

Diagnosis of Diabetes Mellitus in Bengaluru North using Gradient Boosting Algorithm

Dr Tejas K R¹, Dr B Rajalakshmi², Harshitha K³, Hemanth M⁴ and Likith A B⁵

¹⁻⁵Department of Computer Science and Engineering, New Horizon College of Engineering, Bangalore, India
Email: tejas.rayangoudar@gmail.com, hod_cse@newhorizonindia.edu, 1nh19cs058.harshitha@gmail.com,
1nh19cs060.hemanth.m@gmail.com, 1nh19cs088.likith@gmail.com

Abstract—Diabetes mellitus is one of the most distinguished illness all over the world. As the populace has developed to be generally solitary, the impact of diabetes is quickly spreading. It is the metabolic illness where the affected person has excessive blood sugar both because of the fiasco of the body to supply requisite insulin or the fiasco of the body cells to react to the already generated insulin. It may be recognized with the aid of analyzing numerous readings taken from the affected person which includes albumin, creatinine, fasting, glucose, potassium, sodium and plenty more. Due to the complicated interdependence of numerous elements and the fact that diabetes affects human organs including the kidney, eye, heart, nerves, foot, and others, early diabetes prediction is a difficult task for medical professionals.

In the realm of data science, machine learning is a young scientific field that investigates how machines learn via experience. The purpose of this study is to create a system that can accurately and early diagnose diabetes in a patient.

Index Terms— Diabetes, Prediction, Detection, Insulin, Algorithms, Machine Learning.

I. INTRODUCTION

With advancement in technology, the unmanned vehicle and Diabetes is amongst the diverse continual ailments in the recent years. It is a major cause of fatalities in today's world. Thus, an early prediction of diabetes plays a vital role in reducing the mortality rate caused by the disease. A recent survey shows that the number of people being diagnosed with diabetes has tripled. It has proven to be fatal for adults. Diabetes can also cause damages in the nerves, kidney and Alzheimer's disease.

In this paper, the gradient boosting algorithm is used as it has the highest accuracy. This continual disorder takes place when blood glucose is just too excessive, inflicting a number of impediments, in addition to being the main purpose of death. As the superiority of diabetes increases, it becomes important to find a brand-new method for the prevention and control of this disorder. However, modern-day techniques have barriers in which consumer comfort and accessibility are low. As human society has grown to be socially isolated, the impact of diabetes is quickly spreading. It is the metabolic disease where the affected person has excessive blood sugar both because of the failure of the body to supply requisite insulin or the failure of the body cells to react to the produced insulin. The usual symptoms of diabetes consist of development in thirst, hunger, weight loss, frequent urination to name a few. Diabetic people are always at a high risk of obtaining heart related issues. The risk of diabetes can be diminished if early prediction is possible. A recent survey shows that the number of people being diagnosed with diabetes has tripled. It has proven to be fatal for adults. Among diabetic patients, 30% of the

adult population is diagnosed with type 2 Diabetes Mellitus.

II. OBJECTIVES

The objective of this project is to develop a machine learning model that can accurately predict the likelihood of diabetes in individuals belonging to the North Bengaluru population, based on their demographic, clinical, and lifestyle data.

III. ANALYSIS OF PARAMETERS

The fluctuating insulin levels is most commonly observed in diabetic patients. The attributes such as diastolic blood pressure, age, thickness of skin triceps, and insulin levels are usually checked to predict diabetes. Most of the adults are susceptible to this disease. If not detected at an early stage, it is difficult to cure diabetes. It can result in fatal deaths by causing failure of nerval system, heart ailments to name a few.

IV. LITERATURE SURVEY

A. Implementation of Diabetes Incidence Prediction Using a Multilayer Perceptron Neural Network

In this paper [1], a multilayer perceptron neural network is used to create a diabetes incidence prediction tool that allows users to submit basic data and will improve the usability and accessibility of the prediction capabilities. The suggested gadget was tested on diabetes in Pima Indians, and results showed that it performed well at predicting diabetes incidences.

The findings of this study demonstrate that a software tool designed to predict the disease, diabetes, using machine learning algorithms has a higher degree of accuracy and can also improve user accessibility .It is also user friendly. The dataset used in this paper is the PIMA Indian Diabetes dataset. The outliers do not affect the outcome of this model. Data pre processing is also done. This model also shows that application of variables of different diseases will show the presence of that disease.

B. Diabetes Mellitus Prediction using Hybrid Artificial Neural Network

In this paper [2], The data is normalised and extracted from the PIMA Indian diabetes dataset. Following a K-fold cross validation, the neural network's SCG back propagation design is defined. Prediction follows analysis in the next step. The occurrence of diabetes is predicted by 8 distinct characteristics. The next step is to divide the normalised dataset into k separate, mutually exclusive subsets. Scaled conjugate gradient is the algorithm in use. The accuracy range for the model is between 77% and 100%. It has been discovered that accuracy increases to its maximum when there are 16 or more neurons in the hidden layer

C. Diabetes Prediction Using Different Machine Learning Algorithms

In this paper [3], It makes use of the PIMA Indian diabetes dataset. Training and testing versions of the dataset are separated. the forest tree at random. During the training phase, numerous decision trees are built, and the random forest selects the decision made by the majority of the trees to produce the result. The KNN algorithm sorts fresh input data according to how its neighbours are arranged. Data is normalised before being classified using KNN using available cases and similarity cases. In order to predict diabetes, this study used two alternative algorithms. To determine whether diabetes is present, it takes into account 8 factors. The training dataset, which makes up 80% of the dataset, is used to test the two algorithms.

While the KNN model's accuracy is 80.8%, the random forest model's accuracy is 78.4%. At K=11, KNN produced accurate results.

D. Deep Learning based System Design for Diabetes Prediction

In this paper [4], the PIMA Indian diabetes dataset is used. CNN is used for extracting the underlying data pattern. Max pooling is used as it reduces a number of parameters that are dependent on a significant factor. Softmax function helps to produce probability distribution across every level. The input data is sent through a convolutional layer which learns multiple features. This is followed by applying a dot product for weight values and given inputs. The map is then sent to the fully connected layer. Later the softmax function is used to determine if the patient has diabetes or not. Four classification algorithms are used to get the highest accuracy which is 98.07%.

E. Analysis and Prediction of Diabetes Using Machine Models

In this paper [5], we compare 4 machine learning models. SVM, Naive Bayes, KNN, and Gradient Boosting are among them. The Kaggle website provided the dataset. To predict diabetes, 8 characteristics are used. KNN is an algorithm used to solve problems involving classification and regression. To classify the training set, it uses the separation between the data points. SVM is applied to classification. In order to categorise the data, it creates an

ideal hyperplane. Naives Bayes is used to classify situations with a binary class or several classes. Probability and conditional probability are involved. Decision trees are used in gradient boosting. To make the weak model stronger, it incorporates the pattern into the residuals.

From the results, it is shown that Gradient boost has the best accuracy of 91.63%.

F. A Neural Network based Diabetes Prediction on Imbalanced Data

In this paper [6], a technique to classify diabetes is implemented using adaptive sampling synthetic method 4 machine learning models are compared. The data set used is PIMA. It uses class label to see if diabetes is present or not. In the proposed system, missing values are handled by performing data pre-processing. Feature selection of attributes is also performed. Classification is conducted using MLP classifier with and without detection of missing values and outliers. Then we evaluate the results using some metrics like accuracy, recall, precision.

The result shows that the proposed model with MLP has the highest accuracy, precision, recall and f-measure.

G. Deep Belief Neural Network Model for Prediction of Diabetes Mellitus

The PIMA diabetes dataset is utilized in this study [7]. There are 768 samples in it. In this study, DBN is used to predict the development of diabetes mellitus. There are 3 steps in the DBN model. Preprocessing, pretraining, and fine tuning are being done. Min-Max normalization is the preprocessing technique employed. Fine tuning is nothing but the training of the second layer based on the results of pretraining. F1 measure, recall, and precision are some of the several accuracy techniques. The findings of this study demonstrate that, in terms of Recall, Precision, and F1 measure, DBN outperforms other techniques such as naïve Bayes, Decision Tree, LR, RF, and SVM.

H. A Comparative Performance Analysis of Machine Learning Approaches for the Early Prediction of Diabetes Disease

The dataset for this study [8] contains the medical records of 520 individuals who have diabetes. The different machine learning techniques for the early diagnosis of diabetes are compared using this dataset. Logistic Regression, Naive Bayes, Decision Tree, K Nearest Neighbor, SVM, and Random Forest are the many machine learning techniques that are compared. The preprocessing of the dataset involves replacing missing values, eliminating noise, and converting strings to numeric values. Accuracy, F1-score, recall, and accuracy are some of the different evaluation measures. After comparing the six different algorithms, it is observed that SVM and KNN have similar accuracy when compared to other algorithms

I. Early Diabetes Prediction Based on Stacking Ensemble Learning Model

This study [9] aims to use the ensemble stacking method for diabetes early diagnosis and prediction. Learning from multiple different learners is referred to as a stacking algorithm. Random forest, gradient ascent trees, and Adaboost are some of the numerous learning techniques being deployed. The machine repository at UCI is where the dataset was gathered. There are 16 attributes in this dataset. Accuracy rate, recall rate, and ROC curve are the three metrics used to gauge the effectiveness of the model. This experiment's findings show that stacking ensemble approaches outperform single models. Overall performance, recall, and accuracy are all improved.

V. METHODS AND MATERIALS

Data Collection -The data used for this paper was accumulated from precise areas. The resulting phases of exploration was smeared to the poised information: Data Cleaning, Selection, Transformation and Mining.

Data Cleaning- During the information cleaning process, a coherent form of the given data for the data set is developed. Data cleaning helps in correcting absent data, discovering replicated information, and removing evil information. Only necessary attributes required to predict diabetes is condensed. The data set is saved as a .csv extension file.

Data Selection – In the data selection stage, only the data relevant to the analysis is taken from the dataset.

Data Transformation – In the data transformation stage, the designated information is altered into procedures which are suitable for data mining.

Data Mining -The data mining stage is used to analyze the data set. The data-set is divided into training and testing methods. After removing unnecessary attributes, it led to interesting factors used to predict diabetes.

VI. METHODOLOGY

Problem statement - In order to avoid or lessen the effects of diabetes on physical and social well-being, it is necessary to identify and resolve scientific issues associated with diagnosing the disease in North Bengaluru using machine learning.

Algorithm used - Gradient boosting is a machine learning technique that combines weak learners to create a strong learner. To correct the defects in the earlier tree, this ensemble approach incrementally adds decision trees. a loss function that determines the gap between predicted and actual and to accomplish the purpose, actual values must be minimized. Gradient boosting is commonly used in the medical field, especially for diagnosing diseases like diabetes.

Relevant data collection - Relevant data were gathered from Hospital in North Bengaluru on diabetes.

Diabetes dataset - The gathered diabetes data were transformed into a (tabular) format that a machine learning model could recognize. The dataset was collected from one of the hospitals in North Bengaluru. The attributes considered from the dataset are Pregnancy, Glucose, BP, Skin Thickness, Insulin, BMI, Diabetes pedigree function and age.

This project helps in improving the accuracy of predicting diabetes for an individual of that confined area.

VII. IMPLEMENTATION

1. Add the pandas and streamlit Python libraries.
2. Import the dataset for diabetes.
3. The dataset will be shown using the df.describe command.
4. To enter the input data, a slider bar will be provided on the right side of the screen.
5. You'll see the visualisation chart.
6. This creates a graphic that shows the user how their attribute value compares to other values in the dataset.
7. This chart also tells the user the range which is healthy and unhealthy.
8. The user's inputted numbers are subjected to the gradient boosting technique, which produces a report.
9. The report informs the user of their diabetes status.
10. It also demonstrates how accurate the algorithm is.

VII. OUTPUTS

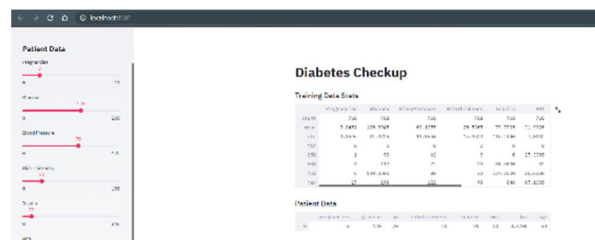


Fig: 1 Input data



Fig: 2 Visualized patient report



Fig: 3 Final Report

VIII. CONCLUSION

The usefulness of the Gradient Boosting algorithm in diagnosing diabetes has been examined in this research report. The study's findings show that Gradient Boosting is an extremely precise and effective algorithm that can be used to diagnose diabetes.

This work has showed that Gradient Boosting can deliver precise predictions of diabetes diagnosis by examining a big dataset of patient information who belong to Bengaluru North.

REFERENCES

- [1] H. Song and S. Lee, "Implementation of Diabetes Incidence Prediction using a Multilayer Perceptron Neural Network," 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2021.
- [2] B. Paul and B. Karn, "Diabetes Mellitus Prediction using Hybrid Artificial Neural Network," 2021 IEEE Bombay Section Signature Conference (IBSSC), 2021.
- [3] S. K. Reddy, T. Krishnaveni, G. Nikitha and E. Vijaykanth, "Diabetes Prediction Using Different Machine Learning Algorithms," 2021 Third International Conference on Inventive Research in Computing applications (ICIRCA), 2021.
- [4] R. Bhargava and J. Dinesh, "Deep Learning based System Design for Diabetes Prediction," 2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), 2021.
- [5] P. Saxena, S. Saha and S. K. Devi, "Analysis and Prediction of Diabetes Using Machine Models," 2022 International Mobile and Embedded Technology Conference (MECON), 2022.
- [6] S. Yadav, Y. P. S. Maravi, J. Agrawal and N. Mishra, "A Neural Network based Diabetes Prediction on Imbalanced Data," 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT), 2021.
- [7] Network Model for Prediction of Diabetes Mellitus, "2019 3rd International Conference on Imaging, Signal Processing and Communication (ICISPC), 2019.
- [8] T. R. Mahesh, V. Vivek, V. V. Kumar, R. Natarajan, S. Sathya and S. Kanimozhi, "A Comparative Performance analysis of Machine Learning Approaches for the early Prediction of diabetes Disease," 2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), 2022.
- [9] J. Liu, L. Fan, Q. Jia, L. Wen and C. Shi, "Early Diabetes Prediction Based on Stacking Ensemble Learning Model," 2021 33rd Chinese Control and Decision Conference (CCDC), 2021.