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 (1/2/2025)

I apply mean normalization and min-max scaling in the feature scaling section to compare the model performance with different preprocessing methods. The hyperparameters are the same as the baseline model. In the testing, random forest performs good which both mean normalization (77% Accuracy) and min-max scaling (78% Accuracy) are better than XG Boost (76% Accuracy) in Pima Indian Diabetes dataset. In 2013-2014 NHANES dataset, standardization is still better than mean normalization and min-max scaling with 88.7% Accuracy and 92% precision.

```
[[80 20]
 [15 39]]
Accuracy Score 0.7727272727272727
              precision    recall  f1-score   support

           0       0.84      0.80      0.82       100
           1       0.66      0.72      0.69        54

    accuracy                           0.77       154
   macro avg       0.75      0.76      0.76       154
weighted avg       0.78      0.77      0.77       154
```

Result (Random Forest, Min-max) (Pima Indian Diabetes dataset)`

```
[[81 19]
 [15 39]]
Accuracy Score 0.7792207792207793
              precision    recall  f1-score   support

           0       0.84      0.81      0.83       100
           1       0.67      0.72      0.70        54

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

Result (Random Forest, Mean normalization) (Pima Indian Diabetes dataset)`

```
[[1661  158]
 [  64   89]]
Accuracy Score 0.8874239350912779
              precision    recall  f1-score   support

         0.0       0.96      0.91      0.94      1819
         1.0       0.36      0.58      0.45       153

    accuracy                           0.89      1972
   macro avg       0.66      0.75      0.69      1972
weighted avg       0.92      0.89      0.90      1972
```

Result (Random Forest, Standardization) (2013-2014 NHANES dataset)

```
[[1650  169]
 [  66   87]]
Accuracy Score 0.8808316430020284
              precision    recall  f1-score   support

         0.0       0.96      0.91      0.93      1819
         1.0       0.34      0.57      0.43       153

    accuracy                           0.88      1972
   macro avg       0.65      0.74      0.68      1972
weighted avg       0.91      0.88      0.89      1972
```

Result (Random Forest, Min-max) (2013-2014 NHANES dataset)

```
[[1648  171]
 [  66   87]]
Accuracy Score 0.8798174442190669
              precision    recall  f1-score   support

         0.0       0.96      0.91      0.93      1819
         1.0       0.34      0.57      0.42       153

    accuracy                           0.88      1972
   macro avg       0.65      0.74      0.68      1972
weighted avg       0.91      0.88      0.89      1972
```

Result (Random Forest, Mean normalization) (2013-2014 NHANES dataset)

Also, I record the performance of Random Forest classifier without feature scaling. The reason I try to remove the feature scaling in Random Forest classifier is that the random forest split the data according to feature values rather than calculating distances between data points, which means it is not sensitive to the scale of the features. In the testing, I found that random forest without feature scaling, which have 78.6% Accuracy that perform good compared to the models which used standardization in Pima Indian Diabetes dataset. For the 2013-2014 NHANES dataset, the model which did the feature scaling (standardization, 88.7% Accuracy) better than the one without feature scaling (88% Accuracy).

```
[[77 23]
 [12 42]]
Accuracy Score 0.7727272727272727
              precision    recall  f1-score   support

           0       0.87      0.77      0.81       100
           1       0.65      0.78      0.71        54

    accuracy                           0.77       154
   macro avg       0.76      0.77      0.76       154
weighted avg       0.79      0.77      0.78       154
```

Result (XG Boost, Standardization) (Pima Indian Diabetes dataset)

```
[[80 20]
 [16 38]]
Accuracy Score 0.7662337662337663
              precision    recall  f1-score   support

           0       0.83      0.80      0.82       100
           1       0.66      0.70      0.68        54

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

Result (Random Forest, Standardization) (Pima Indian Diabetes dataset)

```
[[81 19]
 [14 40]]
Accuracy Score 0.7857142857142857
              precision    recall  f1-score   support

           0       0.85      0.81      0.83       100
           1       0.68      0.74      0.71        54

    accuracy                           0.79       154
   macro avg       0.77      0.78      0.77       154
weighted avg       0.79      0.79      0.79       154
```

Result (Random Forest, No Feature Scaling) (Pima Indian Diabetes dataset)

```
[[1644  175]
 [  61   92]]
Accuracy Score 0.8803245436105477
              precision    recall  f1-score   support

         0.0       0.96      0.90      0.93      1819
         1.0       0.34      0.60      0.44       153

    accuracy                           0.88      1972
   macro avg       0.65      0.75      0.69      1972
weighted avg       0.92      0.88      0.89      1972
```

Result (Random Forest, No Feature Scaling) (2013-2014 NHANES dataset)

```
[[1661  158]
 [  64   89]]
Accuracy Score 0.8874239350912779
              precision    recall  f1-score   support

         0.0       0.96      0.91      0.94      1819
         1.0       0.36      0.58      0.45       153

    accuracy                           0.89      1972
   macro avg       0.66      0.75      0.69      1972
weighted avg       0.92      0.89      0.90      1972
```

Result (Random Forest, Standardization) (2013-2014 NHANES dataset)