IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Special Issue 2, December 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.569

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569| || Volume 10, Special Issue 2, December 2021 ||

4th International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM '2021]

15th & 16th September 2021

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

Performance Analysis of Machine Learning Algorithm for Diabetic Patients

Dr. M Natarajan[#1], Dr. A. Muruganandam[#2] and Dr. N.Thinaharan[#3]

Assistant Professor, Department of Computer Science, Thanthai Hans Roever College (Autonomous), Perambalur, Tamilnadu, India

Principal, Department of Computer Science, Sri Aravindar Arts and Science College, Vanur, Villupuram, TamilNadu, India.

Assistant Professor, Department of Computer Science, Thanthai Hans Roever College (Autonomous), Perambalur, TamilNadu, India

ABSTRACT : Diabetes is a blood associated metabolic disorder which regulates the blood glucose level in our body. Most of the people are affected by the diabetes in the world. Classification techniques will be an efficient method for classifying data of diabetes to make easy decision making. Machine learning is an emergent technology which enables computers to learn without human intervention from previous data. Machine learning uses different algorithms for classification. The Classification algorithm is a Supervised Learning Method that is used to classify the group of new observations on the basis of training data. The main goal of a classification problem is to categorize the class to which a new data will fall under. Naive Bayes and Random Forest algorithms are major efficient technique in the field of classification. In this paper essentially analyze the performance of Naïve Bayes and Random Forest classification algorithm based on accuracy and speed.

KEYWORDS: Diabetes, Classification Algorithms, Naive Bayes, Random Forest, Accuracy and Speed.

I. INTRODUCTION

Diabetes is a blood associated metabolic disorder which occurs due to the lack of insulin in our body. Pancreas secretes and regulates the level of insulin in our body so that the blood glucose level maintains at the normal level. Blood sugar can be regulated by the hormone insulin and if it is not secreted by pancreas, the blood glucose level can be raised as hyperglycemia. It leads to the cellular damage and metabolic associated problems [1].

Machine Learning Classification algorithm has become significant tool in the field of physical condition. Because they can process and examine huge data to mine useful information in decision making. In this paper we examined the accuracy and speed of two machine learning algorithms in the classification of the diabetes. Naive Bayes and Random Forest classification algorithms are evaluated on diabetes dataset using the WEKA tool.

Rest of the paper is organized as follows: Section two describes the concept of diabetes. Section three present the related research to this paper. Section four gives theoretical explanation of machine learning algorithms and its evaluation. Section Five and six analyze the result and conclusion.

II. DIABETES

Diabetes happens when your body is not capable to take up sugar into its cells and use it for energy. This results in a buildup of extra sugar in your bloodstream.

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569| || Volume 10, Special Issue 2, December 2021 ||

4th International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM '2021]

15th & 16th September 2021

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

2.1 Types of Diabetes

Different types of diabetes listed below.

2.1.1 Type I Diabetes

This type of diabetes mellitus considered as Non Insulin Dependent Diabetes Mellitus (NIDDM). Insulin producing pancreatic cells is destroyed so that Insulin cannot be secreted by the pancreas but it is not functioning properly and it is considered as Juvenile Diabetes [2].

2.1.2. Type II Diabetes

According to this type, diabetes mellitus considered as Insulin Dependent Diabetes Mellitus (IDDM). This insulin resistant diabetes present in our body which these cells does not respond for the normal level of insulin. Most of the people affected by type II diabetes mellitus [3].

2.1.3. Prediabetes

Prediabetic stage can be diagnosed before the type II diabetes which the blood glucose raised above the normal level.

2.1.4. Gestational Diabetes

This type of diabetes develops in women during the pregnancy. If the gestational diabetes is not regulated it leads to type II diabetes in future.

2.2. Factors affecting Diabetes

Due to obesity and hereditary factors, the blood glucose level enhanced in our body and it leads to diabetes. The following factors which affect the diabetes mellitus are,

- Obesity
- Heredity
- Life Style
- Gestational diabetes
- Age factor
- Blood Pressure
- Food habits

2.3 Symptoms of Diabetes

The high blood sugar causes several of the warning signs of diabetes listed as below:

- Increased thirst and Dry mouth
- Weak and tired feeling
- Blurred vision
- Numbness
- Slow-healing sores or cuts
- Unplanned weight loss
- Frequent urination
- Frequent unexplained infections

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569| || Volume 10, Special Issue 2, December 2021 ||

4th International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM '2021]

15th & 16th September 2021

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

2.4 Complications of Diabetes

The blood glucose level high over a long period of time, human body tissues and organs can be seriously damaged. These types of complications can be arrived,

- Cardiovascular problem
- Kidney damage (nephropathy)
- Eye damage (retinopathy)
- Foot and nerve damage
- Skin infections
- Erectile dysfunction
- Depression
- Dental problems

III. RELATED WORKS

Many researchers have been carried out in the medical field; they used machine learning algorithms to classify the different types of diabetes.

Robert A. Sowah et al *[4]* to describe Intelligent Meal Recommender Module. This module schedules a diet for diabetes management by generating entire meals for breakfast, lunch, and supper to meet the dietary requirements of the diabetic patient. Here implementing K-Nearest Neighbour (KNN) algorithm to advised meals to diabetic person. The result gives that the classification with 95 % accuracy.

S.Saru et al [5] to predict diabetes and enhance the accuracy using naïve bayes (NB), Decision Tree and KNN and compare their performance. Here, they applied 10 cross of the machine learning classification methods. It provides high accuracy with value of 90.36%.

Ioannis Kavakiotis et al[6] to carry out a systematic review of the applications of machine learning, data mining techniques and tools in the field of diabetes research with respect to Prediction and Diagnosis, Diabetic Complications, Genetic Background and Environment. Support vector machines (SVM) are a successful algorithm and widely used in healthcare field.

Ibrahim Mahmed et al[7] to discussed systems for both diagnosing and treatment but mostly for diabetic type 1. The 90% of diabetic patients have type 2, only diagnosing systems were reported for this type 2. AI and Expert System approaches are implemented for both type-1 and type-2 diabetes.

Rebaz A H Karim et al[8], to calculate the short (two weeks) training sample of a continuous glucose monitor and dietary/insulin log is sufficient to provide accurate predictions. Artificial Neural Network (ANN) approach is used to predict blood glucose level between meals.

IV. MACHINE LEARNING

Most of the real world problems are very complex; we need to perform few predictions. So instead of writing a code, just feed the data to the classification algorithms and with the help of these algorithms, machine builds the logic as per the data and predicts the output. Machine learning has changed the way of thinking about the problem. The block diagram explains the working of Machine Learning algorithm.

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569| || Volume 10, Special Issue 2, December 2021 ||

4th International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM '2021]

15th & 16th September 2021

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

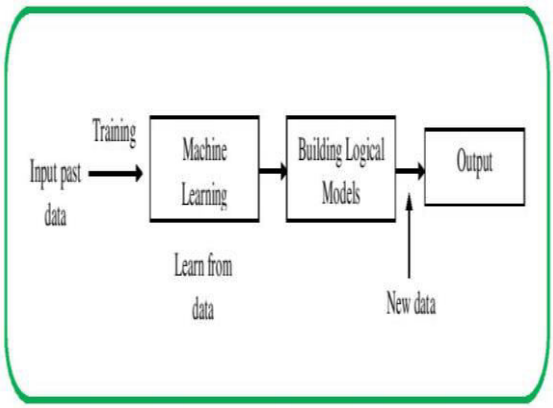


Figure 1: Block diagram of Machine Learning Model

4.1 Proposed Methodology using Machine Learning Algorithm

The proposed Machine learning model takes patient sample (Diabetes) as input and it passes through Naive Bayes/Random Forest classification algorithms which will find the broad category of diabetes then it decides which type of diabetes. Finally, the Machine Learning model generates the output as the class of diabetes.

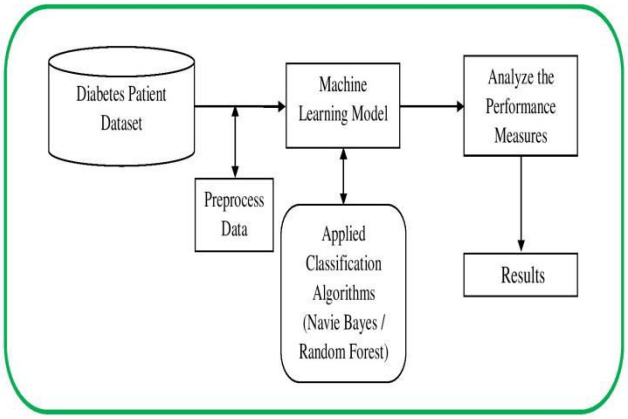


Figure 2: Proposed Machine Learning Model

4.2 Naive Bayes Classifier

Naive Bayes is a classification method with an idea which defines all features is autonomous and distinct to each other. It defines that status of a explicit feature in a class does not change the status of another feature. Since it is based on conditional probability it is measured as a powerful algorithm working for classification reason. It works well for the data with unbalancing problems and missing values. Naive Bayes [9] is a machine learning classifier which workings the Bayes Theorem. Bayes theorem is a mathematical formula used to determine the conditional probability of events. Here, posterior probability P(H|E) can be measured from P(H), P(E) and P(E|H).

Therefore, P(H|E) = (P(E|H) P(H)) / P(E) Where,

- P(H|E) - is the posterior probability of class (target) given predictor (attribute).

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569| || Volume 10, Special Issue 2, December 2021 ||

4th International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM '2021]

15th & 16th September 2021

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

- P(E|H) – is the likelihood which is the probability of predictor given class.
- P(H) – is referred to as the prior probability.
- P(E) – This is the probability of the occurrence of evidence regardless of the hypothesis.

This algorithm used to classify whether patient will be affected by diabetes or not based on blood glucose level. Let's track the following steps to carry out it.

4.2.1 NBC Algorithm

Step 1: Convert the data set into a frequency table

Step 2: Create Likelihood table by finding the probabilities

Step 3: To calculate the posterior probability for each class using Bayesian equation.
Step 4: Class has the highest posterior probability is the outcome of prediction.

4.3 Random Forest Classifier

Random forest is a group classifier which consists of many decision tree and gives class as outputs i.e., the class's output by individual trees. Random forest is given several numbers of classification trees without pruning. Each classification tree is presented a particular number of votes for each class. Among all the trees, the algorithms choose the classification with the most number of votes. Random forest runs powerfully on large datasets but is comparatively slower than naïve bayes algorithms [10]. It can efficiently estimate missing values and hence it is suitable for managing datasets with large number of missing values. The algorithm used to categorize whether patient will be affected by diabetes or not based on blood glucose level. The following steps to carry out it.

4.3.2 RFC Algorithm

Step 1: The 'N' numbers of random records are taken from the data set having 'K' number of records.

Step 2: Separate decision trees are constructed for each sample.

Step 3: Each decision tree will produce an output.

Step 4: Final output is measured based on bulk selection or average for Classification and regression correspondingly.

V. EXPERIMENTAL RESULTS

5.1 Dataset: In this work WEKA tool [11], is used for performing the experiment. WEKA is software which is designed in the country New Zealand by University of Waikato, which includes a set of different machine learning methods for data classification, clustering, regression and visualization. One of the main advantages of using WEKA is that it can be adapted according to the requirements. The main objective of this study is the prediction of the patient affected by diabetes using the WEKA tool by using the medical database from kaggle with dataset of diabetes.csv file.

No. of Samples	Classification Algorithms	Correctly Classified	Incorrectly Classified
768	Naïve Bayes	586	182
	Random Forest	692	76

Table 1: Classified Sample values

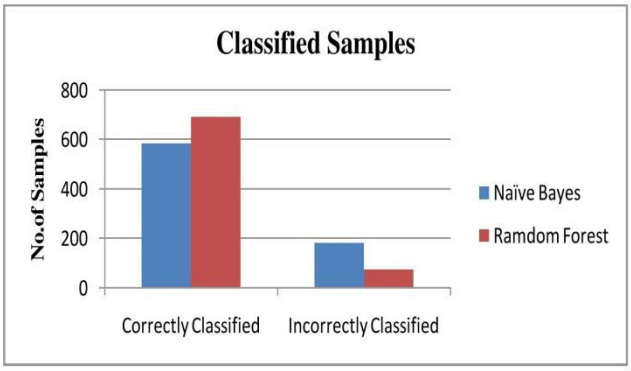
|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569| || Volume 10, Special Issue 2, December 2021 ||

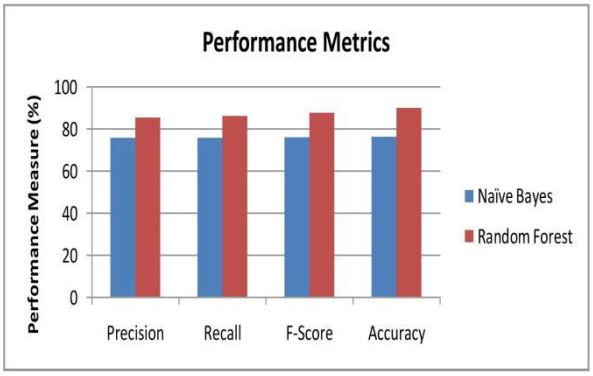
4th International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM '2021]

15th & 16th September 2021

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India





5.2 Accuracy Measures:

Naive Bayes and Random Forest algorithms are used in this research work. To calculate performance of above algorithms using Confusion Matrix[12]. The following parameters are used to measure the metrics such as Precision, Recall, F-Score and Accuracy.

True Positive (TP) is correctly predicting a label, i.e. predicted as 'yes' and the actual is 'yes'

True Negative (TN) correctly predicting the other label, i.e. predicted as 'no' and the label is 'no'

False Positive (FP) falsely predicting a label, i.e. predicted as 'yes', but the actual is 'no'

False Negative (FN) is missing an incoming label, i.e. predicted as 'no', but the actual is 'yes'

To evaluate the more effectiveness of the above algorithms. Which is calculated using the Formulas? The evaluation results are shown in the table 2.

Precision = TP / (TP +FP)

Recall = TP / (TP + FN)

F-Score = 2TP / (2TP + FP + FN)

Accuracy=(TP + TN) / (Total No. of samples)

Measure	Naive Bayes	Random Forest
Precision	75.9	85.5
Recall	76	86.2
F-Score	76.1	87.9
Accuracy	76.3	90.1

Table 2: Performance Metrics and Measure

Figure 4: Performance Result

V. RESULTS

Table 1 represents the classifier performance on the source of classified samples. Performance of individual algorithm is evaluated on the basis of Correctly Classified samples and Incorrectly Classified samples out of a total number of samples. Figure 3 represents the graphical performance of both classification algorithms on the basis of classified samples.

Table 2 shows various performance values of all classification algorithms measured on various metrics. From Table-1, it is analyzed that Random Forest showing the maximum accuracy. So the Random Forest machine learning classifier can predict the chances of diabetes with more accuracy as compared to Naïve Bayes. Performances of both

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569| || Volume 10, Special Issue 2, December 2021 ||

4th International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM '2021]

15th & 16th September 2021

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

classifiers based on various measures are shown via a chart in Figure 4. From Table 1 and Table 2 we can conclude that Random Forest classification algorithm outperforms compare to Naive Bayes algorithm.

VI. CONCLUSION

Diabetic is one of the significant real-world medical problems. In this research, two machine learning classification algorithms are experimented and evaluated on different measures. Experiments are performed on the medical database from kaggle with dataset of diabetes.csv file. Random Forest algorithm is considered as the best resulted machine learning method of this experiment. Random Forest gives 14% improved accuracy when compared to the Naïve Bayes classification algorithm. Naïve Bayes classification algorithm speed is higher when compared to the Random Forest classification, because Random Forest classification uses more subsets for the experiment.

REFERENCES

1. R.Gojka, Diabetes, World Health Org. https://www.who.int/healthtopics/diabetes#tab=overview. 2019.
2. G. Fico, A. Fioravanti, M. Arredondo et al., "Integration of personalized healthcare pathways in an ICT platform for diabetes management: a small-scale exploratory study, IEEE Journal of Biomedical and Health Informatics, vol. 20, no. 1, pp. 29–38, 2016.
3. M. Alotaibi, "A mobile diabetes educational system for fasting type-2 diabetics in Saudi Arabia," 2nd International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE), pp. 173–176, Semarang, Indonesia, October 2015.
4. Robert A. Sowah et al, "Design and Development of Diabetes Management System Using Machine Learning", Research Article, 2020.
5. S. Saru et al, "ANALYSIS AND PREDICTION OF DIABETES USING MACHINE LEARNING", International Journal of Emerging Technology and Innovative Engineering Volume 5, Issue 4, April 2019.
6. Ioannis Kavakiotis et al, "Machine Learning and Data Mining Methods in Diabetes Research", Computational and Structural Biotechnology Journal Volume 15, Pages 104-116, 2017.
7. Ibrahim Mahmed et al, "A study on Expert Systems for Diabetic Diagnosis and Treatment" Recent Advances in Information Science, ISBN: 978-960-474-304-9,
8. Rebaz A H Karim et al, "Improved Methods for Mid-Term Blood Glucose Level Prediction Using Dietary and Insulin Logs" National Centre for Biotechnology Information, 2021.
9. Rish et al, "An empirical study of the naïve Bayes classifier", Research Article, IBM.pp.41–46, 2001.
10. Natarajan M et al, "Mixed Model of Extreme Learn Machine Tree and Random Forest Classifier for Prediction of Oral Cancer", International Journal of Innovative and Creative Engineering, Volume 11, February 2021.
11. Arora, R, Suman et al, " Comparative Analysis of Classification Algorithms on Different Datasets using WEKA", International Journal of Computer Applications, Volume 54, Page 21–25, 2012.
12. Deepti S and Dilip SS, "Prediction of Diabetes using Classification Algorithms", International Conference on Computational Intelligent and Data Science", 2018.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.569

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details