Prediction of Diabetes using Ensemble Techniques

Prema N S, Varshith V, Yogeswar J

Abstract: ***In the recent decades, there has been a significant improvement in the quality and quantity of medical data that is produced by digital devices. This has led to inexpensive and easy production of data. Thus, there has been an increased advantage in the areas of Big Data and Machine Learning. In this paper, various machine learning algorithms are applied to predict diabetes, based on specific attributes. The performances of the algorithms are compared in terms of accuracy, voting based ensemble techniques is applied for the normalized pima diabetes data for which a highest accuracy is achieved.***

Keywords: Diabetes, Ensemble classifier, Voting classifiers, SVM.

I. INTRODUCTION

Diabetes mellitus, most commonly termed as Diabetes, is one of the diseases which affects a very large population of human beings. In 2017, more than 425 million people were affected by Diabetes [1], which is a very huge number. Due to Diabetes and its related complications, around 4 million people died in the same year. Whereas in India 74 million people were affected by diabetes, and India is referred to as the "Diabetes Capital of the World". If this disease is not considered seriously and if no major steps are taken to diagnose and prevent it, the number of people to be affected by Diabetes may increase to more than 629 million people worldwide by 2045 according to an estimation [1].

Diabetes is a state where the blood glucose levels are high which is caused when the body cannot produce the required amount of insulin hormone or when the body is unable to effectively use the produced insulin. Obesity, urbanization, physical inactivity, unhealthy diet, aging, family history of diabetes are the most common causes of Diabetes. If Diabetes is not diagnosed at the right time or if it is not managed properly, it can lead to many complications like cardiovascular problems, kidney related diseases, blindness, neural complications like cerebrovascular accident [2]. In the effective management of Diabetes and its related complications, an early diagnosis is the most important factor, along with a recommended healthy day to day lifestyle [2].

II. LITERATURE REVIEW

Some of the different methods that have been applied on PIMA Indian diabetes dataset are described below with its results.

Rohan bansal and et.al used KNN classifier for the diagnosis of diabetes; the attributes are selected using Particle swarm optimization (PSO) techniques. This method is proved to provide a prediction accuracy of 77% [4]

A Class wise KNN(CKNN) methodology for classification of diabetes dataset was proposed where the preprocessing of the dataset is done using normalization and an improvised model of KNN algorithm, i.e., class wise KNN algorithm is applied on the dataset for classification. This method achieves an accuracy of 78.16% [5].

Lin Li et.al have proposed one of the voting classifier techniques popularly known as weight-adjusted voting technique. This method after implementation on PIMA Indian diabetes dataset gives out a prediction accuracy of 77% [6].

Priyadarshini et.al have used the concept of modified extreme learning machines to predict whether the patient is having diabetes or not basing on the available diabetes dataset. The authors have drawn comparative inferences using neural networks and extreme learning classifiers [7].

A brief review of data mining techniques used for the diagnosis of diabetes is discussed in [9]. in the review the authors have mentioned that many data mining algorithms were used for the diagnosis in that 85% were characterized by supervised learning approaches and 15% by unsupervised ones, and more specifically, association rules. A support vector machine (SVM) is the most successful and widely used algorithm.

In [10-11] different classifiers are used for the Prediction of gestational diabetes mellitus (GDM) and the obtained highest accuracy is 86.7%, the classifiers used are Decision tree(J48), Random forest and Naïve Bayes.

III. MATERIAL AND METHODS

Pima Indian Diabetes Data set

The UCI Machine Learning Repository has a repository of various dataset which is used for the study and application of machine learning algorithms. It has been widely used by researchers, students and educators as a primary source of machine learning data sets. From this repository, we have taken the PIMA Indian Diabetes Dataset [3] for the purpose of our study. This dataset consists of the medical data of 768 patients.

There are 8 attributes in each data point and they are:

1. Number of times pregnant
2. Plasma glucose concentration
3. Diastolic blood pressure

Revised Manuscript Received on April 05, 2019.

Prema N S, Department of Information Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, India (prema.gowda@gmail.com)

Varshith V, Department of Information Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, India

Yogeswar J, Department of Information Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, India

4. Triceps skin fold thickness
5. 2-hour serum insulin
6. Body mass index
7. Diabetes pedigree function
8. Age

The 9th attribute of each data point is the class variable. The outcome will be either 0 or 1 for positive or negative diabetes

Normalization

Usually, any real-world data contains some kind of noise. So, pre-processing of data is done to reduce it. Each attributes of the data may have a very different range of values [8]. For example, in our PIMA Indian Diabetes dataset, the second attribute, i.e. plasma glucose concentration has a range of 44 to 199, whereas the seventh attribute, i.e. diabetes pedigree function has a range of 0.078 to 2.42. This kind of variation in the ranges makes the algorithms which uses distance between data points, to have different weightage for the variance in different attributes. To fix this, Normalization is done. The main aim of normalization is to bring all the attributes under a same scale, that is under a same minimum, maximum and median values, so that the previously mentioned problem is overcome. We have used feature standardization (z-score normalization) which is normalizing the data using mean and standard deviation. It is done using the following formula [16]:

$$z = \frac{x_i - \mu}{\sigma}$$

μ = Mean
σ = Standard Deviation
Z = Normalized attribute value
Xi = Original attribute value

Ensembling

Ensemble is a Machine Learning technique whose methods are meta-algorithms that combine several machine learning techniques into one optimal predictive model in order to reduce variance, bias or improve predictions. This approach enables improved predictive performance when compared to that of a single model.

There are various methods of ensembling such as bagging, boosting, ada-boosting, stacking, voting, averaging etc.

We have applied voting based ensembling method on PIMA Indian diabetes dataset.

The Ensemble Vote Classifier is a meta-classifier which combines similar or conceptually different machine learning classifiers for classification through majority or plurality voting.

There are two types of voting based ensembling methods. They are:

Majority voting.
Weighted voting.

In this Model, we have applied Majority voting classifier.

Majority Voting classifier

Each model makes its own prediction and the output which has received more than half of the votes is considered as the final prediction. We may say that the ensembling method was not able to make a stable prediction when none of the prediction gets more than half of the votes. Although this is the widely used technique, we may sometimes consider the prediction with most votes (even though if it is less than half of the votes) as a final prediction. This method might also be called as "Plural voting". In this research, we have applied majority voting classifier for various methods such as K Nearest Neighbours, Logistic regression, Decision Tree, Random Forest, Naive Bayes, Linear SVM, RBF SVM, Gaussian Proc, AdaBoost, QDA. The prediction made by all these classifiers on a test case is voted and the prediction with the highest votes is considered as a final prediction.

Algorithm for Majority Voting classifier:

```
# votes = List of integer votes
Votes_table = {} # New Hash Table
for vote in votes :
if vote in votes_table:          # To check if key is in hash table
votes_table[vote] += 1   # Increment the counter
else:
votes_table[vote] = 1    # Creating counter for the vote
# Finally find max counter in hash table
return max (votes_table, key = votes_table.get)
```

IV. RESULTS

Various classification techniques are applied to the pima Indian diabetes the results are shown in table 1. The data to the classifiers is sent in two ways one by splitting data into 30% testing and 70% training, the accuracy of the classifiers and voting classifier is shown in fig 1, second 10 tenfold cross validations is done the accuracy of the classifiers is compared in fig 2.

Table1: Performance measures of the classifiers

Classifier/Performance measure	Specificity	Sensitivity (precision)	Recall	F-Measure
KNN	0.51	0.88	0.77	0.82
Logistic Regression	0.54	0.87	0.78	0.82
Decision Tree	0.63	0.77	0.79	0.78
Naive Bayes	0.52	0.85	0.77	0.81

Linear SVM	0.52	0.81	0.76	0.78
RBF SVM	0.52	0.88	0.77	0.82
Gaussian Process	0.52	0.86	0.77	0.81
Ada Boost	0.57	0.82	0.78	0.80
Random Forest	0.63	0.83	0.81	0.82
Voting Classifier(30% test data)	0.60	0.92	0.81	0.86

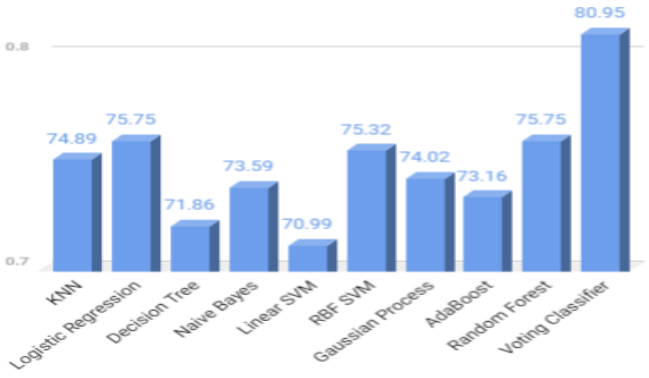

Fig 1: Accuracy of the classifiers with training and testing data

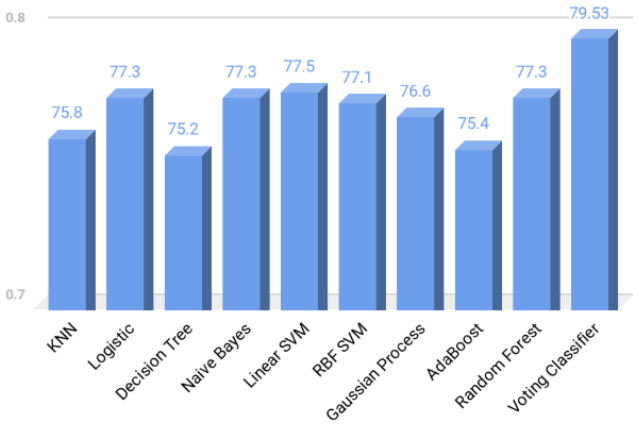

Fig 2: Accuracy of the classifiers with cross validation

V. CONCLUSION

Prediction of diabetes is done using ensemble voting classifiers for pima Indian diabetes dataset, in comparison with different classification algorithms, the highest accuracy of 80% and 81% is achieved for data set by using 10-fold cross validation and by spitting data into 30% testing and 70% training.

ACKNOWLEDGEMENTS

The manuscript is prepared by taking assistance from Accendere Knowledge Management Services Pvt. Ltd, we are thankful to them. We also express our gratitude to our teachers and mentors for guiding us throughout the work.

REFERENCES

1 "IDF DIABETES ATLAS - 8TH EDITION," International Diabetes Federation, 2017. [Online]. Available: https://diabetesatlas.org/. [Accessed: 15-Dec-2018].

2 GLOBAL REPORT ON DIABETES WHO Library Cataloguing-in-Publication Data Global report on diabetes. 2016.

3 "PIMA Indian Diabetes Dataset, An open dataset," UCI Machine Learning Repository. [Online]. Available: http://ftp.ics.uci.edu/pub/machine-learning-databases/pima-indians-diabetes/. [Accessed: 11-Jan-2019].

4 R. Bansal, S. Kumar, and A. Mahajan, "Diagnosis of diabetes mellitus using PSO and KNN classifier," in 2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN), 2017, pp. 32–38.

5 Y. A. Christobel and C. Sivaprakasam, "A NEW CLASSWISE K NEAREST NEIGHBOR (CKNN) METHOD FOR THE CLASSIFICATION OF DIABETES DATASET," Int. J. Eng. Adv. Technol., vol. 2, pp. 396–400, 2013.

6 L. Li, "Diagnosis of Diabetes Using a Weight-Adjusted Voting Approach," in 2014 IEEE International Conference on Bioinformatics and Bioengineering, 2014, pp. 320–324.

7 R. Priyadarshini, N. Dash, and R. Mishra, "A Novel approach to predict diabetes mellitus using modified Extreme learning machine," in 2014 International Conference on Electronics and Communication Systems (ICECS), 2014, pp. 1–5.

8 S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data Preprocessing for Supervised Leaning," World Acad. Sci. Eng. Technol. Int. J. Comput. Electr. Autom. Control Inf. Eng., vol. 1, no. 12, pp. 4091–4096, 2007.

9 Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, and Nicos Maglaveras , "Machine Learning and Data Mining Methods in Diabetes Research", Computational and Structural Biotechnology Journal, vol. 15, pp. 104–116, 2017.

10 Prema, N S and Pushpalatha, M P. "Prediction of gestational diabetes mellitus (GDM) using classification", 2017 IEEE International Conference on science, technology, engineering and management (icstem'17), Coimbatore, 2017.

11 Kotsiantis, Sotiris., Kanellopoulos, Dimitris and Pintelas, P. "Data Preprocessing for Supervised Learning. International Journal of Computer Science", vol. 1, pp. 111-117, 2006.