The Hong Kong Polytechnic University

Department of Electrical and Electronics Engineering

EIE4430 Honours Project

2024-2025 Semester 1

Student Name: Chan Hou Ting Constant (21034774d)

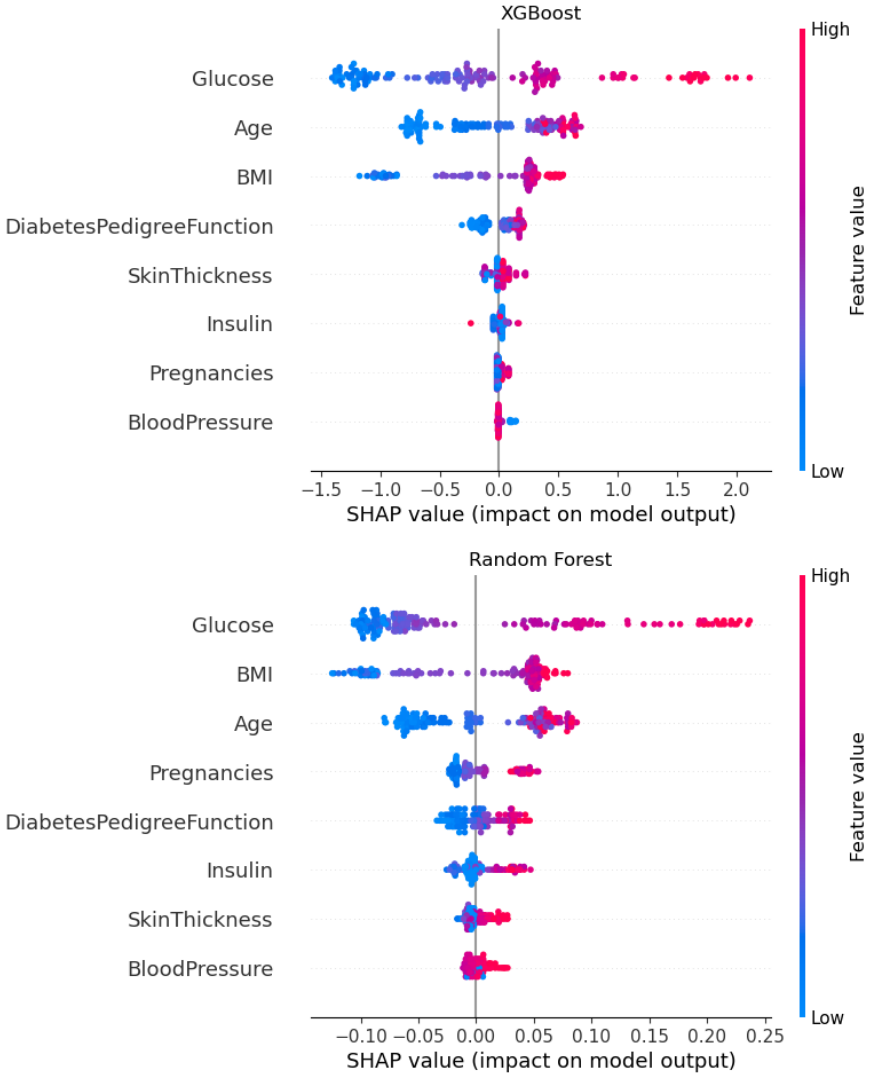
Project Title: **Machine learning model to predict the risk of diabetes**

Progress Report (1/11/2024)

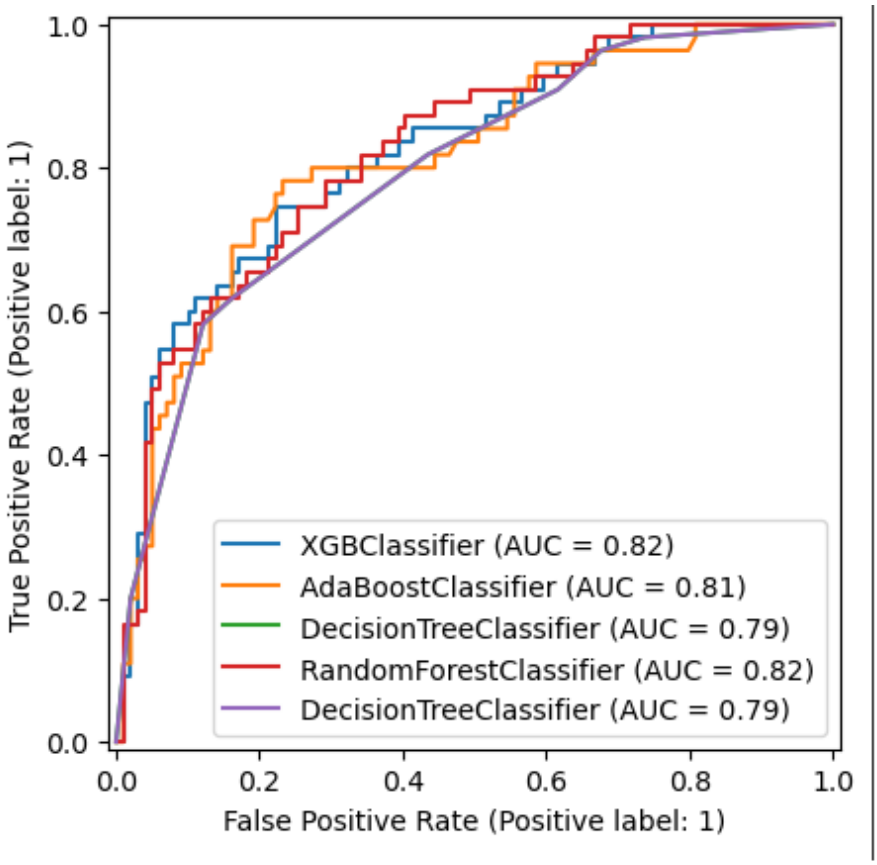
Works did in past month

I added AdaBoost and Decision Tree to the codebook, also applied SHAP value (summary plot) and ROC to each model to see feature importance and performance of the models. There were some troubles when I applied SHAP value to AdaBoost as the shap package did not support the AdaBoost. Then, I sorted the codebook to made it more understandable and applied the model to the dataset for the prediction. I found the accuracy of the prediction on the dataset and test set were close to each other. The following tasks were to solve the problem of applying SHAP value to AdaBoost and found the suitable hypermeters to get the best results.

SHAP value


XGBoost
Glucose
Age
BMI
DiabetesPedigreeFunction
SkinThickness
Insulin
Pregnancies
BloodPressure
−1.5 −1.0 −0.5 0.0 0.5 1.0 1.5 2.0
SHAP value (impact on model output)
High
Feature value
Low
Random Forest
Glucose
BMI
Age
Pregnancies
DiabetesPedigreeFunction
Insulin
SkinThickness
BloodPressure
−0.10 −0.05 0.00 0.05 0.10 0.15 0.20 0.25
SHAP value (impact on model output)
High
Feature value
Low

ROC


True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0
XGBClassifier (AUC = 0.82)
AdaBoostClassifier (AUC = 0.81)
DecisionTreeClassifier (AUC = 0.79)
RandomForestClassifier (AUC = 0.82)
DecisionTreeClassifier (AUC = 0.79)

Accuracy (Prediction vs test set)

XGBoost

```
              precision    recall  f1-score   support

           0       0.81      0.85      0.83        99
           1       0.70      0.64      0.67        55

    accuracy                           0.77       154
   macro avg       0.75      0.74      0.75       154
weighted avg       0.77      0.77      0.77       154

[[84 15]
 [20 35]]
Training Accuracy: 0.8110749185667753
Validation Accuracy: 0.7727272727272727
```

```
Differences between Original Result and Prediction:
    Outcome
       self other
6       1.0   0.0
9       1.0   0.0
15      1.0   0.0
17      1.0   0.0
23      1.0   0.0
..      ...   ...
731     1.0   0.0
739     1.0   0.0
744     0.0   1.0
750     1.0   0.0
756     0.0   1.0

[151 rows x 2 columns]
```

~= 80.36% Accuracy [100-(151(predict error)/769(total count))*100]

Random Forest

```
              precision    recall  f1-score   support

           0       0.78      0.92      0.85        99
           1       0.79      0.55      0.65        55

    accuracy                           0.79       154
   macro avg       0.79      0.73      0.75       154
weighted avg       0.79      0.79      0.77       154

[[91  8]
 [25 30]]
Training Accuracy: 0.7964169381107492
Validation Accuracy: 0.7857142857142857

 Differences between Original Result and Prediction:
     Outcome
        self other
2        1.0   0.0
6        1.0   0.0
9        1.0   0.0
15       1.0   0.0
16       1.0   0.0
..       ...   ...
749      1.0   0.0
750      1.0   0.0
756      0.0   1.0
757      1.0   0.0
766      1.0   0.0

 [158 rows x 2 columns]
```

~= 79.45% Accuracy [100-(158/769)*100]

Logistic Regression

```
              precision    recall  f1-score   support

           0       0.81      0.81      0.81        99
           1       0.65      0.65      0.65        55

    accuracy                           0.75       154
   macro avg       0.73      0.73      0.73       154
weighted avg       0.75      0.75      0.75       154

[[80 19]
 [19 36]]
Training Accuracy: 0.7703583061889251
Validation Accuracy: 0.7532467532467533

 Differences between Original Result and Prediction:
     Outcome
        self other
6       1.0   0.0
7       0.0   1.0
9       1.0   0.0
12      0.0   1.0
15      1.0   0.0
..      ...   ...
744     0.0   1.0
750     1.0   0.0
755     1.0   0.0
757     1.0   0.0
766     1.0   0.0

 [179 rows x 2 columns]
```

~= 76.72% Accuracy [100-(179/769)*100]

AdaBoost

```
              precision    recall  f1-score   support

           0       0.84      0.81      0.82        99
           1       0.68      0.73      0.70        55

    accuracy                           0.78       154
   macro avg       0.76      0.77      0.76       154
weighted avg       0.78      0.78      0.78       154

[[80 19]
 [15 40]]
Training Accuracy: 0.8029315960912052
Validation Accuracy: 0.7792207792207793


Differences between Original Result and Prediction:
    Outcome
       self other
6       1.0   0.0
9       1.0   0.0
15      1.0   0.0
17      1.0   0.0
19      1.0   0.0
..      ...   ...
731     1.0   0.0
739     1.0   0.0
744     0.0   1.0
756     0.0   1.0
757     1.0   0.0

[155 rows x 2 columns]
```

~= 79.84% Accuracy [100-(155/769)*100]

Decision Tree

```
              precision    recall  f1-score   support

           0       0.80      0.84      0.82        99
           1       0.68      0.62      0.65        55

    accuracy                           0.76       154
   macro avg       0.74      0.73      0.73       154
weighted avg       0.76      0.76      0.76       154

[[83 16]
 [21 34]]
Training Accuracy: 0.7768729641693811
Validation Accuracy: 0.7597402597402597

 Differences between Original Result and Prediction:
     Outcome
        self other
6       1.0   0.0
9       1.0   0.0
15      1.0   0.0
16      1.0   0.0
17      1.0   0.0
..      ...   ...
740     1.0   0.0
744     0.0   1.0
756     0.0   1.0
757     1.0   0.0
766     1.0   0.0

[174 rows x 2 columns]
```

~= 77.37% Accuracy [100-(174/769)*100]