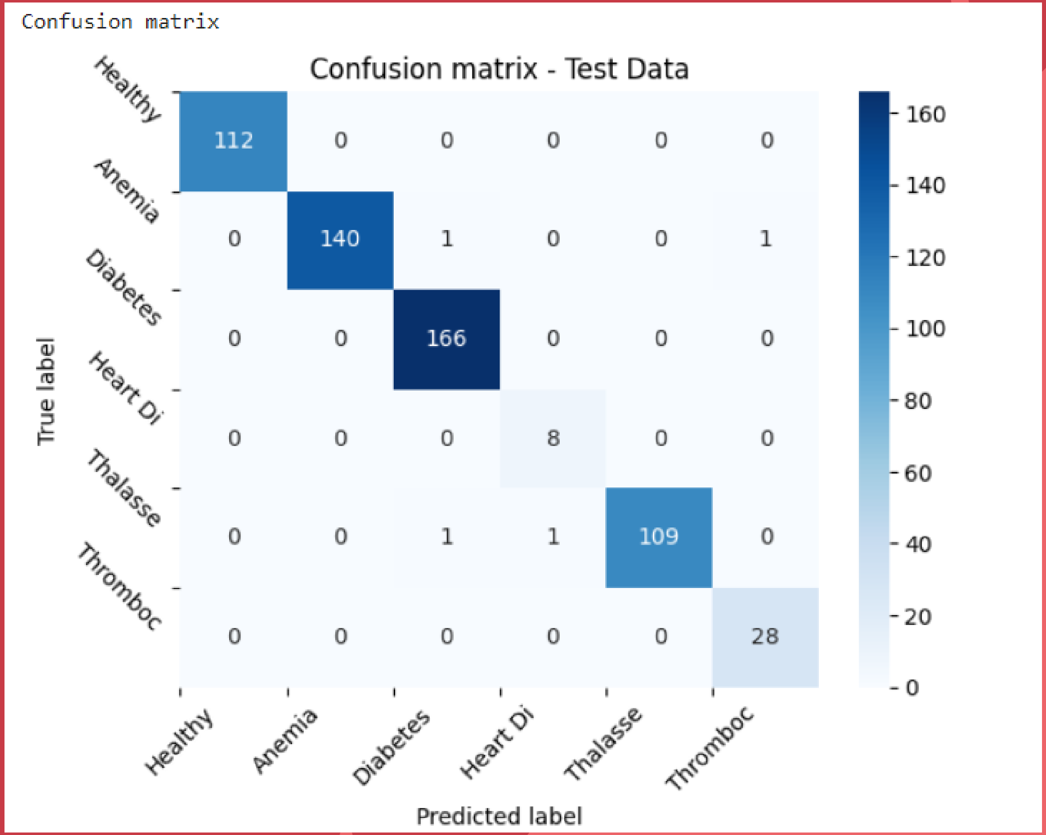


```
Fitting 5 folds for each of 5 candidates, totalling 25 fits
Best Parameters: {'subsample': 0.9, 'n_estimators': 80, 'min_child_weight': 5, 'max_depth': 7, 'learning_rate': 0.2, 'gamma': 0.2, 'colsample_bytree': 0.9}
Best Score: 0.9967577934066207
Test Accuracy: 0.9876543209876543
```

5. Creation of final prediction model

XGBoost after model tuning:

- Accuracy: 99.29 %
-
- Precision score: 99.32 %
- Recall score: 99.29 %
- F1-score: 99.30 %



Confusion matrix

	Healthy	Anemia	Diabetes	Heart Di	Thalasse	Thromboc
Healthy	112	0	0	0	0	0
Anemia	0	140	1	0	0	1
Diabetes	0	0	166	0	0	0
Heart Di	0	0	0	8	0	0
Thalasse	0	0	1	1	109	0
Thromboc	0	0	0	0	0	28

6. Conclusion

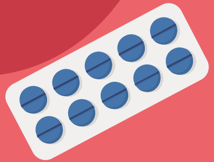
- As the given dataset is imbalanced, oversampling using SMOTE is performed after StratifiedKFold.

- After oversampling, different models are trained on this dataset and performances are observed.
-
- Of all the models, XGBoost performs better.

- Then model is tuned to get best hyper parameters.

- Then finally I created XGBoost model with the found best hyperparameters.

- Test Set Evaluation:
  Accuracy: 99.29 %
  Precision: 99.32 %
  Recall: 99.29 %
  - F1-score: 99.30 %



THANK YOU