The Hong Kong Polytechnic University

Department of Electrical and Electronics Engineering

EIE4430 Honours Project

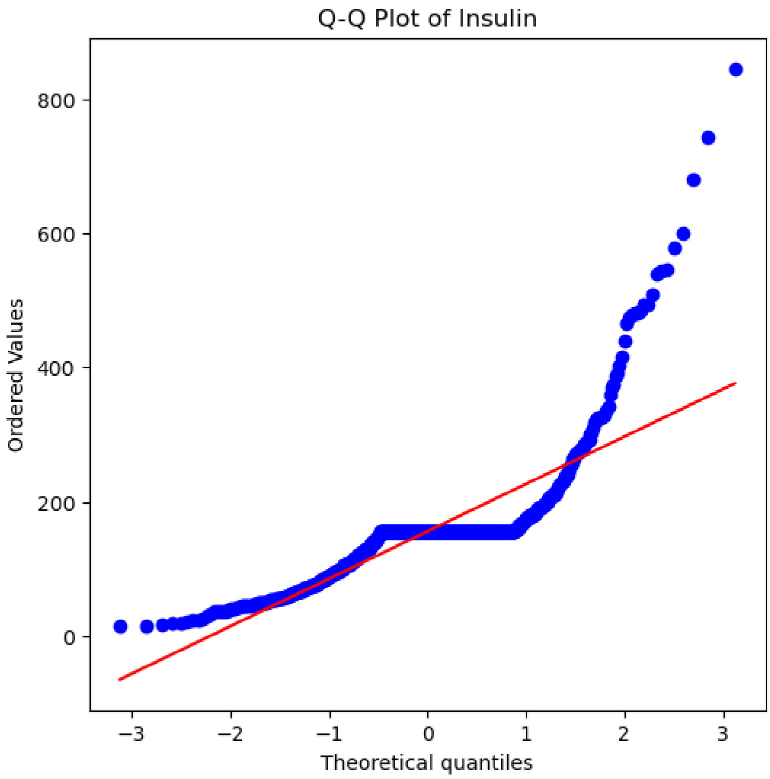
2024-2025 Semester 1

Student Name: Chan Hou Ting Constant (21034774d)

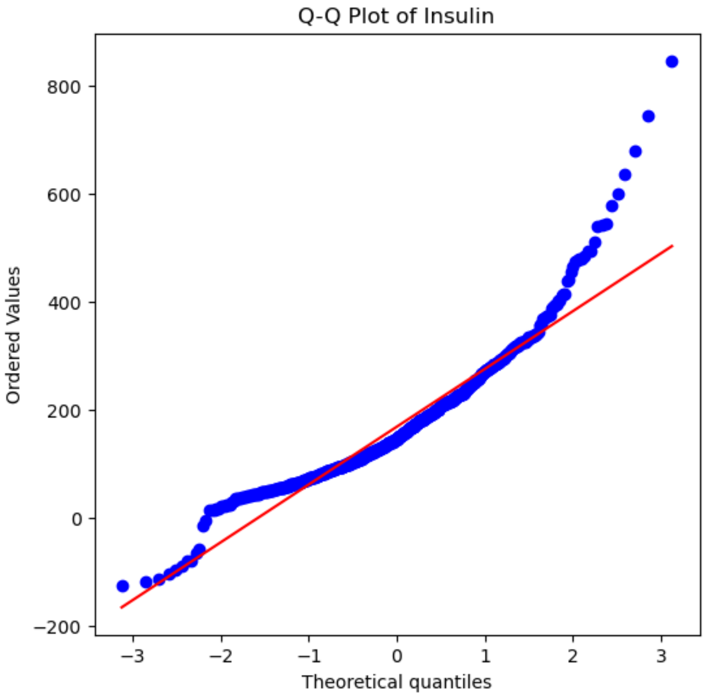
Project Title: **Machine learning model to predict the risk of diabetes**

Progress Report (1/12/2024)

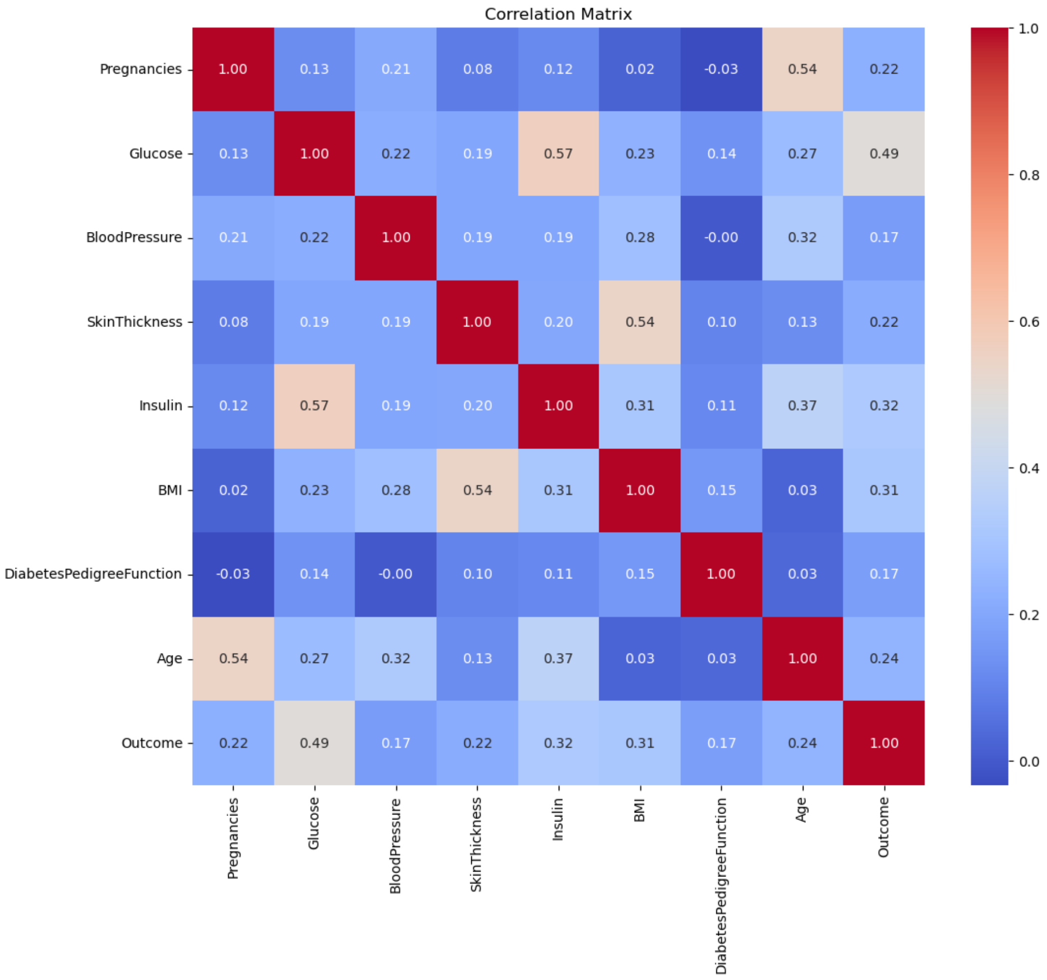
For the dataset, I have modified the existing methods used in the pre-processing section. Most of cases, I filled the missing values by its mean. But I found that Insulin had too much missing value that it constitute around 50% of the total. If filled the missing values on the Insulin by its mean, q-q plot of the Insulin looks weird that the insulin is not a normal distribution. Therefore, I tried the polynomial regression to predict the Insulin' value which are the missing value. Polynomial regression have been used by some paper and it can increase the model performance when applied in the pre-processing. I would try another method like Datawig and compare the result on the q-q plot.



Fill missing value by its mean



Fill missing value by Polynomial Regression



After filled the missing value and compute the correlation matrix, I selected 'Glucose', 'Insulin', 'BMI', 'Age' for doing the prediction because they have high score on ‘Outcome’ class on the correlation matrix. For the Feature Scaling, I tried Standardization and applied SMOTE to process Imbalance data. I would like to applied other Feature Scaling methods like Min-Max Scaling and compare the results.