International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 01 | January - 2023 Impact Factor: 7.185 ISSN: 2582-3930

Diabetes Prediction Using Machine Learning KNN -Algorithm Technique

Nikhil Reddy Nookala (Author) CSE Student
Hyderabad, India

Abstract:- Diabetes is a chronic condition characterized by an excess of glucose in the human body. If diabetes is not treated, it can lead to serious health problems such as renal failure, heart attacks, high blood pressure, eye damage, weight loss, frequent urination, and so on. Insulin, which is generated by the pancreas, is found in the human body. This insulin aids in the entry of glucose into blood cells, where it is converted into energy for the body. Diabetes is classified into two types: Kind 1 and Type 2. Another type is gestational diabetes, which occurs during pregnancy. This can be managed in the early phases of an attack. According to the International Diabetes Federation (IDF), 382 million people have diabetes, with the number expected to treble in the next 20 years to 592 million. To achieve this aim, we may use several machine learning techniques such as Random Forest (RF), K-nearest neighbors (KNN), Decision Trees (DT), and others in this research to predict diabetes in individuals or patients with high accuracy. In this study, however, we are predicting diabetes with the KNN classifier model. As we can see these days, machine learning is a developing technology that is a blessing to many issue solutions.

I. INTRODUCTION

Machine Learning Machine learning (ML) is a subset of artificial intelligence (AI) that enables code applications to get more accurate at predicting outcomes while not being explicitly designed to do so. Machine learning algorithms estimate new output values by using past data as input.

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 thought to be a response condition. This indicates that your system incorrectly assaults and kills the beta cells in your duct gland, which generate hypoglycemic agents. The damage is irreversible. We frequently fail to recognise the symptoms of illness. There are other genetic and environmental elements, as well as modus vivendi considerations, that are considered to have a role.

Type a pair of diabetes: This kind begins with resistance to hypoglycemic agents. Our bodies are unable to respond to a systemic hypoglycemic agent. Because it is harmful to one's health, this regulates the duct gland to give extra hypoglycemic agent. The generation of hypoglycemic agents reduces, resulting in elevated blood glucose level. This type of illness is caused by genetic science, a lack of activity, and 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

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

In this section, we are learning KNN classifier model technique 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.

The 9 th 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

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

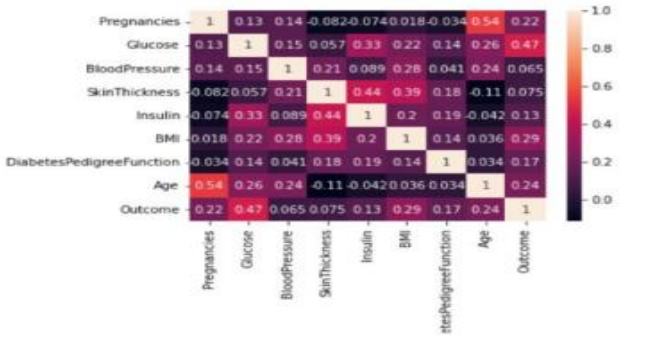
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578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

```
#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.

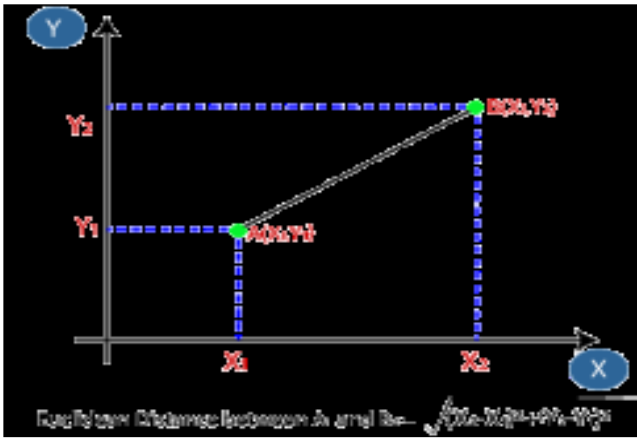
Working of KNN:

• 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.



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.

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
```