Title
Prediction of Pima Indians Diabetes with Machine Learning Algorithms

Author
Huang, Ruodi

Publication Date
2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Prediction of Pima Indians Diabetes with Machine Learning Algorithms

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Applied Statistics

by

Ruodi Huang

2021

ABSTRACT OF THE THESIS

Prediction of Pima Indians Diabetes
with Machine Learning Algorithms

by

Ruodi Huang

Master of Applied Statistics in
University of California, Los Angeles, 2021
Professor Yingnian Wu, Chair

Diabetes mellitus is a chronic disease that occurs when one’s pancreas no longer able to produce enough insulin. The long-term hyperglycemia during diabetes causes chronic damage and dysfunction of various tissues, especially the eyes, kidneys, heart, blood vessels, and nerves. Nowadays, diabetes is a major public health challenge and a worldwide problem. This paper will introduce how to use medical data to predict an individual's diabetes with machine learning tools. The extreme gradient boosting which is the final model suggests that top five important features which lead to high probability of diabetes are ‘DiabetesPedigreeFunction’ ‘Pregnancies’ ‘BMI’ ‘Glucose’ and ‘Insulin’.

Content

List of Figures

List of Tables

CHAPTER 1 Introduction

1.1 Background

Diabetes mellitus is a chronic disease that occurs when one's pancreas no longer able to produce enough insulin. The long-term hyperglycemia during diabetes causes chronic damage and dysfunction of various tissues, especially the eyes, kidneys, heart, blood vessels, and nerves. Nowadays, diabetes is a major public health challenge and a worldwide problem. This study will introduce how to use medical data to analyze the relation between medical indexes and diabetes with machine learning tools. It may be helpful to doctors as a screening tool.

1.2 Data Source

This dataset is taken from UCI Machine Learning Repository and it is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether a patient has diabetes or not, based on certain diagnostic measurements included in the dataset.

Several constraints were placed on the data selection. All patients here are females at least 21 years old of Pima Indian heritage. Pima Indian is 'North American Indians who traditionally lived along the Gila and Salt rivers in Arizona, U.S.'. They are generally considered to be famers who live on crops. And 'some hunting and gathering were done to supplement the diet.'

1.3 Outline

In this study, chapter two will provide data overview. Chapter three is about exploratory data analysis of Pima Indian dataset. Chapter four will illustrate methodology and model training. Chapter five will make evaluation and comparison of modeling results. Conclusion will be discussed in the last chapter.

CHAPTER 2 Data processing

This chapter will make an overview of Pima Indians diabetes dataset. The preprocessing to data contains missing value processing and data scaling. Then the dataset will be split into training set and test set, where random oversampling will be applied to rebalance the training data after.

2.1 Data Overview

The dataset in this study contains 768 observations. Each row represents a patient. The first three rows of the dataset are showed in the Table 2.1.

Table 2.1 Header of Data Type Table

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	DiabetesPedigree Function	Age	Outcome
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1

There are 8 independent variables and one dependent variable. And the independent variables are several medical features of patients. The description of each feature is listed below.

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration after 2 hours in an oral glucose tolerance test

  The glucose tolerance test, also known as the oral glucose tolerance test, measures your body's response to sugar (glucose). The glucose tolerance test can be used to screen for type 2 diabetes.
- BloodPressure: Diastolic blood pressure (mm Hg)

'Blood pressure is the force that moves blood through our circulatory system'1. Both high blood pressure and low blood pressure can have serious consequences, and the disappearance of blood pressure is a precursor to death.

- SkinThickness: Triceps skin fold thickness (mm)

  The variable of 'SkinThickness' means triceps skin fold thickness, it provides a good estimate of obesity and body fat distribution.

- Insulin: 2-Hour serum insulin (mu U/ml)

  Insulin is a peptide hormone produced by beta cells of the pancreatic islets, which "is the main anabolic hormone of the body. It regulates the metabolism of carbohydrates, fats, and protein by promoting the absorption of glucose from the blood into liver, fat, and skeletal muscle cells"[2]. Variable of 'Insulin' in this dataset means 2-Hour serum insulin. Based on one's insulin levels after a meal, we can tell if there is a metabolic disorder and whether there is a defect in islet function which are related with diabetes.

- BMI: Body mass index (Weight/Height, unit in kg/m^2)

  Body mass index (BMI) is a measure of obesity and health, commonly used in statistical analysis. The degree of obesity cannot be judged directly by the absolute value of weight, it is naturally related to height. So, BMI is defined as the body mass divided by the square of the body height, and is expressed in units of kg/m^2.

- DiabetesPedigreeFunction: Diabetes pedigree function

- Age: Age (years)

---

[1] Luscher, Thomas, and F. 'What is a normal blood pressure?.' *European Heart Journal: The Journal of the European Society of Cardiology* 39.24(2018):2233-2240.

[2] "Incretin - WikiMili, The Free Encyclopedia." *WikiMili.com*, wikimili.com/en/Incretin.

- Outcome: Class variable (0 or 1) 268 of 768 are 1, the others are 0

2.2 Data pre-processing

2.2.1 Missing values processing

Table 2.2 shows the data type of each variable. According to table 2.2, all independent variables are numerical. And figure 2.1 shows distribution of each numerical variable. Since it is non-null for each variable, it seems no missing value in the dataset.

Table 2.2 Data Type Table

Name	Quantity	Completeness	Format
Pregnancies	768	non-null	int64
Glucose	768	non-null	int64
BloodPressure	768	non-null	int64
SkinThickness	768	non-null	int64
Insulin	768	non-null	int64
BMI	768	non-null	float64
DiabetesPedigreeFunction	768	non-null	float64
Age	768	non-null	int64
Outcome	768	non-null	object

Basic quantitative analysis is performed in Table 2.3 to calculate the mean, median, etc. From those basic statistics of parameters in table 2.3, many of them have a minimum value of 0, which is counter-intuitive in realistic. For instance, it is impossible for people to have zero blood pressure. Those zero-values in the data are potential the missing values. After investigation, invalid zero values are listed below:

- Glucose
- BloodPressure
- SkinThickness
- Insulin
- BMI

Table 2.3 Summary of different variables in dataset

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768	3.84	3.37	0.00	1.00	3.00	6.00	17.00
Glucose	768	120.89	31.97	0.00	99.00	117.00	140.25	199.00
BloodPressure	768	69.11	19.36	0.00	62.00	72.00	80.00	122.00
SkinThickness	768	20.54	15.95	0.00	0.00	23.00	32.00	99.00
Insulin	768	79.80	115.24	0.00	0.00	30.50	127.25	846.00
BMI	768	31.99	7.88	0.00	27.30	32.00	36.60	67.10
DiabetesPedigree Function	768	0.47	0.33	0.08	0.24	0.37	0.62	2.42
Age	768	33.24	11.76	21.00	24.00	29.00	41.00	81.00
Outcome	768	0.35	0.48	0.00	0.00	0.00	1.00	1.00

If we choose to drop all the observations with invalid zero values, more than three hundred rows would be dropped in total. Such number of rows dropping will make the sample too small to

perform training. It is better to replace invalid zero values with suitable values. To choose suitable values, we need to make an understanding of the distribution of those variables which is shown in Figure 2.2. From the distribution and realistic meaning of the five variables, the missing values of Glucose and Blood Pressure are replaced with the mean of existing values and the missing values of Skin Thickness, Insulin and BMI are replaced with the median of existing values in corresponding column.

2.2.2 Data scaling

In data analysis, when the magnitudes of independent parameters vary, it will become a hinder to correlate different characteristics of parameters when doing classification. Additionally, unscaled data can make it difficult for data visualization. So, it is necessary to perform data scaling (normalization) in preprocessing procedure. In the preprocessing of this project, all predictors will be scaled by Standardize features by removing the mean and scaling to unit variance by the built-in function StandardScaler() in sklearn package in Python. After the scaling, the data into a range of 1 and a standard deviation of 1. By using this package, the scaled data will be stable when there are enough samples existing and not susceptible to outliers.

2.3 Splitting data to training set and testing set

The train-test split procedure is used to evaluate the performance of data analysis when they are used to make predictions on data not used to train the model. This is a fast and easy procedure to perform, the results of which allow people to compare the performance of different algorithms for predictive modeling problem. But the splitting will be limited when the dataset is small or

additional configuration is required. In this project, the data is split into training set and testing set, by a ratio of 75% and 25%.

2.4 Balance checking and Oversampling

An imbalanced classification problem is one type "of classification problem where the distribution of examples across the known classes is biased or skewed"[3]. In this project, the dependent variable 'Outcome' is a binary variable with value of 1 or 0 which represents whether the patient has diabetes or not. The number of diabetic patients is 268 while the number of non-diabetic people is 500. The distribution of the dependent variable is shown in Figure 2.3. According to the plot, the non-diabetes observations make up more than a half of the data, which are almost twice the number of diabetic patients. That means, the classifier of the data is biased. In the other words, this classifiers in the data "have poor predictive performance, specifically for the minority class"[4], diabetic patients.

One way to solve the problem of class imbalance is to resample the training dataset randomly, which includes random oversampling and undersampling. In simple terms, oversampling means duplicate examples from the minority class and undersampling means deleting examples from the majority class.

Since the data is imbalanced and data sample is not enormous, technique of oversampling will be applied to training set. Random data with Outcome equal to 1 are duplicated to make the numbers

[3] [4] Brownlee, Jason. "A Gentle Introduction to Imbalanced Classification." *Machine Learning Mastery*, 14 Jan. 2020, machinelearningmastery.com/what-is-imbalanced-classification/.

of diabetic patients equal to the numbers of non-diabetic patients. Here the number is 500 - 268 = 232.

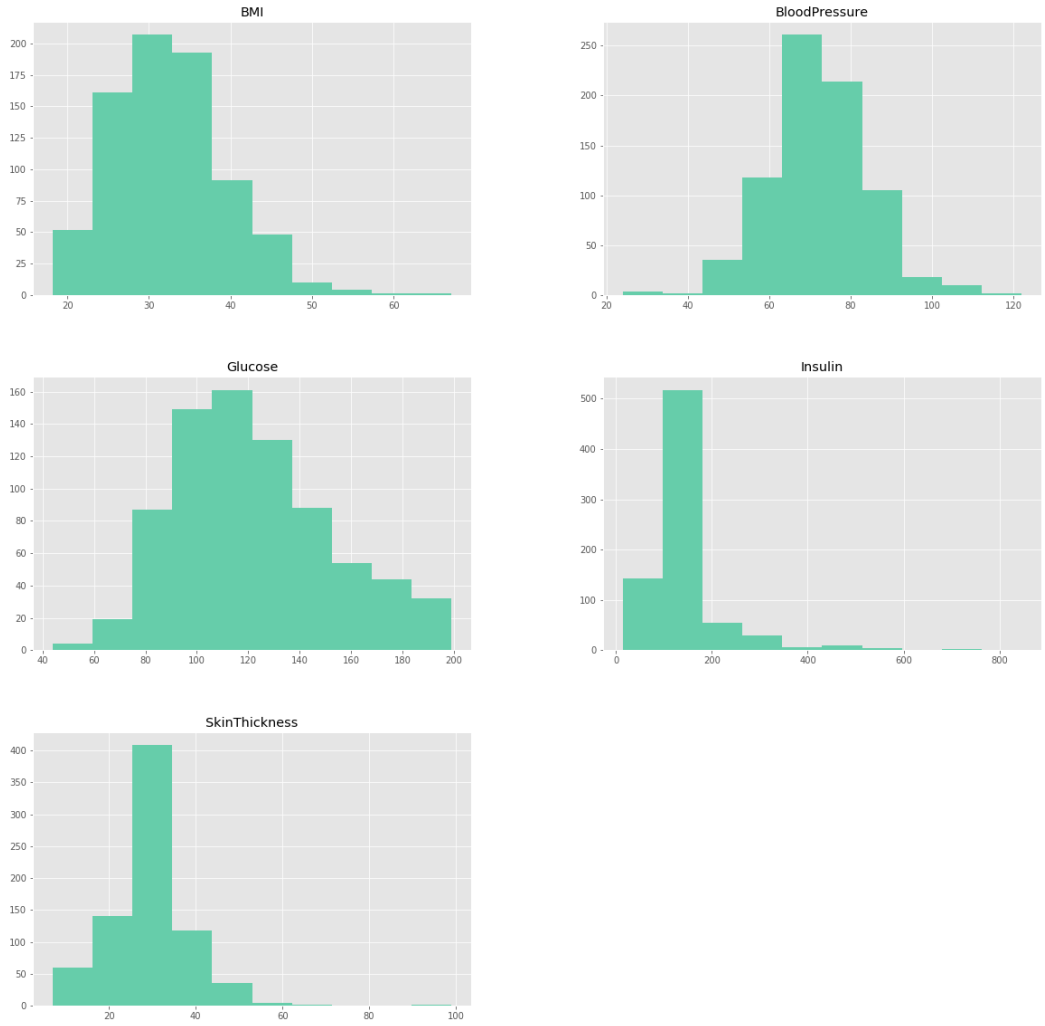

Figure 2.1 Distribution of variables with missing values

Figure 2.2 Density plot of all variables

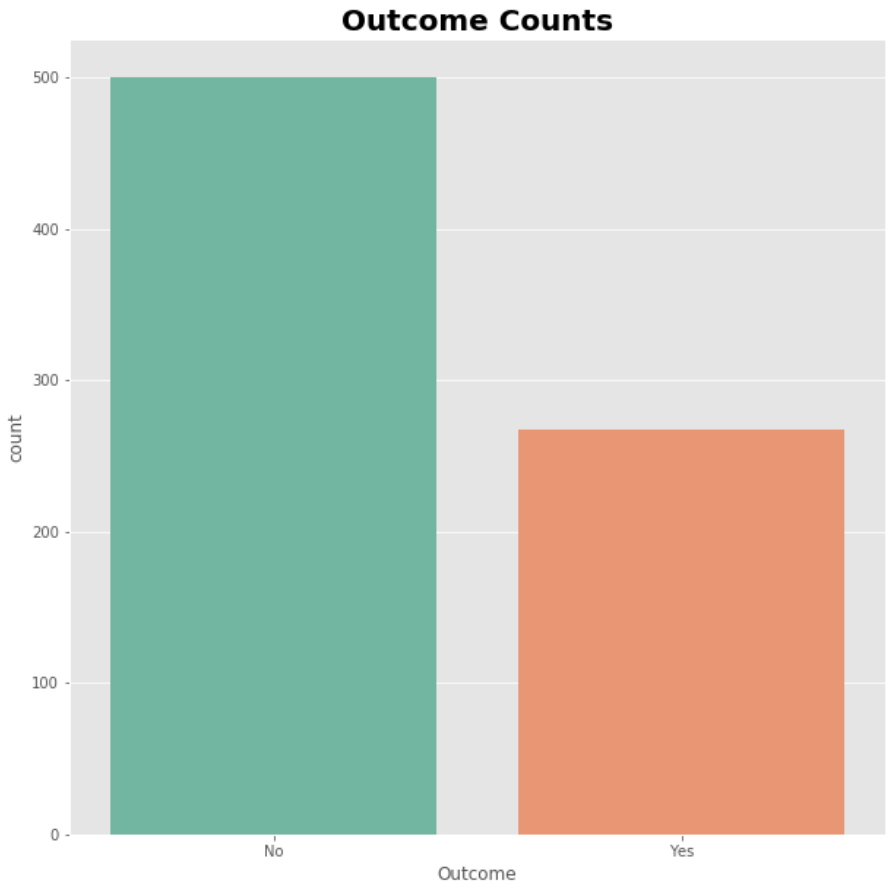

Figure 2.3 Outcome Counts

CHAPTER 3 Exploratory Data Analysis (EDA)

In statistics, Exploratory Data Analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. Basically, the EDA method tries to explore data and achieve the structure or rules of a data based on as less assumptions as possible. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task by graphing, tabling, regression, and characteristic quantities. This chapter will focus on basic EDA approach including correlation check, variable distributions, and factor analysis.

3.1 Correlation Check

In statistics, correlation is used to denote association between two quantitative variables. The degree of association is measured by a correlation coefficient. A correlation coefficient matrix is a simple table to summarize the correlations between all variables. Correlation check will give us a basic understanding of the relationship among variables of the dataset after missing data processing. From the achieved correlation matrix, the parameter pairs with highest correlation coefficient will be emphasized to show the tightest association. Figure 3.1 is the correlation plot which shows the correlation coefficients between all variables related with Pima Indian Diabetes. According to the correlation plot, we found several pairs of variables are highly correlated. For instance, ‘Age’ and ‘Pregnancies’, ‘SkinThickness’ and ‘BMI’ are related pairs. So, it is possible that the there are some underlying, but unobservable, random quantities. These candidates of quantities are defined as factors. We may apply factor analysis to find the relationships among independent variables.

3.2 Variables Distribution

According to the correlation plot, it is found that the associations between target variable ‘Outcome’ and independent variables, such as ‘Insulin’, ‘BMI’ and ‘Age’ are relatively high.

Figure 3.2 shows the distribution of ‘Insulin’. The distribution of ‘Insulin’ of non-diabetes people is roughly of bell shaped, centered at about 100. When value of ‘Insulin’ is higher than 150, patients are more like to be diabetic.

Figure 3.3 shows the distribution of ‘BMI’. The ‘BMI’ distribution of both non-diabetes and diabetes people are roughly close to normal distribution, centered at 26 and 34 respectively. The shapes of both curves are around similar meaning that the sigma and mu value are similar. The right shift of curve clearly shows the influence of the BMI on the possibilities to get diabetes. People with high BMI are more likely to be diabetic. Since BMI is proportional to the ratio between height and weight, assuming distribution of heights in these two group are similar, it means that people have lower weight will be less possible to get diabetes. The result is of common sense but still need the data of height information to confirm.

Figure 3.4 shows the age distribution of non-diabetes and diabetes people. Age distribution of non-diabetes people is left-skewed and centered at 25. Age distribution of diabetes patients is also left-skewed and centered at 35. The left-skewed of both diabetes and non-diabetes is due the non-uniform sampling of the ages in the research sample because in both group the young-aged people (around30) tended to be much more than the elder people. But the difference of the shape still shows the variance. For the distribution of Non-diabetes people, the peak at 25 is very sharp and decrease very fast and get around 0.15 with ages equal to 40. While for the distribution of diabetes

people, although the curve is skewed, it is shifted to right compared to the non-diabetes' curve. The decrease is relative slow, and the absolute value is higher than non-diabetes patient when the age ranging from 30-60. For people older than 60, the possibility of getting diabetes is around 50% for this dataset.

3.3 Factor analysis

Besides exploration of the relationship between different independent variables and the dependent variable, the relationship among independent variables themselves is significant as well. Factor analysis is a statistical method aim to find the relationships among many variables (covariance) in terms of a few underlying, but unobservable, random quantities called factors. There are several approaches to get common factor. The most common approaches are maximum likelihood approach and principal component approach. This study will focus on MLE approach.

The parallel analysis scree plot in Figure 3.5 suggests the number of factors could be three. However, three factors only count 50.7% of total sample variance, while four factors count 60.1%. So, result of four factors is chosen.

With MLE approach, the four factors and their loadings are listed in table 3.1 below:

Table 3.1 Results of factor analysis

```
Loadings:
                          ML1    ML3    ML4    ML2
Pregnancies                      0.715
Glucose                          0.144  0.717  0.152
BloodPressure             0.189  0.202         0.958
SkinThickness             0.630          0.157
Insulin                                  0.639
BMI                       0.972          0.179  0.133
DiabetesPedigreeFunction  0.122          0.223
Age                              0.881  0.159  0.159

                 ML1   ML3   ML4   ML2
SS loadings    1.411 1.361 1.055 0.996
Proportion Var 0.176 0.170 0.132 0.124
Cumulative Var 0.176 0.346 0.478 0.603
             Pregnancies                   Glucose             BloodPressure
               0.5151476                 0.5661977                 0.9950000
           SkinThickness                   Insulin                       BMI
               0.4393439                 0.4154836                 0.9950000
DiabetesPedigreeFunction                       Age
               0.0664863                 0.8301944
```

The first factor is positive correlated with both ‘BMI’ and ‘SkinThickness’, this factor can be named as ‘Obesity factor’. The BMI index is a widely accepted anthropometric method used to discriminate between being overweight and obese. But BMI is not the only measure of obesity. In order to diagnose obesity more accurately, skin thickness should also be referenced.

The second factor only negative correlated with age and pregnancies (these two variables are positive correlated) and blood pressure, so this factor can be called as ‘Pregnancies factor’. Many women during pregnancy suffer from diabetes. When pregnant women with gestational diabetes get pregnant again, the recurrence rate is high. Patients with gestational diabetes have a higher chance of developing diabetes in the long term.

The third factor is related with ‘Insulin’ and’ Glucose’, this factor can be called as ‘Blood glucose metabolism factor’. Glucose tolerance reflects the body's ability to regulate blood glucose

concentration. Healthy people have a strong tolerance for glucose, and the body can maintain blood glucose in a relatively stable range through the flexible cooperation of various organs and hormones, the opening and closing of multiple channels, and the amount of insulin secreted. Once this harmony is broken, the body's ability to regulate blood sugar will be destroyed, and blood sugar may no longer be controlled.

The fourth factor just related with ‘Blood pressure’, this factor can be called as ‘Blood pressure factor’. High blood pressure is a complication of diabetes.

The factors above are intuitively show in the figure 3.6.

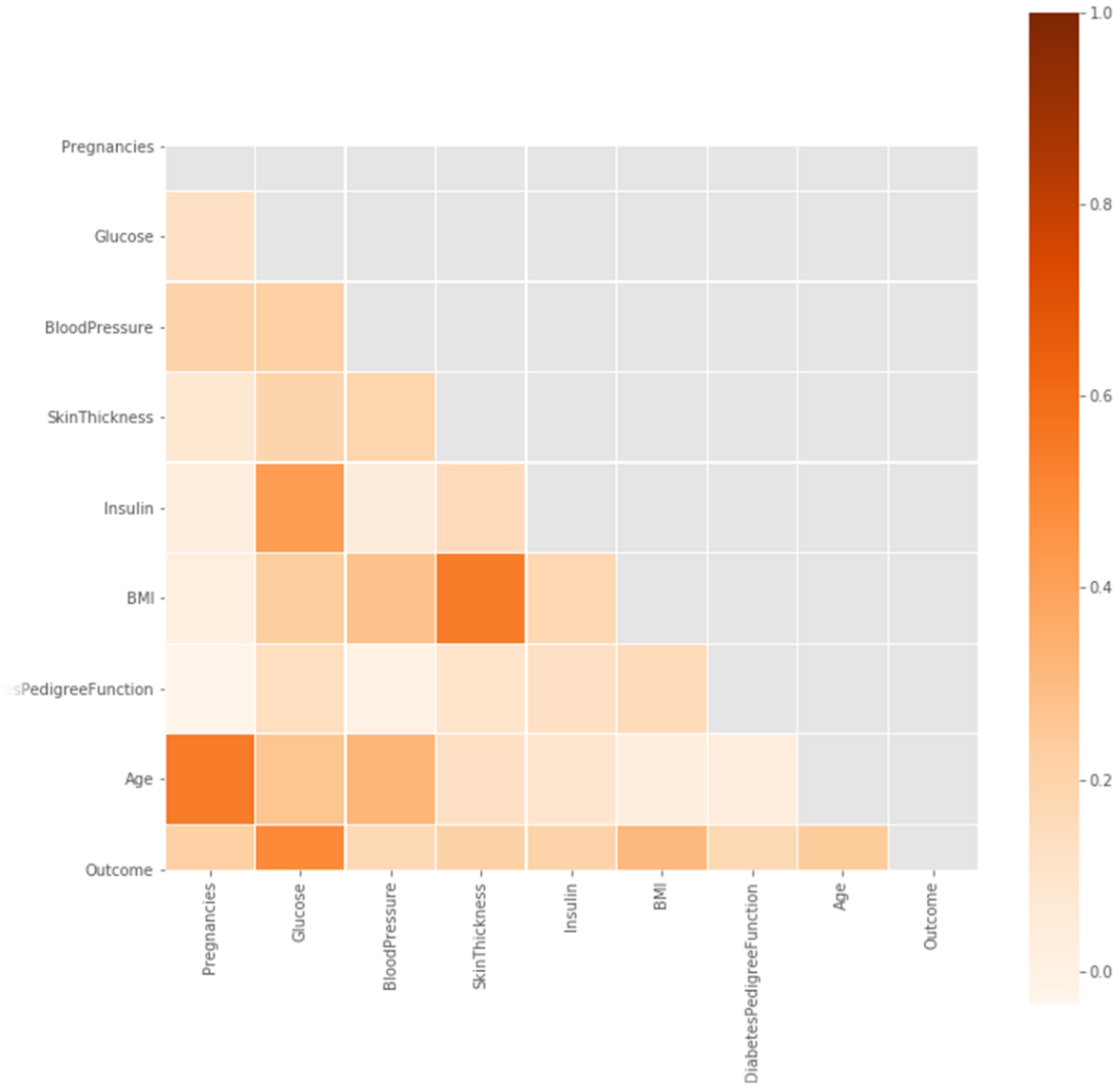

Figure 3.1 Correlation Plot for all variables

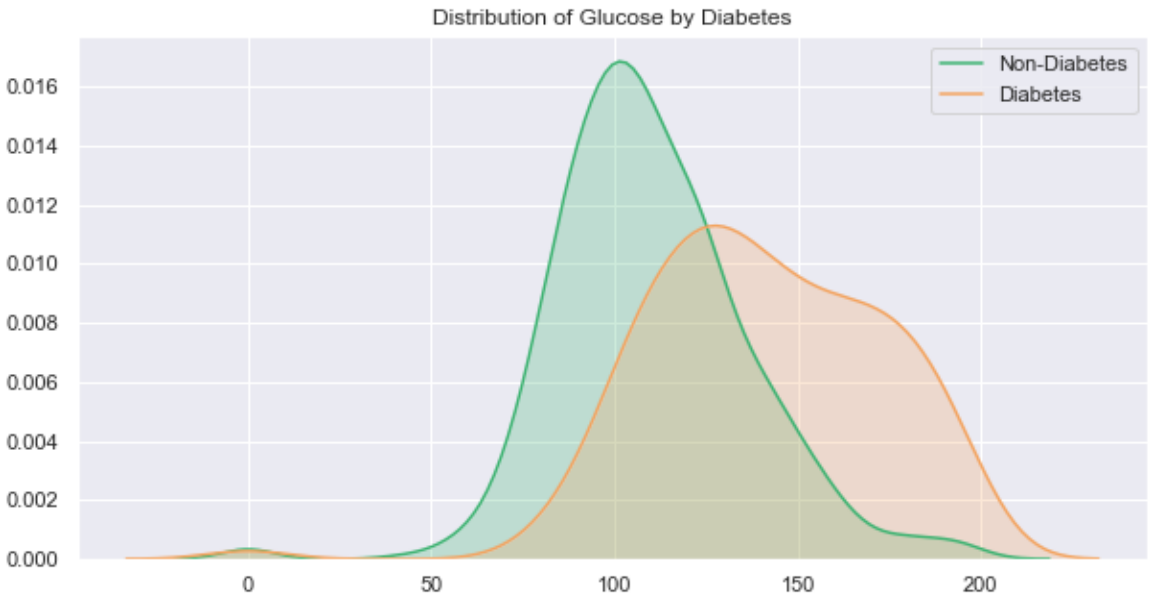

Figure 3.2 Distribution of Glucose by Diabetes

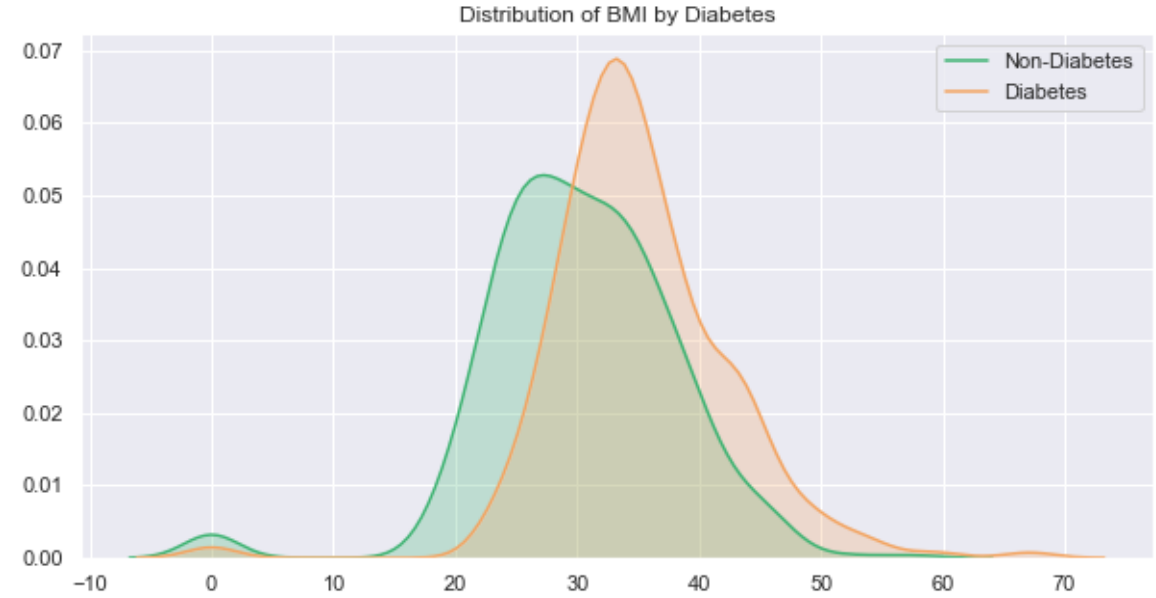

Figure 3.3 Distribution of BMI by Diabetes

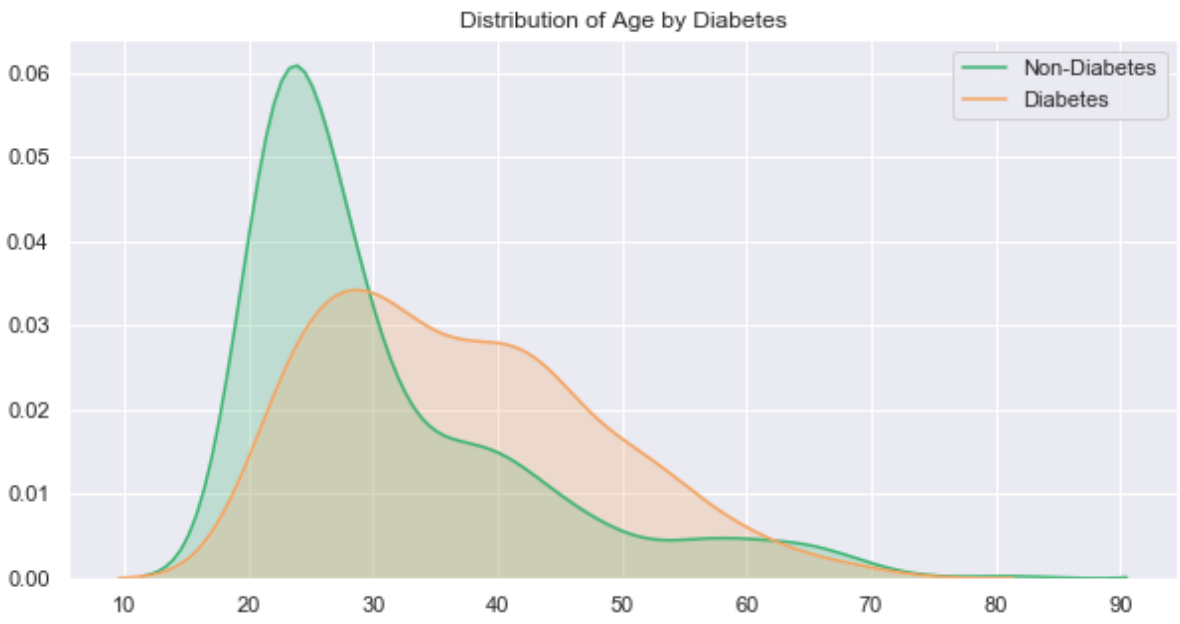

Figure 3.4 Distribution of Age by Diabetes

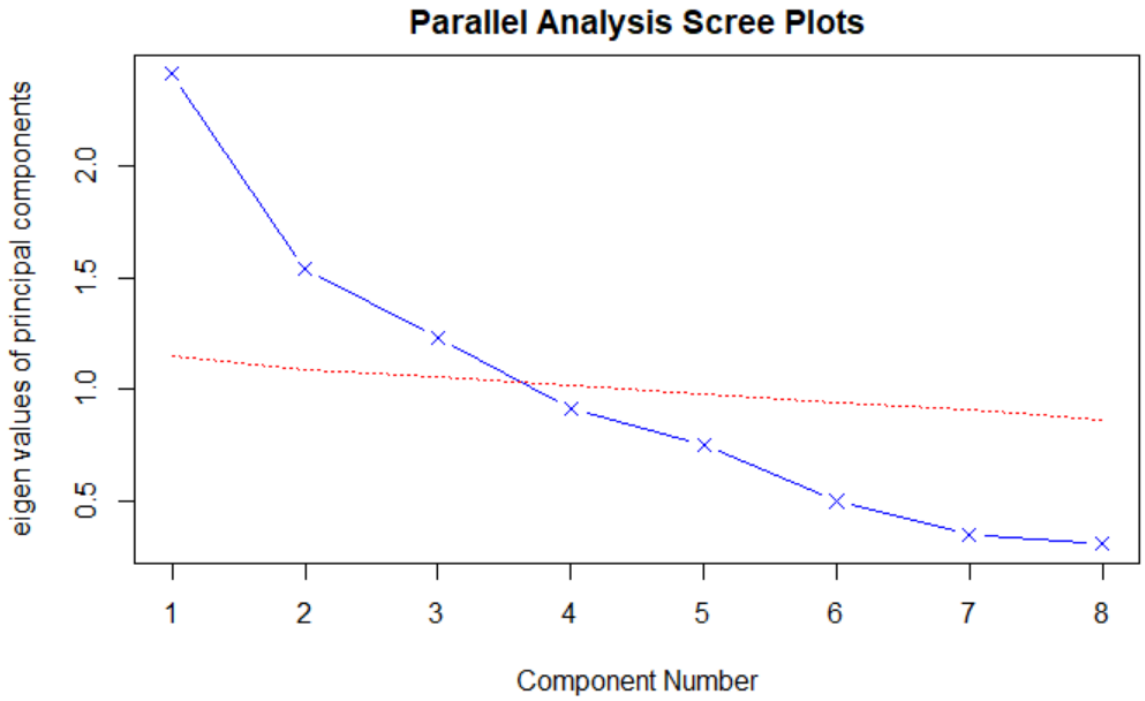

Figure 3.5 Parallel Analysis Scree Plots

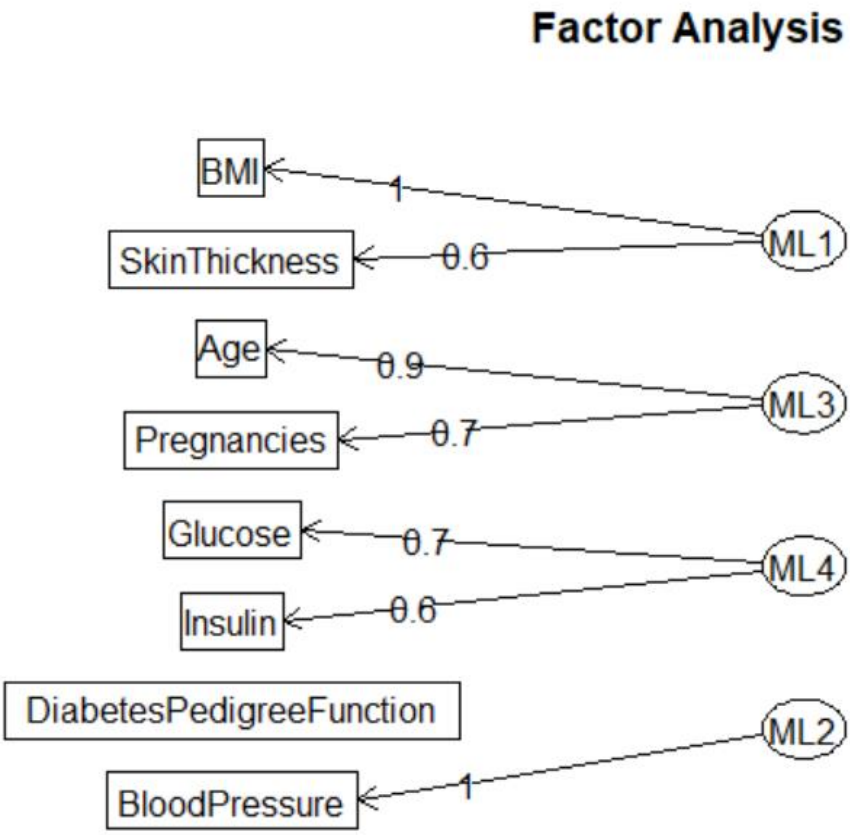

Figure 3.6 Factor Analysis

CHAPTER 4 Methodology

For most machine learning project, after categorizing problem, understanding, and preprocessing the data, we need to find fitted model and algorithm in training process. Some elements can be determined to affect the choice of model including the accuracy, interpretability, and complexity. Even consuming time will be a factor to be considered for the selection. In this project, some potential methodology such as logistic regression, tree searching, and support vector machine are introduced in this chapter.

4.1 Logistic Regression

Since the dependent variable 'Outcome' in this dataset only has the number 0 and 1, logistic regression will be most straightforward method people can think of. Logistic regression is used to predict the probability of conditions of happening event for the yes/no, or A/B situation. It can estimate probability of occurrence of a categorical response based on one or more predictors.

As a method for analyzing possibilities of different categorical-response variables, logistic regression is performing better for its versatility and adaptability for modelling most situations than discriminant analysis. Because discriminant analysis has to be used based on the assumption of normally distribution of all independent variables, but logistic regression does not need that.

In logistic regression, the basic idea is that, a dependent variable Y will have N (usually $N = 2$) unique values. It is then regressed based on a series of p independent variables $X_1, X_2, \dots, X_p$. For instance, Y may be pass or fail of an exam, presence or absence of a disease or winning status of a prize. Logistic regression can be divided into three types, one is binary logistic regression which

only have two types for dependent variables, ‘0’ and ‘1’. Another one is multi-nominal logistic regression, where the dependent variables have three or more types. The last one is ordinal logistic regression, where the dependent variables are not ordered and cannot be expressed by specific types. In this study we only focus on the binary logistic regression because the dependent variable ‘Outcome’ in the dataset only has two types: ‘0’ and ‘1’.

4.2 Decision Tree

In machine learning, a decision tree is a prediction model which is commonly used operations research, specifically in decision analysis. It represents the correlation between the features and values for different variables. Decision tree uses a tree-like model which contains conditional control statements in order to show different conditions and display the possible consequences.

In a decision tree model, each node represents one object and each branch represents the possible values for this object. A decision tree will have three kinds of nodes: decision nodes, chance nodes and end nodes. The path from one decision node to one end node correlates to one possibility, which each variable will have its own value. Going through each node simulate a "test" on an attribute, each branch represents the outcome of the test.

4.3 Random Forest

Decision tree is a popular method in machine learning. But due to the ability of unlimited expansion, the decision tree can have low bias but very high variance, which means overfitting of datasets. Random forests then invented to be a kind of classification method which ensembles multiple decision trees together to correct for decision trees' habit of overfitting to their training set.

In random forests, many decision trees with relatively uncorrelated models are built with a subset of variables of all candidate variables by random sampling. Then the group of decision trees will outperform any of the individual constituent models due to the low correlation between models. The combination of uncorrelated models can produce more accurate predictions than any single prediction because they can protect each other from their individual errors. Even some trees could be wrong, many other trees will be correct to enforce the whole group move in the right direction together.

With the specificity of random forests method, it can produce very high-dimensional (many features) data even without dimensionality reduction or feature selection with relatively fast speed. Additionally, it can differentiate the importance of different features and the mutual influence between different characteristics. It is easy to make a parallel method with decision trees inside forest. More importantly, even if a large part of the features is missing, the accuracy can still be maintained.

4.4 K-Nearest-Neighbors (KNN)

In statistics, the k-nearest neighbors algorithm (k-NN) is a non-parametric classification method which is used for classification and regression. It outputs a class outcome and the property value for the object which is the average of the values of k nearest neighbors. So it is sensitive to the local structure of the data. In k-NN method, the inputs are purely neighbors of each datapoint, and the classification based on a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. Weights can be also assigned to the contributions of the neighbors, which means that the closer neighbors will have more weight to the average than the farther ones. Because this algorithm relies on distance for classification, it is

better to scale the datasets when the magnitude or physical meaning of different parameters vary a lot, which could be helpful to the accuracy dramatically.

Normally k is a positive integer. If we pick a small k number, which means the very few nearest neighbors are picked for the classification, the whole model will become complicated and easy to be overfitting. Oppositely, the larger number of k will make the bias high which means underfitting. So, the proper k value is good the model of classification or regression. Since each new input will be classified by its k nearest neighbors, the original dataset will be treated as training set and there is no extra training step is required.

4.5 Support Vector Machine

A Support Vector Machine (SVM) is a kind of supervised machine learning algorithm which is used to do generalized linear classification, whose decision boundary is maximum-margin hyperplane. With an example of labeled training data with a dependent variable including two types, SVM algorithm can output an optimal hyperplane which classifies new examples. Basically, with a p-dimensional sample space, the classification hyperplane will be p-1 dimension. For instance, the hyperplane is a line in two-dimensional space, dividing a plane in two parts. Whereas in 3D, the hyperplane is a plane to separate two classes where in each class lay in either side.

4.6 Extreme Gradient Boost (XGBoost)

Extreme gradient boosting (XGBoost) is a powerful machine learning technique for regression and classification problems. It is a kind of supervised machine learning algorithm which can be regarded as an improved version of gradient boosting machine. “It uses ‘a more regularized model

formalization to control over-fitting, which gives it better performance,' according to the author of the algorithm, Tianqi Chen. Therefore, it helps to reduce overfitting."[5]

XGBoost produces a prediction model in the form of an ensemble of weak prediction models by Newtons method rather than traditional gradient descent to optimize the loss function. "It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function."[6]

XGBoost uses compound tree model, which is a set of classification and regression trees (CART). In CART, compared to the classification result, the score of each leaf node gives us more explanation. This makes it easier for CART to unify and optimize nodes, which is similar to the method used in random forest method. However, they are different due to the way of training. XGBoost initially uses constants as predictions, and then adds a new learning function (parameter) for each subsequent prediction so that the results are optimized for every step until the best gain is found.

[5] Nishida, Kan. "Introduction to Extreme Gradient Boosting in Exploratory." *Medium*, Learn Data Science, 21 Mar. 2017, blog.exploratory.io/introduction-to-extreme-gradient-boosting-in-exploratory-7bbec554ac7.
[6] "Infinite Script." *Infinite Script RSS*, infinitescript.com/tag/boosting/.

CHAPTER 5 Model Training and Performance

After the methodology investigation, few candidate methods are being picked in the study. And then K-Fold cross validation will be used to estimate the model performance as mentioned above, which is "commonly used in applied machine learning to compare and select a model for a given predictive modeling problem because it is easy to understand, easy to implement, and results in skill estimates that generally have a lower bias than other methods."[7]

5.1 Metrics

To measure the performance of our models above, we need confusion matrix. The confusion matrix is used for checking whether the predictors are truly classified into correct classes.

- True positive (TP): Diabetes patients correctly identified as diabetic
- True negative (TN): Healthy person correctly identified as healthy
- False positive (FP): Healthy person incorrectly identified as diabetic
- False negative (FN): Diabetic patients incorrectly identified as healthy

False Positive (FP) can be defined as Type 1 error and False Negative (FN) can be defined as Type 2 error.

We also need other metrics:

- Accuracy = (TP +TN) / (TP + TN + FP +FN)
- Precision = TP / (TP + FP)

[7] Brownlee, Jason. "A Gentle Introduction to k-Fold Cross-Validation." *Machine Learning Mastery*, 2 Aug. 2020, machinelearningmastery.com/k-fold-cross-validation/.

- Recall (Sensitivity)=TP / (TP + FN)
- F1 score = 2 x ((Precision x Recall) / (Precision + Recall))
- True Negative Rate (TNR) = TN/(TN+FN)
- False Positive Rate (FPR) = 1 – Specificity

5.2 Logistic regression

The Logistic regression is the first method being tested. The coefficients of logistic regression with L2 penalty are listed in the table below. Here we assume the 8 different variables are linear related, which mean that: for the variables $x_1, x_2, x_3, \dots, x_8$, the regression equation should be:

$$y = a_1x_1 + a_2x_2 + \cdots . a_8x_8$$

$a_1, a_2, \dots a_8$ are the coefficient for different variables. After fitting, the estimated coefficients are shown in Figure 5.1, which correlated to the feature importance of different variables in logistic regression. The feature 'DiabetesPedigreeFunction' dominates with around 0.4 coefficient, the top three important features are 'DiabetesPedigreeFunction' 'Pregnancies' and 'BMI'.

The confusion matrix and receiver operating characteristics (ROC) are also calculated and shown in Figure 5.2. From the confusion matrix we can get the accuracy score of tests set equal to:

$$\frac{(\mathrm{TP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})} = \frac{113 + 41}{113 + 41 + 26 + 12} = 80.2\%$$

The precision equal to:

$$\frac{(\mathrm{TP})}{(\mathrm{TP} + \mathrm{FP})} = \frac{41}{41 + 12} = 77.4\%$$

And the True Negative Rate equal to:

$$\frac{TN}{(TN+FN)} = \frac{113}{113+26} = 81.3\%$$

False Positive Rate (FPR) equal to:

$$1 - TNR = 1 - 81.3\% = 18.7\%$$

Additionally, from the ROC curve, we can see that the logistic regression method is relative trustable since most of the data are well defined to the correct value.

Table 5.1 Estimated Coefficients of Logistic Regression

	Feature name	Estimated coefficients
1	Pregnancies	0.11
2	Glucose	0.03
3	BloodPressure	-0.02
4	SkinThickness	0.01
5	Insulin	0
6	BMI	0.05
7	DiabetesPedigreeFunction	0.4
8	Age	0
9	intercept	-5.39958

5.3 Decision Tree

The decision tree is the second model being trained and tested. The max depth of tree is set as 4 to prevent overfitting. Top three important features are ‘Glucose’ ‘BMI’ and “Age” while the feature ‘Glucose’ dominates with over 0.5 coefficient. The confusion matrix and receiver operating characteristics (ROC) are also calculated and shown in Figure 5.3. From the confusion matrix, we

can find this algorithm misclassified 51 out of 192 test instances. So, the accuracy score of tests set equal to:

$$\frac{(\mathrm{TP}\ +\mathrm{TN})}{(\mathrm{TP}\ +\ \mathrm{TN}\ +\ \mathrm{FP}\ +\mathrm{FN})} = \frac{36+105}{36+105+20+31} = 73.43\%$$

The precision equal to:

$$\frac{(\mathrm{TP}\ )}{(\mathrm{TP}\ +\ \ \mathrm{FP}\ )} = \frac{36}{36+20} = 64.3\%$$

The recall equal to:

$$\frac{(\mathrm{TP}\ )}{(\mathrm{TP}\ +\ \ \mathrm{FN}\ )} = \frac{36}{36+31} = 53.7\%$$

And the True Negative Rate equal to:

$$\frac{TN}{(TN+FN)} = \frac{105}{105+31} = 77.2\%$$

False Positive Rate (FPR) equal to:

$$1 - TNR = 1 - 76.7\% = 22.8\%$$

According to the ROC curve, the AUC is 0.69 which is lower than 0.7.

5.4 Random Forest

As for random forest training, the max depth of tree is also set as four. There are 100 trees in the forest. Top three important features of random forest are ‘Glucose’ ‘BMI’ and ‘Age’ which is similar to that of decision tree. While feature of ‘Glucose’ only dominates with 0.25 coefficient. The performance is proved compared to that of decision tree. The confusion matrix and receiver

operating characteristics (ROC) are also calculated and shown in Figure 5.4. From the confusion matrix we can get the accuracy score of tests set equal to:

$$\frac{(\mathrm{TP}\ +\mathrm{TN})}{(\mathrm{TP}\ +\ \mathrm{TN}\ +\ \mathrm{FP}\ +\mathrm{FN})} = \frac{39+114}{39+114+11+28} = 79.7\%$$

The precision equal to:

$$\frac{(\mathrm{TP}\ )}{(\mathrm{TP}\ +\ \ \mathrm{FP}\ )} = \frac{39}{39+11} = 78.0\%$$

The recall equal to:

$$\frac{(\mathrm{TP}\ )}{(\mathrm{TP}\ +\ \ \mathrm{FN}\ )} = \frac{39}{39+28} = 58.2\%$$

And the True Negative Rate equal to:

$$\frac{TN}{(TN+FN)} = \frac{114}{114+28} = 80.3\%$$

False Positive Rate (FPR) equal to:

$$1 - TNR = 1 - 74.5\% = 19.7\%$$

According to the plot of ROC curve, the AUC is about 0.75. The model is better than decision tree at distinguishing between diabetes patients and healthy people.

5.5 KNN

In this study, k=10 is used in model training. Which means the object is simply assigned to the class of that the 10 nearest neighbors. The confusion matrix and receiver operating characteristics

(ROC) are also calculated and shown in Figure 5.5. According to ROC plot, we know that AUC is 0.7. From the confusion matrix we can get the accuracy score of tests set equal to:

$$\frac{(\mathrm{TP}\ +\mathrm{TN})}{(\mathrm{TP}\ +\ \mathrm{TN}\ +\ \mathrm{FP}\ +\mathrm{FN})} = \frac{35+110}{35+110+15+33} = 75.5\%$$

The precision equal to:

$$\frac{(\mathrm{TP}\ )}{(\mathrm{TP}\ +\ \ \mathrm{FP}\ )} = \frac{35}{35+15} = 70.0\%$$

The recall equal to:

$$\frac{(\mathrm{TP}\ )}{(\mathrm{TP}\ +\ \ \mathrm{FN}\ )} = \frac{35}{35+33} = 51.5\%$$

And the True Negative Rate equal to:

$$\frac{TN}{(TN+FN)} = \frac{110}{110+33} = 76.9\%$$

False Positive Rate (FPR) equal to:

$$1 - TNR = 1 - 76.9\% = 23.1\%$$

5.6 Support Vector Machine

The kernel type of linear is used in the algorithm of support vector classifier. Top three important features are ‘DiabetesPedigreeFunction’ ‘Pregnancies’ and ‘BMI’, while the feature ‘DiabetesPedigreeFunction’ dominates with over 0.5 coefficient. The confusion matrix and receiver operating characteristics (ROC) are also calculated and shown in Figure 5.6. From the

confusion matrix, we can find this algorithm misclassified 41 out of 192 test instances. So, the accuracy score of tests set equal to:

$$\frac{(\mathrm{TP}\ +\mathrm{TN})}{(\mathrm{TP}\ +\ \mathrm{TN}\ +\ \mathrm{FP}\ +\mathrm{FN})} = \frac{41+110}{41+110+15+26} = 78.6\%$$

The precision equal to:

$$\frac{(\mathrm{TP}\ )}{(\mathrm{TP}\ +\ \ \mathrm{FP}\ )} = \frac{41}{41+15} = 73.2\%$$

The recall equal to:

$$\frac{(\mathrm{TP}\ )}{(\mathrm{TP}\ +\ \ \mathrm{FN}\ )} = \frac{41}{41+26} = 61.2\%$$

And the True Negative Rate equal to:

$$\frac{TN}{(TN+FN)} = \frac{114}{114+28} = 80.3\%$$

False Positive Rate (FPR) equal to:

$$1 - TNR = 1 - 76.9\% = 19.7\%$$

5.7 XGBoost Classifier

Extreme gradient boosting is the last model trained and tested. The max depth of tree is four, same as that of decision tree. The feature ‘Glucose’ dominates with over 0.25 coefficient and Top three important features are ‘Glucose’ ‘BMI’ and ‘Insulin’.

The confusion matrix and receiver operating characteristics (ROC) are also calculated and shown in Figure 5.7. From the confusion matrix we can get the accuracy score of tests set equal to:

$$\frac{(\mathrm{TP}\ +\mathrm{TN})}{(\mathrm{TP}\ +\ \mathrm{TN}\ +\ \mathrm{FP}\ +\mathrm{FN})} = \frac{51+107}{51+107+18+16} = 82.3\%$$

The precision equal to:

$$\frac{(\mathrm{TP})}{(\mathrm{TP}\ +\ \ \mathrm{FP})} = \frac{51}{51+18} = 73.9\%$$

The recall equal to:

$$\frac{(\mathrm{TP}\ )}{(\mathrm{TP}\ +\ \ \mathrm{FN}\ )} = \frac{51}{51+16} = 76.1\%$$

And the True Negative Rate equal to:

$$\frac{TN}{(TN+FN)} = \frac{107}{107+16} = 87.0\%$$

False Positive Rate (FPR) equal to:

$$1 - TNR = 1 - 76.9\% = 13.0\%$$

In addition, the AUC is 0.8 leads to a good performance of model.

5.8 Model performance summary

There are 6 different modeling approaches are tried to be implemented for this case. The performance scores of each model are list below. According to the table, the extreme gradient boosting model has the highest accuracy score of 82.29%, the highest recall of 0.76 and the highest AUC of 0.8. The recall, also known as sensitivity, “refers to the proportion of people with disease

who have a positive test result"[8]. Since the aim of this study is to predict the diabetes patients, the recall is very essential metric for model evaluation.

[8] *More Details on Sensitivity and Specificity*, ebm-tools.knowledgetranslation.net/resource/sensitivity.

Table 5.2 Summary of Model Performance

Model	Accuracy Score	Recall score	Precision	F1 score	Area under curve(train)	Area under curve(test)
Logistic regression	0.8021	0.61	0.77	0.68	0.6979	0.7580
Decision Tree	0.7552	0.54	0.64	0.58	0.7717	0.6887
Random Forest	0.7760	0.58	0.78	0.67	0.7441	0.7470
KNN	0.7552	0.52	0.70	0.60	0.7383	0.7012
SVM(linear)	0.7865	0.61	0.73	0.67	0.7054	0.7460
XGBoost	0.8229	0.76	0.74	0.75	0.9293	0.8086

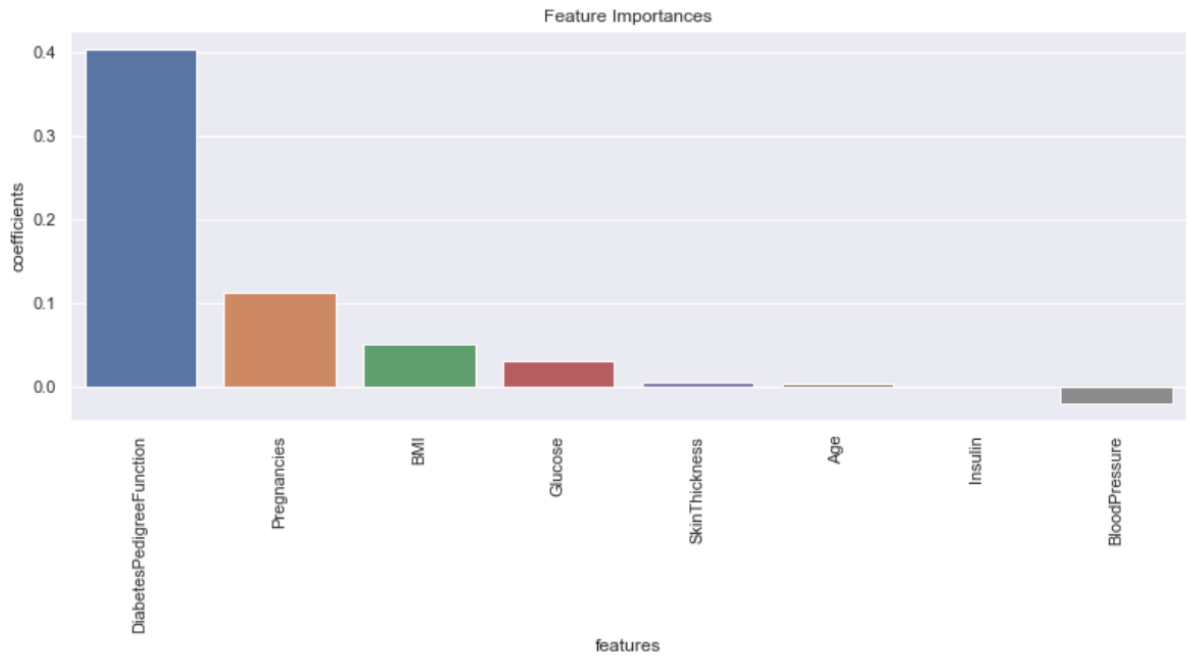

Figure 5.1 Feature importance for Logistic regression

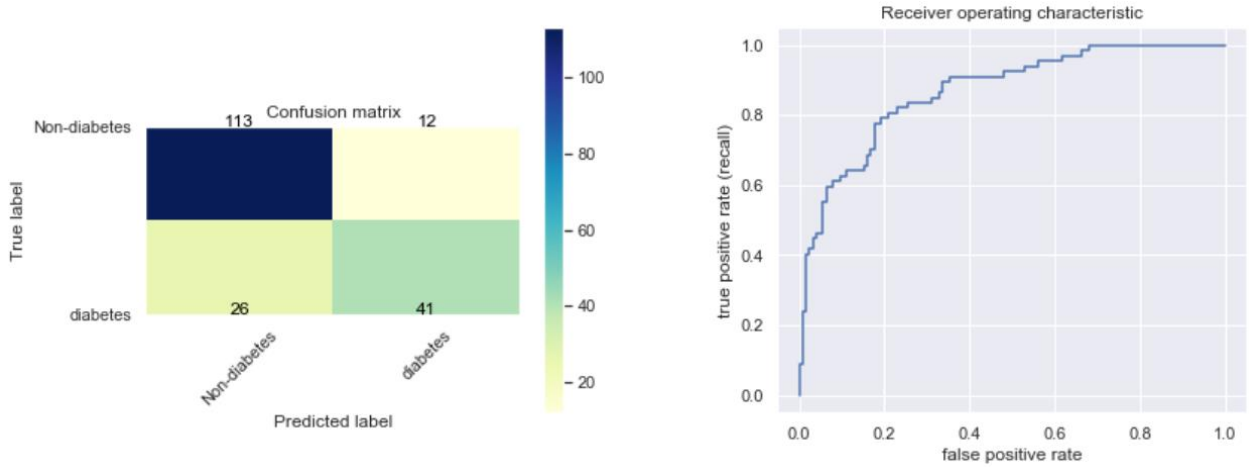

Figure 5.2 Confusion Matrix and ROC of Logistic Regression

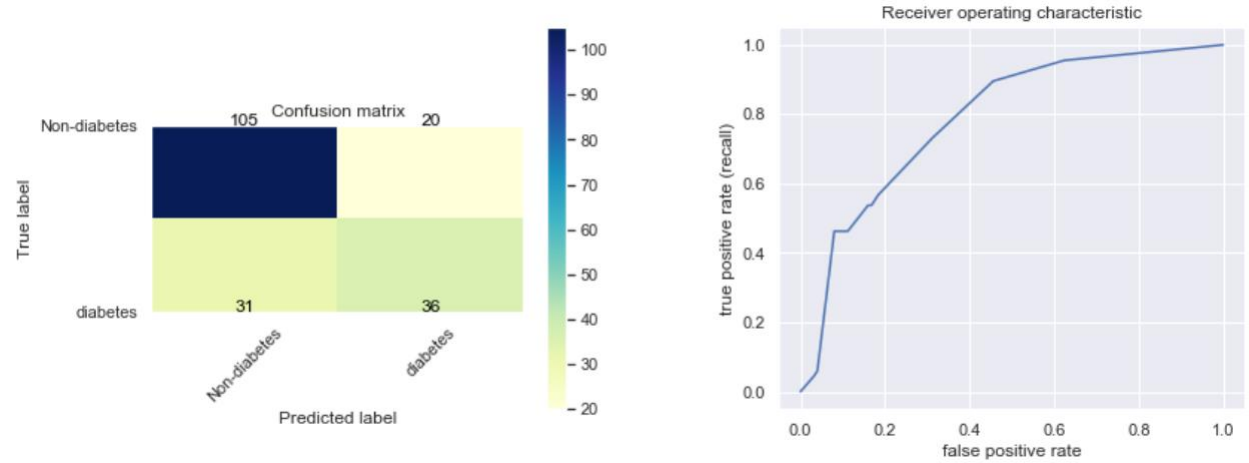

Figure 5.3 Confusion Matrix and ROC of Decision Tree

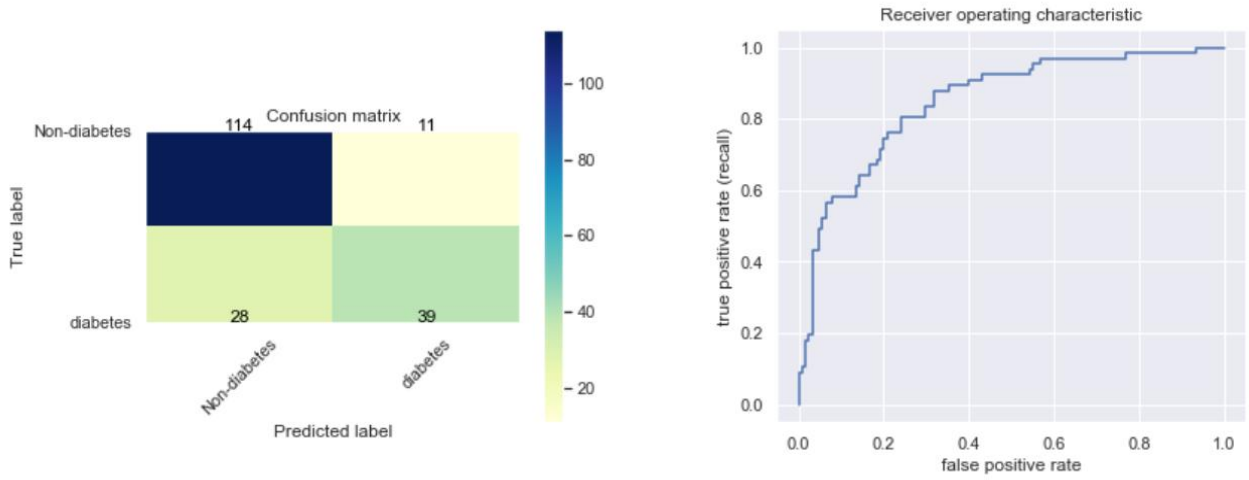

Figure 5.4 Confusion Matrix and ROC of Random Forest

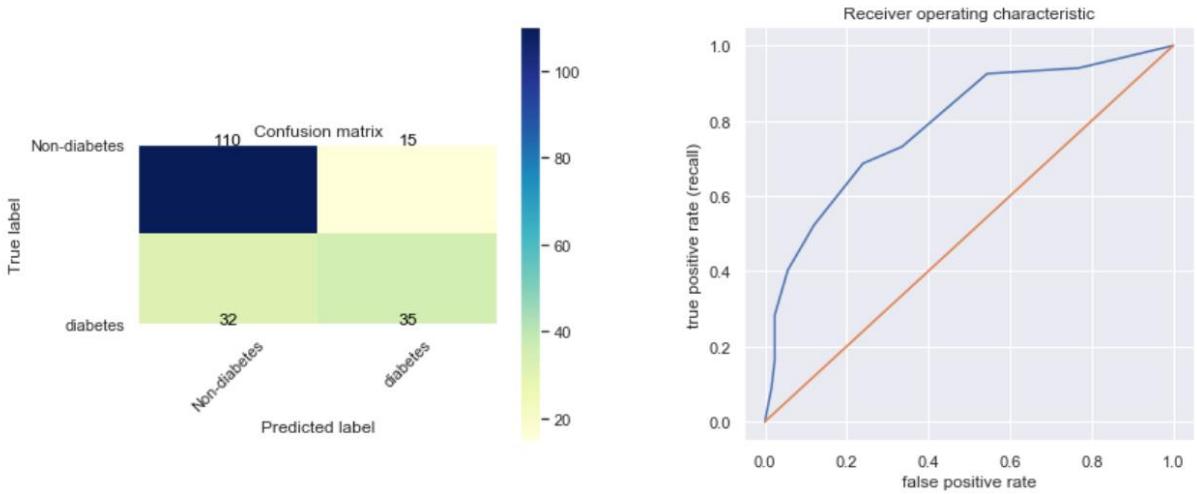

Figure 5.5 Confusion Matrix and ROC of KNN

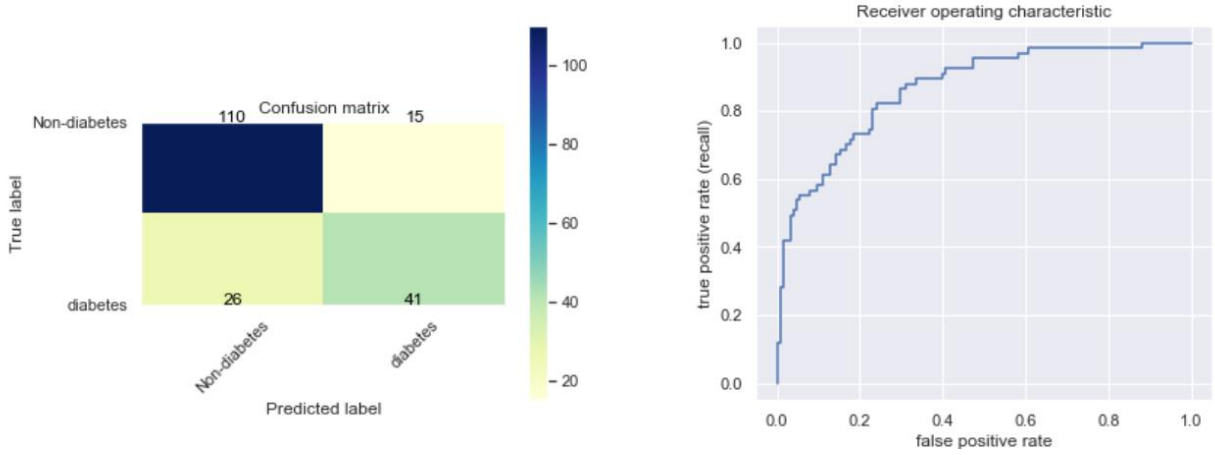

Figure 5.6 Confusion Matrix and ROC of SVM

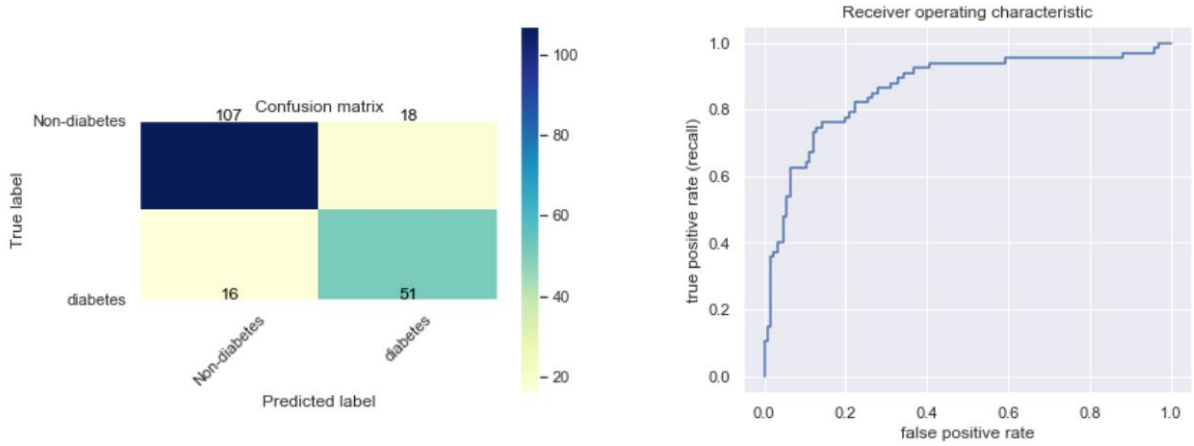

Figure 5.7 Confusion Matrix and ROC of XGBoost

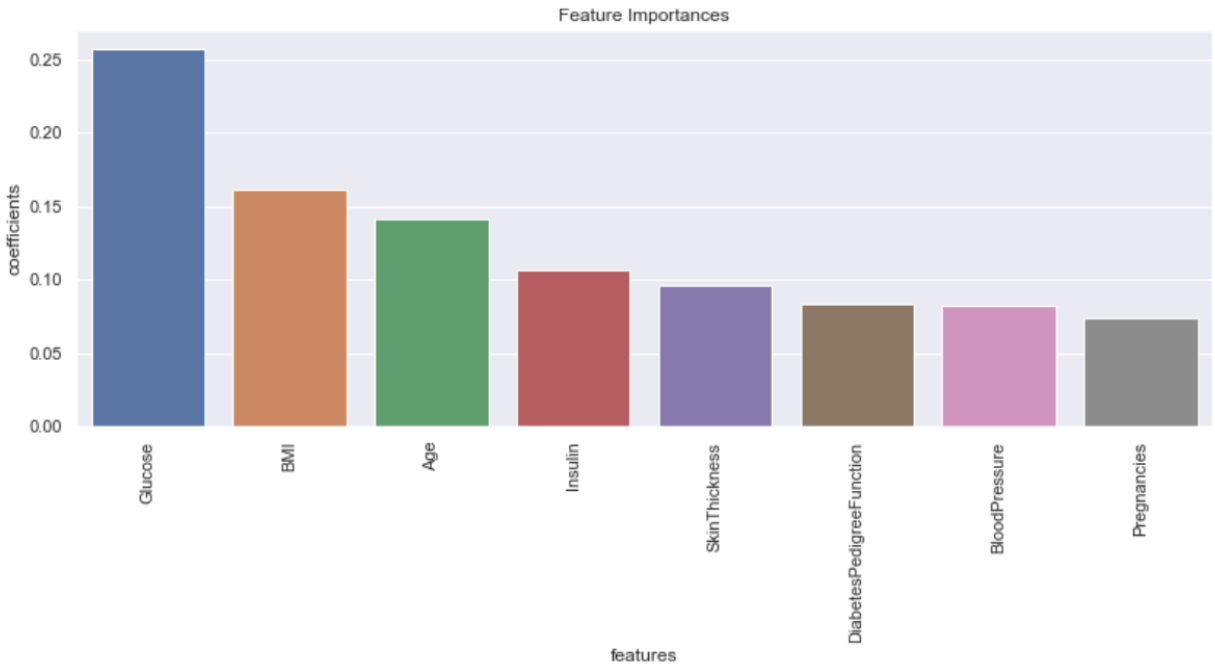

Figure 5.8 Feature importance for XGBoost

CHAPTER 6 Conclusion

The Extreme gradient boosting is chosen as the final model due to its performance. According to the result of final model, the important features which lead to high probability of diabetes are ‘Glucose’ ‘BMI’ ‘Age’ and ‘Insulin’. Both doctor and patients can pay close attention to those medical indexes in order to prevent or diagnose diabetes.

For the further study, the outcome variable of ‘Yes’ could be classified into type 1 diabetes, type 2 diabetes, and other types of diabetes. There are several types of diabetes. Patients with Type 1 diabetes cannot generate insulin. Patients with Type 2 diabetes have difficulty not only in making insulin but also in taking use of insulin. Type 2 diabetes is the most common type which often happened to patients in middle age or older. Gestational diabetes mellitus is a special type of diabetes, it happened to woman during pregnancy. A more detailed classification of the target variable could better reveal the relationship between medical predictors and outcome.

And more predictors could be collected. For example, daily food intake, dietary structure and daily physical exercise are all related with diabetes.

As for data selection, the age and gender of patients could be expanded so that we can get information about children’s diabetes and male’s diabetes.