



Supervised Learning Project

By

Zhiyun(Jenny) Liang



Project Objective

- Use supervised learning techniques to build a machine learning model that can predict whether a patient has diabetes or not, based on certain diagnostic measurements.



Project Flow Structure

- Exploratory data analysis
- Preprocessing and feature engineering
- Training machine learning models
- Results and Discussion



Exploratory Data Analysis

- Exploratory data (null value, outlier, etc.)
- Distribution of each predictor variable
- Correlation between the predictor variables



Preprocessing and Feature Engineering

- Handling missing values
- Handling outliers
- Scaling and normalization variable



Training ML Model

- Training Decision Tree Model
- Training Random Forest Model



Results and Discussion

- 'SkinThickness', 'Insulin' column around 1/3 values are missing, and 'Insulin' has a lot of outliers
- 'SkinThickness', 'BloodPressure' column has a lowest correlation with 'Outcome', which is 0.075, 0.065 respectively
- After Scaling the features, the model performance is improved
- The performance of Random Forest Model is better than Decision Tree Model