IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

DIABETES PREDICTION USING MACHINE LEARNING

S. MANICHARAN, K. VANDANA, N. JAHNAVI, M. BHUVANESHVARI

Final Year – Department of Electronics and Communication Engineering.

Jitty Ramesh, Assistant Professor, Dept. of ECE.

Jyothishmathi Institute of Technology and Science Karimnagar, Telangana.

ABSTRACT

Over Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to the International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Diabetes mellitus or imply diabetes is a disease caused due to the increased level of blood glucose. Various traditional methods, based on physical and chemical tests, are available for diagnosing diabetes However, early prediction of diabetes is quite a challenging task for medical practitioners due to complex interdependence on various factors as diabetes affects human organs such as kidney, eye, heart, nerves, foot etc. Data science methods have the potential to benefit other scientific fields by shedding new light on common questions. One such task is to help make predictions on medical data. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience.

The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques. This project aims to predict diabetes via three different supervised machine learning methods including: SVM, Logistic regression, KNN. This project also aims to propose an effective technique for earlier detection of the diabetes disease using Machine learning algorithms and end to end deployment using flask.

We split the Dataset into: 1) Training set and 2) Testing Set and then perform analysis on them. The Pima Indian dataset was used to study and analyze the data, alongside with data mining techniques. It is the data obtained from the National Institute for Diabetes. Contains several medical predictor variables and one target variable.

1. INTRODUCTION

All around there are numerous ceaseless infections that are boundless in evolved and developing nations. One such sickness is diabetes. Diabetes is a metabolic issue that causes blood sugar by creating a significant measure of insulin in the human body or by producing a little measure of insulin. Diabetes is perhaps the deadliest sickness on the planet. It is not just

a malady yet, but also a maker of different sorts of illnesses like coronary failure, visual deficiency, kidney ailments and nerve harm, and so on.

Subsequently, the identification of such chronic metabolic ailment at a beginning period could help specialists around the globe in forestalling loss of human life. Presently, with the ascent of machine learning, AI, and neural systems, and their application in various domains [1, 2] we may have the option to find an answer for this issue. ML strategies and neural systems help scientists to find new realities from existing well-being-related informational indexes, which may help in ailment supervision and detection. The current work is completed utilizing the Pima Indians Diabetes Database. The point of this framework is to make an ML model, which can anticipate with precision the likelihood or the odds of a patient being diabetic. The ordinary distinguishing process for the location of diabetes is that the patient needs to visit a symptomatic focus.

One of the key issues of bioinformatics examination is to achieve precise outcomes from the information. Human mistakes or various laboratory tests can entangle the procedure of identification of the disease. This model can foresee whether the patient has diabetes or not, aiding specialists to ensure that the patient in need of clinical consideration can get it on schedule and also help anticipate the loss of human lives. Neural networks use neurons to transmit data across various layers, with each node working on a different weighted parameter to help predict diabetes. Presently, with the ascent of machine learning, AI, and neural systems, and their application in various domains [1, 2] we may have the option to find an answer for this issue. ML strategies and neural systems help scientists to find new realities from existing well-being-related informational indexes, which may help in ailment supervision and detection. The current work is completed utilizing the Pima Indians Diabetes Database.

2. LITERATURE SURVEY

K. Vijayakumar proposed the Random Forest algorithm for the Prediction of diabetes to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by using Random Forest algorithm in machine learning technique. The proposed model gives the best results for diabetic prediction and the result showed that the prediction system is capable of predicting the diabetes disease effectively, efficiently and most importantly, instantly.

Nonso Nnameka presented predicting diabetes onset: an ensemble supervised learning approach they used five widely used classifiers are employed for the ensembles and a meta- classifier is used to aggregate their outputs. The results are presented and compared with similar studies that used the same dataset within the literature. It is shown that by using the proposed method, diabetes onset prediction can be done with higher accuracy.

Deeraj Shetty proposed diabetes disease prediction using data mining assemble Intelligent Diabetes Disease Prediction System that gives analysis of diabetes malady utilizing diabetes patient's database. In this system, they propose the use of algorithms like Bayesian and KNN (K-Nearest Neighbour) to apply on diabetes patient's databases and analyse them by taking various attributes of diabetes for prediction of diabetes disease.

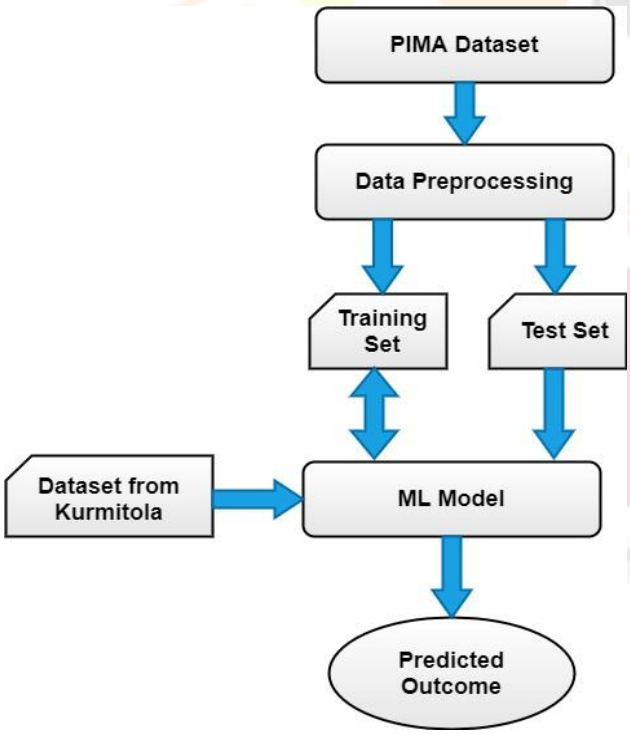
Muhammad Azeem Sarwar proposed a study on prediction of diabetes using machine learning algorithms in healthcare. They applied six different machine learning algorithms. Performance and accuracy of the applied algorithms is discussed and compared. Diabetes Prediction is becoming the area of interest for researchers to

train the program to identify if the patient is diabetic or not by applying a proper classifier on the dataset. Based on previous research work, it has been observed that the classification process is not much improved. Hence a system is required as Diabetes Prediction is an important area in computers, to handle the issues identified based on previous research.

3. PROPOSED SYSTEM

The proposed system employs a multi-step process to predict diabetes in its early stage. Firstly, it preprocesses the raw medical data, handling missing values, outlier detection, and feature normalization to ensure data quality and consistency. Feature selection techniques are then applied to identify the most relevant attributes, reducing the dimensionality of the dataset and enhancing the model's efficiency. To provide a user-friendly interface, the proposed system incorporates a web or mobile application that allows individuals to input their health information and receive a personalized diabetes risk assessment. The application utilizes the trained machine learning models to analyze the user's data, generating a prediction of the likelihood of developing diabetes in the near future. This information empowers individuals to take proactive measures, such as lifestyle modifications, dietary changes, and seeking medical advice, to mitigate their diabetes risk.

Diabetes Prediction in Early Stage using Machine Learning harnesses the potential of advanced algorithms to accurately predict the risk of developing diabetes. By combining cutting-edge technology with user-friendly interfaces, this system aims to improve early detection rates and enhance diabetes management, ultimately leading to better health outcomes for individuals at risk of this chronic condition.



4. IMPLEMENTATION

4.1 METHODOLOGY

Firstly, we pre-process two separate datasets. In the pre-processing stage, correlation between attributes of the datasets is analysed for finding useful features in detecting diabetes. After that, the data is divided into two sets: training and testing. The training set is utilized to develop predictive ML models using a variety of machine

learning algorithms. Next, we assess the proposal's performance with respect to different metrics. Finally, the best ML model is deployed in a web application using flask. Following this, we describe the workflow of each part briefly:

Data Collection : We collected two alternative datasets, each with a different number of factors or features, to ensure the model's robustness. The datasets were compiled from a wide variety of sources, including diabetes statistics and health characteristics obtained from people around the world and from various health institutes.

Data Analysis and Data Pre-processing: Several pre-processing techniques are applied on the datasets before feeding these datasets into the machine learning model so that the performance of the model is improved. The pre- processing tasks include removing outliers and dealing with missing values, data standardization, encoding, and so on.

● Outliers Removal: Attributes values that are beyond acceptable boundaries and have high variation from the rest of the respective attribute's value might be present in the dataset. Such attributes' value might degrade the machine learning algorithm's performance. To eliminate such outliers, we applied the IQR (Interquartile Range) approach.

● Missing value Handling: To improve model performance, the mean value of each attribute was employed for handling the missing values.

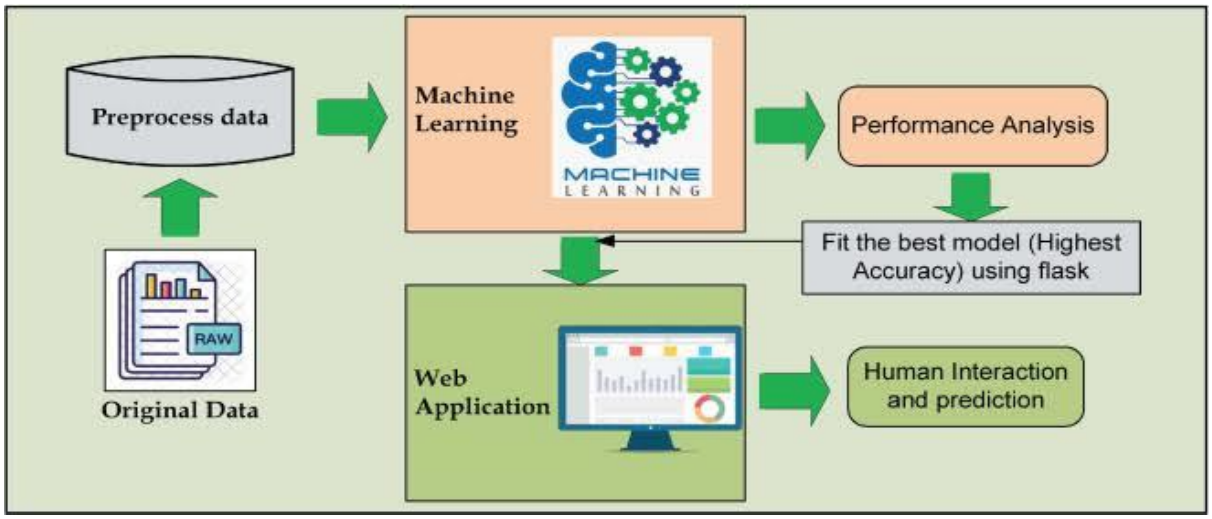
● Label Encoding: Label encoding is the process of converting the labels of text/categorical values into a numerical format that ML algorithms can interpret.

Model Construction and Prediction: To construct the predictive model, 80% of the pre-processed data has been used for training while the remaining 20% data is used for the testing purpose.

Performance Analysis : We have analysed the results of the proposed model in terms of several performance metrics. The algorithm that provides the highest prediction accuracy is selected as the best algorithm for web application development.

Performance Comparison: In this step, the accuracy of the proposal has been compared with some recent works related to diabetes prediction. The performance results indicate that the proposal can improve the performance compared to the recent related research.

Web Application Development: To develop a smart web application, we have used the Flask micro-framework and integrated the best model. To predict diabetes, a user is required to submit a form with necessary numbers of diabetes related parameters. The application uploaded in a server predicts the results using the adopted machine learning model. We describe the adopted machine learning algorithms in the following sections.

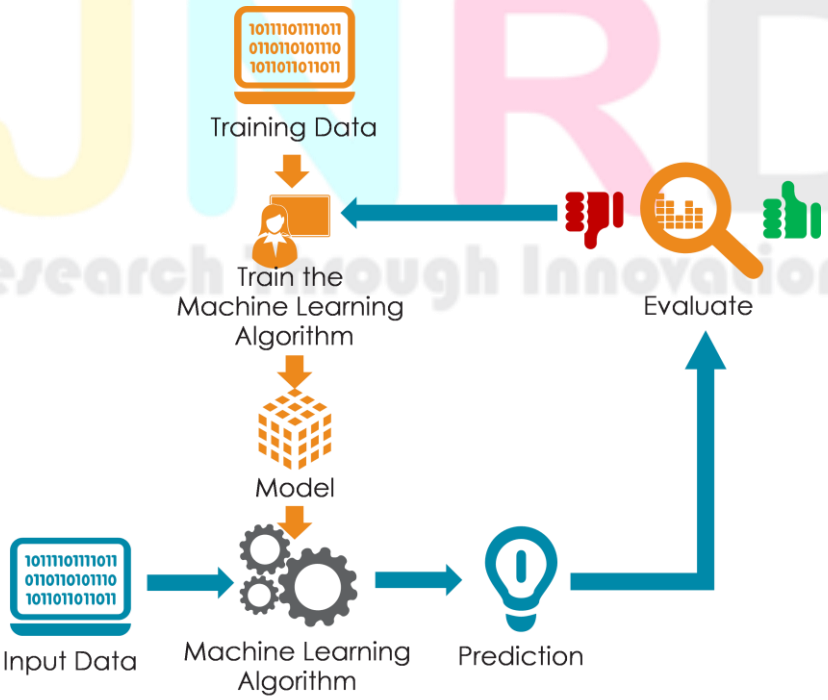


WORKING:

First, we procure the dataset, which is the PIMA Indian dataset. It is a dataset which is used mainly for diabetes prediction. The dataset contains up to 1000 rows and mainly depicts the features required for the prediction of diabetes. We split the dataset into training and testing data where part of the dataset is trained and part of the dataset is used for testing.

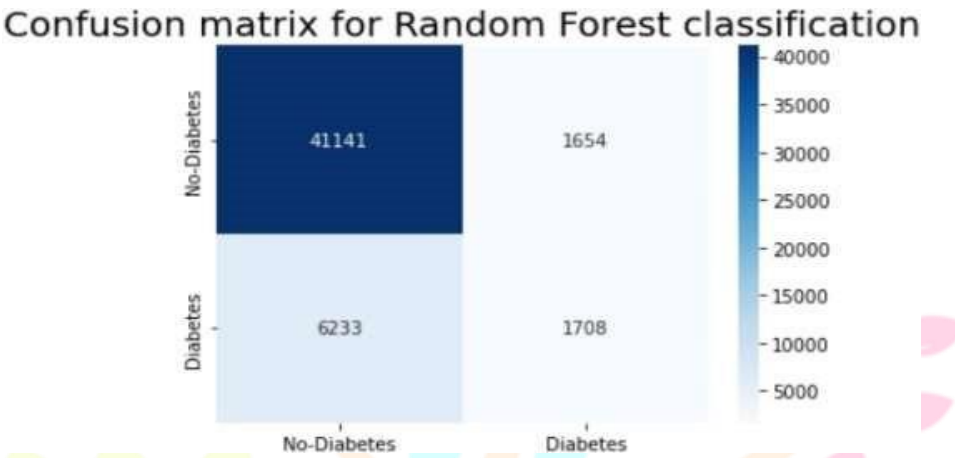
We train the dataset in order to find the accuracy of the percentage of people having and not having diabetes. Many methods are used for the purpose of the prediction of diabetes such as Naïve Bayes, Random Forest, Logistic Regression, Decision Trees etc.

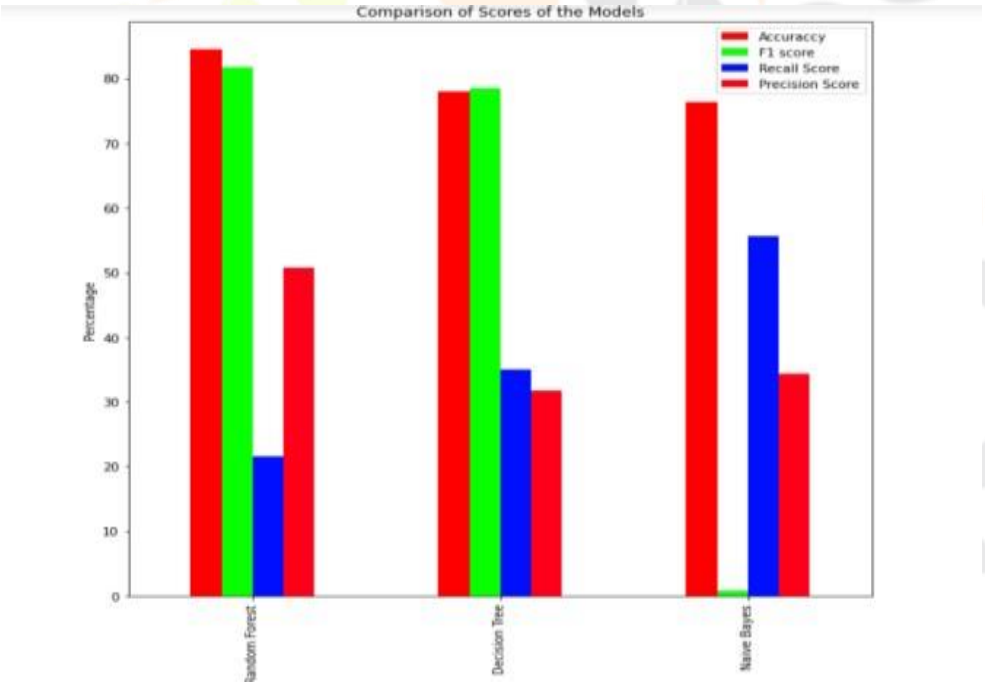
We mainly focus on Naïve Bayes and Random Forest as these two are the most efficient in getting an efficient result for the prediction. We perform Naïve Bayes and Random Forest on the training and testing data and find the accuracy percentage of both the data for finding the best evaluation method among the two for the analysis of the data set.



RESULT

To get the output we used, 70% of data for training and 30% of data for testing. In this ratio of data splitting here we found that Random Forest Classifier predicted with 99% of accuracy the highest accuracy for the dataset.





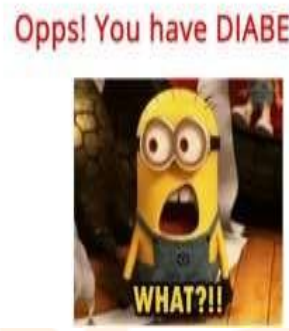
Graphical Representation of 3 algorithms result

ANALYSIS:

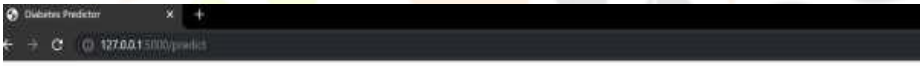
The project predicts the onset of diabetes in a person based on the relevant medical details collected. When the person enters all the relevant medical data required in the online Web portal, this data is then passed on

to the trained model for it to make predictions whether the person is diabetic or non-diabetic the model then makes the prediction with an accuracy of 98%, which is good and reliable.

Following figure shows the basic UI form which requires the user to enter the specific medical data fields. These parameters help determine if the person is prone to develop diabetes our research has the added benefit of an associated Web app, which makes the model more user friendly and easily understandable for a novice.



Prediction for diabetic person



Prediction for non-diabetic person

CONCLUSION

The objective of the project was to develop a model which could identify patients with Diabetes who are at high risk of hospital admission. Prediction of risk of hospital admission is a complex task. Many factors influence this process and the outcome. There is presently a serious need for methods that can increase healthcare institution's understanding of what is important in predicting the hospital admission risk.

The project predicts the onset of diabetes in a person based on the relevant medical details that are collected using a Web application. When the user enters all the relevant medical data required in the online Web application, this data is then passed on to the trained model for it to make predictions whether the person is

diabetic or non-diabetic. The model is developed using an artificial neural network consisting of a total of six dense layers. Each of these layers is responsible for the efficient working of the model. The model makes the prediction with an accuracy of 98%, which is good and reliable.

FUTURE SCOPE

In the future, Machine learning models can be trained on large datasets of patient information, including medical records, lifestyle factors, and genetic data, to identify patterns and risk factors associated with diabetes. By analyzing these patterns, machine learning algorithms can predict the likelihood of developing diabetes in individuals at an early stage. This early detection can enable healthcare professionals to intervene with preventive measures, lifestyle modifications, or medical interventions to reduce the risk of developing diabetes. Overall, the future scope of diabetes prediction using machine learning is vast and holds great potential for improving early detection, personalized treatment, real-time monitoring, population health, and research advancements. As technology continues to advance and more data becomes available, machine learning algorithms will play an increasingly important role in diabetes care and management.

REFERENCES

1. Abdelghani Bellaachia and Erhan Guven (2010), "Predicting Breast Cancer Survivability Using Data Mining Techniques", Washington DC 20052, vol. 6, 2010, pp. 234-239.
2. Oyelade, O. J, Oladipupo, O. O and Obagbuwa, I. C (2010), "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance", International Journal of Computer Science and Information Security, vol. 7, 2010, pp. 123-128.
3. N.Umapathi., elt .,(2017), "Optimizing link quality and bandwidth estimation for dynamic source routing". **IEEE** international conference on algorithms, methods, model and application in emerging technology, at Bharath institute of higher education and research, held on 16th to 18th feb 2017. 10.1109/ICAMMAET.2017.8186636
4. N. Umapathi, N. Ramaraj and R. Adlin Mano, (2012) A Proactive Ant Colony Algorithm for Efficient Power Routing using MANET, International Journal of Computer Applications, 58(20):33-36.
5. Azhar Rauf, Mahfooz, Shah Khusro and Huma Javed (2012), "Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity", Middle-East Journal of Scientific Research, vol. 12, 2012, pp. 959-963.
6. Kajal C. Agrawal and Meghana Nagori (2013), "Clusters of Ayurvedic Medicines Using Improved K-means Algorithm", International Conf. on Advances in Computer Science and Electronics Engineering, vol. 23, 2013, pp. 546-552.
7. Min Chen, Yixue Hao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang (2017), "Disease Prediction by Machine Learning over Big Data from Healthcare Communities", 2017, IEEE, vol. 15, 2017, pp- 215-227
8. Akhilesh Kumar Yadav, Divya Tomar and Sonali Agarwal (2014), "Clustering of Lung Cancer Data Using Foggy K-Means", International Conference on Recent Trends in Information Technology (ICRTIT), vol. 21, 2013, pp.121-126.

9. Sanjay Chakrabotry, Prof. N.K Nigwani and Lop Dey (2014), “Weather Forecasting using Incremental K-means Clustering”, vol. 8, 2014, pp. 142-147.
10. Chew Li Sa, BtAbang Ibrahim, D.H., Dahliana Hossain, E. and bin Hossin, M. (2014), "Student performance analysis system (SPAS)", in Information and Communication Technology for The Muslim World (ICT4M), 2014 The 5th International Conference on, vol.15, 2014, pp.1- 6.