ISSN No:-2456-2165

Diabetes Prediction Using Machine Learning KNN -Algorithm Technique

Dr. B. Premamayudu (Professor)[1] , K. Muralikrishna[2] , K. Pramodh[3] ,
[1]Professor, Department of Information Technology, Vignan's Foundation for Science, Technology & Research , Guntur AP-522213, India
[2,3] UG student, Department of Infotmation Technology Vignan's Foundation for Science, Technology & Research, Guntur AP-522213, India

Abstract:- Diabetes is a chronic disease caused due to high amount of glucose present in the human body. If this diabetes is ignored, this may lead to severe health problems such as kidney failure, heart attacks, blood pressure, eye damage, weight loss, frequent urination, etc. Basically, human body contains Insulin which is produced by pancreas. This insulin helps to enter glucose in to blood cells in order to generate energy to the body. There are types in diabetes Type1 and Type 2 other form is gestational diabetes which is caused during pregnancy. This can be controlled in the earlier stages of the attack. According to International Diabetes Federation (IDF) 382 million people are suffering with diabetes and by next 20years the count will be doubled as 592 million. To accomplish this goal, in this project we can do early prediction of diabetes in humans or patients for good accuracy through applying various machine learning techniques such as Random Forest (RF), K-nearest neighbors (KNN), Decision Trees (DT), etc. However, in this project we are predicting diabetes using KNN classifier model. As we see now a days machine learning is an emerging technology and boon to many problem solutions.

I. INTRODUCTION

➢ *Machine Learning*

Machine learning (ML) could be a variety of AI (AI) that enables code applications to become additional correct at predicting outcomes while not being expressly programmed to try and do, therefore. Machine learning algorithms use historical knowledge as input to predict new output values.

➢ *Types of Learning:*

1. Supervised learning.
2. Unsupervised learning.
3. Reinforcement learning

In this project, we have a tendency to square measure victimization supervised learning classifier technique. i.e., KNN algorithmic rule to search out the accuracy of predicting the new outcomes.

In this project we tend to use some datasets to predict the attack of polygenic disorder to the folks. Diabetes could be a fast-growing sickness in folks even in kids too. It's a gaggle of sickness during which blood doesn't turn out enough quantity of hypoglycemic agent, doesn't properly use the hypoglycemic agent that's created. The body is unable to urge sugar from the blood into the cells. that results in increase in blood glucose levels. Glucose, the shape of sugar found in your blood, is one amongst your main energy sources. There area unit three main kinds of polygenic disorder they're

1.Type one polygenic disorder.
2.Type a pair of polygenic disorder.
3.Gestational polygenic disorder.

➢ *Type one diabetes:*

It is believed to be reaction condition. this suggests your system erroneously attacks and destroys the beta cells in your duct gland that produces hypoglycemic agent. The harm is permanent. we tend to cannot realize the prompts of sickness simply. There could also be each genetic and environmental reasons and modus vivendi factors thought to play a job.

➢ *Type a pair of diabetes:*

This type starts as hypoglycemic agent resistance. Our body cannot respond for systematic hypoglycemic agent. That regulates duct gland to supply additional hypoglycemic agent because it isn't sensible for health. hypoglycemic agent production decreases and results in high blood glucose levels. The reason behind this kind sickness is genetic science, lack of exercise, being overweight.

➢ *Gestational diabetes:*

This is thanks to hypoglycaemic agent obstruction hormones created by throughout maternity. this kind of sickness happens solely throughout maternity solely.

➢ *Symptoms:*

- Blood pressure downside repeated elimination
- Dry and fidgety skin
- Visionary issues
- Slow recovery of health conditions

II. LITERATURE REVIEW

1. KM. Jyothirani aims to apply 5 machine learning classification algorithms to predict diabetes and compare each to find which algorithm gives accurate target outcomes. In her research PIMA datasets were used and the study concluded that Decision trees gave 98% accuracy score.
2. Avantika Nahar had applied the KNN algorithm for classification and prediction of diabetes using trained data and predicts the time of getting diabetes also. This project

result is based on YES or NO. if the result is NO then time prediction module is used. Else we use just prediction of diabetes and accuracy of the KNN algorithm.

3. Umatejaswi and P. Suresh Kumar had talked about algorithms such as Support Vector Machine, NaiveBias, Decision Trees in order to find diseases through data mining technique.

III. METHODOLOGY

In this section, we are learning KNN classifier model used in machine learning to predict diabetes. We shall also explain our proposed methodology to improve the accuracy of finding the targeted outcomes.

A. Dataset Description: -

This data is collected from UCI repository which is named as PIMA Indian diabetes dataset. The dataset has many attributes of 768 patients.

Table 1: Dataset Description

S. No	Attributes
1	Pregnancy
2	Glucose
3	Blood pressure
4	Skin thickness
5	Insulin
6	BMI (body mass index)
7	Diabetes pedigree function
8	age

The 9th attribute is class variable of each data points. This class variable shows the outcomes 0 & 1 for diabetes which indicates non-diabetic & diabetic.

	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

Dataset head part

```
#describing the statistical values of dataset
dataset.describe()
```

	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

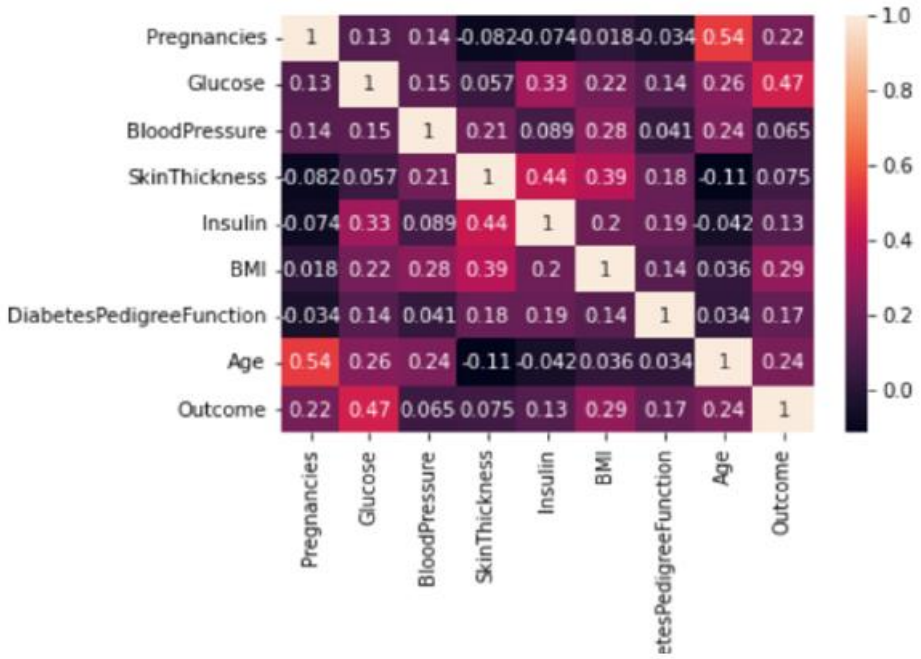
Above values are the statistical values of the dataset which we have used.

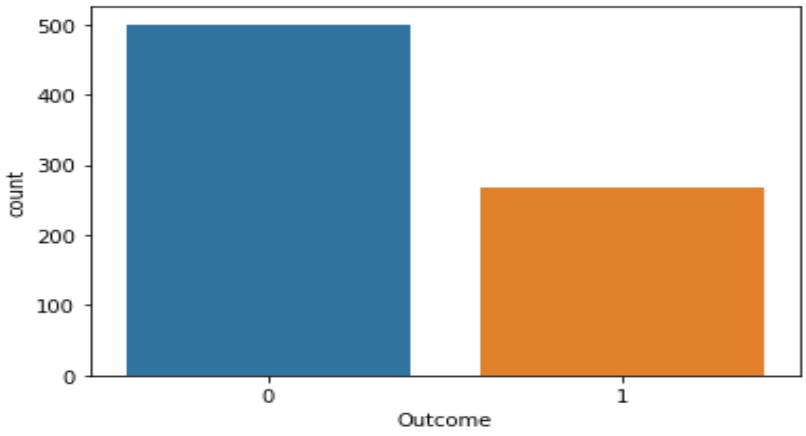
Here, from this correlation matrix we came to know that pregnancies and glucose columns are very important to predict the output. These two columns played key role.

➢ *Distribution of diabetic patient*

This model is made to predict how many numbers of patients are having diabetes. In this below outcome we can see **0's** label contains 500 classes and **1's** label contains 268 classes.

```
#correlation matrix
# this describes about the relation between each relation
corr_mat=dataset.corr()
sns.heatmap(corr_mat,annot=True)
plt.show()
```





B. Data pre-processing:

This is the most crucial process. Mostly healthcare related data may contain many missing values and many mistakes which might cause for low effective of data.so to improve the quality and effectiveness data processing should be done. This process is more essential to get good accuracy. There are mainly two steps in this data pre-processing they are

1.Missing values removal.

2.Splitting of data into training and testing sets.

C. Applying classifier technique:

After training and testing datasets are separated and without null values in the dataset, we can now apply the machine learning classifier technique to the dataset.

We have many classification techniques such as support vector machine (SVM), random forest, decision trees, KNN algorithm etc. However here in this project we are using K nearest neighbour's classifier technique only.

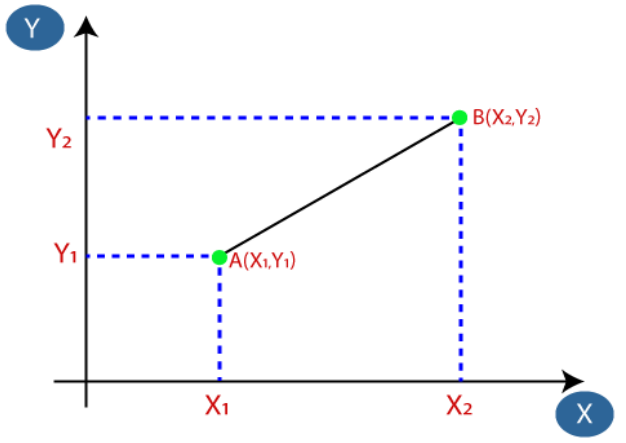
➢ *KNN Classifier*

- It is one in every of the best machine learning algorithms supported supervised learning techniques.
- It could be a non-parametric rule, which implies it doesn't build any assumption on underlying knowledge.
- It is additionally known as as lazy learner rule as a result of it doesn't learn from the coaching set quickly.

- At the coaching part simply stores the knowledge set and once it gets new data, then it classifies that knowledge into class that's abundant the same as the new knowledge.

➢ *Working of KNN:*

- Select the amount K of the neighbours.
- Calculate the euclidian distance of K variety of neighbors.
- Take the K neighbors as per the calculated euclidian distance.
- Among these K neighbors, count the amount of the info points in every class.
- Assign the new knowledge points there to class that the amount of neighbors is most.
- our model is prepared to use.



$$EculideanDistance = \sqrt{\sum_{i=1}^{y}\sum_{j=1}^{m}\sum_{l=1}^{n-1}(R_{(j,l)} - P_{(i,l)})^2}$$

IV. EXPERIMENTAL RESULTS

In this task different stages were performed. This approach used K- Nearest Neighbors (KNN) classifier technique. Using this machine learning technique, we find the accuracy of predicting diabetes using KNN algorithm. And we have got the accuracy score of 79% which is better to apply for prediction. Overall, study states that we can use this KNN algorithm for achieving high performance accuracy. There are many variants in KNN algorithm, all those variants may give different accuracy scores compared to the accuracy which we got now.

V. CONCLUSION

The main target of this project was to find whether KNN classifier algorithm is suitable for prediction or not. This we can see by checking the performance analysis, which we had get 79%. To find this accuracy we use the library called sci-kit learn in python. This accuracy is good to apply for prediction. The experimental results can be helpful in healthcare to predict and make early decision-making to cure the diabetes and save the lives of humans. if we would apply this pattern in finding diabetics in patients it would be really helpful for all the humans and hospital management as well. We can find the results fast.

➢ *Output accuracy*

```
[19] from sklearn.metrics import accuracy_score
     accuracy_score(y_test , y_pred)

     0.7922077922077922
```

REFERENCES

[1]. Mitushi soni, Dr. Sunitha Varma “Diabetes Prediction victimization Machine Learning Techniques” IJERTV9ISO90496.
[2]. Avantika Nahar, Dr. Ajay Lala, Saurabh Sharma, “Diabetes Prediction victimization Machine Learning” ISSN 2347-6435.
[3]. k. Jyothi aristocrat “Diabetes Prediction victimization Machine Learning” IJSRCSEIT206463, ISSN: 2456-3307.
[4]. Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabetes Prediction victimization data processing "(ICIIECS), 2017
[5]. Vijayakumar, Lavanya, I. Nirmala, Sofia Carolingian, "Random Forest algorithmic program for the Prediction of polygenic disorder “, 2019.
[6]. Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importances for polygenic disorder Prediction victimization Machine Learning". IEEE, pp 942-928, 2018.
[7]. A.K., Dewangan, and P., Agrawal, Classification of diabetes victimization Machine Learning Techniques, International Journal of Engineering and Applied Sciences, vol. 2, 2015.
[8]. Nahla, Andrew – “Intelligible support vector machines for diagnosis of diabetes mellitus.” “Information technology in biomedicine” IEEE transactions. (July,2010),1114-20.