IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DIABETES PREDICTION USING DATA MINING TECHNIQUES

VIGNESH S

Master of Computer Application
Coimbatore Institute of Technology
Coimbatore,TamilNadu,India
vikiwillrock@gmail.com

JEYA SANKAVI P

Master of Computer Application
Coimbatore Institute of Technology
Coimbatore,TamilNadu,India
jeyasankavi09@gmail.com

ABRNA N S

Master of Computer Application
Coimbatore Institute of Technology
Coimbatore,TamilNadu,India
abrnans@gmail.com

Dr.J.B.JONA

Associate Professor,
Department of Computer Application
Coimbatore Institute of Technology
Coimbatore, Tamilnadu, India
jona@cit.edu.in

Abstract

Diabetes is powerful disease in the world. There are many proposed solutions for solving the problem and find whether a person is having diabetes or not. In this paper, various methods are being proposed on how to find the person is having diabetes using A diabetes dataset by implementing various machine-learning algorithms. A comparison study has been carried out for the machine learning algorithms to verify which algorithm is performing well with better accuracy.

Keywords: Prediction, SVM-Algorithm, Decision tree Algorithm, Gradient Boosting Algorithm, Accuracy.

I. INTRODUCTION

Diabetes is also known as Diabetes mellitus. It usually occurs when pancreases don't not produce sufficient insulin in our body. This leads damage in body part like heart and blood vessels, eyes, kidneys and nerves. People get diabetes mostly between age (17-70). In the past three years, the diabetes cases have increased enormously. Worldwide there are 430 million people having diabetes. The number of people with diabetes in India has increased from 30 million to 75 million. According to 2018 National diabetes center

research data, 15.8% people died due to diabetes disease. This means India actually has the highest number of diabetes than any other country in the entire world.

There are three categories in diabetes,Category-1: Insulin is not produced at pancreases. Category-2: Insulin is produced but not at a sufficient level. Category-3: Gestational diabetes occurs in the short term on a pregnant woman. Category 1 and 2 are caused by improper intake of food and overweight.

II. LITERATURE REVIEW

According to the literature survey that has been carried out, [Velmurugan ,K Saravananathan-10] Diabetes mellitus prediction is carried out for classification and prediction techniques. There are many ways prediction can be done and different results can be produced. Many of them use the PIMA(Participant Identification and Messaging Address) dataset to detect whether the diabetes result is positive value or negative value. For prediction, the data mining model is used because it is very adaptive and it can be used to test more than one dataset. For preprocessing the dataset, WEKA Tool is used and various filters are applied. The data is transformed from the noisy data to pure data for applying correct data mining techniques and K-mean algorithm is used for clustering the data which further is used as input for the next level. Then logistic regression is used to classify the data and the model is verified using the K-fold process.

[Abdulhakim Salum etal.,[1] proposed the dataset has been collected from USA hospital for analysis and prediction. For analysis,Navies Bayes and Random Forest methods are used and the accuracy is compared for both the methods. For Random forest, accuracy is 69.23% which is better than the Naive Bayes algorithm.

[Aishwarya Mujumdar etal.,[2]] In this work, preprocessing of the dataset is carried out to remove the duplicate and null values. For study a Hybrid classification model was built using the cooperative technique. A study is performed on various data mining algorithm and techniques for analysis and prediction of diabetes at the beginning stage. The accuracy is determined by using the logistic regression technique. It provides high accuracy in analyzing the disease.

[Misba Reyaz, etal.,[3]] The author here used the dataset collected from Kaggle (PIMA) for prediction and implemented the SVM algorithm, random forest and decision tree. Among all technique SVM shows the best accuracy of 75.45%. This paper gives a solution to diabetes if they take care of food and proper exercise diabetes can be controlled. Various mining methods and probabilty methods are used to calculate diabetes.

[Minyechil Alehegn etal.,[4]]The author describes an android application to overcome the deficiency caused by diabetes and awareness about diabetes and the effect of diabetes at the adult stage. The application was developed by using the data mining algorithm and they used the decision tree classifier to predict diabetes for the patient. The app also provides information about diabetes. The app uses the PIMA dataset for analysis of diabetes and adds a risk analysis feature to detect the level of diabetes.

[Deepti Sisodiaa-5]In the designed system for the prediction of diabetes, genetic programming has been used for training and loading the dataset to the database. The programming also proposes a solution for diabetes using a machine learning algorithm. By applying the classification algorithm a model is designed to predict diabetes. By analyzing the dataset of diabetes taken from Kaggle, it gives an optimal accuracy as compared to other methods.

[Ms. Nilam chandgude(Author) etal.,[6]]A Hybrid model is developed by author which is

used to detect the diabetes at early stage by applying the mathematical model and machine learning model. For this model PIMA dataset is being used and the accuracy is improved. There is another type in diabetes that is called as diabetes retinopathy which mainly affects the blood vessels in our body. So a neural network is developed using advanced technology to predict diabetes at early stage.

After conducting a literature survey, in this paper, three machine learning algorithms has been implemented and they are Decision tree, Gradient Boosting and Support Vector Machine.

III METHODOLOGY

This process transforms the data without any missing value and null value and finding the hidden pattern and relation between the pattern in the large dataset. These are steps that has been followed while find accuracy and prediction for given dataset.

- Data cleaning
- Data integration
- Data transformation
- Exacting pattern
- Visualization the data

Dataset

PIMA dataset is initially collected from National Institute of Diabetes and Kidney diseases in India. The main point of dataset to predict whether the diabetes patient having diabetes or not based on certain variables like age,inslulin included in the dataset which are presented below.

Variables in Participant Identification and Messaging Address dataset

- Body mass index
- Insulin values
- Classic pregnancy
- Ages
- 0-Non Target
- 1-Target

3.1.Machine Learning Algorithms used

Here in this section the Machine learning algorithms such as, Decision tree, Gradient Boosting and Support Vector Machine that are going to be implemented on the PIMA dataset are discussed.

Decision Tree: A decision tree is a predictive modeling approach that is used in machine learning. It makes use of a decision tree to give the output from a set of observations. It is a supervised learning algorithm. Here decision making is represented in the form of a tree-like structure.

GradientBoostingAlgorithm:Gradient Boosting is a group learning the method that is commonly used for classification problems and regression related problem, which work by building a large no of decision_trees and random forest tree at the training time and produces an output in class that is the mode of the classes or mean/average prediction of individual trees from the given dataset.

Support vector machine (SVM): SVM is a collaborative learning method that is commonly used for classification and regression related problem. Where hyperplane a line drawn between the points and finding distance point to the line. Here SVM is represented in the form of a Graph points.

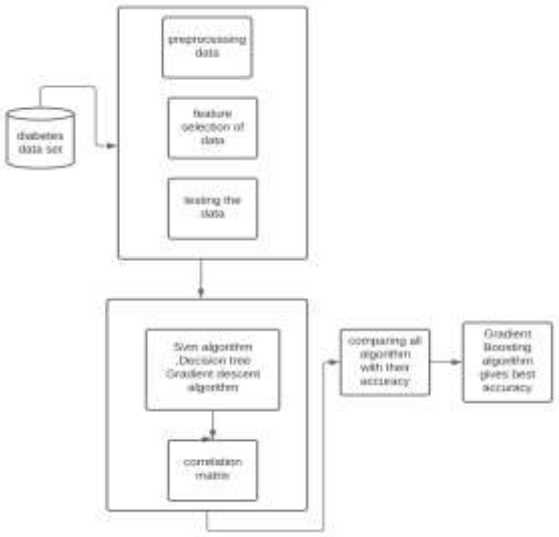

Fig. 1. Architecture Diagram of proposed system

IV RESULTS AND DISCUSSIONS

Before choosing the right machine learning algorithm for our project we must have a complete idea about the algorithms that we are about to use. We have to find the error matrices, accuracy, and confusion matrix so that we can have the best result for our project.

Confusion matrix: Confusion matrix is used which is a better way to evaluate a performance of a classification algorithm**.**

Accuracy is an indicator for evaluating classification models. Informally, precision is part of the correct prediction of our model. Formally, precision has the following definition.

Accuracy = number of correct predictions / total number of predictions

The Error metrics, accuracy, and confusion matrix for the three algorithms Decision tree, Support Vector Machine and Gradient Boosting on the PIMA dataset are depicted in Fig-3 toFig-8.

```
Mean Absolute Error: 0.3246753246753247
Mean Squared Error: 0.3246753246753247
Root Mean Squared Error: 0.5698028822981898
Accuracy for desicison tree: 0.6753246753246753
```

Fig. 2. Error metrics and accuracy for Decision Tree.

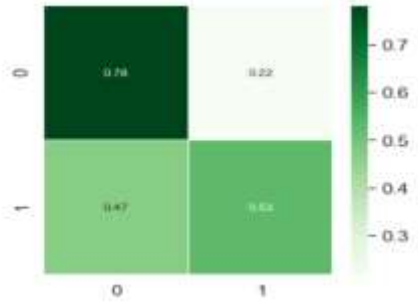

Fig. 3. Confusion_matrix for Decision Tree using PIMA Dataset

```
Mean Absolute Error: 0.23376623376623376
Mean Squared Error: 0.23376623376623376
Root Mean Squared Error: 0.48349377841522817
Accuracy for SVM tree: 0.7662337662337663
```

Fig. 4. Error metrics and accuracy for Support Vector Machine

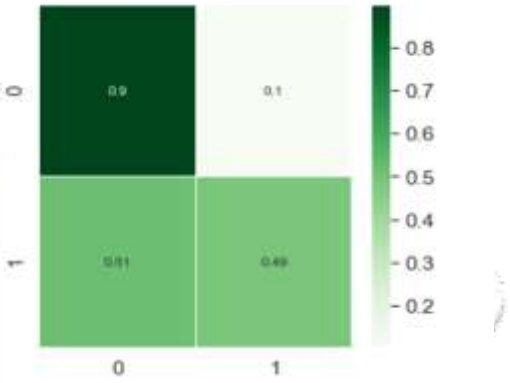

Fig. 5. Confusion matrix for Support vector Machine.

```
Mean Absolute Error: 0.04329004329004329
Mean Squared Error: 0.04329004329004329
Root Mean Squared Error: 0.20806259464411975
Accuracy for Gradient tree: 0.9567099567099567
```

Fig. 6. Error metrics and accuracy for Gradient Boosting.

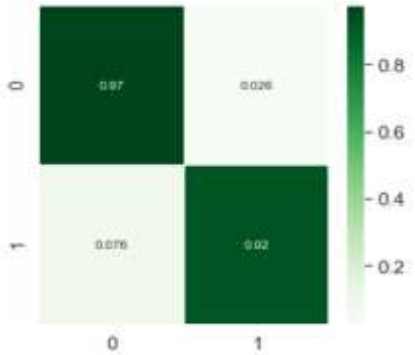

Fig.7.Confusion matrix for Gradient Boosting

Accuracies of all the three algorithms has been presented in Table1. By comparing all the outcomes, Gradient Boosting algorithm is

found to provide the highest accuracy than other two methods.

Method	Model Accuracy
Gradient Boosting	0.95%
Support vector Machine	0.76%
Decision tree	0.67%

Table1 Accuracy comparison

V CONCLUSION

For prediction of diabetes, various machine-learning model have been used. The models like Support Vector Machine and Gradient Boosting and Decision tree are executed to Participant Identification and Messaging Address dataset. Comparing all models, it has been found that Gradient Boosting model gives better accuracy among the three methods. In future, the other areas of machine learning could be explored with high dimensional datasets to predict the performances of the algorithms in terms of faster prediction as well as better accuracy.

REFERENCES

1. Han Wu, Shengqi Yang , Zhengyi Huang, Jian He, Xiao Wang, “Type 2 diabetes mellitus prediction model based on data mining”, Informatics in Medicine Unlocked, 2017.

2. Nahla Barakat ,Raj prakash, Andrew P. Bradley ,Mohamed Nabil H. Baraka. "Intelligible Support Vector Machines For Diagnosis of Diabetes Mellitus IEEE transaction (2010)", International Conference on Computational_Intelligence and Data Science(ICCIDS 2018), 2018

3. Abdullah A. Aljumah, Mohammed Gulam Ahamad, Mohammad Khubeb Siddiqui, " Application of data mining: Diabetes health care in young and old patients ", Journal of King Saud University – Computer and Information Sciences, 2012.

4. Aishwarya Mujumdar, Dr. Vidani "Accuracy improvement for diabetes disease classification: A case on a public medical dataset", Fuzzy Information and Engineering. Elsevier,vol(9),345-357,2017.

5. Deepti Sisodiaa A. Adekunle , Adnan Khashman.K,M.Ebenezer,O.Olaniyi,Oyebade ,K.OYEDOTUN,"DiabeticRetinopathy Diagnosis Using Neural Network Arbitration", Bulletin of the Transilvania_University of Braşovstanley • Vol 10(59), No. 1 – 2017

6. Ms. Nilam chandgude(Author), Prof. Suvarna pawar, "A survey on diagnosis of diabetes using various classification algorithm", International Journal on Recent and Innovation Trends in Computing and Communication Volume: 30 Issue: 12 ISSN: 2221-8169 7979 - 6710

7. DR. M. Mayilvaganan, R. Deepa, P. Nandakumar, "A study on data mining and statistical methods used in diabetes mellitus diagnosis", International Journal of Advanced Research (2016), Volume 4, Issue 7, 447-452

8. Minyechil Alehegn, Rahul Joshi "A review of ensemble machine learning approach in prediction of diabetes diseases", International journal of university for datanalaysis & Communication Engineering Volume: 4 Issue: 3 ISSN: 2454-4248 463 – 466, 2018

9. Misba Reyaz, Gagan Dhawan, "Various Data Mining (IJTSRD) International university of japan ISSN No: 2456 - 7989 | | Volume - 2 | Issue – 4, 2014

10. Velmurugan , S. Jeyalatha and Ronak Sumbaly, "Diagnosis of diabetes using classification mining techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP), Vol.5,No.1,January 2015.