International Journal For Technological Research In Engineering
Volume 9, Issue 11, July-2022 *ISSN (Online): 2347 - 4718*

DEVELOPMENT OF MACHINE LEARNING MODEL FOR PREDICTION OF DIABETIC IN PATIENTS

1Satyaveer Saini, 2Irfan Khan
[1]PG Scholar, [2]Assistant Professor,
[1,2] Department Computer Science And Engineering
Shekhawati Institute Of Engineering And Technology
SIKAR (RAJASTHAN)

Abstract: - The health care industry has become a major source of big data because of the digitization revolution. The examination of these data could be a fantastic resource for gaining fresh perspectives and raising people's awareness of health issues. Worldwide, diabetes and its complications are acknowledged as the biggest danger to public health. Diabetes is currently one of the most prevalent, chronic, and deadly diseases in the world due to various complications. In present work, healthcare data of female diabetic patient has been studied and derive the various insights from it. Machine learning algorithm such as logistic regression, decision tree, random forest classifier, Support Vector Machine (SVM) and K-Nearest Neighbour (KNN) classification has been used for prediction and modelling of the data. Python programming has been used as a tool for data analysis and various exploratory data analysis has been performed to determine the structure of the data. The predicted model found good relation with the collected data and gives a higher accuracy. This model help medical person to predict the risk associated for prediction of diabetic in female patients.

Keywords: - Machine Learning, Python programming, Healthcare, Data science

1. INTRODUCTION

In present work, data of women has been collected from hospital, which includes various categories. The objective of the collection of data is to predict a model, which can show women, has diabetic or not. The data include 768females with 09 different category. It include the month of pregnancies a women has, their glucose and blood pressure level, their skin thickness, whether they are taking insulin or not, their Body Mass Index (BMI), their diabetes pedigree function, their age and finally actual condition of women means is she a diabetic or not in terms of outcome.

After collecting the data, we have used python programming for building the machine-learning model. Python is open source software free available free from google. It is a high level, interpreted and general-purpose language.

Jupyter notebook is used for preparing or developing the code. This can be done on google colab available online or installing in personal computer. After installing the Jupyter notebook, we need to follow some standard procedure. Given below is the procedure needed to perform the coding in python programming.

2. METHODOLOGY

This dataset is originally from the National Institute of Diabetes, Digestive, and Kidney Diseases. The objective of the dataset is to diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

2.1 Problem Statement:
Build a model to accurately predict whether the patients in the dataset have diabetes or not?

2.2 Dataset Description:
The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

Pregnancies: Number of times pregnant
Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
Blood Pressure: Diastolic blood pressure (mm Hg)
Skin Thickness: Triceps skin fold thickness (mm)
Insulin: 2-Hour serum insulin (mu U/ml)
BMI: Body mass index (weight in kg/(height in m)^2)
DiabetesPedigreeFunction: Diabetes pedigree function
Age: Age (years)
Outcome: Class variable (0 or 1) 268 of 768 are 1, the others are 0

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunctio
0	6	148	72	35	0	33.6	0.62
1	1	85	66	29	0	26.6	0.35
2	8	183	64	0	0	23.3	0.67
3	1	89	66	23	94	28.1	0.16
4	0	137	40	35	168	43.1	2.28

Figure 1 first five Data set for analysis purpose

Total data are 768 rows and 9 columns including Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, DiabetesPedigreeFunction, Age and Outcome

2.3 Approach:
Following pointers will be helpful to structure your findings.

1. Perform descriptive analysis. It is very important to understand the variables and corresponding values. We need to think through - Can minimum value of below listed columns be zero (0)? On these columns, a value of zero does not make sense and thus indicates missing value.
- Glucose
- Blood Pressure
- Skin Thickness
- Insulin
- BMI

How will you treat these values?

2. Visually explore these variables; you may need to look for the distribution of these variables using histograms. Treat the missing values accordingly.

3. We observe integer as well as float data-type of variables in this dataset. Create a count (frequency) plot describing the data types and the count of variables.

4. Check the balance of the data by plotting the count of outcomes by their value. Describe your findings and plan future course of actions.

5. Create scatter charts between the pair of variables to understand the relationships. Describe your findings.

6. Perform correlation analysis. Visually explore it using a heat map.

(Note: Do not focus on visualization aspects when working with SAS)

7. Devise strategies for model building. It is important to decide the right validation framework. Express your thought process. Would Cross validation be useful in this scenario?

(Note: if you are working with SAS, ignore this question and perform stratified sampling to partition the data. Create strata of age for this.)

8. Apply an appropriate classification algorithm to build a model. Compare various models with the results from KNN.

(Note: if you are working with SAS, ignore this question. Apply logistic regression technique to build the model.)

9. Create a classification report by analysing sensitivity, specificity, AUC(ROC curve) etc. Please try to be as descriptive as possible to explain what values of these parameter you settled for? any why?

10. Create a dashboard in tableau by choosing appropriate chart types and metrics useful for the business. The dashboard must entail the following:

a) Pie chart to describe the diabetic/non-diabetic population
b) Scatter charts between relevant variables to analyse the relationships
c) Histogram/frequency charts to analyse the distribution of the data
d) Heat map of correlation analysis among the relevant variables
e) Create bins of Age values – 20-25, 25-30, 30-35 etc. and analyse different variables for these age brackets using a bubble chart.

3. RESULT AND DISCUSSION

Visually explore these variables using histograms. Treat the missing values accordingly.

3.1 Exploration for Glucose

Here we can see the plasma Glucose concentration level approximately of 165 people are varies b/w 80 to 137 in huge proportion. The highest range of people approximately 212 people's Glucose level are varies b/w 98 and 120. As we can see here, few of the values are zero (0) in glucose columns. Our problem statement says zero (0) indicates like a missing value it does not make sense here, so our first priority to treat this missing values. In the data description, part we have seen median is 117 and mean is 120.89 for glucose means mean is greater than median i.e. positive skewed means right skewed. Now we can treat our zero (0) values by mean or median.

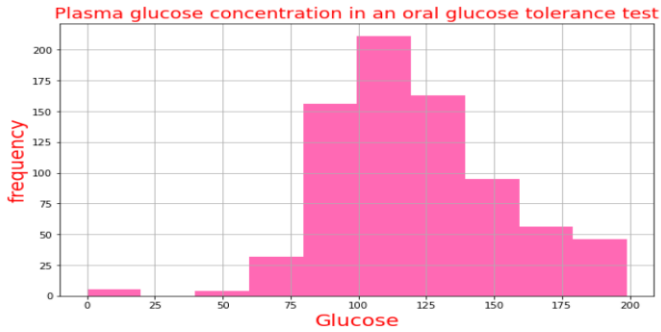

Figure 2 relation of glucose with respect to frequency at plasma glucose concentration

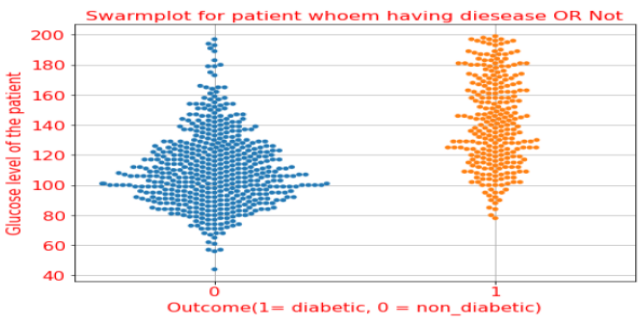

Figure 3 Swarm plot for detection of diabetic or non-diabetic

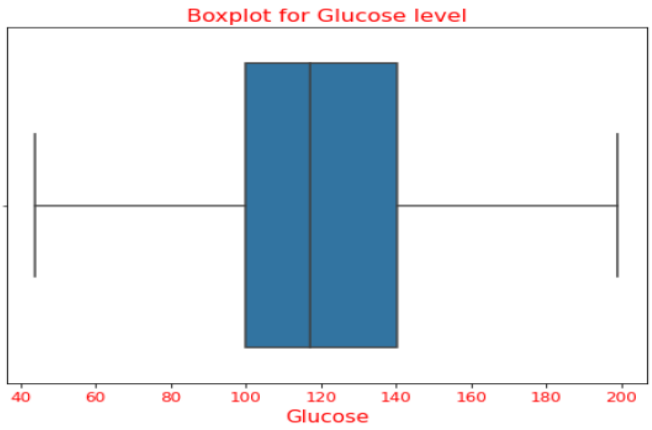

Figure 4 Box plot for determination of outlier in case of glucose

We can infer here that increasing glucose levels are likely to lead to a diabetic patient.

Exploration for Blood Pressure

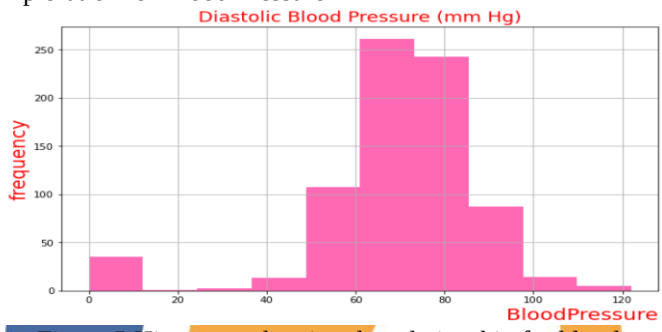

Figure 5 Histogram showing the relationship for blood pressure and number of patients

We can see here, approximately 260 Peoples blood pressure are lying b/w 62 to 77. As we can see here, few of the values (i.e approximately 35 values) are zero (0) in Blood Pressure columns. Our problem statement says zero (0) indicates like a missing value it does not make sense here, so our first priority to treat these missing values. We have seen in the data description part median is 72 and mean is 69.10 for Blood Pressure. Here we can see mean is less than median so this is left skewed means negatively skewness in data. Now we can treat our zero (0) values by mean or median.

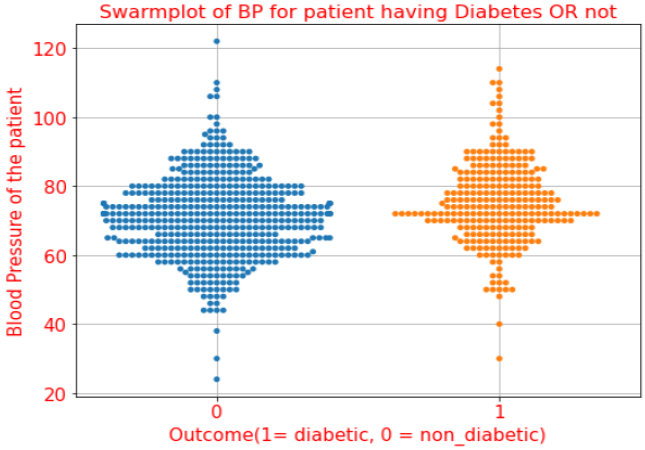

Figure 6 Swarm plot for detection of diabetic or non-diabetic in case of blood pressure

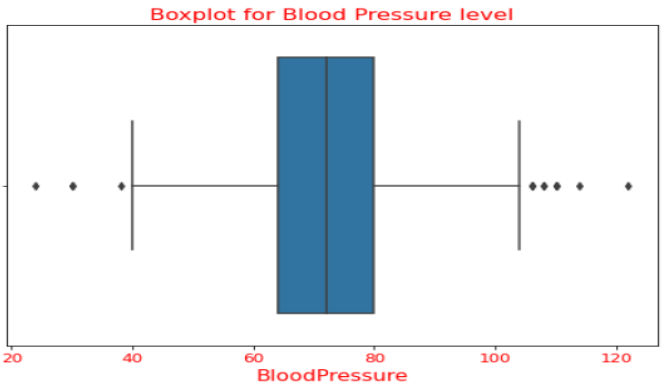

Figure 7 Box plot for determination of outlier in case of blood pressure

In figure, we can see that blood pressure record some values look like outliers here. However, changing these values indicates that we are making changes to the data. Having outliers can reduce the statistical power and affect the model equation, but this is part of the data so I do not want to treat these values as outliers.

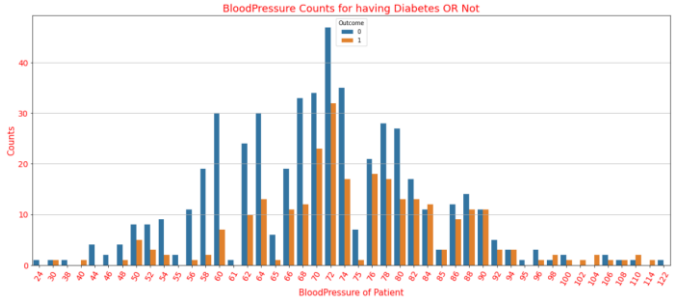

Figure 8 bar chart of relationship for blood pressure of patient with respect to number of patients

3.2 Exploration for Insulin

This is a right skewed histogram. Most of the values (i.e 374 values out of 768 records) in insulin columns are zero (0). As we can see here, most of the values are zero (0) in insulin columns. Our problem statement says zero (0) indicates like a missing value it does not make sense here, so our first priority to treat this missing values. In the data description, part we have seen median is 30.50 and mean is 79.73 for insulin Columns. Means mean is even more than double of median that indicates histogram is highly positive skewed (right skewed). Now we can treat our zero (0) values by median.

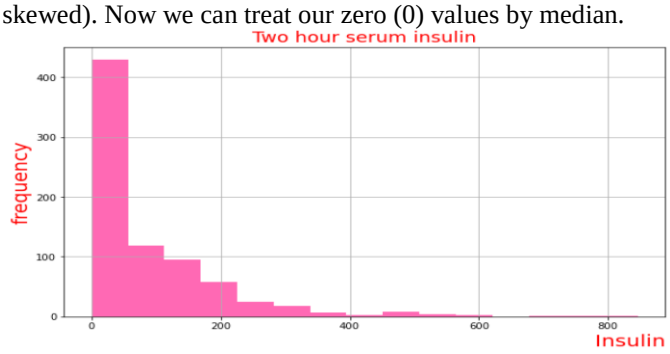

Figure 9 Histogram showing the relationship for insulin and number of patients.

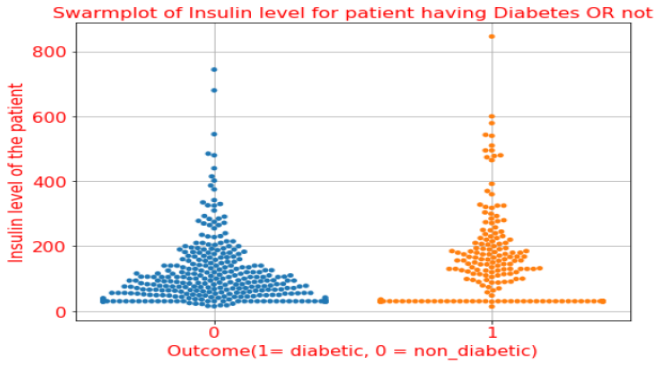

Figure 10 Swarm plot of insulin level of patient showing diabetes or not

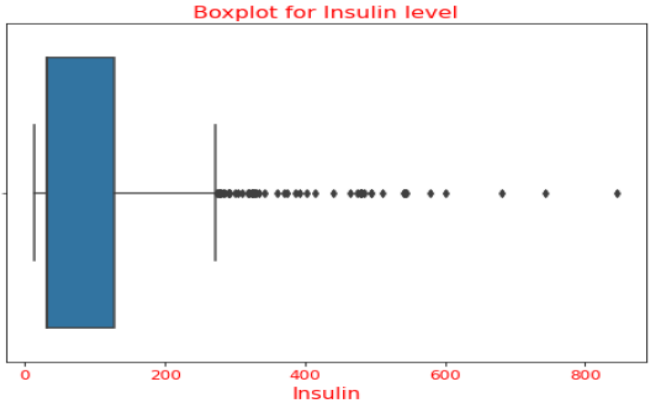

Figure 11 Box plot for determination of outlier in case of insulin

We can see here, out of 768 records only 49 values exceed the upper whisker (upper bound) which is only 6.38%. But we would not like to entertain these values in order to maintain the originality in the data.

3.3 Exploration for Body Mass Index

This histogram indicates that there are approximately 140 values of BMI between 33 and 36 and 110 records lie b/w BMIs of 23 and 37. This histogram also shows like some outlier's values on the left and right sides of the histogram, which we can see and treat later by using boxplot. As we can see here, few of the values are zero (0) in BMI columns. Our problem statement says zero (0) indicates like a missing value it does not make sense here, so our first priority to treat these missing values. In the data description, part we have seen median is 32.00 and mean is 31.99 for BMI Columns. Means mean and median both are pretty equal and lying at the same points. Now we can treat our zero (0) values by median or mean.

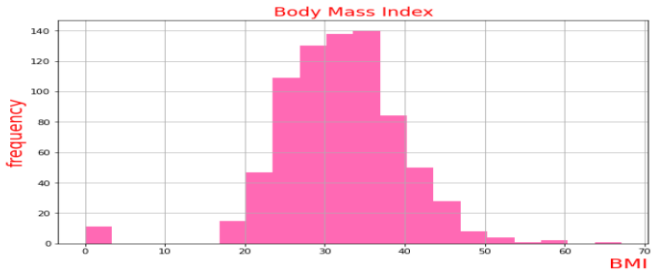

Figure 12 Histogram showing the relationship for body mass index with respect to frequency

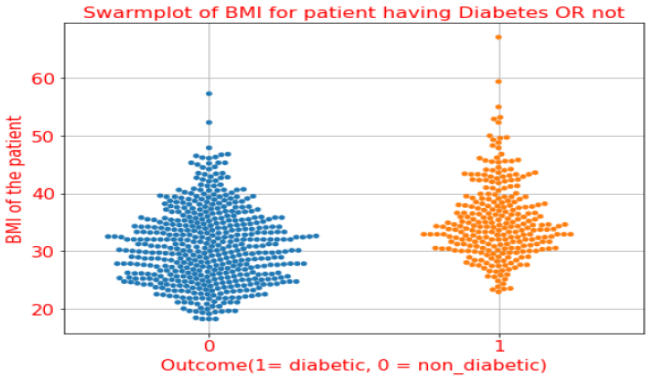

Figure 13 Swarm plot of BMI of patient showing diabetes or not

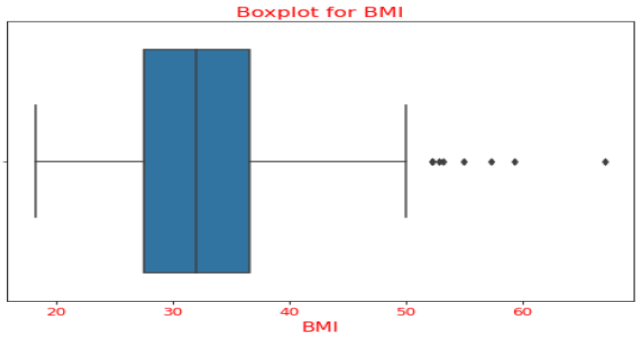

Figure 14 Box plot for determination of outlier in case of BMI

We can see here that 08 values out of 768 are outliers, which is about 1% as compared to total records. Which is negligible, we want to keep data as it is so we are not going to treat these values.

Exploration for Skin Thickness My observation: This histogram indicates that there approximately 200 values shows skin Thickness zero (0) and approximately 75 records are having skin thickness around 15 to 40 mm. This histogram also shows like a single outlier values, which is shows skin thickness of 100 mm that would be clearer when we will use boxplot and we can treat these values there. As we can see here, few of the values are 0 in Skin Thickness columns. Our problem statement says zero (0) indicates like a missing value it does not make sense here, so our first priority to treat these missing values. In the data description, part we have seen median is 23.00 and mean is 20.53 for Skin Thickness Columns. Means mean is less than median that is represent data is slightly left skewed (negatively skewed). Now we can treat our zero (0) values by median or mean.

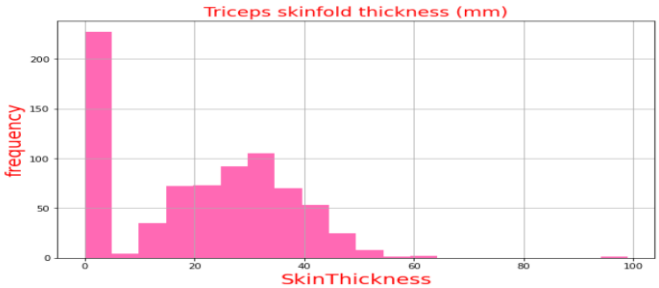

Figure 15 Histogram showing the relationship of a skin thickness

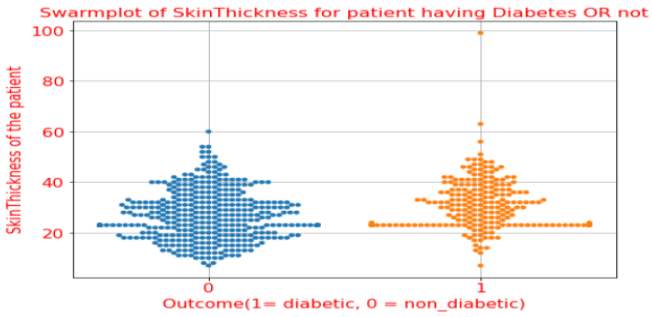

Figure 16 Swarm plot of skin thickness

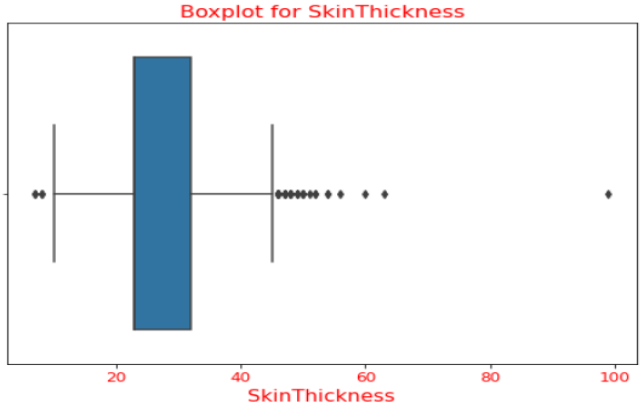

Figure 17 Box plot for determination of outlier in case of BMI

We can see here that out of 768 records only 35 values are outliers, which is only 4.5% of total records. However, changing these values is going to change the original records so we want to keep these values same. Most records for skin thickness are between 21 and 25 mm in about 300 people.

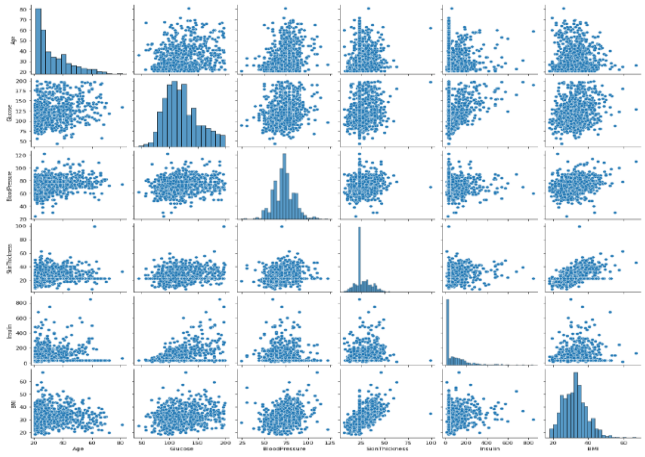

Figure 18 combined charts between all the parameters

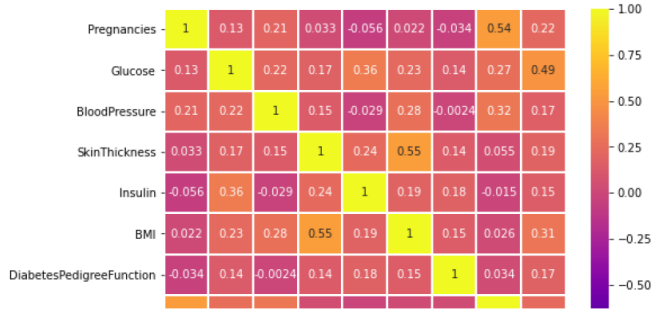

Figure 19 Heat map to show the co-relationship between the all parameters

3.4 Machine learning algorithm

3.4.1 Applying Logistic Regression Algorithms

Figure 20 ROC curve for all the parameters

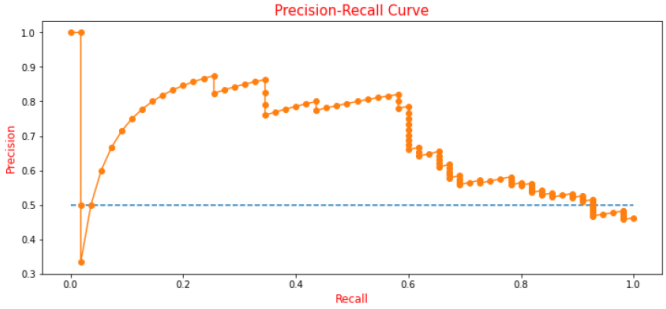

Figure 21 Precision recall curve for data

3.4.2 Apply Decision Tree Model

Decision Tree classifier performs with accuracy of 70.77% on this test data

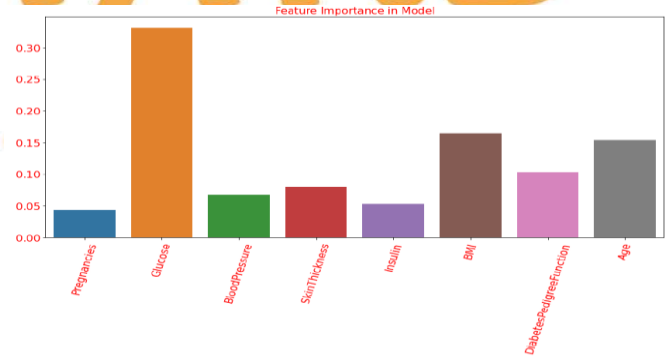

Figure 22 Feature engineering model using decision tree model

3.4.3 Apply Random Forest Classifier

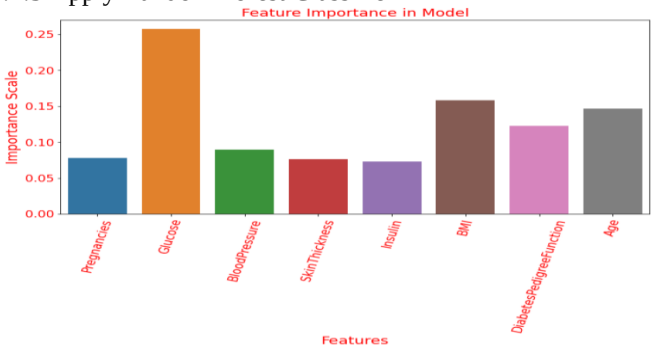

Figure 23 Feature engineering model using random forest classifier

3.4.4 K-Nearest Neighbour (KNN) Classification

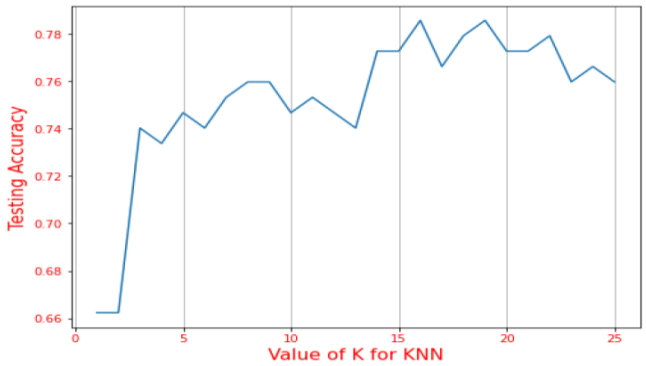

Figure 24 Testing of KNN model

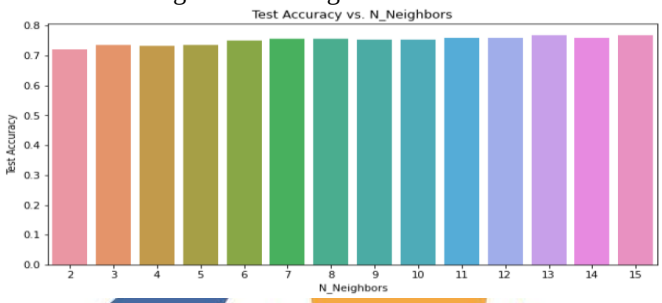

Figure 25 Test accuracy of KNN model

3.4.5 Ensemble Learning, Boosting Technique, Adaptive Boosting

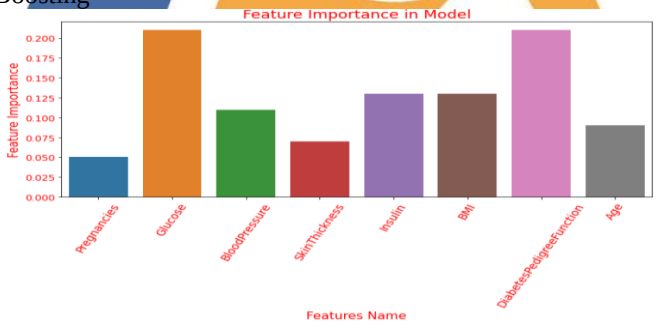

Figure 26 Feature engineering using

Sr No	Algorithm	Parameter/ Variable	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Train Accuracy - Test Accuracy
1	Logistic		0.71	0.51	0.59	0.76	0.73	0.6	0.66	0.78	-0.02
2	DT	Default	1	1	1	1	0.6	0.73	0.66	0.73	0.27
		max_depth = 5	0.77	0.73	0.75	0.83	0.71	0.67	0.69	0.79	0.04
3	RF	Default	1	1	1	1	0.66	0.64	0.65	0.75	0.25
		max_depth=7, min_samples_leaf=5, n_estimators=50	0.86	0.76	0.8	0.87	0.67	0.65	0.66	0.76	0.11
4	SVM	default	0.78	0.58	0.66	0.8	0.71	0.58	0.64	0.77	0.03
		SVC(C=20, gamma=0.005, probability=True)	0.73	0.49	0.59	0.76	0.79	0.6	0.68	0.8	-0.04
5	Knn	(n_neighbors=3)				0.85				0.74	0.11
		(n_neighbors=16)				0.78				0.79	-0.01
		(n_neighbors=15)				0.79				0.77	0.02
6	AdaBoost	(n_estimators=100)				0.85				0.74	0.11
		(n_estimators = 200)				0.89				0.71	0.18

Figure 27 comparison of all model

4. CONCLUSION

In this project, we collect the data form medical agency of female patients who are suffering from diabetics as well as having pregnancy. Data are analysed using python code and model has been developed using machine learning algorithm. Healthcare is one of the emerging and required fields for the development of humankind and science and technology. Machine learning is found one of the emerging area in the technology. The project discus about the detail of the all input and output parameters and developed a model with the help of that one can predict the diabetic problem in a pregnant women. We can observe here in logistic regression model, test accuracy is greater than train accuracy that is indicating model is little over fitted here. Therefore, the difference of training and testing accuracy is only 2%. As per the confusion matrix we can see how this model is predicting here, That means 33 patients are truly predicted diabetic, And 12 patient are wrongly predicted Diabetic (False Positive), 87 patient are non-diabetic (means True negative), And 22 patient are wrongly predicted negative but in reality they are positive(mean false negative). We can see here this model predict large no. of false negative and false positive so this is not going to be good for this diabetic prediction. In terms of accuracy model is giving good accuracy on SVM and then for KNN and adaboost.

REFERENCES

1. Dwivedi AK (2016) Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Comput & Applic 12:1–9.
2. Patidar S, Pachori RB, Rajendra Acharya U (2015) Automated diagnosis of coronary artery disease using tunable-Q wavelet transform applied on heart rate signals. Knowl-Based Syst 82:1–10.
3. Sanz JA, GalarM, Jurio A, BrugosA, PagolaM, Bustince H (2014) Medical diagnosis of cardiovascular diseases using an intervalvalued fuzzy rule-based classification system. Appl Soft Comput 20:103–111.
4. Acharya U, Rajendra KSV, Ghista DN, Lim WJE, Molinari F, Sankaranarayanan M (2015) Computer-aided diagnosis of diabetic subjects by heart rate variability signals using discrete wavelet transform method. Knowl-Based Syst 81:56–64.
5. Bashir S, Qamar U, Khan FH (2015) BagMOOV: a novel ensemble for heart disease prediction bootstrap aggregation with multi objective optimized voting. Australas Phys Eng Sci Med 2:305–323.
6. Sergi G, Veronese N, Fontana L, De RuiM, Bolzetta F, Zambon S, CortiM-C et al (2015) Pre-frailty and risk of cardiovascular disease in elderly men and women: the pro. VA study. J Am Coll Cardiol 10:976–983.
7. Shao YE, Hou C-D, Chiu C-C (2014) Hybrid intelligent modelling schemes for heart disease classification. Appl Soft Comput 14:47–52.
8. Acharya UR, Faust O, Vinitha S, Swapna G, Martis RJ, Kadri NA, Suri JS (2014) Linear and nonlinear analysis of normal and CAD affected heart rate signals. Comput Methods Prog Biomed 1:55–68.
9. Samuel OW, AsogbonGM, Sangaiah AK, Fang P, Li G (2017) An integrated decision support systembased on ANN and Fuzzy_AHP for heart failure risk prediction. Expert Syst Appl 68:163–172.
10. Sabahi F (2018) Bimodal fuzzy analytic hierarchy process (BFAHP) for coronary heart disease risk assessment. J Biomed Inform 83:204–216.
11. Uyar K, İlhan A (2017) Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks. Procedia Comput Sci 120:588–593.

12. Tayefi M, Tajfard M, Saffar S, Hanachi P, Amirabadizadeh AR, Esmaeily H, Taghipour A, Ferns GA, Moohebati M, Ghayour-Mobarhan M (2017) Hs-CRP is strongly associated with coronary heart disease (CHD): a data mining approach using decision tree algorithm. Comput Methods Prog Biomed 141:105–109.
13. Joshi, Sujata, and Mydhili K. Nair. "Prediction of heart disease using classification based data mining techniques". In Computational Intelligence in Data Mining Springer, New Delhi 2, (2015)503–511.
14. Anooj PK (2012) Clinical decision support system: risk level prediction of heart disease using weighted fuzzy rules. J King Saud Univ Comput Inf Sci 24(1):27–40.
15. Kim HC, Greenland P, Rossouw JE, Manson JAE, Cochrane BB, Lasser NL, Limacher MC, Lloyd-Jones DM, Margolis KL, Robinson JG (2010) Multimarker prediction of coronary heart disease risk: theWomen's Health Initiative. J Am Coll Cardiol 55(19): 2080–2091.
16. Ouwerkerk W, Voors AA, Zwinderman AH (2014) Factors influencing the predictive power of models for predicting mortality and/or heart failure hospitalization in patients with heart failure. JACC: Heart Fail 2(5):429–436.
17. Chandel K, Kunwar V, Sabitha S, Choudhury T, Mukherjee S (2016) A comparative study on thyroid disease detection using Knearest neighbor and naive Bayes classification techniques. CSI Trans ICT 4(2–4):313–319.
18. Gao R, Yang Y, Han Y, Huo Y, Chen J, Yu B, Su X et al (2015) Bio restorable vascular scaffolds versus metallic stents in patients with coronary artery disease: ABSORB China trial. J Am Coll Cardiol 66(21):2298–2309.
19. Fleisher LA, Fleischmann KE, Auerbach AD, Barnason SA, Beckman JA, Bozkurt B, Davila-Roman VG et al (2014) ACC/ AHA guideline on perioperative cardiovascular evaluation and management of patients undergoing noncardiac surgery: a report of the American College of Cardiology/American Heart Association task force on practice guidelines. J Am Coll Cardiol 64(22):e77–e137.
20. Nørgaard BL, Leipsic J, Gaur S, Seneviratne S, Ko BS, Ito H, Jensen JM et al (2014) Diagnostic performance of non invasive fractional flow reserve derived from coronary computed tomography angiography in suspected coronary artery disease: the NXT trial (Analysis of Coronary Blood Flow Using CT Angiography: Next Steps). J Am Coll Cardiol 63(12):1145–1155.
21. Park J, Bhuiyan MZA, Kang M, Son J, Kang K (2018) Nearest neighbor search with locally weighted linear regression for heartbeat classification. Soft Comput 22(4):1225–1236.

