Assignment 2

Data Modelling and presentation

To accurately predict whether or not a patient is having diabetes or not depending on various medical diagnostic measurements

Amal Joy: S3644794
s3644794@student.rmit.edu.au

Pooja Dandge: S3698457
s3698457@student.rmit.edu.au

22/5/18

Table of Contents

1. Executive Summary

The project was aimed to build a model to predict if a person is diabetic or not based on his medical diagnostics. The experiment was conducted using the pima Indians diabetics data. Numerous issues were present in the data including impossible values and missing values. These issues were handles appropriately by imputing class average in most of the cases. Normalization of the variables were applied on the data. A thorough analysis of the features was conducted and the most important features were selected for building the models. Different models like KNN, decision trees and Random forest were built on the data. Due to the issue of data imbalance, oversampling using SMOTE was also tried on KNN model. But it was later found that the oversampling is actually overfitting the data and the model will not be useful. Important features were identified from the random forest model and the new data set was made using only the important features. Comparison of the models were also done and found that random forest model gave the best results. Parametric tuning was also performed on all the models. Random search grid was applied on the decision tree and the random forest model to tune the parameters. Even though the parametric tuning didn't have much effect on the random forest model it had a 3.75% improvement on the decision tree model. Cross validation was applied on all the models and Random Forest model was found to be more stable than any other model in the experiment with a 10-fold cross validation accuracy of 87.11%.

2. Introduction

Problem statement

To predict if a person is diabetic or not using various medical attributes provided.

Assumptions

1. The attributes provided in the dataset is enough to accurately predict if the patient is diabetic or not.
2. The data set has enough data points (786) to split the data and predict the diabetic condition of the patient accurately.

The aim of this project is to build a machine learning model to accurately predict whether or not a patient is having diabetes or not depending on various diagnostic measurements including the number of times the patient is pregnant, Plasma glucose concentration from a 2 hour in an oral glucose tolerance test, Diastolic blood pressure and so on. We are planning to make 3 models to predict if the patient is having diabetics or not. Including K- Nearest Neighbours, Decision Tree and Random Forest. We will also try to do some parameter tuning for all of the models. We will compare the models and then choose the best model for the prediction.

3. DATASET

The dataset used in this project is downloaded from GitHub (GitHub, n.d.). The original source of the dataset is from the National Institute of Diabetes and Digestive and Kidney Diseases. All of the patients considered in this project are females above 21 years of age and from Pima Indian Heritage. The dataset consists of several medical diagnostic measurements including the number of times the patient is pregnant, Plasma glucose concentration a 2 hour in an oral glucose tolerance test, Diastolic blood pressure, Triceps skinfold thickness, 2-Hour serum insulin, Body mass index (BMI), Diabetes pedigree function and age of the patient. All these variables are the predictor variables for the model and the outcome of the diabetics check is the target variable. An instance of the table is given below.

	Pregnancy	Glucose	BloodPressure	SkinfoldThickness	Insulin	BodyMassIndex	DiabetesPedigreeFunction	Age	Class
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Table 1 First 5 rows of the Pima Indans Diabetes table

The variables in the dataset is explained below:

1. pregnancies - Number of times pregnant
2. Glucose - Plasma glucose concentration a 2 hour in an oral glucose tolerance test
3. BloodPressure - Diastolic blood pressure (mm Hg)
4. SkinThickness - Triceps skin fold thickness (mm)
5. Insulin - 2-Hour serum insulin (mu U/ml)
6. BMI - Body mass index (weight in kg/(height in m)^2)
7. DiabetesPedigreeFunction - Diabetes pedigree function
8. Age - Age (years)
9. Class - (0 or 1) class value 1 is interpreted as "tested positive for diabetes"

The data is highly imbalanced. It has 500 observations of non-diabetic cases and 268 observations of diabetic cases.

4. METHODOLOGY

The first step in the model building is to understand the distribution and the structure of the data using various summary functions and visualizations.

4.1. DATA EXPLORATION

It is seen from the 'info()' function that that the data consists of all integers and float values. Description of the numerical attributes of the data is given below in the following table.

	Pregnancy	Glucose	BloodPressure	SkinfoldThickness	Insulin	BodyMassIndex	DiabetesPedigreeFunction	Age	Class
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Table 2 Summary of the numerical variables

The check for the null values showed that there are no null values in the dataset. But from the summary of the numerical variables, we can see that there are variables with zero values. As per the check with the medical data, we found that some of these variables cannot have zero values in this dataset or in another way, zeros for some of these variables are meaningless. For example, glucose level in an alive human's body cannot be zero at any time. Similarly, the case of triceps skinfold thickness and Body Mass Index (BMI). But the case of insulin is different. People can be alive with having zero insulin (Dubois, 2013). But such people are categorized as diabetic and needs treatment to produce insulin in their body. Since this data also includes patients already diagnosed with Diabetics, this is quite a possibility.

The maximum value of triceps skinfold thickness is found to be 99mm. According to various internet sources and articles (Ramírez-Vélez, et al., 2016), triceps skinfold thickness varies between 10 to 60 mm in normal human beings. So, value of 99mm is for sure a typo error. We can impute this impossible value with the average of the class or the average of the entire column depending on its importance in the model. We may also have to check for other impossible values in the column below 99. But the next highest value is found to be 63mm.

4.2. Feature Selection

All of the eight features in the dataset was analyzed and their connection with the prediction class was identified. It was found that all of these features are important causes of diabetes and were included in the analysis.

Number of times pregnant

When women get pregnant, the placenta produces hormones that will inhibits the effects of insulin in the body. So more and more a women gets pregnant, there is more chance that she may be having diabetics as her body becomes weak and insulin become less effective in delivering sugar to the cells.

Glucose

Glucose is the sugar in the body which stores energy. This glucose is delivered to the cells with the help of insulin. When a patient is diabetic, it means that his effectiveness of delivering glucose to the blood has reduced for some reason and there are sugars circulating through the blood without being delivered.

Diastolic blood pressure

About 25% of people with Type 1 diabetes and 80% of people with Type 2 diabetes have high blood pressure (bloodpressureuk, n.d.). Blood pressure is the pressure of the blood in the arteries as the heart pumps it around the body. Diabetes changes the body chemistry in a way that increases the risk of blood pressure (diabetesaustralia, n.d.).

Triceps skin fold thickness

Skin fold thickness measurement provides an estimated size of the subcutaneous fat, which is the layer of subcutaneous tissue and composed of adipocytes. Subcutaneous fat is the major determinant of insulin sensitivity and has a strong association with insulin resistance. However, evidence to predict, the effect of duration of diabetes on skin fold thickness remains unclear (Selvi, et al., 2016).

Insulin

Insulin is required to transfer the glucose from the blood to the cells that require glucose. Pancreas is the gland that produces insulin in the body. If insulin is not produced enough in the body, glucose in the blood starts accumulating in the blood and cause diabetes.

BMI - Body mass index

An increase in body fat is generally associated with increased risk of metabolic diseases such as type 2 diabetes mellitus, hypertension and dyslipidemia. Data from two national surveys reported the common clinical observation that patients with higher BMI are at higher risk for having diabetes (Bays, et al., 2007).

Diabetes pedigree function

It provided some data on diabetes mellitus history in relatives and the genetic relationship of those relatives to the patient. It utilizes information from a person's family history to predict how diabetes will affect that individual.

Age

As people get aged their capacity to produce more insulin will reduce and their cells are ageing which require more energy to do work. In this case the sugar will be produced more in the body but the current level of insulin production may not be able to handle it. This leads to more sugar in the blood which may lead to diabetes.

4.3. Data Visualization

Data visualizations helped us to understand various issues with the data. It also helped in understanding the distribution of the data. Feature selection is done by finding the importance of various features and how they are related to the target class.

Histograms - Analysis of the Data

Histograms for all the attributes of the dataset was generated and given below.

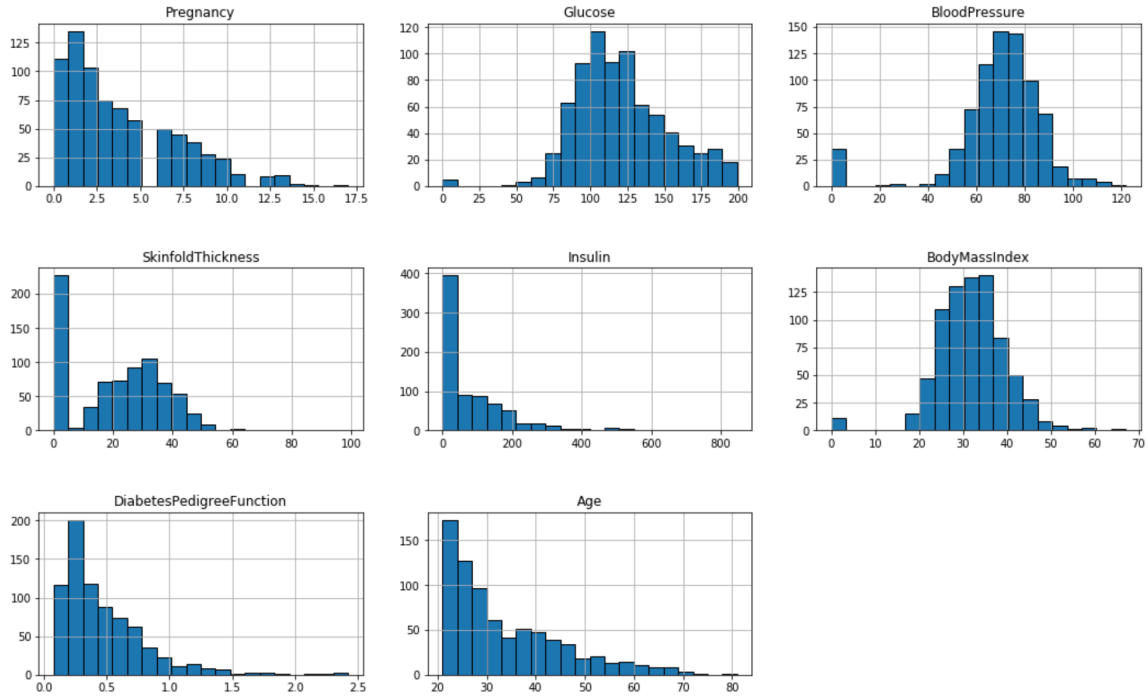

Figure 1 Histograms for the entire data

It is seen that the data have major issues. There are few values in 'Glucose' with zero value. There are many zero values present in case of 'Blood Pressure'. There are more than 200 observations with zero skinfold thickness. Also, there is an outlier towards the 100 in case of Skinfold thickness. Data of insulin is highly skewed. There are nearly 400 values with zero values. This can be either a missing value or a real value. Only patients identified with diabetics can have a zero value for insulin. To identify that we may have to plot the histogram for the cases with diabetics. The few outliers in the Body Mass Index is the missing values as BMI value for a person can never be zero. Even though the Diabetics pedigree function is highly skewed, all of the values are in the range. Same is the case with age.

Histograms - Analysis of the Diabetic cases

Histograms for all the cases identified as diabetic is plotted below.

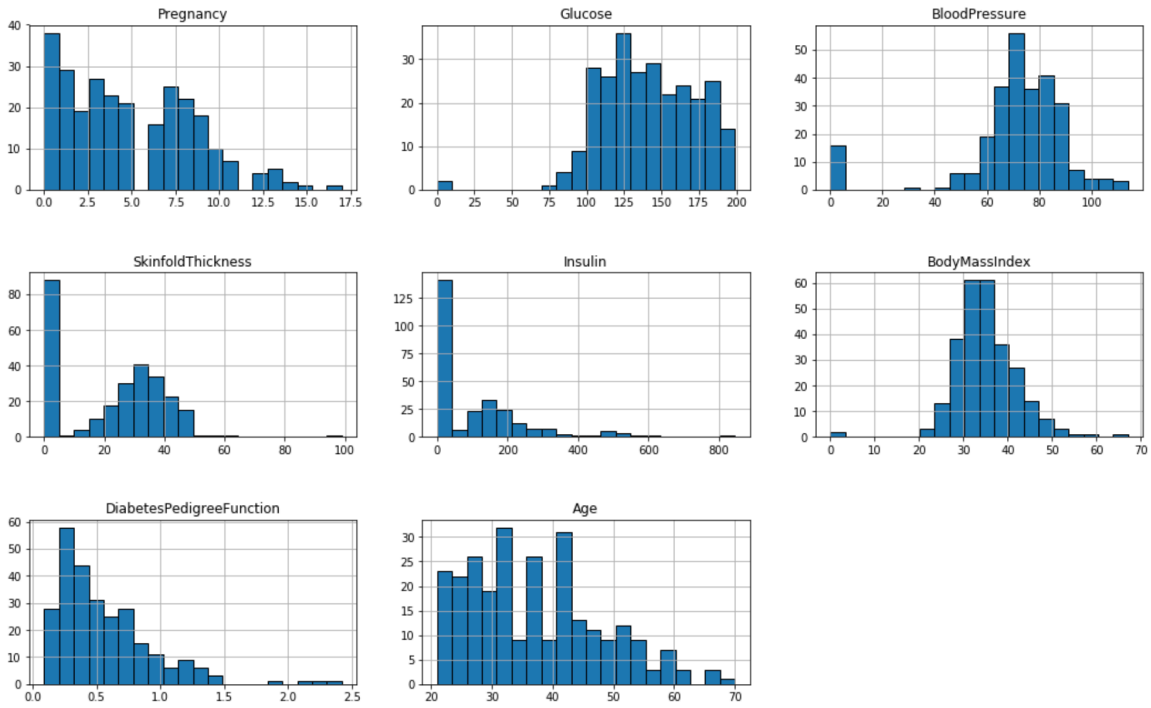

Figure 2 Histograms for the diabetic cases

Most of the missing values in 'blood pressure' comes from the diabetic cases. There are only few cases in diabetics where insulin is zero. Rest every data lies in the non-diabetic cases. This means that the zero values in insulin is not actually real cases and its data missing. Since insulin being zero is a very rare case of diabetics we can without doubt classify this as missing value.

Box Plots

Box plots were plotted for all the variables in the dataset. It is depicted below:

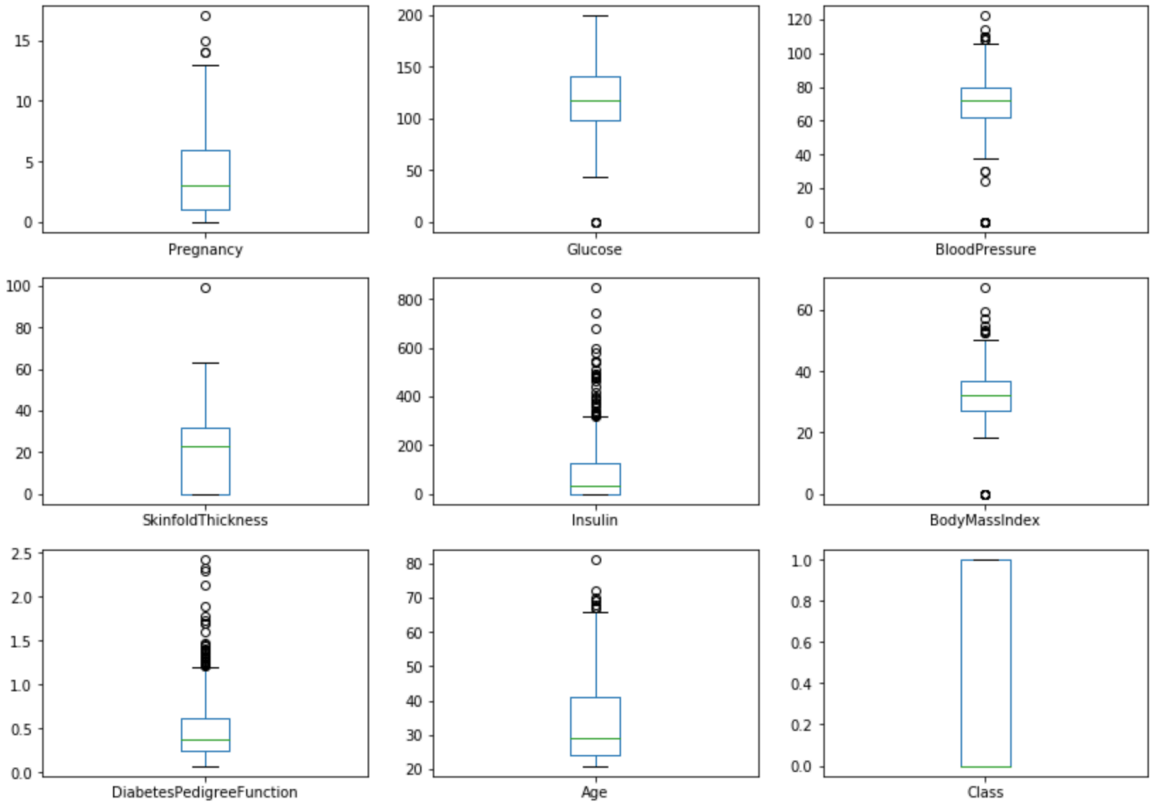

Figure 3 Box plots of the variables

There are few cases of pregnancy there it is more than 10. Even though these cases are rare, we can't remove them from the data as we are concerned about higher rate of pregnancy leading to diabetics. Some values of glucose being zero is clearly visible in the boxplot. There are some rare cases for extremely high blood pressure and some with low blood pressure. There is an outlier in the case of skinfold thickness as seen before in the data summary. The zero values in insulin is affecting the data a lot. The box plot function clearly shows that. The mean value is very less. So, we have to treat the insulin properly to get better results.

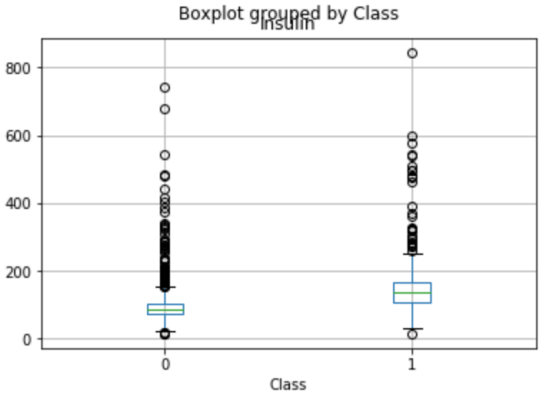

Figure 4 Box Plot of insulin by class

Multi-variate plots

Side by side box plot of insulin grouped by class is given here. It is seen that people who are diagnosed with diabetics are having more insulin level compared to the people with non-

diabetic cases. This is a contradiction to already known fact that in most of the cases of diabetics there is deficiency of insulin. This may be due to the data missing in the dataset.

Side by side box plot of pregnancy grouped by class is given below in figure 5. Women who had more number of pregnancies are having more cases of diabetics being reported. According to the dataset, rare cases of women being pregnant 13 times are having diabetics for sure.

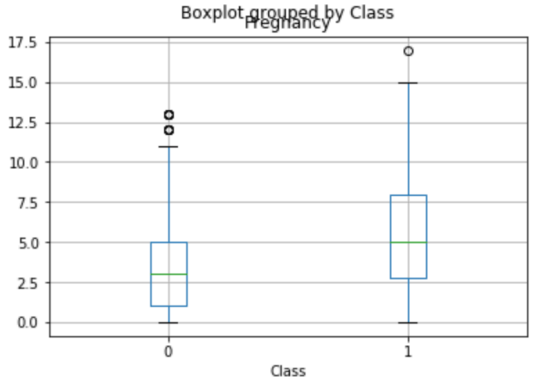

Figure 6 Box plot of pregnancy grouped by class

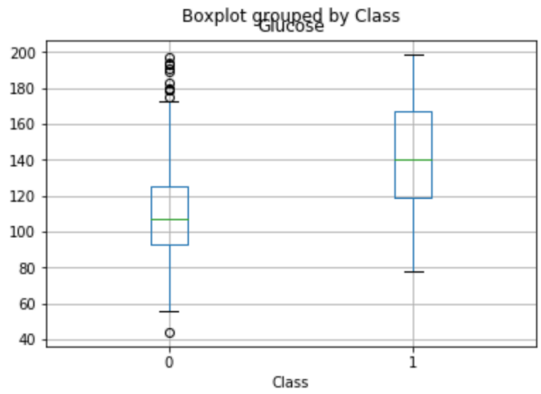

Figure 5 Box plot glucose grouped by class

Side by side box plot of pregnancy grouped by class is given above in figure 6. Diabetic Patients are found to be having more levels of glucose compared to other class. So, this is also an important class to determine the diabetic susceptibility of a person.

The below diagram shows the histogram of insulin values grouped by class. It is seen that people who are diagnosed with diabetics are having more insulin level compared to the people with non-diabetic cases.

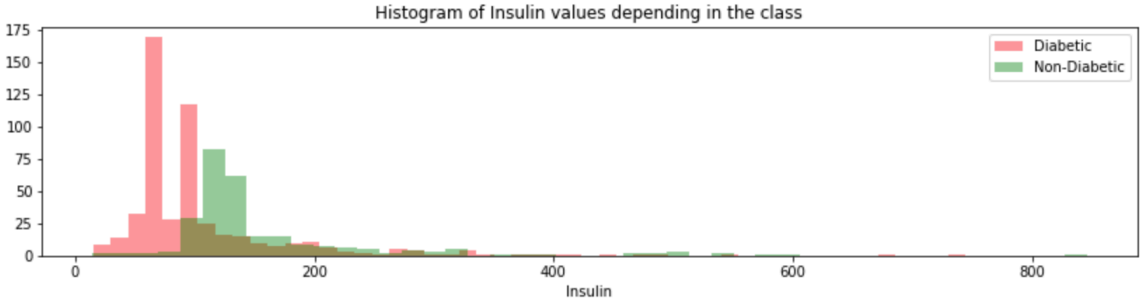

Figure 7 Histogram of insulin values grouped by class

The percentage chance of being diagnosed with diabetics are shown below in figure 8. More the level of glucose in the body there is more chance that one will be a diabetic patient.

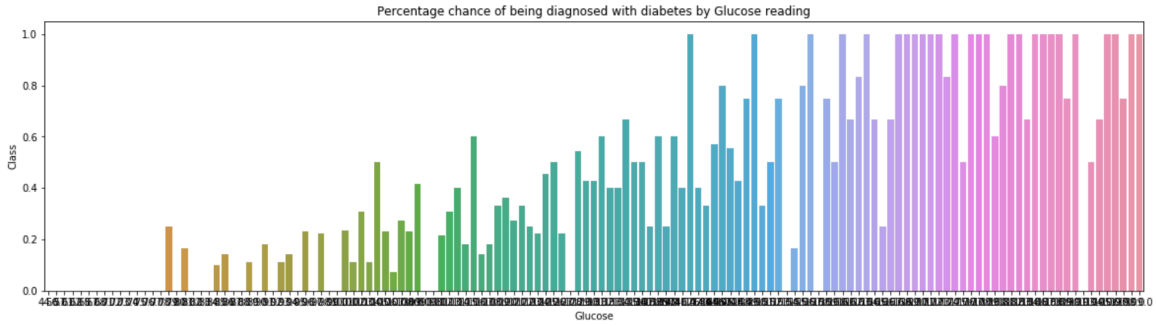

Figure 8 Chance of getting Diabetics

Correlation Diagram

Figure 9 Corrrelation diagram

There is not much correlation between the variables. So, no need to worry about leaking in duplicate information in the data.

Pair Plot

Figure 10 Pair plot of the dataset

4.4. Data Cleaning

In case of 'Glucose', there are 3 data points in non-diabetic cases and 2 data points in diabetic cases which are missing. We imputed the missing values in each class with the average value of 'Glucose' for each class. In case of 'Blood Pressure', there are 19 data points in non-diabetic cases and 16 data points in diabetic cases which are missing. We imputed the missing values in each class with the average value of 'Blood Pressure' for each class. In case of 'Skinfold Thickness', there are 121 data points in non-diabetic cases and 73 data points in diabetic cases which are missing. We imputed the missing values in each class with the average value of

'Skinfold Thickness' for each class. There is typo in 'Skinfold Thickness', and this value of 99 is replaced with average value of the class which is '26.77'. In case of 'Insulin', there is only 1 data point in non-diabetic cases which is missing. We imputed the missing value with the average value of 'BMI'.

After cleaning the data, we will check the summary of the again.

	Pregnancy	Glucose	BloodPressure	SkinfoldThickness	Insulin	BodyMassIndex	DiabetesPedigreeFunction	Age	Class
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	4.121662	121.691999	72.267826	26.770604	124.771038	32.400962	0.471876	33.240885	0.348958
std	3.206121	30.461151	12.115948	9.144460	91.935806	6.978689	0.331329	11.760232	0.476951
min	0.000000	44.000000	24.000000	7.000000	14.000000	0.000000	0.078000	21.000000	0.000000
25%	1.750000	99.750000	64.000000	20.371904	71.691720	27.500000	0.243750	24.000000	0.000000
50%	3.324384	117.000000	72.000000	23.404712	100.000000	32.050000	0.372500	29.000000	0.000000
75%	6.000000	141.000000	80.000000	32.000000	136.297569	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	63.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Table 3 Summary of the dataset after data cleaning

Box plot for all of the variables after data cleaning.

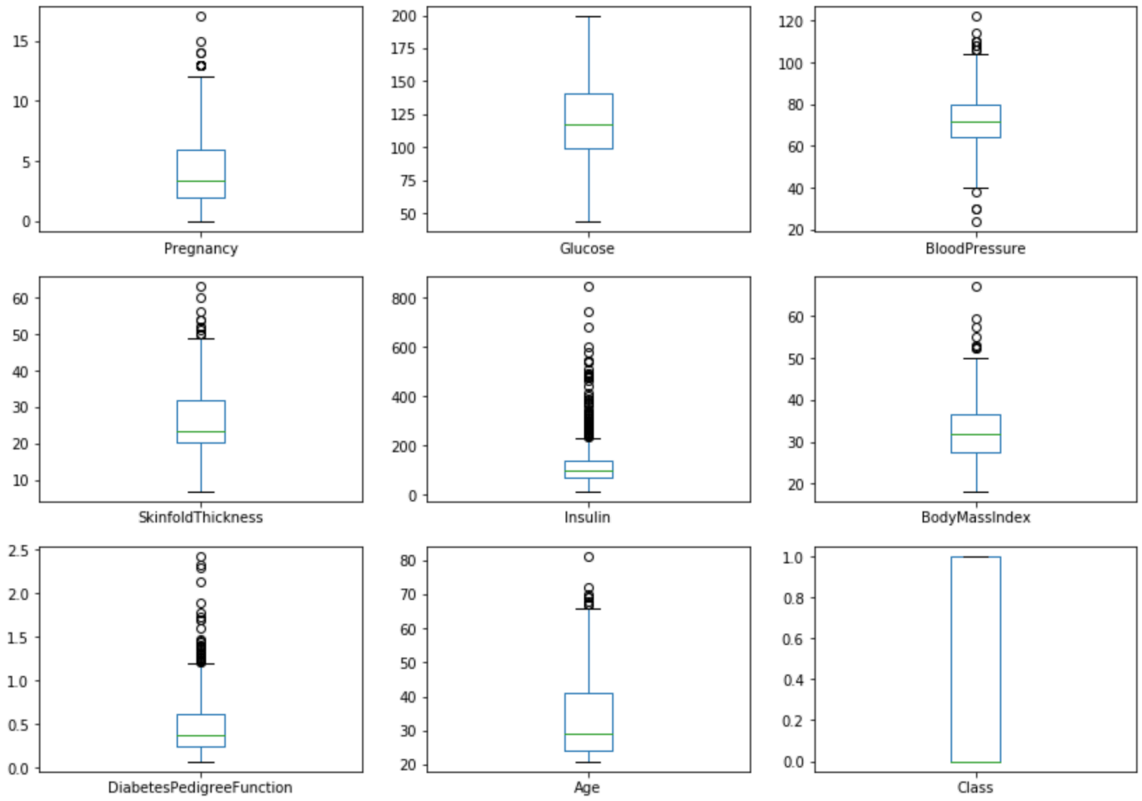

Figure 11 Box plot

The box plots are now clearer and we feel confident to proceed with the modelling process.

4.5. Data Modelling

Since there is high data imbalance and we are using KNN classifier as one of the model, we may have to deal with oversampling of the data for KNN model. We have used **SMOTE**: Synthetic Minority Oversampling Technique to oversample the data. Decision tree and Random Forest is by default dealing with the data imbalance and doesn't need to do it separately. So, we are supplying Decision tree classifier and random Forest Classifier with data before applying the SMOTE function. The data was split into 20:80 ratio for test and train.

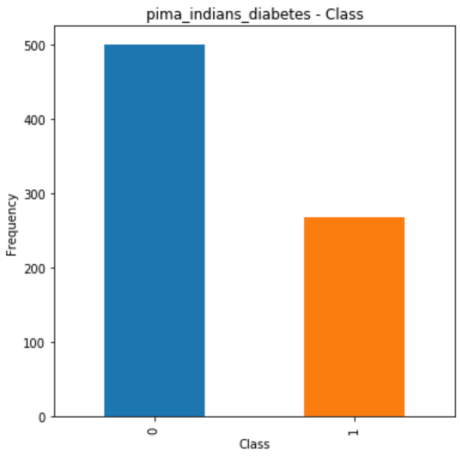

Figure 12 Balance of the data

4.5.1. K-Nearest Neighbor Classifier

To begin with we chose the value of K to be 3.

1. **With SMOTE**

The confusion matrix obtained for the KNN model was:

```
[81 22]
[ 9 88]
```

The classification report obtained for the model is:

	precision	recall	f1-score	support
0	0.90	0.79	0.84	103
1	0.80	0.91	0.85	97
avg / total	0.85	0.84	0.84	200

About 84% of the data was recalled by the model and the precision of the model was found to be 85%. The weighted average of the precision and the recall was found to be 84%. The number of samples of the true response that lie in that class are 200.

Normalizing

Since there are many attributes with different variance, we are normalizing the data. Scale function from the preprocessing package was used to normalize the data. After normalizing the SMOTE model is found to be overfitting the data.

The confusion matrix obtained for the KNN model was after normalizing

```
[102   1]
[  0  97]
```

The classification report obtained for the model is:

	precision	recall	f1-score	support
0	1.00	0.99	1.00	103
1	0.99	1.00	0.99	97
avg / total	1.00	0.99	1.00	200

The model is overfitting the data.

K-value optimization

For loop function is used to find the optimum k value for the model.

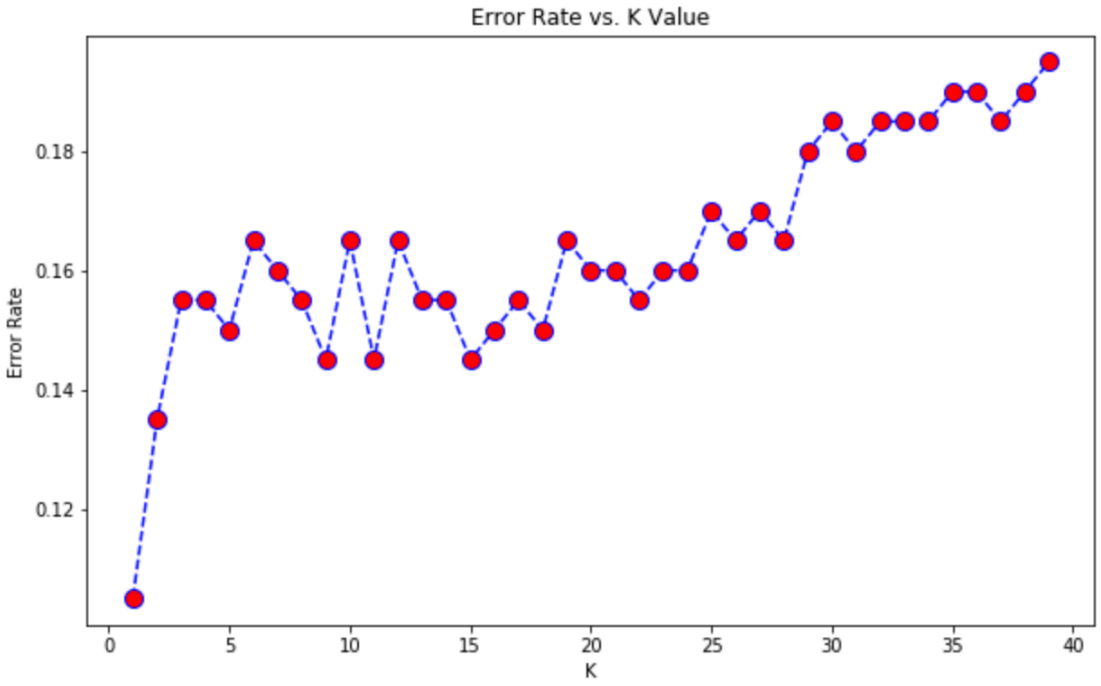

Figure 13 K value optimisation

We found that the k value of 1 is the best k value for better accuracy. So we rerun the model with new k value.

The confusion matrix obtained for the new KNN model is:

```
[90 13]
[ 2 95]
```

The classification report obtained for the model is:

	precision	recall	f1-score	support
0	0.98	0.87	0.92	103

```
          1       0.88      0.98      0.93        97

avg / total       0.93      0.93      0.92       200
```

The model had improved a lot with the new K value. About 93% of the data was recalled by the model and the precision of the model was found to be 93%. The weighted average of the precision and the recall was found to be 92%. The number of samples of the true response that lie in that class are 200.

Feature Selection using Hill Climbing method

The hill climbing method selected 4 features and the output was:

```
Score with 1 selected features: 0.6493506493506493
Score with 2 selected features: 0.7922077922077922
Score with 3 selected features: 0.8051948051948052
Score with 4 selected features: 0.8181818181818182
```

The value optimization was done for the new model as well.

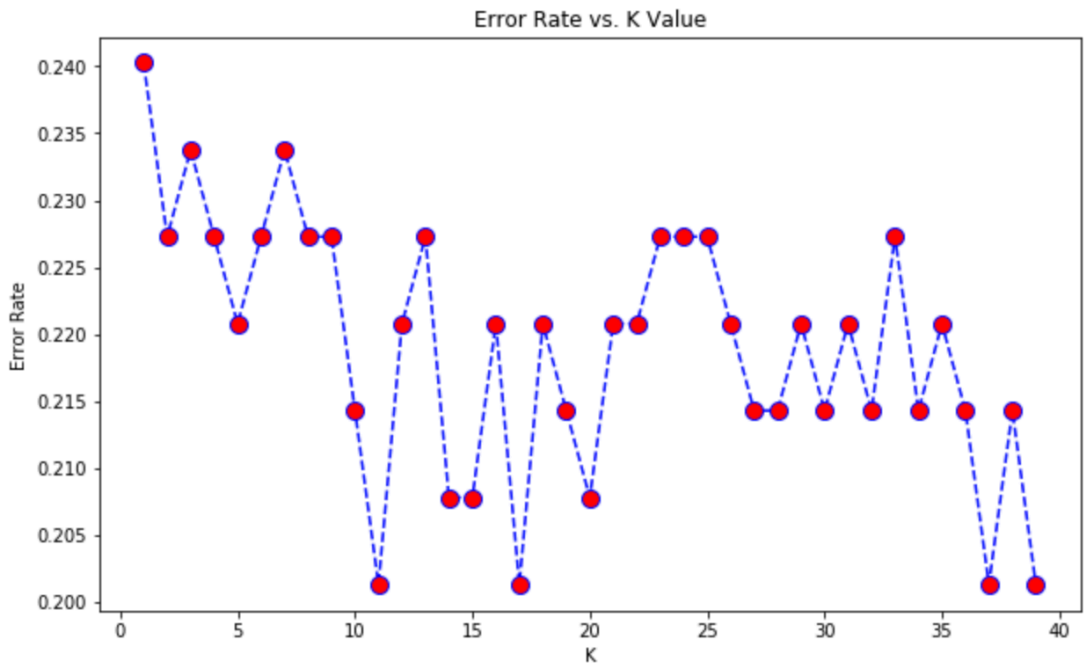

Figure 14 K value optimization for model obtained after removing attributes from hill climbing method.

New k value is found to be 11. The confusion matrix obtained for the new KNN model is:

```
[90 11]
[16 37]
```

The classification report obtained for the model is:

	precision	recall	f1-score	support
0	0.85	0.89	0.87	101
1	0.77	0.70	0.73	53
avg / total	0.82	0.82	0.82	154

The model accuracy had decreased with the Hill Climbing method. About 82% of the data was recalled by the model and the precision of the model was found to be 82%. The weighted average of the precision and the recall was found to be 82%. The number of samples of the true response that lie in that class are 154. So, we will not go with this model.

2. Without SMOTE

K- Nearest Neighbor without SMOTE is considered for the next analysis. The K-value optimization is done for this model.

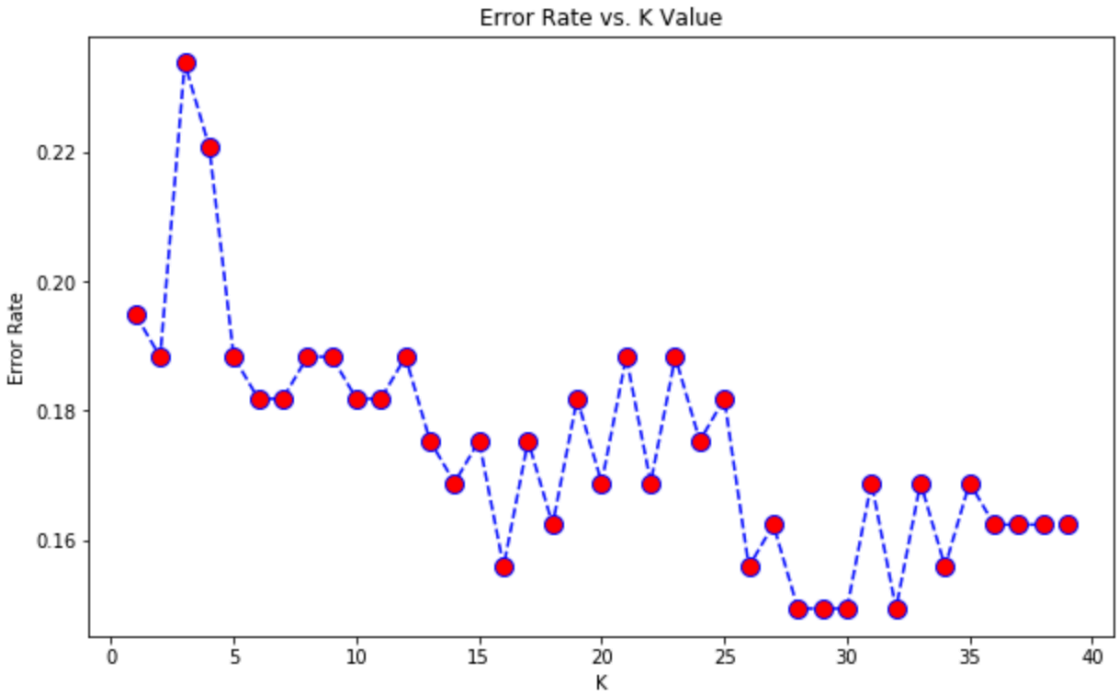

Figure 15 K value optimzation for model without SMOTE

The best k value is found to be 30. The confusion matrix obtained for the new KNN model is:

```
[86 14]
[ 8 46]
```

The classification report obtained for the model is:

	precision	recall	f1-score	support
0	0.91	0.86	0.89	100
1	0.77	0.85	0.81	54
avg / total	0.86	0.86	0.86	154

About 86% of the data was recalled by the model and the precision of the model was found to be 86%. The weighted average of the precision and the recall was found to be 86%. The number of samples of the true response that lie in that class are 154.

4.5.2. Decision Tree Classifier

The accuracy for decision tree classifier with default parameters is found to be 83.12%.

The confusion matrix obtained for the Decision Tree model is:

```
[86 14]
[12 42]
```

The classification report obtained for the model is:

	precision	recall	f1-score	support
0	0.88	0.86	0.87	100
1	0.75	0.78	0.76	54
avg / total	0.83	0.83	0.83	154

About 83% of the data was recalled by the model and the precision of the model was found to be 83%. The weighted average of the precision and the recall was found to be 83%. The number of samples of the true response that lie in that class are 154.

4.5.3. Random Forest Classifier

Initial parameters for the random forest classifier is found given to be 100 decision tree estimators and maximum depth of any given decision tree as 5.

The confusion matrix obtained for the Decision Tree model is:

```
[91  9]
[ 5 49]
```

The classification report obtained for the model is:

	precision	recall	f1-score	support
0	0.95	0.91	0.93	100
1	0.84	0.91	0.88	54
avg / total	0.91	0.91	0.91	154

This is the best model we have achieved. About 91% of the data was recalled by the model and the precision of the model was found to be 91%. The weighted average of the precision and the recall was found to be 91%. The number of samples of the true response that lie in that class are 154.

Feature importance

It is found that 'Insulin' is the most important feature for this model by collecting the ranking for each feature for predicting the target. The feature importance graph is given below.

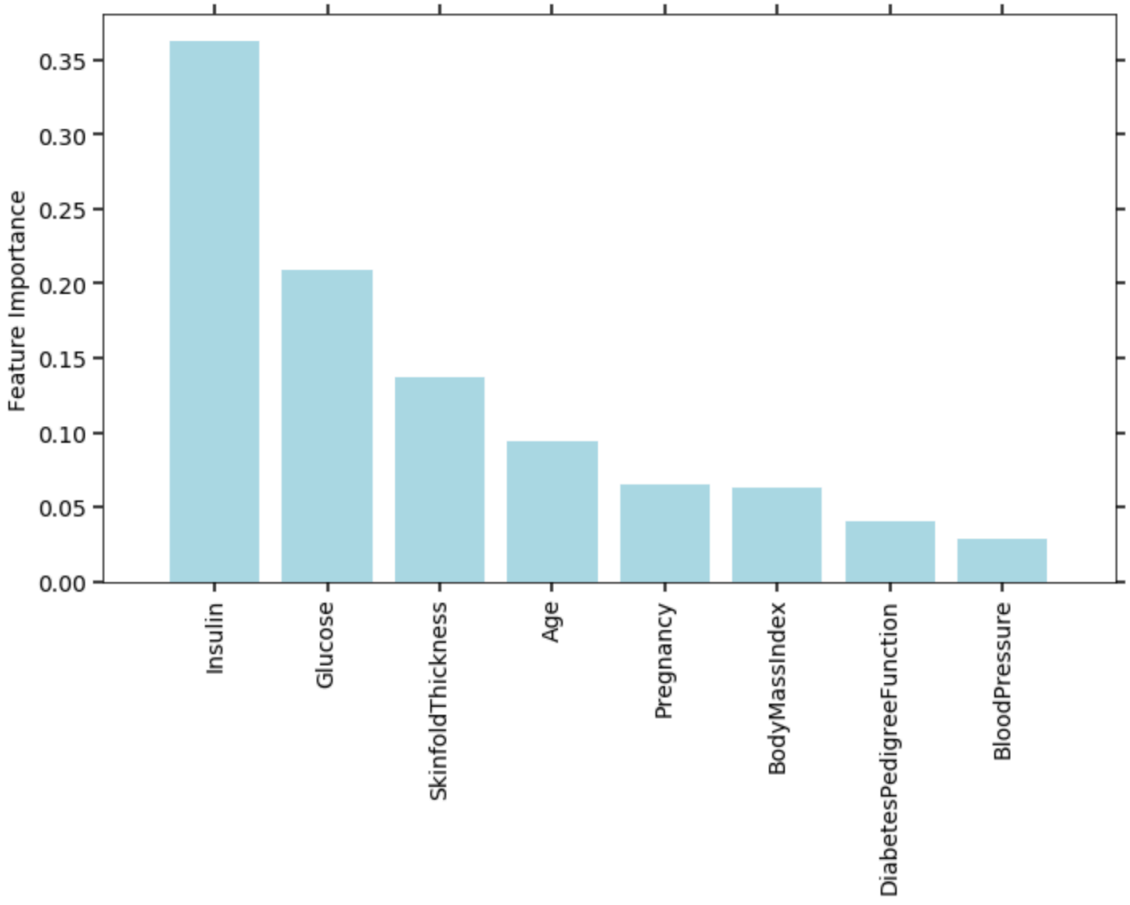

Figure 16 Feature importance

The most 6 important features identified to be insulin, glucose, Skinfold thickness, age, pregnancy and Body mass index.

5. Results

All the models obtained till now are compared against each other to identify the best model among them. Accuracy of the models are considered to be the basic criteria to identify the best model.

	Accuracy
Decision Tree	0.831169
Random Forest	0.909091
KNN	0.818182
KNN (Smote)	0.925000

Table 4 Accuracy of the basic models

5.1. Using only the important features

Using the most important 6 features identified from the model, that is insulin, glucose, Skinfold thickness, age, pregnancy and Body mass index.

	Accuracy
Decision Tree	0.850649
Random Forest	0.902597
KNN	0.850649
KNN (Smote)	0.880000

Table 5 Accuracy of the models with only important features

The accuracy of all other model except Random forest have increased after selecting only the important features.

5.2. Cross Validation

10-fold cross validation is applied on all of the model. The results are plotted below.

	CV Mean
KNN	0.753794
KNN (SMOTE)	0.910000
Decision Tree	0.802085
Random Forest	0.871138

Table 6 10-fold Cross validation

It is found that except Random Forest, all other models decreased in their accuracy after the cross validation. The box plot of the cross-validation result is given below.

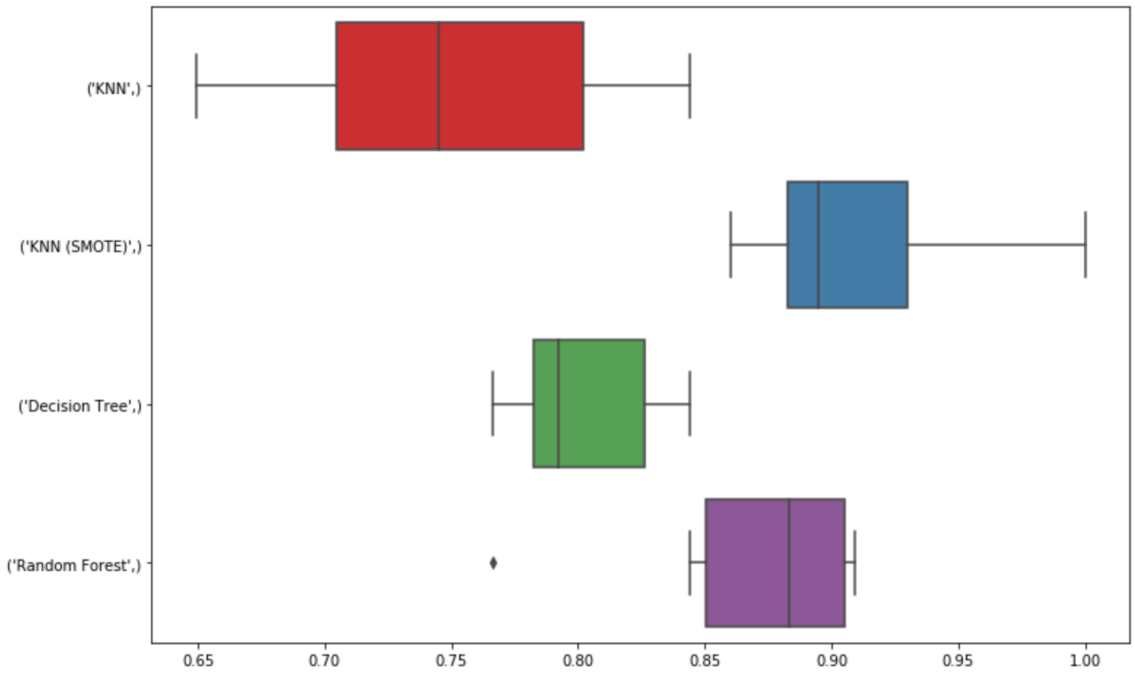

Figure 17 Box plot of the cross validation result

It can be seen that the KNN with SMOTE is almost overfitting the data at some validations as there is a case of KNN nearly 100% accuracy. Random Forest is not having much variation and is therefore a stable model.

6. Parametric tuning

Parametric tuning is applied on Random forest model and the Decision Tree model to improve the accuracy of the model.

6.1. Tuning Random Forest

The method used was Random Search grid and range of parameters provided for the search grid is given below:

'n_estimators': [200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000]

'max_features': ['auto', 'sqrt']

'max_depth': [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None]

'min_samples_split': [2, 5, 10]

'min_samples_leaf': [1, 2, 4]

'bootstrap': [True, False]

The parameters obtained for the random forest model is given below:

'bootstrap'	True,
'max_depth'	100,
'max_features'	'sqrt',
'min_samples_leaf'	1,
'min_samples_split'	5,
'n_estimators'	400

Table 7 Tuned Parameters for Random Forest Model

With the new parameters, Cross validation was done on the new Random Forest model. Accuracy of the base model was 87.01%. The accuracy of the model with tuned parameters were 87.66%. There was an improvement of 0.75% in accuracy.

6.2. Tuning Decision Tree

Range of parameters provided for the search grid is given below:

'max_depth': [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None],

'max_features': ['auto', 'sqrt'],

'min_samples_leaf': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],

'min_samples_split': [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16]

The parameters obtained for the decision tree model is given below:

'max_depth'	None
'max_features'	'sqrt'
'min_samples_leaf'	5
'min_samples_split'	8

Table 8 Tuned Parameters for decision tree model

With the new parameters, Cross validation was done on the new Decision tree model. Accuracy of the base model was 82.47%. The accuracy of the model with tuned parameters were 85.06%. There was an improvement of 3.15% in accuracy, which is really good. The decision tree obtained with this model are given below.

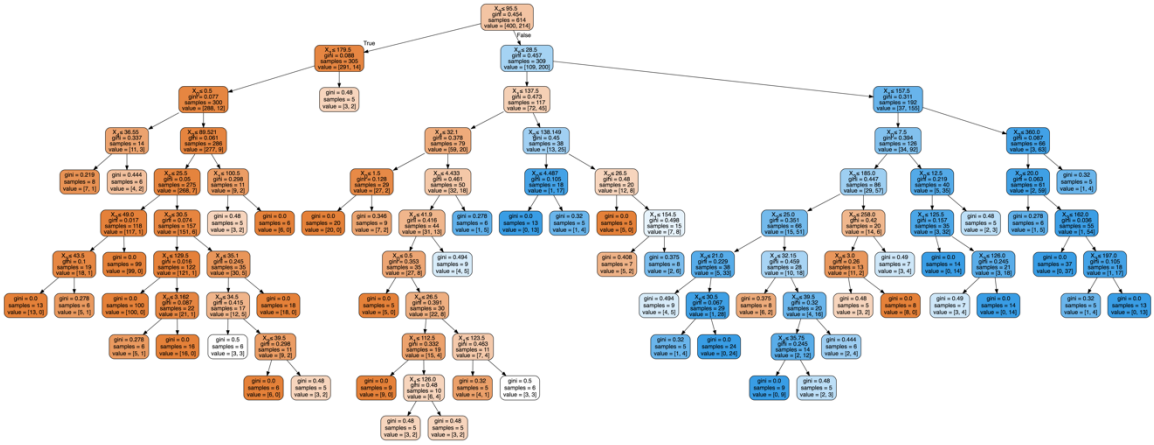

Figure 18 Decision Tree obtained for the final tuned decision tree model.

6.3. Cross Validation

Cross validation result for new decision tree and random forest model is given below along with other models.

	CV Mean
KNN	0.753794
KNN (SMOTE)	0.910000
Decision Tree	0.820215
Random Forest	0.871121

Figure 19 Cross validation result for tuned decision tree and random forest model

The cross-validation result dint showed much improvement with the decision tree. But the random forest is still stable with good accuracy. The box plot of the cross-validation result is shown below.

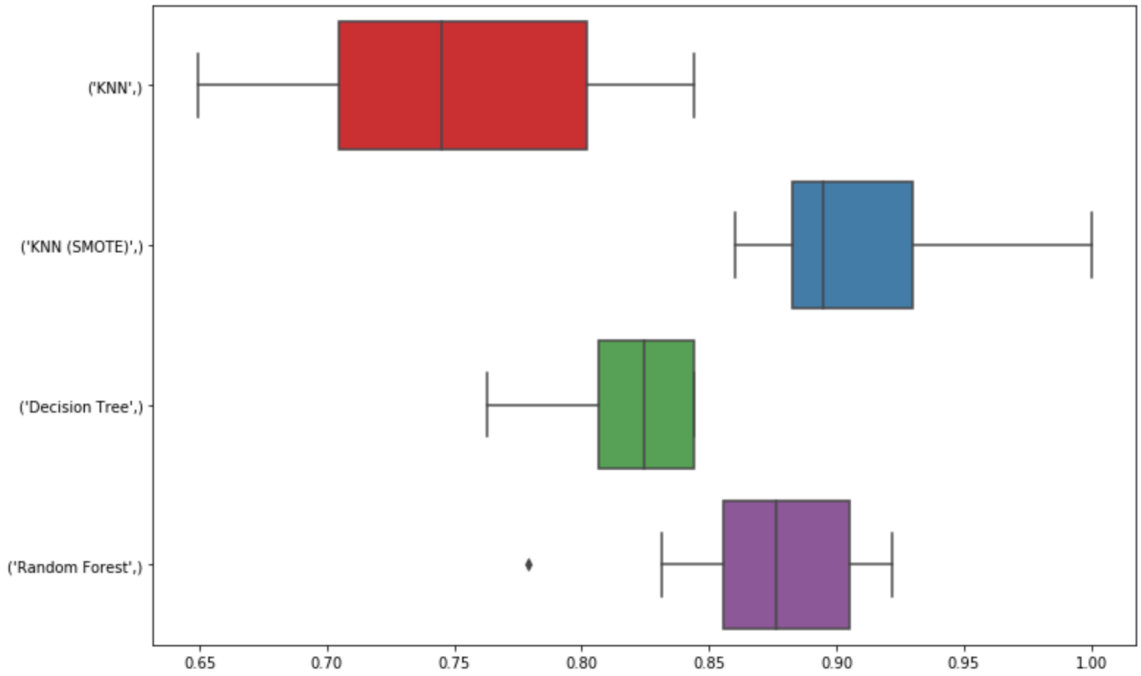

Figure 20 The box plot of cross-validation result after parameter tuning

According to this cross-validation result, KNN with SMOTE is having the highest accuracy among all the model. But it is also overfitting the model at some combinations of split. Random forest is having the next best accuracy followed by decision tree and KNN without SMOTE.

7. Discussion

Even though we found that KNN with SMOTE is having the highest accuracy, the best model according to me is the Random Forest model having only important features and with tuned parameters. Oversampling of the data would have adverse effect on the model as this is not really the original data and SMOTE function is actually extrapolating the feature variables to make more samples of the data. But since we have provided the Random forest model only with the original data, result of random forest is more assuring. Also there are chances of overfitting of the model with oversampling of the data as is seen in some cases of the KNN model. So the final model is chosen as Random Forest.

8. Conclusion

It was found that the Random Forest model with maximum number of levels in a tree as 100, maximum features as square root, minimum samples in a leaf as 1, minimum samples in a node to split the data as 5 and number of trees as 400 was the best model to predict if a person had diabetics or not. The data fed to the model was normalized and only the most important 6 features; 'Pregnancy','Glucose','SkinfoldThickness','Insulin','BodyMassIndex' and 'Age' were only considered for the final model. The model gave an accuracy of 87.11 at a 10-fold cross

validation process. The model was stable in all 10 times of the validation and proves to be the best model for the data provided.

9. BIBLIOGRAPHY

Bays, H. E., Chapman, R. H. & Grandy, S., 2007. The relationship of body mass index to diabetes mellitus, hypertension and dyslipidaemia: comparison of data from two national surveys. *International Journal of clinical practice,* may, 61(5), p. 737–747.

bloodpressureuk, n.d. *Diabetes and high blood pressure.* [Online]
Available at: http://www.bloodpressureuk.org/BloodPressureandyou/Yourbody/Diabetes

diabetesaustralia, n.d. *diabetesaustralia.* [Online]
Available at: https://www.diabetesaustralia.com.au/blood-pressure

Dubois, W., 2013. *DIABETES MINE.* [Online]
Available at: https://www.healthline.com/diabetesmine/ask-dmine-lifespan-sans-insulin#1

GitHub, n.d. *GitHub.* [Online]
Available at: https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.data.csv

Ramírez-Vélez, R. et al., 2016. Triceps and Subscapular Skinfold Thickness Percentiles and Cut-Offs for Overweight and Obesity in a Population-Based Sample of Schoolchildren and Adolescents in Bogota, Colombia. *PMC,* 8 Oct.Volume 595.

Selvi, C., Pavithra.N & Saikumar.P, 2016. Skin Fold Thickness in Diabetes Mellitus: A Simple Anthropometric Measurement May Bare the Different Aspects of Adipose Tissue. *IOSR Journal of Dental and Medical Sciences (IOSR-JDMS) ,* Nov, 15(11), pp. 07-11.