INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 8, Issue 3 - V8I3-1251)
Available online at: https://www.ijariit.com

Prognosticate Diabetic Mellitus in Women by using Performance Evaluation and Classification Algorithms

Patnayakuni Pragathi
ppatnaya@gitam.in
Gandhi Institute of Technology and Management, Visakhapatnam, Andhra Pradesh

Komali Yasudha
ykomali@gitam.edu
Gandhi Institute of Technology and Management, Visakhapatnam, Andhra Pradesh

Maddila Suresh Kumar
smaddila@gitam.edu
Gandhi Institute of Technology and Management, Visakhapatnam, Andhra Pradesh

ABSTRACT

Diabetes is a persistent disorder that takes place while the pancreas does now no longer produces sufficient insulin or while the frame can't use the insulin it produces Diabetes is known as one of the deadliest and most chronic diseases that cause blood sugar levels to rise. Many headaches arise if diabetes stays untreated and unidentified. Early prediction of diabetes can save life. In our undertaking, prediction of diabetes for women in the age between 30 and 80 through the use of classification algorithms. We used diverse Machine Learning classification algorithms like Logistic Regression, Decision Tree and Random Forest on various attributes like Glucose, Blood Pressure, Skin thickness, Insulin, BMI, Diabetes pedigree function, Age, Pregnancies and discover goal variable i.e., outcome. Finally, different classification algorithms along with their comparison of performances with the use of Confusion Matrix, Accuracy, F-Measure, and Recall.

Keywords: *Logistic Regression, Decision Tree, Random Forest*

1. INTRODUCTION

Diabetes may be a condition where the body doesn't properly process food to be used as energy. Most of the food we eat is changed into glucose, or sugar, for our bodies to use for energy. The pancreas, an organ that lies near the stomach, makes a hormone called insulin to assist glucose get into the cells of our bodies. But, sometimes your body does not make enough or any insulin or does not use insulin well. Glucose will remain in your blood and not reach your cells.. Having an uncontrolled amount of glucose in your blood can cause health problems. For someone with diabetes, there's less or no insulin within the body. Thus, blood sugar isn't transferred to the cells and remains within the blood. A number of it is eliminated by the kidneys as urine.

India has an estimated 77 million diabetics, making it the world's second-largest diabetic population behind China. One in every six diabetics in the planet (17%) belongs in India. This creates the urge to detect diabetes at an early stage to prevent it from worsening. It is also important to learn the factors that are mainly causing diabetes. It's also crucial to determine which model performs best in order to achieve better results. In this paper Logistic Regression, Random Forest, Decision Trees were compared in terms of performance.

Different machine learning techniques can be used on various data. Predictive analysis in the healthcare industry is the subject of this study. For analysis, machine learning algorithms are applied to healthcare data sets. In this investigation, the focus is on gestational diabetes. On the Pima Indian Diabetes Database (PIDD) data set, Decision Tree, logistic regression, and random forest ML approaches are used to study diabetes prediction .Several indications, such as glucose, blood pressure, and BMI, are measured during the test.

2. FRAMEWORK AND METHODOLOGY

2.1 DATASET

The Pima Indian Diabetes Database is a well-known and widely utilised data set for diabetes prediction. There are 768 rows and 9 columns in this data collection. Glucose, pregnancies, skin thickness, blood pressure, BMI, insulin, age, and outcomes are all mentioned in the column. The patient's diabetes positive or diabetic-negative status is predicted by the outcome variable. The CSV file is read using the Pandas function, and the data set file is in excel format.

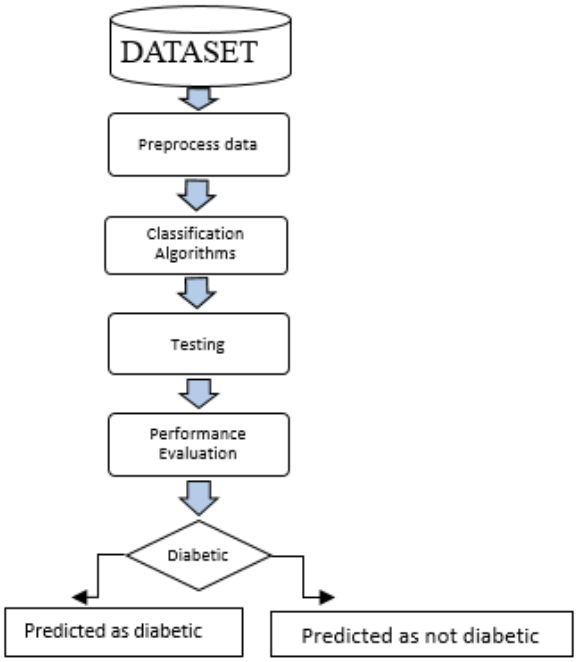

Fig. 1: Methodology

Table 1: Pima Indian Dataset Attributes

S.NO	ATTRIBUTE
1.	Pregnancy
1.	Glucose
2.	Blood Pressure
3.	Skin Thickness
4.	Insulin
5.	BMI
6.	Diabetes Pedigree function
7.	Age
8.	Outcome

2.2 Pre-Process Data

The data that was processed was utilized to create a model. Before applying classifiers to the data index, the data should be pre-processed and appropriately organized. Before linking, certain data should be handled with care.

Data that is inconsistent is analyzed and deleted in this phase, resulting in more exact and dependable results. This data collection contains missing values. Because some properties, such as blood pressure, skin thickness, glucose level, and BMI, cannot have null values, they are allocated with missing values. The data set was then scaled to normalize all values.

2.3 Classification Algorithms

The Scikit-learn Python Toolkit is used to apply ML classifiers after the data has been pre-processed. The data was analyzed using Numpy, Pandas, Scikit, and Matplotlib Scikit is a simple toolkit for data processing and analysis. These tool sets are used for the majority of the work. The data set is divided into training and testing data sets first, using a function like the model selection train test split. Approximately 80% of the data set is used for training, while the remaining 20% is used for testing by selecting data at random.

The diagnosis of diabetes is then carried out using various classifiers, such as machine learning algorithms. Because of their simplicity and popularity, ML classifiers are widely used. The classification algorithms we utilized in this research are as follows:

1. Logistic Regression
2. Decision Tree
3. Random Forest

2.4 Performance Evaluation

Execution measurements such as confusion matrix, precision and test score will be used to evaluate the performance of the logistic regression classifier. After that, the obtained result is compared to the relevant job in order to do result analysis.

3. RESULTS

Several measures were made in this project. The proposed method employs a variety of classification algorithms written in Python. These are common Machine Learning strategies for extracting the most accuracy from data. We can observe that the random forest classifier excels the others in this study. Overall, we employed the greatest Machine Learning approaches to

forecast performance and obtain excellent accuracy. The results of these Machine Learning approaches are depicted in the diagram.

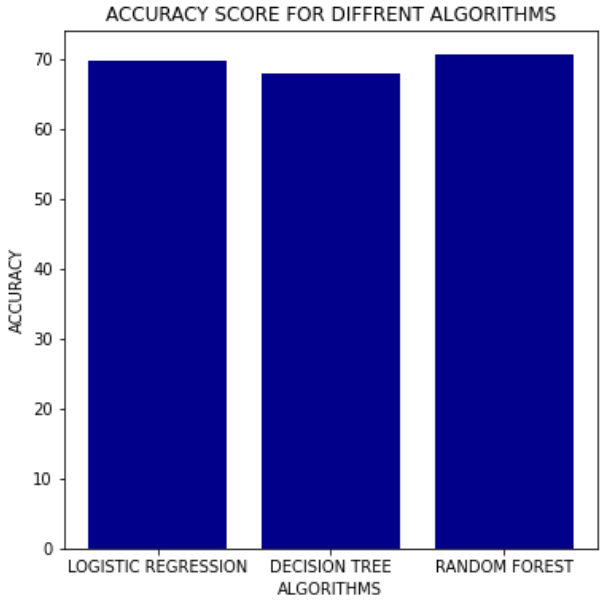

Fig. 2: Accuracy Score

We can see that the True Positives and True Negatives are high in Random Forest compared to Logistic Regression and Decision Tree from the following figures (Fig 3, Fig 4, Fig 5). So with the help of confusion Matrix we can say that Random Forest give accurate result than Logistic Regression and Decision Tree.

TP – True Positives is the number of correct predictions that an instance is positive.

TN – True Negatives is the number of correct predictions that an instance is negative.

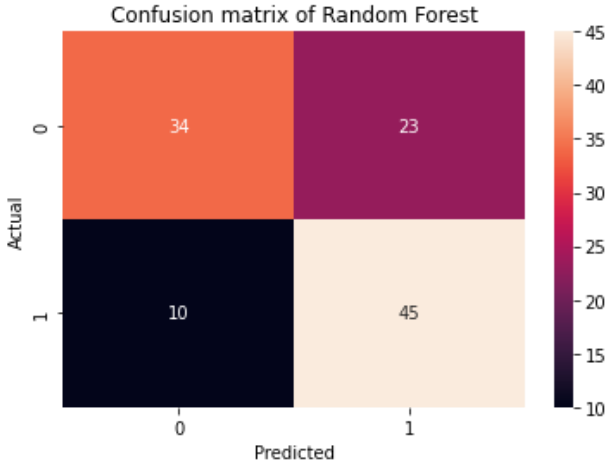

Fig. 3: Confusion Matrix for Random Forest

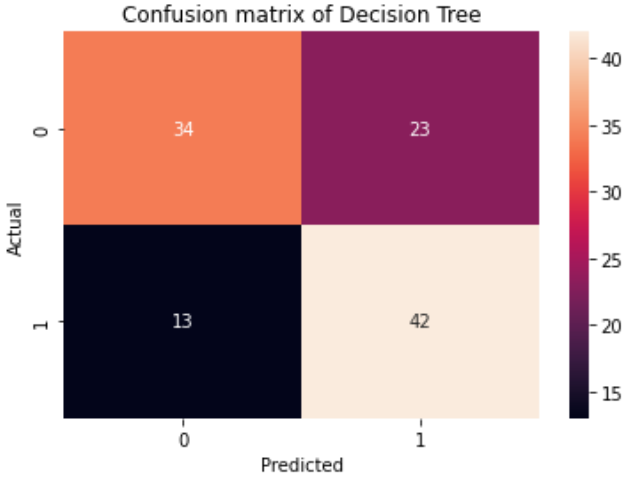

Fig. 4: Confusion Matrix for Random Forest

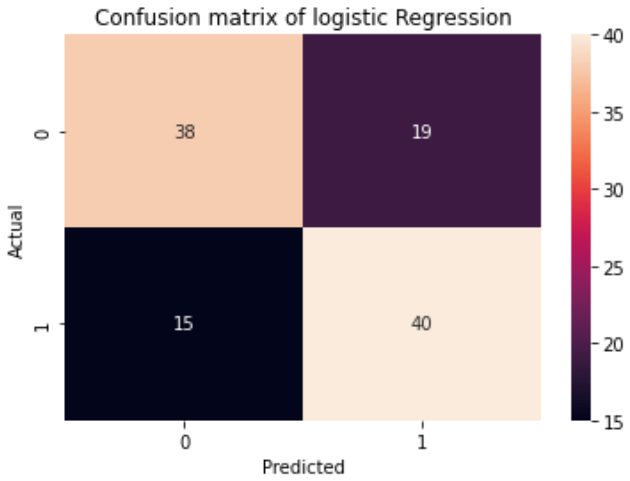

Fig. 5: Confusion Matrix for Logistic Regression

4. DIFFERENCES IN ACCURACY

Table 2: Accuracy Differences

Algorithms	Training Accuracy	Test Accuracy
Logistic Regression	72.9729	69.6428
Decision Tree	100	67.8571
Random Forest	100	70.5357

5. CONCLUSION

The early identification of diabetes is one of the most pressing real-world medical issues. In this work, systematic attempts have undertaken the development of a system that predicts diabetes. Three machine learning classification algorithms are investigated and assessed on several measures as part of this project. On the Diabetes Database, experiments are carried out. Using the Random Forest algorithm, the experimental results determine the suitability of the constructed system, with an accuracy of 70%. The developed approach, together with the machine learning classification techniques used, could be used to predict or detect other diseases in the future. The work can be expanded and enhanced for diabetes analysis automation, as well as some other machine learning techniques.

6. REFERENCES

[1] Diabetes, World Health Organization (WHO): 30 Oct 2018 https://www.who.int/health-topics/diabetes#tab=tab_1
[2] Kaggle Dataset https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
[3] Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineering Research and Application, Vol. 8, Issue 1, (Part -II) January 2018, pp.-09-13
[4] A.K., Dewangan, and P., Agrawal, "Classification of Diabetes Mellitus Using Machine Learning Techniques," International Journal of Engineering and Applied Sciences, vol. 2, 2015.
[5] Orabi, K.M., Kamal, Y.M., Rabah, T.M., 2016. Early Predictive System for Diabetes Mellitus Disease, in: Industrial Conference on Data Mining, Springer. Springer. pp. 420–427.
[6]Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I., 2017. Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal 15, 104–116. doi:10.1016/j.csbj.2016.12.005.
[7]Orabi, K.M., Kamal, Y.M., Rabah, T.M., 2016. Early Predictive System for Diabetes Mellitus Disease, in: Industrial Conference on Data Mining, Springer. Springer. pp. 420–427
[8] American Diabetes Association website [cited 2020 Dec 18, [Online]. Available from: http://www.diabetes.org/diabetes-basics/symptoms/.
[9] Ramachandran A. Know the signs and symptoms of diabetes. Indian J Med Res. 2014; 140: 579-81.
[10] Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2018 American Diabetes Association Diabetes Care 2018; 41(Supplement 1): S13–S27. https://doi.org/10.2337/dc18-S002.
[11]Centres for Disease Control and Prevention. National Diabetes Statistics Report. Atlanta: Centers for Disease Control and Prevention, US Department of Health and Human Services; 2017.
[12]World Health Organization. Diagnostic criteria and classification of hyperglycaemia first detected in pregnancy. World Health Organization. 2013. http://www.who.int/iris/handle/10665/85975.

BIOGRAPHY

Patnayakuni Pragathi
PG Student
Final Year PG student from GITAM (Deemed to be University), Visakhapatnam. This work is a part of my project in my final year .My area of interest is Machine Learning.

Komali Yasudha
Assistant Professor
Working as Assistant Professor in the Dept. of Computer Science, GITAM (Deemed to be University). Areas of research are Deep learning and Machine Learning.

Maddila Suresh Kumar
Assistant Professor
Working as Assistant Professor in the Dept. of Computer Science, GITAM (Deemed to be University). Areas of research Cloud Security.

