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.

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

Reproduce Result (AUC)

一張含有 文字, 螢幕擷取畫面, 字型, 文件 的圖片

自動產生的描述

Paper Result (AUC)

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

Paper Result (XGB+ADASYN)

Reproduce Result (XGB+ADASYN)

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

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

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

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

Preliminary Result (XG Boost) (NHANES dataset)

Preliminary Result (Random Forest) (NHANES dataset)