The Hong Kong Polytechnic University

Department of Electrical and Electronics Engineering

EIE4430 Honours Project

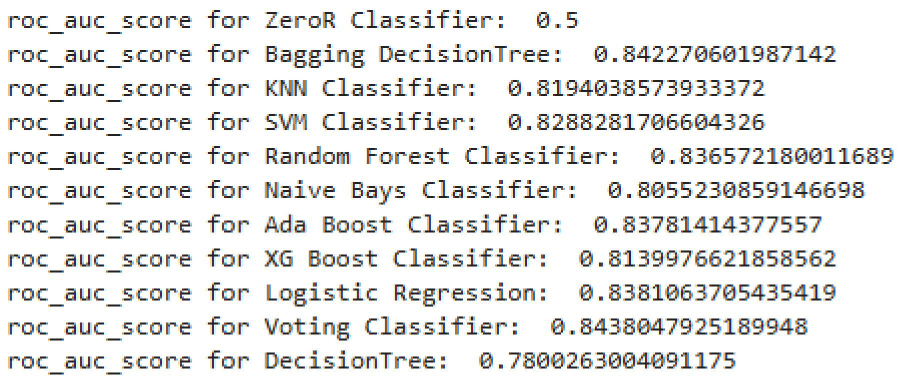
2024-2025 Semester 1

Student Name: Chan Hou Ting Constant (21034774d)

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

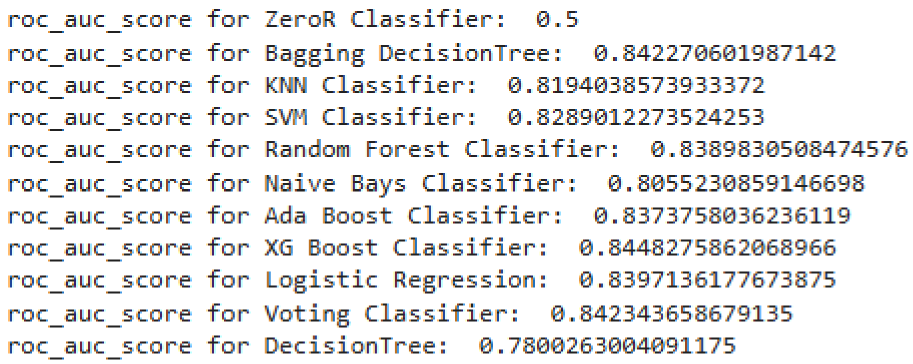
Progress Report (1/1/2025)

I have done the preprocessing on the dataset called "2013-2014 NHANES dataset". The reason I used this dataset is I found that most of research paper used mainly Pima Indian Diabetes dataset and their prepared dataset (which is usually private that are not open access for public). Pima Indian Diabetes dataset and 2013-2014 NHANES dataset are used to make a comparison of different datasets such as the model performance. The difficulty of preprocessing on 2013-2014 NHANES dataset is it is divided into 5 raw data and they have lots of features in each raw data. In addition, the selected features I picked are similar between these datasets to try to make a fair comparison. Also, I found that the result of baseline model that wrote in the paper is different from the baseline model that I reproduced, and only XG Boost have this situation. In this stage, I found that Random Forest perform better in Pima Indian Diabetes dataset and XG Boost perform well in 2013-2014 NHANES dataset.



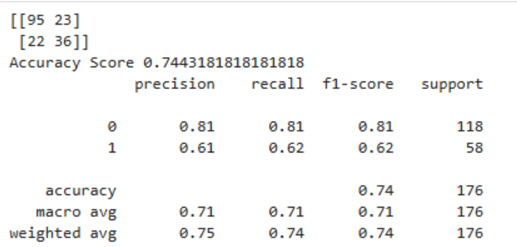
```
roc_auc_score for ZeroR Classifier:  0.5
roc_auc_score for Bagging DecisionTree:  0.842270601987142
roc_auc_score for KNN Classifier:  0.8194038573933372
roc_auc_score for SVM Classifier:  0.8288281706604326
roc_auc_score for Random Forest Classifier:  0.836572180011689
roc_auc_score for Naive Bays Classifier:  0.8055230859146698
roc_auc_score for Ada Boost Classifier:  0.83781414377557
roc_auc_score for XG Boost Classifier:  0.8139976621858562
roc_auc_score for Logistic Regression:  0.8381063705435419
roc_auc_score for Voting Classifier:  0.8438047925189948
roc_auc_score for DecisionTree:  0.7800263004091175
```

Reproduce Result (AUC)



```
roc_auc_score for ZeroR Classifier:  0.5
roc_auc_score for Bagging DecisionTree:  0.842270601987142
roc_auc_score for KNN Classifier:  0.8194038573933372
roc_auc_score for SVM Classifier:  0.8289012273524253
roc_auc_score for Random Forest Classifier:  0.8389830508474576
roc_auc_score for Naive Bays Classifier:  0.8055230859146698
roc_auc_score for Ada Boost Classifier:  0.8373758036236119
roc_auc_score for XG Boost Classifier:  0.8448275862068966
roc_auc_score for Logistic Regression:  0.8397136177673875
roc_auc_score for Voting Classifier:  0.842343658679135
roc_auc_score for DecisionTree:  0.7800263004091175
```

Paper Result (AUC)



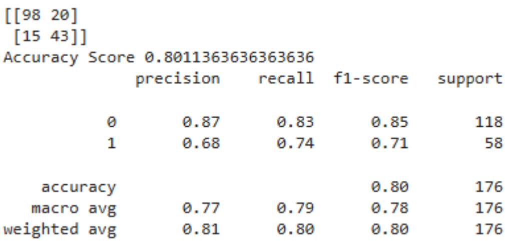
```
[[95 23]
 [22 36]]
Accuracy Score 0.7443181818181818
              precision    recall  f1-score   support

           0       0.81      0.81      0.81       118
           1       0.61      0.62      0.62        58

    accuracy                           0.74       176
   macro avg       0.71      0.71      0.71       176
weighted avg       0.75      0.74      0.74       176
```

Reproduce Result

(XGB+ADASYN)

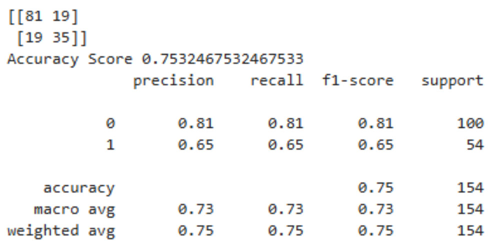


```
[[98 20]
 [15 43]]
Accuracy Score 0.8011363636363636
              precision    recall  f1-score   support

           0       0.87      0.83      0.85       118
           1       0.68      0.74      0.71        58

    accuracy                           0.80       176
   macro avg       0.77      0.79      0.78       176
weighted avg       0.81      0.80      0.80       176
```

Paper Result (XGB+ADASYN)

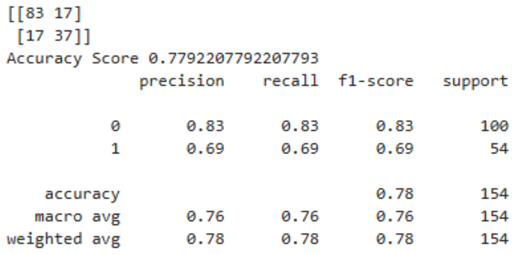


```
[[81 19]
 [19 35]]
Accuracy Score 0.7532467532467533
              precision    recall  f1-score   support

           0       0.81      0.81      0.81       100
           1       0.65      0.65      0.65        54

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

Preliminary Result (XG Boost) (Pima Indian Diabetes dataset)

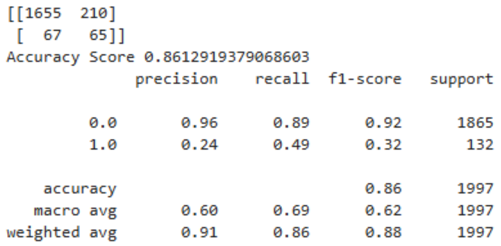


```
[[83 17]
 [17 37]]
Accuracy Score 0.7792207792207793
              precision    recall  f1-score   support

           0       0.83      0.83      0.83       100
           1       0.69      0.69      0.69        54

    accuracy                           0.78       154
   macro avg       0.76      0.76      0.76       154
weighted avg       0.78      0.78      0.78       154
```

Preliminary Result (Random Forest) (Pima Indian Diabetes dataset)`

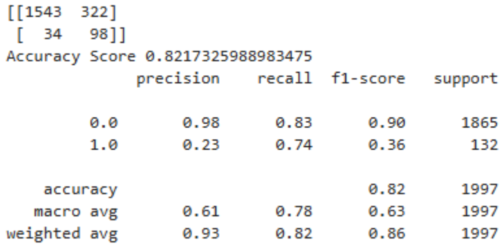


```
[[1655  210]
 [  67   65]]
Accuracy Score 0.8612919379068603
              precision    recall  f1-score   support

         0.0       0.96      0.89      0.92      1865
         1.0       0.24      0.49      0.32       132

    accuracy                           0.86      1997
   macro avg       0.60      0.69      0.62      1997
weighted avg       0.91      0.86      0.88      1997
```

Preliminary Result (XG Boost) (NHANES dataset)



```
[[1543  322]
 [  34   98]]
Accuracy Score 0.8217325988983475
              precision    recall  f1-score   support

         0.0       0.98      0.83      0.90      1865
         1.0       0.23      0.74      0.36       132

    accuracy                           0.82      1997
   macro avg       0.61      0.78      0.63      1997
weighted avg       0.93      0.82      0.86      1997
```

Preliminary Result (Random Forest) (NHANES dataset)