Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 09, September 2022

Detailed Analysis of Classifiers for Prediction of Diabetes

Sanjay Kumar
Department of CSE
BIT MESRA ,RANCHI

Ekta Kumari Gupta
Jharkhand Rai University

Vandana Bhattacharjee
Department of CSE
BIT MESRA ,RANCHI

***Abstract* - Diabetes mellitus is commonly known as diabetes, it falls in noxious diseases in the world. It is mainly a metabolic disease that can cause high blood sugar. Diabetes can occur when either the pancreas doesn't produces enough insulin or say that body cannot effectively uses the insulin it produces. Hyperglycemia, or raised blood sugar, is mainly a common factor of uncontrolled diabetes and over time it leads to serious damage to many organs in the body systems, especially the nerves and blood vessels. Therefore, the objective of this paper is to analyze different classification algorithm such as SVM, KNN, decision tree, random forest to detect diabetes at very early stage based on different parameter.**

Keywords: Diabetes, Machine Learning, SVM, decision tree, Random forest.

I. INTRODUCTION

Diabetes is a chronic disease that can be characterized by abnormally high levels of glucose (blood sugar). The people suffering from diabetes, their body is unable to properly process food for use as energy. The pancreas make a hormone called 'Insulin' helps glucose to penetrate into the cells of the Body, At times, the body doesn't make enough or any insulin. As a result, the glucose (or sugar) stays in the blood and a over a time period it causes health problems. That is why many people refer this disease as "Sugar".

The diabetes can be categorized in three types:-

- Type1
- Type 2
- Type 3

Type 1 diabetes were earlier known as insulin dependent diabetes mellitus (IDDM). In people suffering from Type 1 diabetes does not make sufficient quantities of Insulin. Type 1 diabetes is mainly diagnosed in children and younger adult, although it can occur in any age. People in Type 1 diabetes needs to take insulin every day to keep going. Type 1 account holds for 5-10% of all diagnosed cases of diabetes. Type 2 diabetes, earlier known as Non-Insulin dependent diabetes mellitus (NIDDM). It occurs when the body doesn't produce sufficient insulin or when the glucose stays in the blood and the body fails to utilize it as fuel for energy. Type 2 diabetes is often say linked in obesity. However other risk factor include old age, impaired glucose tolerance, physical inactivity, family history of diabetes is hyperglycemia with blood glucose values above normal but below those diagnostic of diabetes. Gestational diabetes occurs during pregnancy. These women and probably their children also at increased risk of type 2 in future. According to WHO over 429 million people across the globe are suffering from diabetes. According to Indian Medical Association, India has the second highest diabetes population in the world and are four times more likely to develop this disorder. One person dies every eight seconds because of diabetes and its related diseases says IMA early predictions of disease can control and for saving human life. For this objective we used PIMA Indian diabetes dataset for applying various machine learning classifications and various techniques for prediction of diabetes. Various techniques of Machine learning is now capable of doing predictions however it is really difficult to choose the best techniques. Thus for this purpose we apply popular classification and ensemble methods on PIMA dataset for prediction.

II. LITERATURE REVIEW

Mitushi Soni, Dr. Sunita Verma et al. [1] proposed various classification methods to predict diabetes. They also explained that diabetes is noxious diseases in the world. They used Pima Indian diabetes set. They have used various classification methods like KNN, logistics regression, decision tree, SVM, Gradient boosting and Random Forest. Their prediction suggested that among all the classification methods, Random Forest gives the highest accuracy. Aishwarya majumdar, Dr.vaidehi v et al. [2] explained that around 425 million people suffer from diabetes according to 2017 statistics. Current medical diagnosis methods , there are 3 types of errors-(1)the false negative (patient is already diabetic but the results show negative) (2) The false positive (non-diabetic but result positive)(3) unclassifiable (cannot diagnosed). Sonal Kumari ,Dr. Vandana Bhattacharjee et al. [3] used data mining and Machine Learning algorithm to predict the capability to manage the data that has been used and the classification accuracy is based upon the training time and prediction time, they suggested Neural Networks gives the best accuracy among SVM, Decision Tree, KNN and Neural Networks classification model . Tejas N. Joshi ,Prof. pramila M. Chawan et al.[4]used SVM, Logistic Regression and ANN just to predict that the patient is suffered from diabetes or not ,they used supervised learning method for their research. K. Rajesh and V.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 09, September 2022

Sangetha et al.[5] used Naive Bayes data mining classifier technique which produces an optimal prediction model with minimum training set. They predict error rate by different classifier and then the accuracy rate is predicted by different algorithm. Their accuracy report was 91% by c4.5 algorithm among CRT ,CS-RT,IDE , KNN ,LDA, Naive Bayes, PLS-DA,SVM,RND Tree. Priyanka Sonar, Prof. K. Jayamalini et al. [6] model development is based on categorization methods as Decision Tree ,ANN , Naive Bayes and SVM algorithms .Their machine learning matrix used precision, recall and f1-score to calculate the accuracy. They got highest accuracy in ANN classification. Amani Yahyaoui, Jawed Rasheed, Akhtar Jamail, Mirsat Yesiltepe et al.[8] predicted Random Forest as the best classification model for prediction of diabetes. The performance of the proposed method was evaluated in terms of overall accuracy (OA), kappa coefficient (KC), precision (P), recall (R) and f-measure (F) . Muhammad Azeem Sarwan, Nasir Kamal ,Wajeeha Hamid, Munam Ali shah et al.[7] uses predictive analysis in healthcare, six different Machine Learning algorithm are applied on the dataset. They reported that among all the 6 classification SVM and KNN is appropriated for predicting the diabetes disease. Their accuracy is about 77%. MD.Kamrul Hasan ,MD.Ashraful Alam,Dola Das ,Eklas Hossain,Mahmudul Hasan et al.[9] proposed ba robust framework for diabetes prediction where the outlier rejection, filling the missing values,data standardization, feature selection,K-fold cross validation and different Machine Learning classifier (KNN, Decision Tree, Random Forest, Ada Boost, Naive Bayes and XGBoost) and multilayer perceptron (MLP) we're used. Dost Muhammad Khan ,Nawaz Mohanmudally et al.[10] used to integrate the K-means clustering algorithm using intelligent agent, called LIA(learning intelligent agent) ,which is capable to perform, classification, clustering, and interpretation tasks on the dataset. This provide more sophisticated and powerful tools for data mining. Dr. Saravana Kumar, Eswari, Sampath and lavanya et al. [11] uses predictive analysis algorithm in Hadoop /Map reduce environment to predict the diabetes types prevalent, complications associated with it and types of treatment to be provided. They explained when large amount of diabetes data were analyzed using Hadoop and the final results were spread across different servers and replicated over a number of nodes depending upon the location. Utilizing appreciate electronic communication technology to share patient data between health care facilities will result in to received affordable, appropriate care in far-off places at the appropriate time at low cost.

III. MATERIALS AND METHODS

Our goal of this paper is to predict the diabetes, based on different classification methods and ensemble algorithm to find the better accuracy and for prediction of diabetes. We can briefly discuss the phases given below.

A. Data set description

We used PIMA Indians diabetes databased of National Institute of Diabetes and Digestive and kidney diseases from UCI machine learning repository. The database has many attributes of 768 patients.

Table 1: Dataset Description

S.NO	Attributes
1	Pregnancy
2	Glucose
3	Blood Pressure (BP)
4	Skin Thickness
5	Insulin
6	Body Mass Index (BMI)
7	Diabetes Pedigree Function
8	Age

The 9th attribute is class variable of each data prints. This class variable shows 0 and 1 outcome for diabetes which indicates positive or negative for diabetes.

B. Data Pre - Processing:-

Data Processing is the most important process for validating the data because it contains so many missing values and impurities. To use machine learning process on datasets. Effectively this process is essential for the accuracy of results and successful prediction. We need to perform pre-processing in 2 different stages.

1. Missing Values Removal :

Removing all the instances having 0 values as worth is not possible at all. Therefore we normalise these instances by taking mean of those attributes. It makes the data efficient and helps to predict more accurate results.

2. Splitting of the data :

After cleaning the desired dataset , the data is said to be normalized and then the training and testing of the dataset will begin. When the data is splited then we train algorithm on those training dataset and keeping that test data aside. The training process those produce the training model based on the logic and algorithms and values of the feature in training data. Basically, we can says that the normalization is to be bring all the attributed under the same scale.

C. Applying various Machine Learning techniques

Diabetes data once ready, we applied various Machine Learning techniques. We use various classification methods and ensemble techniques, to predict the diabetes. Our main objective of this research is to investigate the performance of those methods and to search out the accuracy of them, and also to work out the main responsible or say important feature which can predict the diabetes more effectively. The various techniques are as follows:-

1. Support Vector Machine : Support Vector Machine is known as SVM is a supervised machine learning algorithm that can be used for both classification and regression. It is the most popular classification techniques. The main objective of SVM is to find out a hyperplane in n- dimensional space that distinctly classifies the various data points. The hyperplane is used for classification or regression.we choose the best hyperplane which represents the largest separation or margin between the 2 classes.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 09, September 2022

Algorithm used for SVM :

- Selecting a hyperplane which divides the classes effectively.
- It is always better to search out a hyperplane which is used to calculate the gap between the planes and therefore the data which is known as margin.
- If the space between the classes is low then we will say the chances of miss conception is high and vice-versa so we want to search out the distance.
- Select the classes which has shows the highest margin among all .
  Margin= distance to positive point + distance to negative point.

2. K-nearest neighbor : An additional supervised machine learning approach is K-nearest neighbor (KNN) .KNN AIDS in the resolution of the classification and regression issues. KNN assumes that the related things are located close to one another.In many cases, comparable data points are very close proximity. KNN facilitate grouping new work according to a similarity index. KNN algorithm records each and every records and sort them based on how similar they are measured . Uses are made of distance measurements between sites . Tree like architecture to forecast a new set of data point. The algorithm identifies it's nearest neighbors or data points in the training data set that are closest to it. Where K is the number. It's always a positive integer, close neighbors. Neighbors value is selected from a class collection.

Algorithm for KNN:

- Take the PIMA Indian diabetes dataset which contains columns and rows of data.
- Create a test dataset with rows and attributes.
- Determine the Euclidean/Manhattan distance between them.
- Next ,choose a random inter for K, where K is the number of proximate neighbors.
- Next, using these minimal distance and find out the nth columns of the distance in euclidean/ Manhattan terms each.
- Determine the identical output values. The patient has diabetes if the levels are the same, otherwise,they are not.

3. Decision Tree : A fundamental method of classification is Decision Tree.This type of learning is supervised. For categorical response variable,a decision tree is employed. The decision tree has a model with a tree like structure that defines how features in the input are used to classifies objects. Input parameters are any type , including text, graphs, discrete and continuous data etc.

Decision Tree algorithm:

- Creating a tree by using node as the input feature.
- Decide which feature will provide the greatest information gain when predicting the output from the input feature.
- Calculate to yield the highest information gain are each attribute in a tree node.
- To create a subtree using the feature, repeat step 2 which the above node does not use.

4. Random Forest : It belongs to the category of ensemble learning methods and is employed in classification and regression application. These techniques were developed by Leo Bremen . It is an effective methods for group learning. Through the reduction of variation, Random Forest enhances the performance of Decision Tree. It works by building many decision tree during training time and produces the class that is the mode of the classes or classification or mean prediction (regression) of the individual trees.It gives the highest performance among all the classification model.

Algorithm for decision tree:

- Step 1: K data points are choosen at random from the training set.
- Step 2: Create the Decision Tree linked to the chosen data point (subset)
- Step 3: Select the decision tree N that you want to construct.
- Step 4: Repeat step 1 and 2.
- Step 5: Discover each decision tree's predictions for any new data points, then place them in the category that receives the most votes.

D. Model building of proposed methodology

The most important phase of our research work is to include model building to predict the diabetes based on various Machine Learning algorithm.

Procedure of our proposed methodology-

- Step 1. Import PIMA diabetes dataset and import all the required libraries.
- Step 2. Pre-processing data is required to get rid of all the missing data
- Step 3. Selecting the machine learning algorithm. i.e. SVM, KNN, Decision Tree and Random Forest.
- Step 4. Building the classification model for the machine learning algorithm supported training set.
- Step 5. Testing is performed based on the given dataset.
- Step 6. Performing the comparison evaluation of experimental performance results obtained for every classifier.
- Step 7. After analyzing all the classification model we select the most effective performing algorithm.

IV. RESULTS AND DISCUSSION

In the research work various classification model were applied on the PIMA dataset with varying parameter value. Table 1 represent the results of SVM classification with Linear, polynomial and rbf kernel. Table 2 represent the KNN classifier results with different number of neighbors and distance measure techniques. Table 3 represent the results of decision tree with different depth of the tree and

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 09, September 2022

attribute selection measures .Table 4 represents the Random Forest classifier with different estimators.

Table 1 : SVM accuracy on different kernels

Types of kernel	Accuracy
Poly(polynomial)	0.72395
Linear	0.7916
Rbf(Gaussian/radial basis function)	0.7552

Table 2: accuracy value with KNN

Distance Measure	Number of neighbors			
	25	50	75	100
P=1 Euclidean	0.78125	0.78125	0.7656	0.7760
P=2 Manhattan	0.79785	0.76041	0.76041	0.74479

Table 3: accuracy value with DT Classifier

Attribute selection measure	Max Depth Value					
	3	4	5	6	7	8
Gini	0.7343	0.682	0.76041	0.7291	0.7604	0.7968
Entropy	0.75	0.75	0.7708	0.7968	0.77	0.77083
Gini –Gini index; Entropy: information gain						

Table 4: accuracy value with random forest Classifier

No of estimators	100	200	300	400
	0.8181	0.81168	0.81168	0.83116

As seen from Table 1, for the SVM classifier the accuracy obtained with polynomial kernel is 72.39%, for the Linear kernel it is 79.16% and for Gaussian kernel is 75.52%.Similarly for Table 2, the KNN classifier include euclidean distance which is 78.12% and Manhattan distance is 79.78% accuracy report.In Table 3, accuracy report for Decision Tree classifier shows 79.68% for Gini and 79.68% for entropy , based on different depth values.Table 4 shows Random Forest classifier based on the number of estimator, it shows 83.11% accuracy which shows the best estimator among all the different classifier.

IV. CONCLUSION

The primary goal of this research was to create, implement and analyze the performance of approaches for predicting diabetes using machine learning. We have achieved successful results. We have used pima Indian diabetes dataset for our research. The suggested strategy makes advantage of numerous categorization and ensemble learning approach in which the classifier SVM ,KNN , Decision Tree and Random Forest are employed. The categorization accuracy rate was 83% in Random Forest classifier, which gives the better accuracy among all the classifier. The experimental results can help health care providers make early prediction and decisions to treat diabetes and save lives of people.

V. REFERENCES

[1] Mitushi Soni and Dr.Sunita Verma,"Diabetes Prediction Using Machine Learning Techniques", International Journal of Engineering Research and Technology (IJERT) ISSN:2278-0181,Vol.9 Issue 09, September-2020

[2] Aishwarya Majumdar, Dr. Vaidehi v " Diabetes Prediction using Machine Learning Algorithms", International Conference on recent trends in advanced computing 2019, ICRTAC 2019

[3] Sonal Kumari and Vandana Bhattacharjee , " An In depth experiment with classifiers for prediction of diabetes", International Journal of Engineering Sciences and Research Technology . ISSN:2277-9655, Impact factor:5.164, CODE:IJESS7

[4] Tejas N. Joshi, Prof Pramila M. Chawan,"Diabetes Prediction using Machine Learning Technique" S.Dewargan.et.al. Int. Journal of Engineering Research and Application ISSN:2248-9622 Vol.8, Issue 1,(part-2) January 2018, pp-09-13

[5] K. Rajesh and V. Sangeetha,"Application of Data Mining Methods and Techniques for Diabetes Diagnosis", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.

[6] Priyanka Sonar and Prof K.Jaya Malini,"Diabetes Prediction using different Machine Learning Approaches", Proceeding of the Third International Conference on Computing Methodology and Communication (ICC MC 2019) IEEE Xplore Part Number: CFPI9K25-ART ; ISBN : 978-1-5389-7808-4

[7] Muhammad Azim Sarwar, 2 Nasir Kamal, 3 Wajeeha Hamid, 4 Munam Ali Shah, " Prediction of the 24th International Conference on Automation and Computing, Newcastle University, Newcastle upon Tyre UK, 6-7 September 2018

[8] Arani Yahyaoui, Akhtar Jamil, Jawad Rasheed ,Mirsat Yelliepe," Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques ",978-1-72816 3992-0/19/$31.00 ,2019 IEEE.

[9] Md. Kamrul Hasan 1, MD. Ashartul Alan 1, Dola Das 2, Eklas Hossain 3(Senior Member IEEE), Marmudul Hasan 2, " Diabetes Prediction Using Ensembling of Different Machine Learning Classifier", DOI 10.1109/ACCESS.2020.298957,IEEE Access.

[10] Dost Muhammad Khan1 , Nawaz Mohamudally 2, "An Integration of k-means and Decision tree (ID3) towards a more efficient Data Mining Algorithm", Journal of Computing, Volume 3 , Issue 12, December 2011.

[11] Dr. Sarvana Kumar N M , Eswari T , Sampath P, and Lavanya S ," Predictive Methodology for Diabetic Data Analysis in Big Data ", 2nd International Symposium on Big Data and Cloud Computing ,2015.