International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 02 | Feb 2020 www.irjet.net p-ISSN: 2395-0072

PREDICTION AND DETECTION OF DIABETES USING MACHINE LEARNING

B.Selvaraj[1], S.V.Pavithra[1], A.S.Nithya Rak[1], M.Jeyaselvi[2]

[1]*UG Student,Department of CSE,Agni College of Technology, Chennai,Tmail Nadu,India*

[4] *Sr. Assistant Professor,Dept of Computer Science and Engineering,Agni College of Technology,Tamil Nadu,India*

Abstract - *It is apparent from the truth that the occurrence of diabetes mellitus is high and the complication in the prevention of diabetes also increases. Thus, there are many patients who need the required knowledge and skills to enrich their health. In such cases, the patients are needed to visit the diagnostic center for their treatment. Because of this, they lost their time and expenses. In this paper, we are using the Machine Learning algorithms to predict the level of diabetes with future risk and to determine the medications. This idea projected in the paper is to determine the best prediction algorithm with higher accuracy and combine the entire algorithm using voting classifier.*

***Key Words*: Diabetes Mellitus, Machine learning algorithms, Medications, Confusion Matrix, Accuracy Score, Voting Classifier, Precision and Recall.**

1. INTRODUCTION

Diabetes Mellitus is a common disease because of high glucose level, genes, obesity and environmental factors. There are two types of diabetes. Type 1 diabetes occur when the immune system, body's system for fighting infection, attacks and destroys the insulin producing beta cells of the pancreas, it begins before 40 years of age. Type 2 diabetes develops when the pancreas is unable to produce enough insulin. Normally it occurs at any age.

1.1. Type 1 Diabetes

Type 1 is a condition in which your immune system can destroys insulin making cells in our pancreas. The cells are called "Beta cells". This condition is usually diagnosed in children and young people so it is called as "Juvenile diabetes". There is no way to prevent this type1 diabetes. It affects both male and female equally. Only 5% people affected by type1 diabetes.

Symptoms:

1. Extreme thirst
2. Dry mouth
3. Increased hunger
4. Fatigue

1.2 Type 2 Diabetes

Type 2 diabetes is a lifelong disease that keeps our body from using insulin. People with type 2 diabetes are said to have insulin resistance. People who are middle aged or older are mostly get this kind of diabetes so it is also called as adult-onset diabetes. But this type also affects the kids and teenagers because of childhood obesity.

Symptoms:

1. Weight loss without trying
2. Dark rashes around neck
3. Blurry vision
4. Cranky

2. RELATED WORK

In this section, we have studied the prediction of diabetes with different algorithms.

A. Muhammad Azeem Sarwar, Nazir Kamal, Wajeeha Hamid, Munam Ali Shah proposed: "Prediction of diabetes using machine learning algorithm in healthcare" here SVM and KNN gives the highest accuracy of diabetes. By using 768 records it gives 77% accuracy.

B. P. Suresh Kumar, P. Pranavi proposed "Performance analysis of Machine learning algorithm on diabetes dataset using big data analysis" It is used to predict more correctly and accurately and give comprehensive comparative study on different machine learning algorithm.

C. Sofia Benbelkacem, Baghdad Atmani proposed" Random forest for diabetes diagnosis" Based on Pima basis it obtain the good result. It is also used to assist and managing the paediatric emergencies.

D. K.VijayaKumar,B.Lavanya,I.Nirmala,S.Sofia Caroline proposed" Random forest algorithm for the prediction of algorithm "It is used for the early prediction of diabetes with the help of machine learning by the high accuracy.

3. DATASETS

We searched the dataset in many ways but it is difficult to find it. At the end we find related datasets for our project in the Kaggle website. Datasets are collected from the Kaggle website. Kaggle is one of the popular websites to collect the datasets. From this website we can collect various information and used for the appropriate projects. In the Kaggle many of them upload the brief view about the project details like datasets and related information of the project. From this information we can gain knowledge and get idea how to do. The name of the dataset is the pima- Indian-diabetes.

Fig: a

	A	B	C	D	E	F	G	H	I
1	Pregnanci	Glucose	BloodPres	SkinThicki	Insulin	BMI	DiabetesF	Age	Outcome
2	2	138	62	35	0	33.6	0.127	47	1
3	0	84	82	31	125	38.2	0.233	23	0
4	0	145	0	0	0	44.2	0.63	31	1
5	0	135	68	42	250	42.3	0.365	24	1
6	1	139	62	41	480	40.7	0.536	21	0
7	0	173	78	32	265	46.5	1.159	58	0
8	4	99	72	17	0	25.6	0.294	28	0
9	8	194	80	0	0	26.1	0.551	67	0
10	2	83	65	28	66	36.8	0.629	24	0
11	2	89	90	30	0	33.5	0.292	42	0
12	4	99	68	38	0	32.8	0.145	33	0
13	4	125	70	18	122	28.9	1.144	45	1

4. ALGORITHM

In our project we used some algorithms for the prediction and detection of diabetes they are listed below.

1. Naive Bayes Theorem

2. Support Vector Theorem

3. Gradient Boosting Algorithm

1. Naive Bayes Theorem

This theorem is used for prediction of diabetes and name in 18th century by Thomas Bayes. It gives a compressive review of Naive Bayesian network to predict the disease. It gives the accurate result and fast and makes a stable decision. This theorem is used predict cancer, diabetes and so on. Comparing to other algorithm it is simple to predict the disease.

2. Support Vector Theorem

It was introduced by Cortes and Vapnik in 1990s.It has two research communities such as statistical and machine learning. It is used to predict medication and used to improve the predication of disease also. It produces a high performance in medical field.

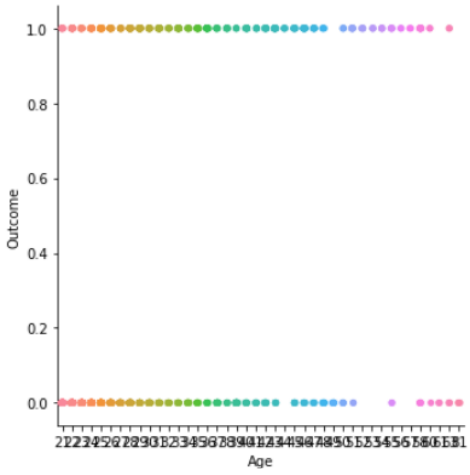
3. Gradient Boosting Algorithm

It is a combination of regression and classification problem which is used to predict the disease. It is in the form decision tree. This algorithm plays an important
role in a clinical research. This can solve the complex data structure including high order iterations.

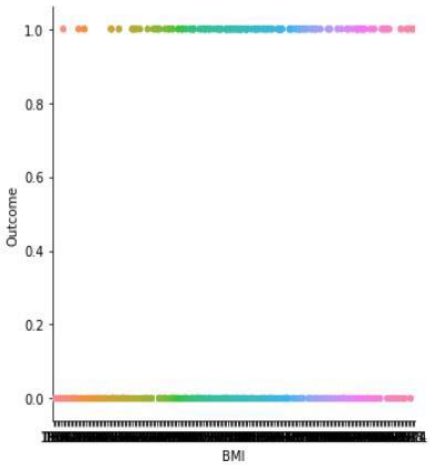
5. PROPOSED SYSTEM

Analysis of patient data

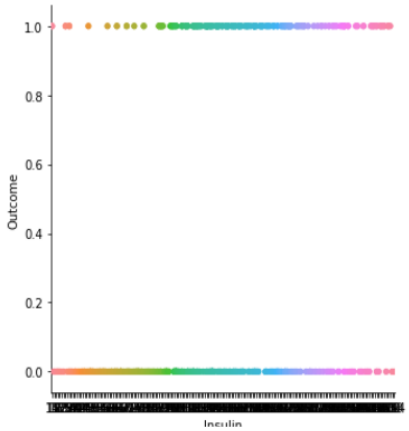
Using our application required information are gathered from the individuals. From the gathered data the required database is prepared for the particular.
The analysis also includes the symptoms in human body for diabetes like frequent urination, excessive thirst, unexplained weight lost, extreme hunger, sudden vision changes, etc., These data are treated under medical conditions such as, appropriate glucose level, finite blood pressure.



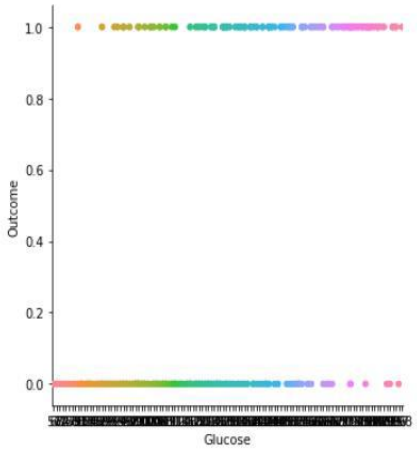
Here the age plot outcome of the individual is detected.



The BMI of the patient are analyzed and determined the occurrence of diabetes.



The insulin outcome is detected.



The glucose outcome is charted.

0.0- Denotes the absence of diabetes.

1.0- Denotes the absence of diabetes.

Here we have discussed three types of algorithms and each of the algorithm shows the different accuracy rate.

The 80% of datasets are used for training and 20% of the datasets are used for testing the algorithms.
The accuracy rate by using Naive Bayes Theorem is given below.

```
Training set score : 0.7620772946859904
Testing  set score : 0.8357487922705314
Accuracy 0.8357487922705314 ROC 0.79340357306459
```

The accuracy rate by using Support Vector Machine is given below

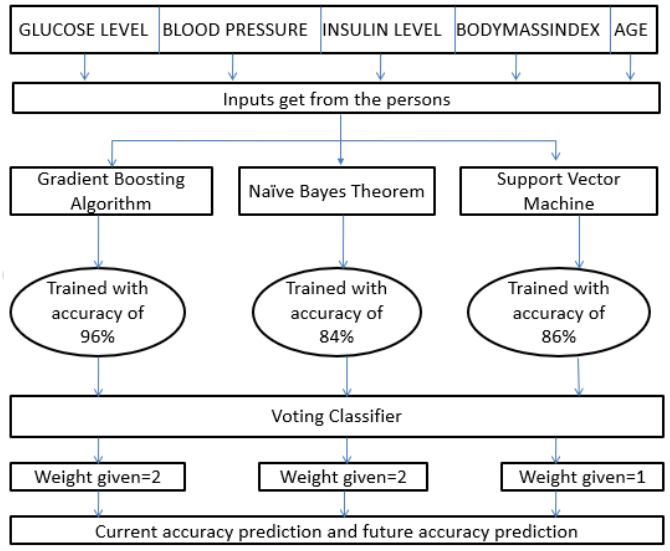
```
Training set score : 0.7789855072463768
Testing  set score : 0.8599033816425121
Accuracy 0.8599033816425121 ROC 0.8169076751946607

Process finished with exit code 0
```

The accuracy rate by using Gradient Boosting algorithm is given below

```
Training set score : 0.9589371980676329
Testing  set score : 0.9371980676328503
Accuracy 0.9565217391304348 ROC 0.9421865474497052
```

Hence the highest accuracy rate (96%) is given by Gradient Boosting algorithm. This is one of the best prediction algorithms for diabetes.



Architecture Diagram

Voting classifier algorithm

It is the simplest way for combing all the predictions of varying machine learning algorithm. It is not a actual classifier but it is a wrapper for different training data sets. We can train the different dataset and ensemble them by using this algorithm. It also includes the hard and soft voting of classifiers.
From the above diagram, we have given a weight age of 2 for Gradient Booster Algorithm and we have given a weight age of 2 for Naïve Bayes alg orithm and for Support Vector

Machine the weight age is 1 because as it's a both classification and regression based algorithm.
Naïve Bayes is a classification algorithm; hence it has a weight age of 1.

Confusion Matrix

The confusion matrix in machine learning is a table that is used to display the performance of the algorithm. The performance is determined by testing the input dataset which is given by the user. So here also we described our performance of the algorithm using this matrix.

The below table shows the predicted and the actual values.
TP- The predicted value is positive and it is true
TN- The predicted value is negative and it is true
FP - The predicted value is positive and it is false
FN- The predicted value is negative and it is false

Predicted Values \ Actual Values	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

Recall
In all predicted classes, the amount which we are predicted correctly, as high as possible.

$$Recall = \frac{TP}{TP + FN}$$

Precision
In the prediction result, there are no of positive classes, precision is to determine how much we are predicted correctly and how many are actually positive.

$$Precision = \frac{TP}{TP + FP}$$

F-Measure
It is used to compare the precision and recall. F-Score is used to measure recall and precision at the same time.

$$F\text{-}measure = \frac{2*Recall*Precision}{Recall + Precision}$$

Accuracy Score
The accuracy score in machine learning is the fraction of our prediction model. Thus, the performance is determined by using this accuracy score. The set of labels predicted for our sample should exactly match the corresponding set of labels under multi label classifications.

$$\textbf{Accuracy} = \frac{\textbf{No of correct predictions}}{\textbf{Total no of predictions}}$$

Information of the dataset before the EDA

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 9 columns):
Pregnancies                 2000 non-null int64
Glucose                     2000 non-null int64
BloodPressure               2000 non-null int64
SkinThickness               2000 non-null int64
Insulin                     2000 non-null int64
BMI                         2000 non-null float64
DiabetesPedigreeFunction    2000 non-null float64
Age                         2000 non-null int64
Outcome                     2000 non-null int64
dtypes: float64(2), int64(7)
memory usage: 140.7 KB
```

Information of the diabetes after the EDA

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	1035.000000	1035.000000	1035.000000	1035.000000	1035.000000	1035.000000	1035.000000	1035.000000	1035.000000
mean	3.185507	122.842512	70.803865	29.263768	153.946860	33.307633	0.522839	30.670531	0.325604
std	3.167531	30.653888	12.336324	10.558989	111.489069	7.097899	0.332292	10.047212	0.468827
min	0.000000	56.000000	24.000000	7.000000	14.000000	18.200000	0.085000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	21.000000	76.500000	28.500000	0.282000	23.000000	0.000000
50%	2.000000	120.000000	70.000000	29.000000	126.000000	33.300000	0.452000	27.000000	0.000000
75%	5.000000	143.000000	78.000000	37.000000	190.000000	37.400000	0.684500	36.000000	1.000000
max	17.000000	198.000000	110.000000	63.000000	744.000000	67.100000	2.420000	81.000000	1.000000

6. OUTPUT

We have manually inputted the values as same as dataset already having the value. The output is derived from PyCharm framework and we have coded for both who is suffering from diabetic patient and normal people.

Fig a: Sample output 1

```
INPUTED LEVEL OF GLUCOSE 83
INPUTED LEVEL OF BLOODPRESSURE 65
INPUTED LEVEL OF INSULIN 66
INPUTED LEVEL OF BMI 36.8
INPUTED LEVEL OF AGE 24
STATUS PREDICTIED BY GRADIENT BOOSTER ALGORITHM IS: [0]
STATUS PREDICTED BY NAIVE BAYESIAN ALGORITHM IS: [0]
STATUS PREDICTED BY SUPPORT VECTOR MACHINE IS: [0]
[[0.96156825 0.03843175]]
PREDICTED OCCURENCE PERCENTAGE OF PERSON HAVING DIABETES USING GRADIENTBOOSTER IS  4 %
PREDICTED OCCURENCE PERCENTAGE OF PERSON NOT HAVING DIABETES USIN GRADIENTBOOSTER IS  96 %
PREDICTED OCCURENCE PERCENTAGE OF PERSON HAVING DIABETES USING NAIVESBAYES IS  2 %
PREDICTED OCCURENCE PERCENTAGE OF PERSON NOT HAVING DIABETES USING NAIVESBAYES IS  98 %
PREDICTED OCCURENCE PERCENTAGE OF PERSON  HAVING DIABETES USING SUPPORTVECTORMACHINE IS  7 %
PREDICTED OCCURENCE PERCENTAGE OF PERSON NOT HAVING DIABETES USING SUPPORTVECTORMACHINE IS  93 %
OVERALL PREDICTED PERCENTAGE OF PERSON HAVING DIABETES USING VECTOR CLASSIFIER IS 4 %
OVERALL PREDICTED PERCENTAGE OF PEARSON NOT HAVING DIABETES USING VECTOR CLASSIFIER IS 96 %

Process finished with exit code 0
```

Fig a: Sample output 2

```
INPUTED LEVEL OF GLUCOSE 195
INPUTED LEVEL OF BLOODPRESSURE 70
INPUTED LEVEL OF INSULIN 145
INPUTED LEVEL OF BMI 25.1
INPUTED LEVEL OF AGE 55
STATUS PREDICTIED BY GRADIENT BOOSTER ALGORITHM IS: [1]
STATUS PREDICTED BY NAIVE BAYESIAN ALGORITHM IS: [1]
STATUS PREDICTED BY SUPPORT VECTOR MACHINE IS: [1]
PREDICTED OCCURENCE PERCENTAGE OF PERSON HAVING DIABETES USING GRADIENTBOOSTER IS  95 %
PREDICTED OCCURENCE PERCENTAGE OF PERSON NOT HAVING DIABETES USIN GRADIENTBOOSTER IS  5 %
PREDICTED OCCURENCE PERCENTAGE OF PERSON HAVING DIABETES USING NAIVESBAYES IS  99 %
PREDICTED OCCURENCE PERCENTAGE OF PERSON NOT HAVING DIABETES USING NAIVESBAYES IS  1 %
PREDICTED OCCURENCE PERCENTAGE OF PERSON  HAVING DIABETES USING SUPPORTVECTORMACHINE IS  93 %
PREDICTED OCCURENCE PERCENTAGE OF PERSON NOT HAVING DIABETES USING SUPPORTVECTORMACHINE IS  7 %
OVERALL PREDICTED PERCENTAGE OF PERSON HAVING DIABETES USING VECTOR CLASSIFIER IS 96 %
OVERALL PREDICTED PERCENTAGE OF PEARSON NOT HAVING DIABETES USING VECTOR CLASSIFIER IS 4 %
```

7. CONCLUSIONS

In this paper, we proposed prediction and detection of diabetes using machine learning which will be useful for people who wants to know their medical condition for instance. As we are using above mentioned algorithm is much better than existing algorithm used. It gives more accuracy to predict the diabetes disease

REFERENCES

[1] C. D. Mathers and D. Loncar, "Projections of Global Mortality and Burden of Disease from 2002 to 2030," *PLoS Medicine*, vol.3, no.11,p.e442,Nov.2006. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[2] M. Wei, S. P. Gaskill, S. M. Haffner, and M. P. Stern, "Effects of Diabetes and Level of Glycemia on All-Cause and Cardiovascular Mortality: The San Antonio Heart Study," *Diabetes Care*, vol. 21, no. 7

[3] Christopher M. Bishop, Pattern recognition and machine learning,Springer,NewYork,1st edition, 2006.

[4] RichardDuda, Peter Hart, and David Stork, Pattern Classification, Wiley, NewYork,2nd edition, 2001.

[5] *International Journal of Advanced Computer and MathematicalSciences*. Bi Publication-BioIT Journals, 2010.

[6] S. Das and A. Thakral, "Predictive analysis of dengue and malaria," in *2016 International Conference on Computing, Communication and Automation (ICCCA)*, 2016, pp. 172–176.

[7] S., Hina, A., Shaikh, and S., Abul Sattar, "Analyzing Diabetes Datasets using Data Mining," Journal of Basic & Applied Sciences, vol. 13, pp. 466–471, 2017.

[8] Prof. Dhomse Kanchan B. and Mr. Mahale Kishor "Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis". International Conference on Global Trends in Signal Processing, Information Computing and Communication 2016.

[9] DeerajShetty, Kishor Rit , SohailShaikh and Nikita Patil "Diabetes Disease Prediction Using Data Mining". International Conference on Innovations.Embedded and Communication Systems, 2016.

BIOGRAPHIES

B.Selvaraj,
Final Year,
Agni College of Technology.

S.V.Pavithra,
Final Year CSE,
Agni College of Technology.

A.S.Nithya Rak,
Final Year CSE,
Agni College of Technology.