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 (1/4/2025)

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

Table 1: Baseline VS Proposed framework (Pima Indian Diabetes dataset)

The proposed Framework used in the Pima Indian Diabetes dataset uses Polynomial Regression to predict missing values of “SkinThickness” and “Insulin”, and the rest of the features are filled by their mean, with no class imbalance technique used.

As mentioned last month, adjustment on “scale_pos_weight” has a noticeable improvement in the model performance after removing SMOTE (a technique used for process class imbalance). I confirmed that adjusting “scale_pos_weight” can be substituted with using class imbalance methods as I think the function of “scale_pos_weight” and class imbalance methods because they are overlapping that both can control the balance of positive and negative weights to process the class imbalance problem in machine learning, but “scale_pos_weight” performs better. Therefore, using the class imbalance technique (e.g., SMOTE) is another possible option if machine learning does not provide the hyperparameters that can control the balance of positive and negative weights to process the class imbalance.

| | **Accuracy (%)** | **Precision(%)** | **F1 Score (%)** | **AUC (0–1)** |
|---|---|---|---|---|
| Baseline (CatBoost, Qin et al. [2]) | 82% (+11.94%) | 82% (+11%) | 82% (+11%) | 0.83 (+0.08) |
| My Proposed framework (CatBoost) | 92.34% (+1.6%) | 92 (+1%) | 92 (+1%) | 0.88 (+0.3) |
| My Proposed framework (XGBoost, Finalized) | **93.94%** | **93%** | **93%** | **0.91** |
| My Proposed framework (1/3/2025, Random Forest) | 91.24% (+2.7%) | 92% (+1%) | 92% (+1%) | 0.86 (+0.05) |

Table 2: Baseline VS Proposed framework (2013-2014 NHANES dataset)

The proposed framework used in the 2013-2014 NHANES dataset uses Mean to predict all features of the missing values, and their Mean fills the rest of the features. No class imbalance technique was applied.

For the NHANES dataset, I added CATBoost to the proposed framework to compare model performance. In Table 1, XGBoost performs better than it gains 93.94% accuracy with an AUC of 0.91. Also, I found that the proposed framework using CATBoost is better than the baseline model proposed by Qin et al. [2], which has 92.34% Accuracy and an AUC of 0.88. **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