Pattern Recognition and Machine Learning (11482) Final Report

Using Machine Learning to Predict Diabetes

Terence Lam (u3206488)
Uyen Nguyen(u3206201)

Project and Problem Summary

- Diabetes is a common disease in nowadays society
- As in 2019, Approximately 1.5 million people die to diabetes (WHO 2021)
- Objective: use machine learning to diagnose and predict existing or potential diabetes disease on a patient

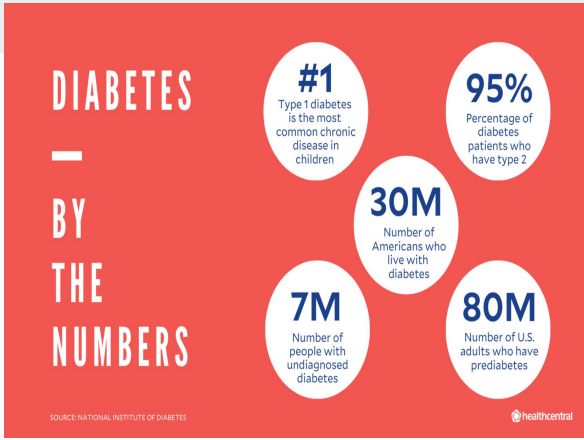

Figure 1: Diabetes by the numbers (WriterMarch 10 et al., n.d.)

Dataset

- Pima Indians Diabetes Database
- From the National Institute of Diabetes and Kidney Digestive
- 786 instances
- 9 attributes (features): pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age, and outcome (label - 0, 1)
- Clean up the dataset

Methodology

1. Input the dataset
2. Pre-processing
3. EDA
4. Scale the dataset
5. Experiment the training dataset with models
6. Evaluate the score with evaluation strategies

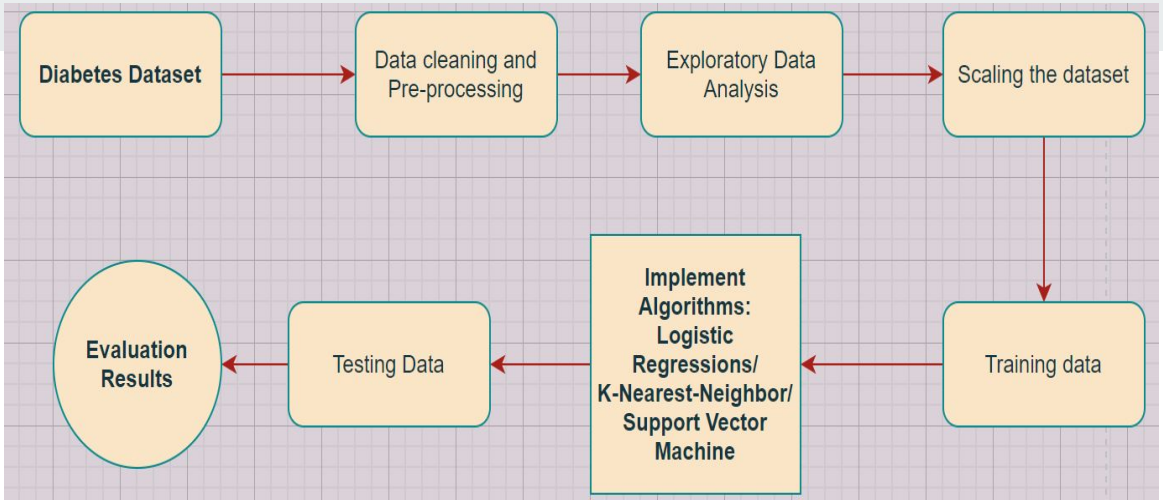

Figure 2: Methodology and procedure of the project

Evaluation Strategies:

- Accuracy
- ROC curve - AUC
- F1-score

7. Compare and analyse the results

Pre-processing

The modification process:

- Depending on the values distribution of each variables:
- Replace “0” values with mean in Glucose, Blood Pressure, Skin Thickness, BMI.
- Replace “0” values with median as for Insulin.
- Cut down the variable named Diabetes Pedigree Functions
- Getting rid of the outliners

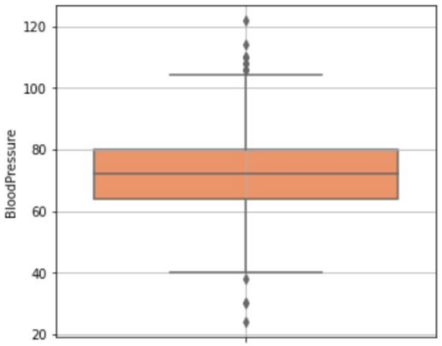

Figure 3

Exploratory Data Analysis

Relationship interpretation:

- The person who has high level of BMI, Glucose, Skinthickness can get diabetic.

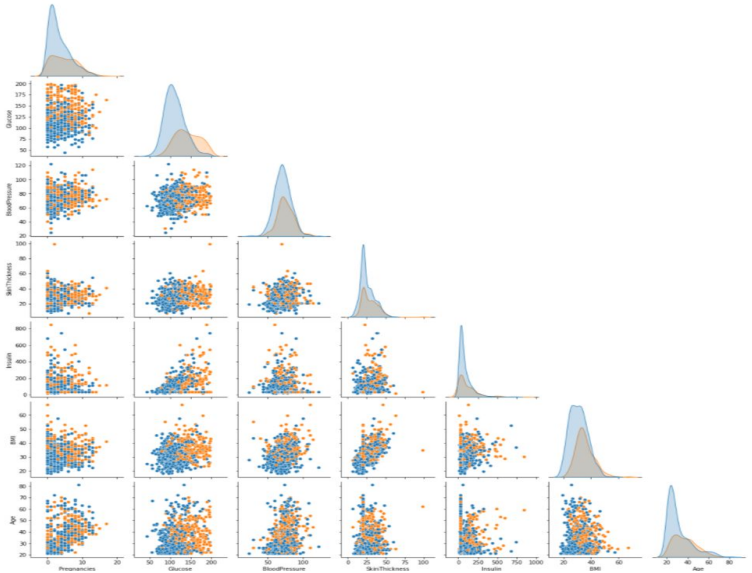

Figure 4

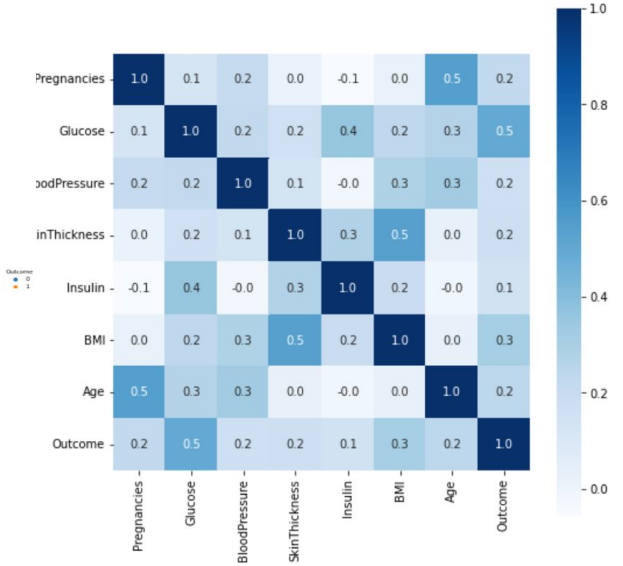

Figure 5

Processing the Dataset

- Standard Scale the dataset to standardize the variables
- Processing the data with 80% training and 20% testing

Models implementation and Hyperparameters tuning

Classification report

Accuracy

Precision

Recall

F1-score

AUC

Logistic Regression	C=10
K-Nearest-Neighbor	{'metric': 'euclidean', 'n_neighbors': 25}
Support Vector Machine	{'C': 100, 'gamma': 0.001, 'kernel': 'rbf'}

Figure 6

Model	Accuracy	F1-score	AUC-score
Logistic Regression	0.7237	0.72	0.84
KNN	0.7960	0.80	0.79
SVM	0.7829	0.78	0.82
Tuning LR	0.7649	0.80	0.7397
Tuning KNN	0.7549	0.75	0.6857
Tuning SVM	0.7516	0.80	0.7447

Figure 7

Contribution and future works

Contributions

- People can improve their health by applying their variables statistic on the predictive models
- Helpful in future research in the medical field

Future works

- Lack of time and experience
- Experiment on other learning models and evaluation strategies
- Try different variations of learning models and evaluation strategies used in this project e.g. time-dependent ROC curve, amalgam KNN
- Experiment on different datasets

Thank you for listening!

Teamwork Contributions

Uyen: Data processing and model learning.

Terence: Data cleaning and formatting.

The whole team: Powerpoint and report preparation.

Reference

https://www.healthcentral.com/condition/diabetes

https://towardsdatascience.com/how-and-why-to-standardize-your-data-996926c2c832