ISSN No:-2456-2165

Prediction of Diabetics based on Machine Learning

[1]Gayatri Tatikonda, [2]Geethika Mannam, [3]Jothsna Bhavani Tirumalasetty
[1]Assistant Professor, [2]Student of B.tech, [3]Student of B.tech, [1],[2],[3]Department of IT
[1],[2],[3]Vignan's foundation for science and technology & research(deemed to be university), Guntur, Andhra Pradesh

Abstract:- Diabetes is a chronic disease that occurs when the blood sugar levels of a human are high. When we ate, body turns the food into sugar(glucose). Big Data Analytics plays important role in the care industries. It can help in identifying the right treatment for people with diabetes. Care industries have massive volume of databases. Kidneys are mainly damaged by the diseases called diabetes damage, blindness, heart failure. Normally pancreas is supposed to release insulin. The future scientific field in the course of data science which deals with the ways to learn the information from the given content is Machine Learning. The ways to project the diabetes in the starting stage in order to control it by taking the several results obtained by the machine learning techniques and comparing them with each other to get the most accurate decision are such as K nearest neighbor, random forest, decision tree, logistical regression are used. By using these kind of algorithms we can calculate the accuracy of the algorithms.

Keywords:- *Symptoms, Types, Random forest ,Decision tree, Logistic regression, KNN.*

I. INTRODUCTION

The most occuring disease now a days is likely to be the Diabetes, Even the youngsters are also effected with this disease. Main reason for this is the food we are taking contains mostly Carbohydrates. Carbohydrates are mainly composed of sugar and that sugar enters our body through the food items like bread, Rice, Pasta, Fruits and vegetables etc..,. Glucose is generated by the break down particles of those food items in the body. The blood helps the Glucose to move across the body no matter what the organ the blood reaches. Insulin acts as a key to the door which means it is unique. If the insulin produced by the pancreases is not sufficient or otherwise the insulin produced is not sufficiently used by the body then the glucose level in the blood increase which results in the increase of chances to get the diabetes . There is a special case called Mellitus in which the sugar level in both the blood and the urine is above the normal level.

II. LITERATURE REVIEW

It makes a specality of numerous predictive analysis techniques and it's far utilizing the premature approximation of a more instances of diabetes from affected person file. The exceptional method analytics strategies are carried out in fitness data subject of diabetes and to notice of virtual stages to handle them in higher way. [2]Diabetes prediction is performed using ensemble voting technique for different diabetes data set, in compared with distinctive category techniques, and the highest accuracy of eighty% and eighty one% is to reach for facts set by the use of 10-fold cross validation and by means of spitting data into thirty% checking out and seventy% training.[3]The performance gadget gaining knowledge of algorithms were in comparison and regular based totally on their perfection(validity). The validity of the approach is vary from earlier than pre purifying and after pre purifying as they diagnosed in this paper. This shows the inside the projection of ailments the pre purifying of facts set has its very own effect on the overall system and validity of the data.[4] In this paper researchers used distinct data mining techniques to project the diabetic illnesses using real international datasets by means of accumulating statistics through dispensed questioner .In this take a look at weka tool have been used for records evaluation and projection respectively and differentiate 3 techniques KNN, Logistic regression, and j48 .subsequently it changed into finish as j48 machine learning algorithm that available systematic and good results.[5]The proposed system targets at supplying sufficient hybrid class framework for projecting and tracking the Diabetes disorder. The principle object of this study is to perceive and assemble models that could help clinical practitioners in an sufficient way via the way by the way benefits of human beings to reap longer life in this world.[6]Analyses about the 3 forms of diabetes and their reasons. It additionally uses the projection, category method. This gives the more result for the sickness projection.[7]we have used Matlab tool for evaluation and finished contrast of decided on category techniques. After the differentiate evaluation we conclude that neural network technique is greater correct and has few error charge. Our intersection also offers the user the option of choosing appropriate projection set of rules. We examine that KNN has greater clarity than different fashions.

A. *Symptoms of diabetes*

- Tiredness/Sleepy
- loosing weight
- Uncleared vision
- Mood swings
- frequent infections

B. *Types of diabetics*

- Type-1 is represented as Insulin-Dependent Diabetes Mellitus (IDDM).Generates un sufficient insulin to the human body so they have to inject insulin directly into their bodies
- Type-2 also known as Non-Insulin-Dependent Diabetes Mellitus (NIDDM). Body generates sufficient insulin but it dose not use it properly
- Type-3 is approximately Gestational Diabetes, increase the glucose level of a pregnant woman and it will become normal after the birth of a baby.

C. Algorithms

- **Random Forest:**It could be used for each type and Regression problem strategies in gadget learning .it's far a classifier that which includes a variety of selection trees on diverse subsets of the given information set and takes the average to enhance the predictive the accuracy of that given facts set. The huge variety of trees in the area ends in high accuracy and prevents the matter of over fitting
- .**Decision Tree :**It is a graphical illustration for obtaining all the possible solutions of a problem based on the given conditions. There are two nodes, they are Decision and Leaf Node**.** Decision nodes which have multiple branches and used to make decision, whereas Leaf node is the output of those decisions and do not have any further branches . So it is a Supervised learning technique**.**
- **K Nearest Neighbour :**KNN is a non-parametric rule algorithm, which means it doesn't create any assumption of underlying the information and This rule assumes compare the similarity among the new record and to be had instances and put that new case into the case that which is most familiar to the usage cases present.
- **Logistic Regression:**The result should be a categorical or different value. That it can be both true or false and zero, one and yes or no, etc but in preference to giving the exact cost as zero and one, it will give probabilistic profit that which lies in the middle of zero and one. It cab be used for resolution of the classification issues.
- **Flowchart**

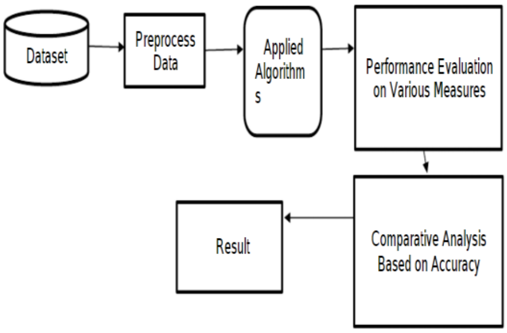
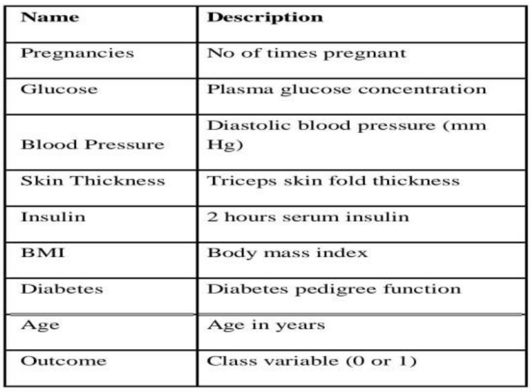

Dataset
Preprocess Data
Applied Algorithm s
Performance Evaluation on Various Measures
Comparative Analysis Based on Accuracy
Result

Fig. 1

- **Proposed system**

Classification is one among the most vital process decision making techniques in lots of actual global problem. The main theme of the model is to achieve high accuracy .For numerous classification the higher wide variety of samples selected however it doesn't ends in higher type accuracy. The survey has assumed that various classification algorithms diabetes and non-diabetic knowledge. Thus, it is observed by the techniques like random forest, decision tree are most suitable for implementation to the system of prediction of diabetes.

D. Attributes



Name	Description
Pregnancies	No of times pregnant
Glucose	Plasma glucose concentration
Blood Pressure	Diastolic blood pressure (mm Hg)
Skin Thickness	Triceps skin fold thickness
Insulin	2 hours serum insulin
BMI	Body mass index
Diabetes	Diabetes pedigree function
Age	Age in years
Outcome	Class variable (0 or 1)

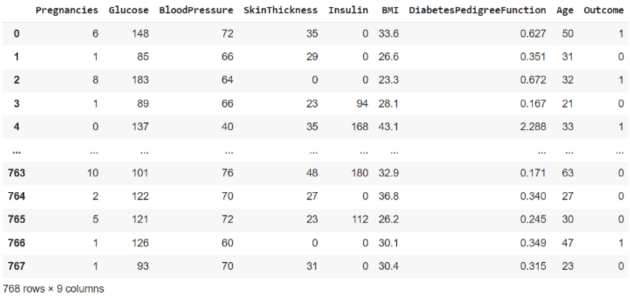
Table 1

III. METHODOLGY

Data set contains seven hundred observations with nine credits. The credits are mentioned below with the Different prediction algorithms like Decision tree, KNN ,Random forest and Logistic regression algorithms are carried out to the data .One of the glucose crucial element to guess the diabetes. The predictions are derived using different algorithms. The information set is to sickness projection the patient is suffering with diabetes or not.

IV. RESULTS

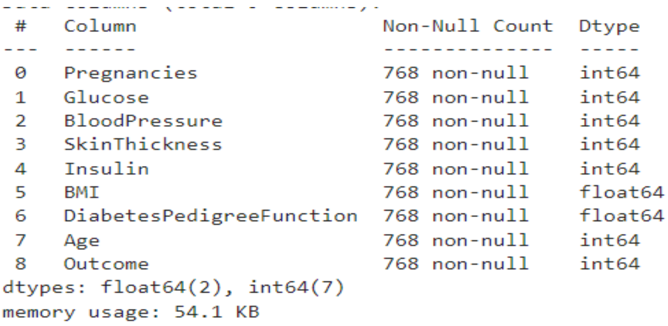
- **Importing the data and reading the data**



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
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
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

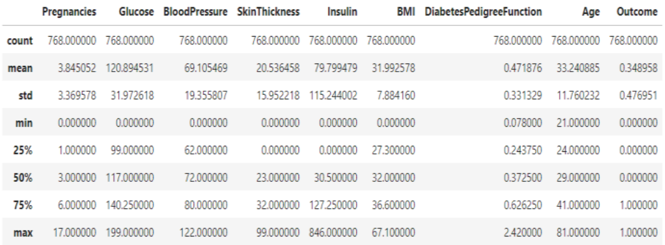
768 rows × 9 columns

Fig. 2



```
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Pregnancies               768 non-null    int64  
 1   Glucose                   768 non-null    int64  
 2   BloodPressure             768 non-null    int64  
 3   SkinThickness             768 non-null    int64  
 4   Insulin                   768 non-null    int64  
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64  
 8   Outcome                   768 non-null    int64  
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Fig. 3

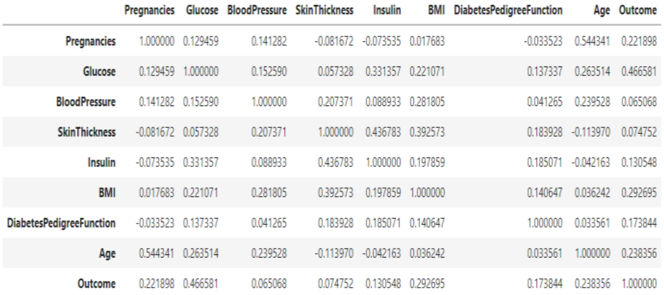


	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	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

Fig. 4

There is no null values in the data set

Pregnencies	0
Glucose	0
Blood pressure	0
Skin thickness	0
Insulin	0
Bmi	0
Diabetics pedigree function	0
Age	0
Outcome	0
Dtype	int 64



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.000000	0.129459	0.141282	-0.081672	-0.073535	0.017683	-0.033523	0.544341	0.221898
Glucose	0.129459	1.000000	0.152590	0.057328	0.331357	0.221071	0.137337	0.263514	0.466581
BloodPressure	0.141282	0.152590	1.000000	0.207371	0.088933	0.281805	0.041265	0.239528	0.065068
SkinThickness	-0.081672	0.057328	0.207371	1.000000	0.436783	0.392573	0.183928	-0.113970	0.074752
Insulin	-0.073535	0.331357	0.088933	0.436783	1.000000	0.197859	0.185071	-0.042163	0.130548
BMI	0.017683	0.221071	0.281805	0.392573	0.197859	1.000000	0.140647	0.036242	0.292695
DiabetesPedigreeFunction	-0.033523	0.137337	0.041265	0.183928	0.185071	0.140647	1.000000	0.033561	0.173844
Age	0.544341	0.263514	0.239528	-0.113970	-0.042163	0.036242	0.033561	1.000000	0.238356
Outcome	0.221898	0.466581	0.065068	0.074752	0.130548	0.292695	0.173844	0.238356	1.000000

Fig. 5

- **Histogram analysis for variable pregnancies**

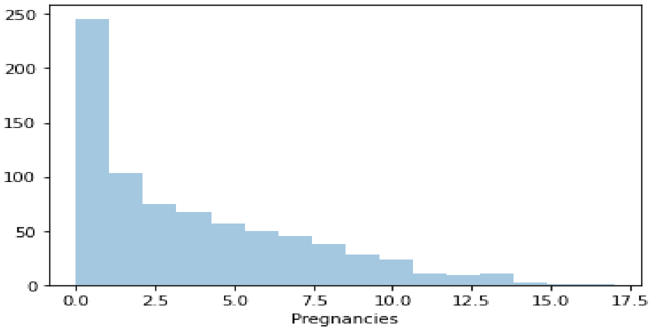


Fig. 6

- **Boxplot analysis for variable pregnancies**

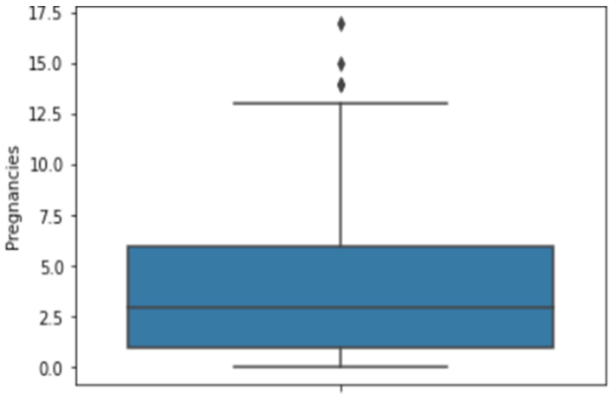


Fig. 7

- **Histogram analysis for variable glucose**

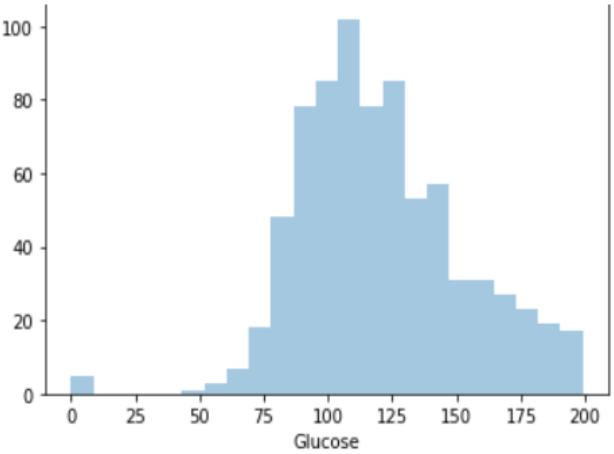


Fig. 8

- **Histogram analysis for variable bp**

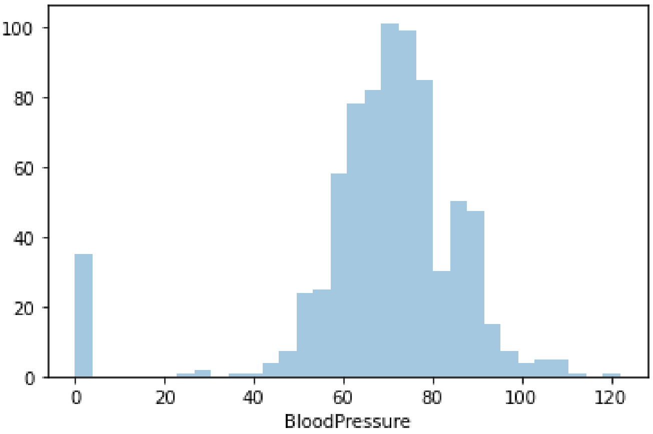
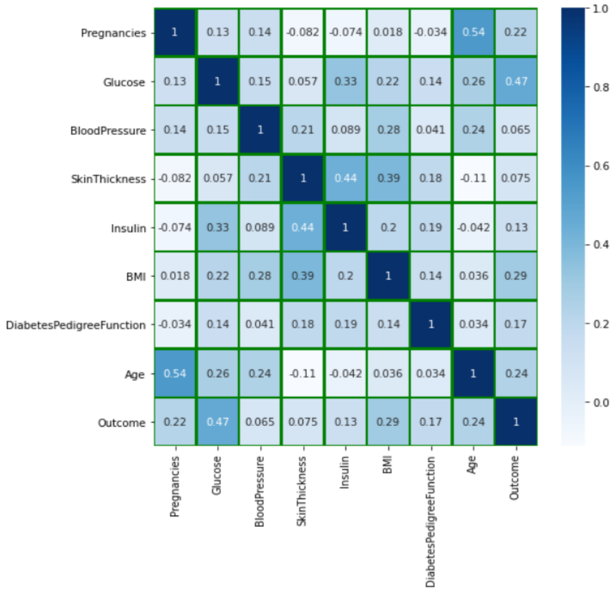


Fig. 9

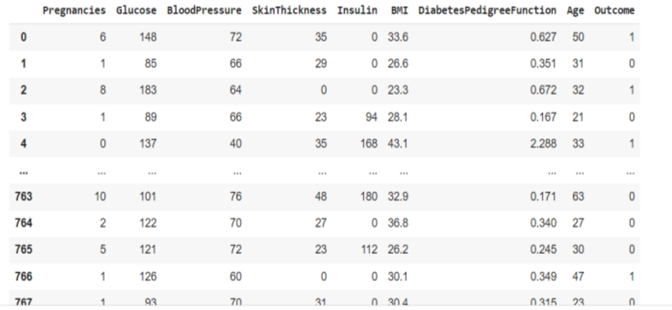
- **Correlation matrix**



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1	0.13	0.14	-0.082	-0.074	0.018	-0.034	0.54	0.22
Glucose	0.13	1	0.15	0.057	0.33	0.22	0.14	0.26	0.47
BloodPressure	0.14	0.15	1	0.21	0.089	0.28	0.041	0.24	0.065
SkinThickness	-0.082	0.057	0.21	1	0.44	0.39	0.18	-0.11	0.075
Insulin	-0.074	0.33	0.089	0.44	1	0.2	0.19	-0.042	0.13
BMI	0.018	0.22	0.28	0.39	0.2	1	0.14	0.036	0.29
DiabetesPedigreeFunction	-0.034	0.14	0.041	0.18	0.19	0.14	1	0.034	0.17
Age	0.54	0.26	0.24	-0.11	-0.042	0.036	0.034	1	0.24
Outcome	0.22	0.47	0.065	0.075	0.13	0.29	0.17	0.24	1

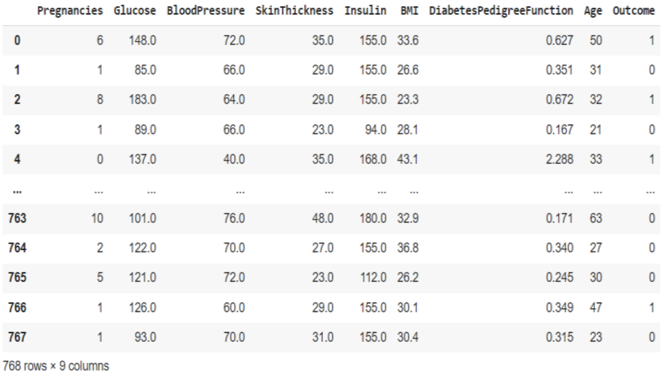
Fig. 10

- **Building the model using knn classifier**



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
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
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

Fig. 11



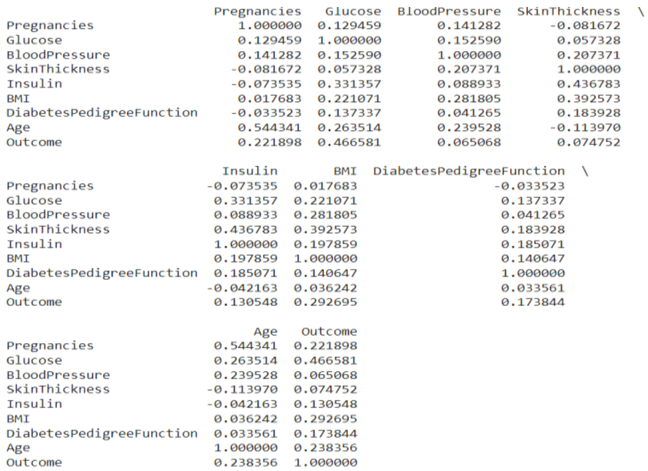
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148.0	72.0	35.0	155.0	33.6	0.627	50	1
1	1	85.0	66.0	29.0	155.0	26.6	0.351	31	0
2	8	183.0	64.0	29.0	155.0	23.3	0.672	32	1
3	1	89.0	66.0	23.0	94.0	28.1	0.167	21	0
4	0	137.0	40.0	35.0	168.0	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101.0	76.0	48.0	180.0	32.9	0.171	63	0
764	2	122.0	70.0	27.0	155.0	36.8	0.340	27	0
765	5	121.0	72.0	23.0	112.0	26.2	0.245	30	0
766	1	126.0	60.0	29.0	155.0	30.1	0.349	47	1
767	1	93.0	70.0	31.0	155.0	30.4	0.315	23	0

768 rows × 9 columns

Fig. 12

R squared value for train from KNN Classifier = 0.7839851024208566
R squared value for test from KNN Classifier = 0.70995670995671

Finding Correlation
corr_matrix=data.c.corr()
print(corr_matrix)



```
                          Pregnancies   Glucose  BloodPressure  SkinThickness  \
Pregnancies                  1.000000  0.129459       0.141282      -0.081672
Glucose                      0.129459  1.000000       0.152590       0.057328
BloodPressure                0.141282  0.152590       1.000000       0.207371
SkinThickness               -0.081672  0.057328       0.207371       1.000000
Insulin                     -0.073535  0.331357       0.088933       0.436783
BMI                          0.017683  0.221071       0.281805       0.392573
DiabetesPedigreeFunction    -0.033523  0.137337       0.041265       0.183928
Age                          0.544341  0.263514       0.239528      -0.113970
Outcome                      0.221898  0.466581       0.065068       0.074752

                           Insulin       BMI  DiabetesPedigreeFunction  \
Pregnancies              -0.073535  0.017683                 -0.033523
Glucose                   0.331357  0.221071                  0.137337
BloodPressure             0.088933  0.281805                  0.041265
SkinThickness             0.436783  0.392573                  0.183928
Insulin                   1.000000  0.197859                  0.185071
BMI                       0.197859  1.000000                  0.140647
DiabetesPedigreeFunction  0.185071  0.140647                  1.000000
Age                      -0.042163  0.036242                  0.033561
Outcome                   0.130548  0.292695                  0.173844

                               Age   Outcome
Pregnancies               0.544341  0.221898
Glucose                   0.263514  0.466581
BloodPressure             0.239528  0.065068
SkinThickness            -0.113970  0.074752
Insulin                  -0.042163  0.130548
BMI                       0.036242  0.292695
DiabetesPedigreeFunction  0.033561  0.173844
Age                       1.000000  0.238356
Outcome                   0.238356  1.000000
```

Fig. 13

Count plot specifying the amount individuals suffering by diabetics

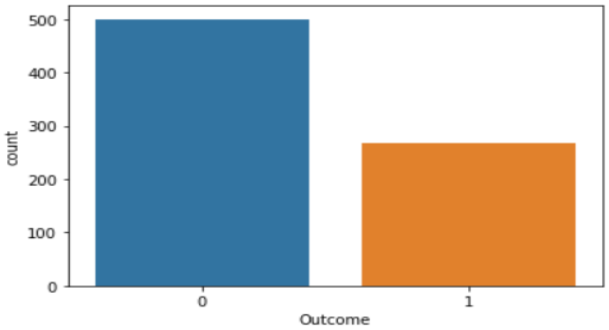


Fig. 14

It is clear from the above graph that there are two outputs referring the diabetic and non-diabetic patients. In the graph blue data presents the record of non-diabetic patients whereas orange record represents the diabetic patients data. When compared to the above values it is clear that the diabetic patients are far more behind the count of the non-diabetic patients.

- **Machine learning algorithms part**
- **K Nearest neighbour classifier**

The most simple and preferable machine learning method is KNN Algorithm. The ways to store the training data is consider as the important task in the building the KNN model. In order to predict the data that has been missing will be obtained by the most recent and the most near value to the missing data, simply called as the "nearest neighbor".

We can now clearly observe that the training data set on the Y-Axis and the near neighbors on the X-Axis. In any case if we select only one of the nearest neighbor the probability of the prediction on the training data is said to be in ideal position. But when we increase the neighbors the probability to predict the training data set will be dropped by a particular percentage. Which states that using only one neigbhor is better option to any of the advanced techniques.

Score accuracy of coaching data is: 0.7914338919925512
Score accuracy of Test data is: 0.71861471861471

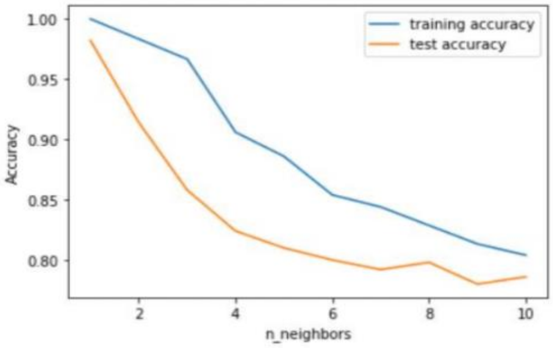
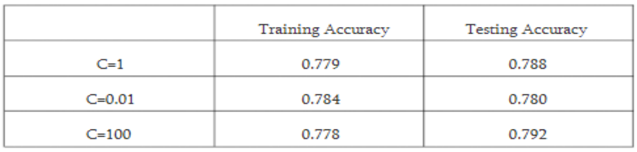


Fig. 15

- **Logistic regression**

Logistic regression is one of all the foremost common classification algorithms



	Training Accuracy	Testing Accuracy
C=1	0.779	0.788
C=0.01	0.784	0.780
C=100	0.778	0.792

Fig. 16

It is clear that there is 77% accuracy on predicting the training dataset and also in accordance with the 78% accuracy on the test dataset where the C value is assigned as 1.So that is is proven that the regulation and using the complex model does not act that much smarter than the default value settings . So here we are using same value which is used in the first case that is c equals to 1.

Accuracy score of coaching data is: 0.776536312849162
Accuracy score of Test data is 0.7662337662337663

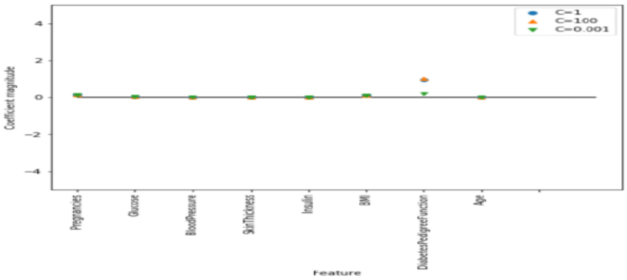


Fig. 17

- **Decision Tree Classifier**

Score accuracy of coaching data is: 1.0
Score accuracy of Test data is: 0.7536796536796536

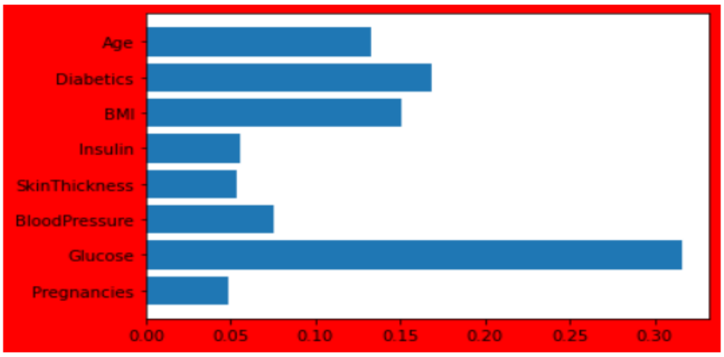


Fig. 18

- **Random Forest**

The score accuracy in the Training dataset is: 1.0
The score accuracy in the Test dataset is: 0. 7992207792207793

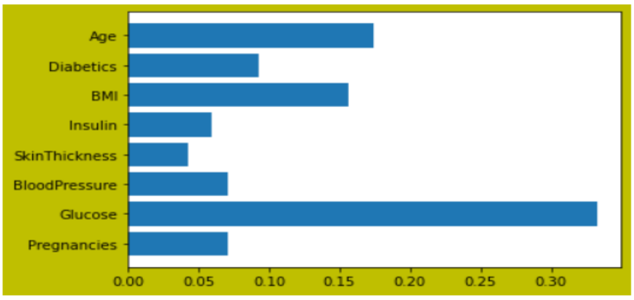


Fig. 19

- **Comparison Table**

Algorithms	Training Accuracy	Testing Accuracy
KNN	0.79	0.71
RANDOM FOREST	1.0	0.80
DECISION TREE	1.0	0.75
LOGISTIC REGRESSION	0.77	0.76

Table 2

V. CONCLUSION

After observing all the obtained values we decided to get into the conclusion that Random forest that is having additional advantage in prediction with the rate of 80% comparison to remaining algorithms namely decision tree, KNN ,random forest of our data set. The technique that is used to know whether the patient is having diabetes or not is fast and simple in the presence of the knowledgable models. It aims to improve the efficiency of the diagnosis of diabetes patients.

REFERENCES

[1.] Sonali Vyas, Navdeep Singh, Arohan Mathur(.2019) " Prediction avaluate of Analysis Strategies for Diabetes Chance."

[2.] Prema N, V Yogeswar J.(2019). "Prediction of Diabetes usage of Ensemble strategies" from world wide Magazine of Recent strategy and Engineering (IJRTE)

[3.] Pradeep, K. R., & Naveen, N. C.(2016)."Prediction analysis of diabetes usageg of J48 technique of class algorithms" In (IC3I), 2nd International Conference on (pp. 347-352). IEEE.

[4.] Huang, Y. X., Rao, D. P., Zhang, Q., & Liu, Q. (2013)."Differentiate of 3 data mining strategies for projection of diabetes"The Kaohsiung journal of medical sciences, 29(2), 93-99.

[5.] N.Deepika, Dr.S.Poonkuzhali. (2018)."Design of Hybrid Classifier for Prediction of Diabetes " from International Journal of Revolution Science & Technology, Vol. 2 Issue 10.

[6.] K. Sharmila and S. Manicka.(2015). "Effective Prediction and Classification of Diabetic Patients" Worldwise magazine of Advanced Engineering studies and Science, vol. 2.

[7.] 7.Musavir Hassan, Muheet Ahmad Butt and Majid Zaman Baba(2017)."The Best Result in the Diabetes projection" from Computer Science and Technology, ISSN: 2249-0701.

[8.] https://github.com/Ravjot03/Diabetes-Prediction.