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 (1/3/2025)

To obtain a better result on the proposed framework on the Pima Indian Diabetes dataset, I adjusted the preprocessing stage. I found that the performance is improved when SMOTE is removed. Therefore, I used this adjustment in the proposed framework.

Also, I compared different hyperparameter settings on XGBoost, in which the baseline model was also used XGBoost. The previous result was 77%, which will be the reference for the following test to see which settings positively affect the model's performance. The yellow highlighted in the following tables are the selected hyperparameters with the highest improvement.

	0.6	**0.7**	**0.8**	**0.9**	**1**
colsample_bytree	0%	0%	0%	+4.8%	+3.5%

Table 1: XGBoost hyperparameter setting for improving model performance (1)

	0	**1**
gamma	0%	+4%

Table 2: XGBoost hyperparameter setting for improving model performance (2)

	0.05	**0.1**	**0.15**	**0.2**
learning_rate	+3.5%	+5.5%	+5.5%	+4.8%

Table 3: XGBoost hyperparameter setting for improving model performance (3)

	3	**7**	**8**	**9**	**15**
max_depth	+1.5%	+3.5%	+5.5%	+4.8%	+2.8%

Table 4: XGBoost hyperparameter setting for improving model performance (4)

	1	**2**	**3**
min_child_weight	+1.5%	0%	0%

Table 5: XGBoost hyperparameter setting for improving model performance (5)

	100	**200**	**250**	**500**	**1000**
n_estimators	+5.5%	+2.8%	+2.2%	+1.5%	0.9%

Table 6: XGBoost hyperparameter setting for improve model performance (6)

	0.8	**0.9**	**1**
subsample	+1.5%	-0.4%	+1.5%

Table 7: XGBoost hyperparameter setting for improving model performance (7)

	1	**2**	**3**
scale_pos_weight	+4%	0%	+1.5%

Table 8: XGBoost hyperparameter setting for improving model performance (8)

	Accuracy (%)	**Precision(%)**	**F1 Score (%)**	**AUC (0–1)**
Baseline (Tasin et al. [1])	81% (+1.46%)	81% (+1.46%)	81% (+1.46%)	0.84 (+0.02)
My Proposed framework (version 3)	82.46%	82%	82%	0.86
My Proposed framework (version 1)	77% (+5.46%)	79% (+3%)	78% (+4%)	0.81 (+0.05)

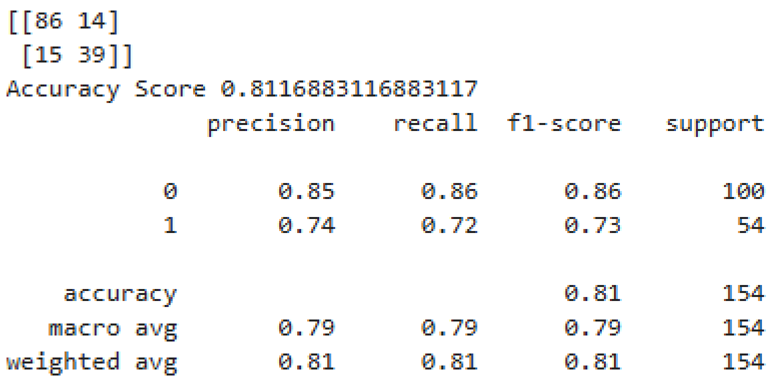
Table 9: Baseline VS Proposed framework (Pima Indian Diabetes dataset)

From the experiment, I observed that **"learning_rate"** (Table 3), **"max_depth"** (Table 4), and **"n_esimators"** (Table 6) have the most significant effect on optimizing the model, which increased by around 5% on evaluation metrics overall (Table 9).

Moreover, adjustment on **"scale_pos_weight"** (Table 8) has a noticeable improvement in the model performance after removing SMOTE. I think it is because of **"scale_pos_weight"** that it can control the balance of positive and negative weights to process the class imbalance problem in machine learning, and it is not necessary to apply the class imbalance method in the data preprocessing stage.

"

Green highlighted is the adjustment. Orange highlighted is the previous adjustment.

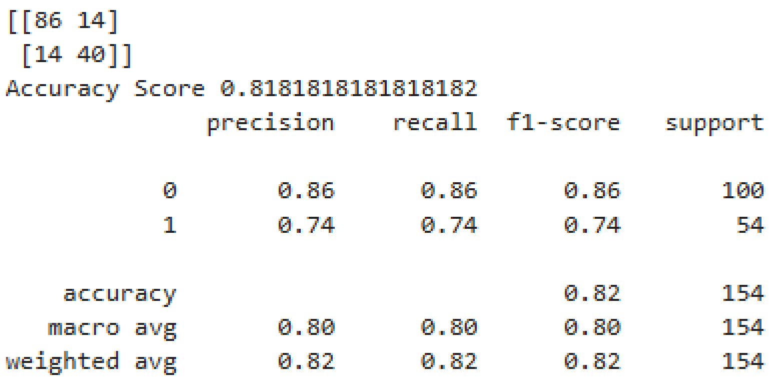


```
[[86 14]
 [15 39]]
Accuracy Score 0.8116883116883117
              precision    recall  f1-score   support

           0       0.85      0.86      0.86       100
           1       0.74      0.72      0.73        54

    accuracy                           0.81       154
   macro avg       0.79      0.79      0.79       154
weighted avg       0.81      0.81      0.81       154
```

Fig 1: Result (version 1) – Pima Indian Diabetes dataset

	Hyperparameter settings
colsample_bytree	0.8
gamma	0
learning_rate	0.2
max_depth	8
min_child_weight	1
n_estimators	100
subsample	0.8
scale_pos_weight	1

Table 10: Hyperparameter settings on Fig. 1

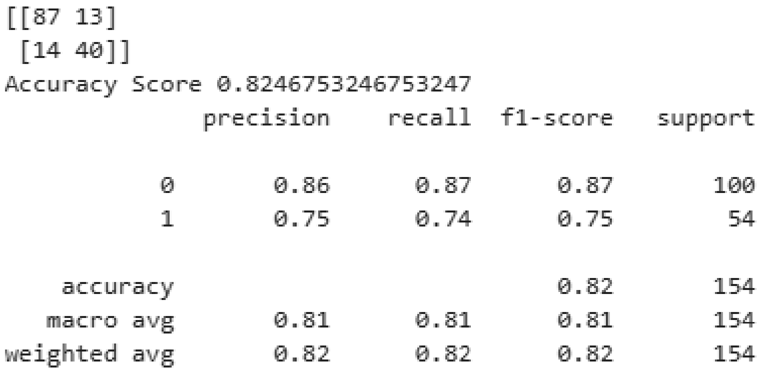


```
[[86 14]
 [14 40]]
Accuracy Score 0.8181818181818182
              precision    recall  f1-score   support

           0       0.86      0.86      0.86       100
           1       0.74      0.74      0.74        54

    accuracy                           0.82       154
   macro avg       0.80      0.80      0.80       154
weighted avg       0.82      0.82      0.82       154
```

Fig 2: Result (version 2) – Pima Indian Diabetes dataset

	Hyperparameter settings
colsample_bytree	0.8
gamma	1
learning_rate	0.2
max_depth	8
min_child_weight	1
n_estimators	100
subsample	0.8
scale_pos_weight	1

Table 11: Hyperparameter settings on Fig. 2



```
[[87 13]
 [14 40]]
Accuracy Score 0.8246753246753247
              precision    recall  f1-score   support

           0       0.86      0.87      0.87       100
           1       0.75      0.74      0.75        54

    accuracy                           0.82       154
   macro avg       0.81      0.81      0.81       154
weighted avg       0.82      0.82      0.82       154
```

Fig 3: Result (version 3) – Pima Indian Diabetes dataset

	Hyperparameter settings
colsample_bytree	0.8
gamma	1
learning_rate	0.1
max_depth	8
min_child_weight	1
n_estimators	100
subsample	0.8
scale_pos_weight	1

Table 12: Hyperparameter settings on Fig. 3

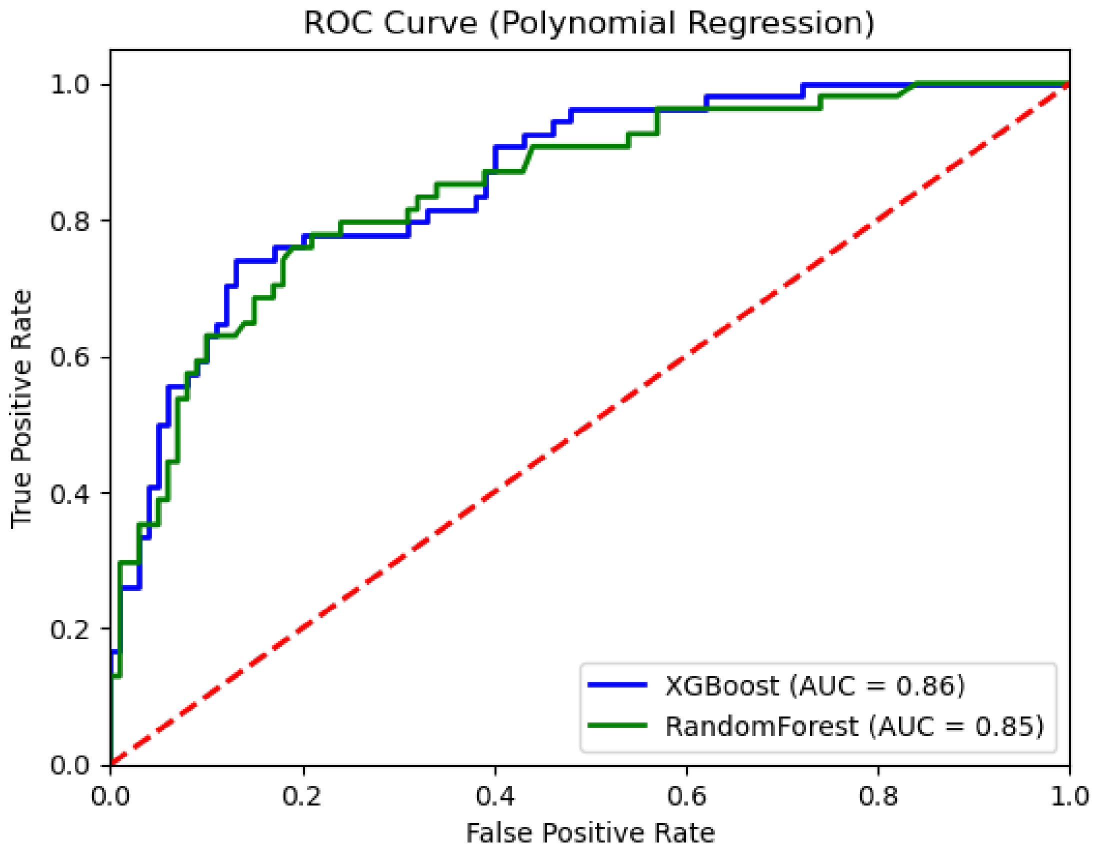


Fig 4: AUC Result (version 3) – Pima Indian Diabetes dataset

	Accuracy (%)	**Precision(%)**	**F1 Score (%)**	**AUC (0–1)**
Baseline (Qin et al. [2])	82% (+9.24%)	82% (+10%)	82% (+10%)	0.83 (+0.03)
My Proposed framework	91.24%	92%	92%	0.86
My Proposed framework (Random Forest)	89% (+2.24%)	92% (0%)	90% (+2%)	0.89 (-0.03)

Table 13: Baseline VS Proposed framework (2013-2014 NHANES dataset)

For the NHANES dataset, I added CATBoost to the proposed framework to compare model performance. In Table 13, I found that the proposed framework using CATBoost has the highest evaluation metrics, with around 91% Accuracy and an AUC of 0.86. Although the AUC is decreased compared to the proposed framework with Random Forest, the model performance is improved overall.

Reference

[1] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, “Diabetes prediction using machine learning and explainable AI techniques,” Healthcare technology letters, vol. 10, no. 1–2, pp. 1–10, 2023, doi: 10.1049/htl2.12039

[2] Y. Qin et al., “Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type,” International journal of environmental research and public health, vol. 19, no. 22, pp. 15027-, 2022, doi: 10.3390/ijerph192215027