IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DIABETES PREDICTION USING MACHINE LEARNING

[1]Gaanavi H N, [2]Madanika G, [3]Prof. SumaRani H, [4]Dr. Buddesab

[1,2]Student, [3]Assistant Professor, [4]Associate Professor

[1,2,3,4] Department of Artificial Intelligence and Machine Learning

Cambridge Institute of Technology
K R Puram, Bangalore

Abstract: In this paper we aim to develop an prediction system using machine learning to detect and classify the presence of diabetes in e-healthcare environment using Ensemble Decision Tree Algorithms for high feature selection. A significant attention has been made to the accurate detection of diabetes which is a big challenge for the research community to develop a diagnosis system to detect diabetes in a successful way in the e-healthcare environment. In this paper we aim to develop an prediction system using machine learning to detect and classify the presence of diabetes in e-healthcare environment using Ensemble Decision Tree Algorithms for high feature selection. A significant attention has been made to the accurate detection of diabetes which is a big challenge for the research community to develop a diagnosis system to detect diabetes in a successful way in the e-healthcare environment. The existing diagnosis systems have some drawbacks, such as high computation time, and low prediction accuracy. To handle these issues, we have proposed diagnosis system using machine learning methods, such as preprocessing of data, feature selection, and classification for the detection of diabetes disease in e- healthcare environment. Model validation and performance evaluation metrics have been used to check the validity of the proposed system. We have proposed a filter method based on the Decision Tree algorithm for highly important feature selection. Two ensemble learning Decision Tree algorithms, such as Ada Boost and Random Forest are also used for feature selection and compared the classifier performance with wrapper based feature selection algorithms also. Machine learning classifier Decision Tree has been used for the classification of healthy and diabetic subjects. The experimental results show that the Decision Tree algorithm based on selected features improves the classification performance of the predictive model and achieved optimal accuracy. Additionally, the proposed system performance is high as compared to the previous state-of-the-art methods. High performance of the proposed method is due to the different combinations of selected features set. Furthermore, the experimental results statistical analysis demonstrated that the proposed method would be effectively detected diabetes disease.

Index Terms **- Ensemble learning Decision Tree algorithms, such as Ada Boost and Random Forest**

I. INTRODUCTION

In recent years, the integration of machine learning (ML) techniques into healthcare has opened up new avenues for predictive analysis and personalized medicine. One of the most pressing challenges in healthcare today is the early detection and management of chronic diseases such as diabetes. With its debilitating effects on individuals' health and the healthcare system at large, diabetes demands innovative solutions for early diagnosis and intervention. Various techniques of Machine Learning can capable to do prediction, however it’s tough to choose best technique. Thus for this purpose we apply popular classification and ensemble methods on dataset for prediction. Diabetes prediction using machine learning

involves leveraging algorithms to analyze data and identify patterns that can indicate the likelihood of an individual developing diabetes. By utilizing features such as age, weight, diet, exercise habits, and genetic predispositions, machine learning models can help predict the risk of diabetes onset, enabling early intervention and personalized healthcare strategies.

II. LITERATURE REVIEW

N. H. Barakat, et al[1] methods have been used for the diagnosis, prognosis, and management of diabetes. In this paper, we propose utilizing support vector machines (SVMs) for the diagnosis of diabetes. In particular, we use an additional explanation module, which turns the "black box" model of an SVM into an intelligible representation of the SVM's diagnostic (classification) decision. Results on a real-life diabetes dataset show that intelligible SVMs provide a promising tool for the prediction of diabetes, where a comprehensible ruleset have been generated, with prediction accuracy of 94%, sensitivity of 93%, and specificity of 94%. Furthermore, the extracted rules are medically sound and agree with the outcome of relevant medical studies.

A. D. Association[2] The basis of the abnormalities in carbohydrate, fat, and protein metabolism in diabetes is deficient action of insulin on target tissues. Deficient insulin action results from inadequate insulin secretion and/or diminished tissue responses to insulin at one or more points in the complex pathways of hormone action. Impairment of insulin secretion and defects in insulin action frequently coexist in the same patient, and it is often unclear which abnormality, if either alone, is the primary cause of the hyperglycemia.

C. D. Mathers and D. Loncar[3] Global and regional projections of mortality and burden of disease by cause for the years 2000, 2010, and 2030 were published by Murray and Lopez in 1996 as part of the Global Burden of Disease project. These projections, which are based on 1990 data, continue to be widely quoted, although they are substantially outdated; in particular, they substantially underestimated the spread of HIV/AIDS. To address the widespread demand for information on likely future trends in global health, and thereby to support international health policy and priority setting, we have prepared new projections of mortality and burden of disease to 2030 starting from World Health Organization estimates of mortality and burden of disease for 2002. This paper describes the methods, assumptions, input data, and results.

K. Kayaer and T. Yıldırım[4]The performance of recently developed neural network structure, general regression neural network(GRNN), is examined on the medical data. Pima Indian Diabetes (PID) data set is chosen to study on that had been examined by more complex neural network structures in the past. The results of early studies and of the GRNN structure presented in this paper is compared. Close classification accuracy to the reference work using ARTMAP-IC structured model, which is the best result obtained since now, is achieved by using GRNN, which has a simpler structure. The performance of the standard multilayer perceptron (MLP) and radial basis function (RBF) feed forward neural networks are also examined for the comparison as they are the most general and commonly used neural network structures. The performance of the MLP was tested for different types of backpropagation training algorithms.

III. METHODOLOGY

In this section we shall learn about the various classifiers used in machine learning to predict diabetes. We shall also explain our proposed methodology to improve the accuracy. Five different methods were used in this paper. The different methods used are defined below. The output is the accuracy metrics of the machine learning models. Then, the model can be used in prediction.

A. Dataset Description

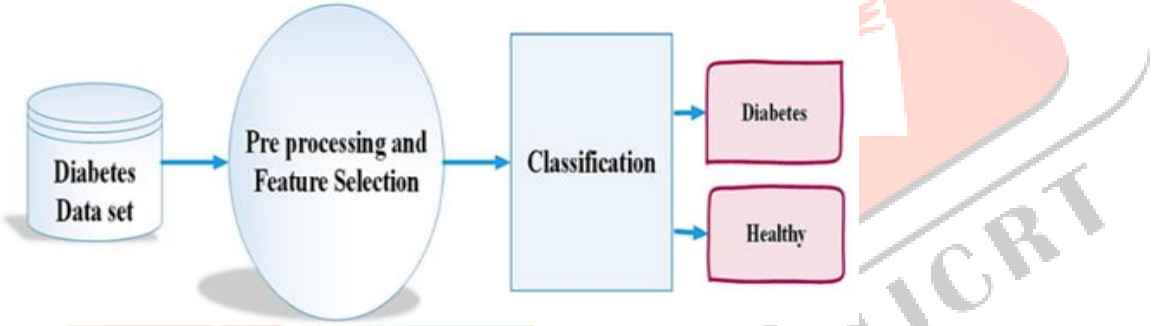
The Diabetes data set was originated from https://www.kaggle.com/datasets/madanikag/diabetes-prediction. Diabetes dataset containing 768 cases. The data is gathered from UCI repository which is named as Pima Indian Diabetes Dataset. The dataset have many attributes of 768 patients. Diabetes dataset containing 2000 cases. The objective is to predict based on the measures to predict if the patient is diabetic or not.

	Id	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	1	6	148	72	35	0	33.6	0.627	50	1
1	2	1	85	66	29	0	26.6	0.351	31	0
2	3	8	183	64	0	0	23.3	0.672	32	1
3	4	1	89	66	23	94	28.1	0.167	21	0
4	5	0	137	40	35	168	43.1	2.288	33	1

→ The diabetes data set consists of 2000 data points, with 9 features each.
→The 9th attribute is class variable of each data points.
→ "Outcome" is the feature we are going to predict, 0 means No diabetes, 1 means diabetes.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               2000 non-null   int64
 1   Glucose                   2000 non-null   int64
 2   BloodPressure             2000 non-null   int64
 3   SkinThickness             2000 non-null   int64
 4   Insulin                   2000 non-null   int64
 5   BMI                       2000 non-null   float64
 6   DiabetesPedigreeFunction  2000 non-null   float64
 7   Age                       2000 non-null   int64
 8   Outcome                   2000 non-null   int64
dtypes: float64(2), int64(7)
memory usage: 140.8 KB
```

→There is no null values in dataset.



Proposed Model Diagram

B. Data Preprocessing

Data pre-processing is most important process. Mostly healthcare related data contains missing vale and other impurities that can cause effectiveness of data. To improve quality and effectiveness obtained after mining process, Data pre-processing is done. To use Machine Learning Techniques on the dataset effectively this process is essential for accurate result and successful prediction. For Pima Indian diabetes dataset we need to perform pre-processing in two steps.

1. **Missing Values removal**

Remove all the instances that have zero (0) as worth. Having zero as worth is not possible. Therefore this instance is eliminated. Through eliminating irrelevant features/instances we make feature subset and this process is called features subset selection, which reduces dimensionality of data and help to work faster.

2. Splitting of data

After cleaning the data, data is normalized in training and testing the model. When data is spitted then we train algorithm on the training data set and keep test data set aside. This training process will produce the training model based on logic and algorithms and values of the feature in training data. Basically aim of normalization is to bring all the attributes under same scale.

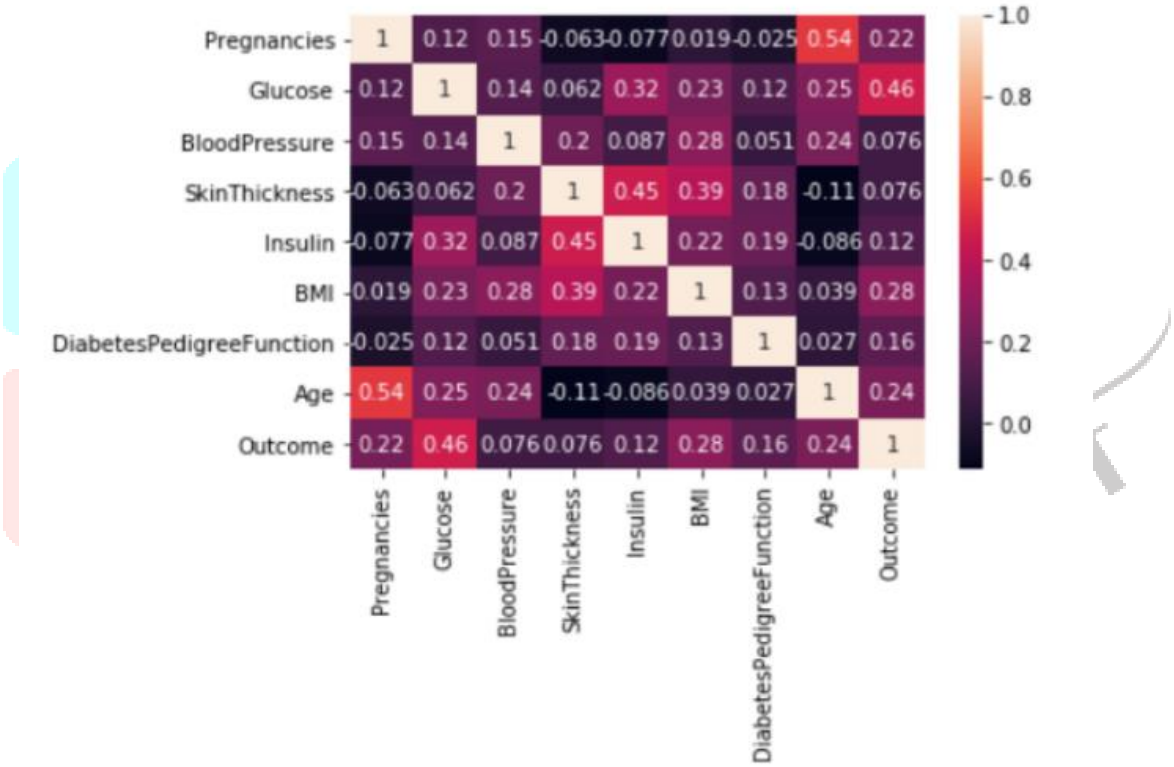
C. Apply Machine Learning

When data has been ready we apply Machine Learning Technique. We use different classification and ensemble techniques, to predict diabetes. The methods applied on Pima Indians diabetes dataset. Main objective to apply Machine Learning Techniques to analyze the performance of these methods and find accuracy of them, and also been able to figure out the responsible/important feature which play a major role in prediction.

IV. CONCLUSION AND DISCUSSION

1) Correlation Matrix

It is easy to see that there is no single feature that has a very high correlation with our outcome value. Some of the features have a negative correlation with the outcome value and some have positive.



2) Ensembling

Ensembling is a machine learning technique Ensemble means using multiple learning algorithms together for some task. It provides better prediction than any other individual model that's why it is used. The main cause of error is noise bias and variance, ensemble methods help to reduce or minimize these errors. There are two popular ensemble methods such as – Bagging, Boosting, ada -boosting, Gradient boosting, voting, averaging etc. Here In these work we have used Bagging (Random forest) and Gradient boosting ensemble methods for predicting diabetes.

3) Precision and recall

precision measures the accuracy of positive predictions, while recall gauges the ability to identify all relevant instances. Achieving high precision ensures accurate positive predictions, while high recall ensures minimal false negatives, crucial for effective early detection and intervention in diabetes.

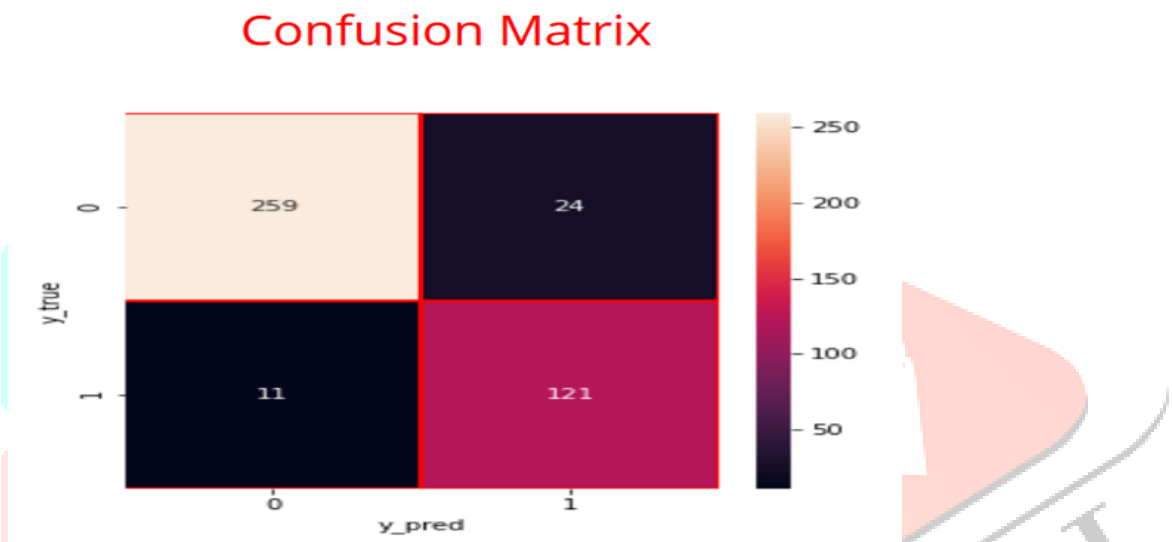
PERFORMANCE ANALYSIS

Precision and recall

	Recall	Precision
Negative(0)	0.92	0.96
Positive(1)	0.92	0.83

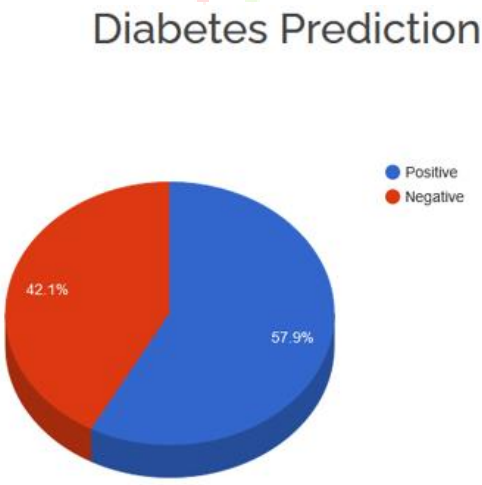
4) Confusion Matrix

A confusion matrix summarizes model performance by comparing actual diabetes status with predictions. It categorizes results into true positives, true negatives, false positives, and false negatives. This aids in assessing model accuracy and identifying areas for improvement in diabetes risk prediction algorithms.



5) Diabetes Prediction

Machine learning facilitates diabetes prediction by analyzing diverse health data to assess an individual's risk of developing diabetes. Through automated algorithms and user-friendly interfaces, it offers personalized risk assessments, integrates with existing systems, and enables early detection, enhancing healthcare efficiency and improving patient outcomes.



V. MODEL BUILDING

This is most important phase which includes model building for prediction of diabetes. In this we have implemented various machine learning algorithms which are discussed above for diabetes prediction.

Procedure of Proposed Methodology

Step1: Import required libraries, Import diabetes dataset.
Step2: Pre-process data to remove missing data.
Step3: Perform percentage split of 80% to divide dataset as Training set and 20% to Test set.
Step4: Select the machine learning algorithm i.e. KNearest Neighbor, Support Vector Machine, Decision Tree, Logistic regression, Random Forest and Gradient boosting algorithm.
Step5: Build the classifier model for the mentioned machine learning algorithm based on training set.
Step6: Test the Classifier model for the mentioned machine learning algorithm based on test set.
Step7: Perform Comparison Evaluation of the experimental performance results obtained for each classifier.
Step8: After analyzing based on various measures conclude the best performing algorithm.

VI. CONCLUSION AND FUTURE WORK

One of the important real-world medical problems is the detection of diabetes at its early stage. In this study, systematic efforts are made in designing a system which results in the prediction of diabetes. During this work, five machine learning classification algorithms are studied and evaluated on various measures. Experiments are performed on john Diabetes Database. Experimental results determine the adequacy of the designed system with an achieved accuracy of 99% using Decision Tree algorithm. In future, the designed system with the used machine learning classification algorithms can be used to predict or diagnose other diseases. The work can be extended and improved for the automation of diabetes analysis including some other machine learning algorithms.

VII. ACKNOWLEDGMENT

We have completed this work under the mentorship of Dr. Buddesab (Project co-ordinator & Associate Professor) & Prof. SumaRani H (Assistant Professor), Department of Artificial Intelligence and Machine Learning at Cambridge Institute of Technology, Bangalore. We would like to express our special thanks to both of my mentors for inspiring us to complete the work & write this paper. Without their active guidance, help, cooperation & encouragement, we would not have our headway in writing this paper. We are extremely thankful for their valuable guidance and support on completion of this paper. We extend our gratitude to “Cambridge Institute of Technology” for giving us this opportunity. We also acknowledge with a deep sense of reverence, our gratitude towards our parents and member of our family, who has always supported us morally as well as economically. Any omission in this brief acknowledgement does not mean lack of gratitude.

REFERENCES

[1] Amin Ul Haq, Jian Ping Li, Jalaluddin khan, Muhammad Hammad Memon, Shah Nazir, Sultan Ahmad, Ghufran Ahmad khan, Amjad Ali, “A New Intelligent Approach for Effective Recognition of Diabetes in the IoT E-HealthCare Environment”, Preprints 2020.
[2] H. E. Massari, Z. Sabouri, S. Mhammedi and N. Gherabi, "Diabetes Prediction Using Machine Learning Algorithms and Ontology," in Journal of ICT Standardization, vol. 10, no. 2, pp. 319-337, 2022, doi: 10.13052/jicts2245-800X.10212.
[3] Z. Ye, J. Wang, H. Hua, X. Zhou and Q. Li, "Precise Detection and Quantitative Prediction of Blood Glucose Level With an Electronic Nose System," in IEEE Sensors Journal, vol. 22, no. 13, pp. 12452-12459, 1 July1, 2022, doi: 10.1109/JSEN.2022.3178996.
[4] S. K. Sharma et al., "A Diabetes Monitoring System and Health-Medical Service Composition Model in Cloud Environment," in IEEE Access, vol. 11, pp. 32804-32819, 2023, doi: 10.1109/ACCESS.2023.3258549.
[5] C.kalaiselvi, G.M.Nasira, “A New Approach of Diagnosis of Diabetes and Prediction of Cancer using ANFIS”, IEEE Computing and Communicating Technologies, pp 188-190, 2014.
[6] E. L. Litinskaia, N. A. Bazaev, K. V. Pozhar and V. M. Grinvald, “Methods for improving accuracy of non-invasive blood glucose detection via optical glucometer”, 2017 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), St. Petersburg, (2017), pp. 47- 49.