Diabetes prediction using machine learning techniques

Mr.A.Hemantha Kumar[1] Ms.R.Swetha[2]

[1]Associate Professor, Dept of CSE, Audisankara College of Engineering and Technology (AUTONOMOUS), Gudur, AP, India

[2] PG Scholar, Dept of MCA, Audisankara College of Engineering and Technology (AUTONOMOUS), Gudur, AP, India.

Abstract: -

Diabetes is a disease that develops when the body's glucose levels are too high. Diabetes should not be disregarded; if left untreated, it can result in serious complications for a person, including damage to the eyes, blood pressure, kidneys, heart, and other body organs. If diabetes is identified earlier, it can be managed. By utilising a variety of machine learning techniques, we will do early diabetes prediction in a human body or patient for a higher degree of accuracy. approaches for machine learning Better prediction outcomes can be achieved by building models from patientcollected datasets. In this work, we'll apply ensemble techniques and machine learning classification to a diabetes prediction dataset which include Gradient Boosting (GB), Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Logistic Regression (LR), and Random Forest (RF). When compared to other models, each model's accuracy varies. The project work reveals that the model is capable of accurately predicting diabetes with an accuracy of 95% or higher. Our findings demonstrate that Random Forest outperformed other machine learning methods in terms of accuracy.

Key Word:

DIABETES, MACHINE LEARNING, PREDICTION, DATASET, ENSEMBLE ARE SOME OF THE KEYWORDS.

I. INTRODUCTION

Diabetes is one of the worst diseases there is. Obesity, a high blood glucose level, and other factors can cause diabetes. It alters the function of the hormone insulin, which causes crabs to have an irregular metabolism and raises blood sugar levels. When the body does not produce enough insulin, diabetes develops. The World Health Organization estimates that 422 million people worldwide have diabetes, primarily in low- and middle-income nations. And up until the year 2030, this might be increased to 490 billion. However, diabetes is more common in many nations, including Canada, China, and India, among others. With a current population of over 100 million, India actually has 40 million diabetics. The leading cause of death worldwide is diabetes. Diabetes, for example, can be managed and controlled early on, saving lives. This study investigates diabetes prediction using a variety of diabetes disease-related factors in order to achieve this. The Pima Indian Diabetes Dataset

is used for this purpose, and several machine learning classification and ensemble techniques are used to forecast diabetes. Machine learning is a technique used to intentionally train computers or other machines. By creating various categorization and ensemble models from the obtained dataset, various machine learning techniques efficiently capture knowledge. such a calm Data can help in diabetes prediction. Many machine learning techniques are capable of making predictions, but selecting the right method can be challenging.

2. LITERATURE REVIEWS

K.Vijiya Kumar et al. [11] proposed the random Forest algorithm for the prediction of diabetes. The goal was to create a system that could more accurately detect diabetes in patients early on. The results indicated that the prediction system is able to forecast the diabetes disease effectively, efficiently, and most significantly, quickly. The suggested model yields the best results for diabetic prediction. Predicting diabetes onset: an ensemble supervised learning strategy was described by Nonso Nnamoko et al. [13]. A meta-classifier is utilised to combine the results of the ensembles' five extensively used classifiers. The findings are discussed and contrasted with similar research that made use of the same dataset. among the written works. It is demonstrated that onset diabetes can be predicted more accurately when the suggested strategy is used. Diabetes Prediction Using Machine Learning Techniques, published by Tejas N. Joshi et al. [12], tries to predict diabetes using three different supervised machine learning approaches, including SVM, Logistic regression, and ANN. In this project, an efficient method for identifying diabetes early on is proposed. Deeraj Shetty et al. [15] proposed a system for predicting the development of diabetes using data mining to create an intelligent diabetes illness prediction system that analyses the condition using a database of diabetes patients. They suggest using Bayesian and KNN (K-Nearest Neighbor) algorithms in this system to apply to a database of diabetes patients and assess them using different attributes. with the purpose of predicting the development of diabetes. Six distinct machine learning algorithms were used in the Muhammad Azeem Sarwar et al. [10] proposed study on diabetes prediction using machine learning algorithms in healthcare. The effectiveness and precision of the used algorithms are explored and contrasted. The algorithm that is most effective for diabetes prediction may be determined by comparing the various machine learning approaches employed in this study. Researchers are growing more interested in diabetes prediction in order to train the software to determine if a patient has diabetes or not by applying the appropriate classifier to the dataset. According to earlier study, the classification process has not significantly improved. Therefore, a system is needed to tackle the problems identified based on prior research, as Diabetes Prediction is a crucial topic in computers.

3.PROPOSED METHODOLOGY,

The purpose of the paper is to look into models that can more accurately forecast diabetes. To forecast diabetes, we tested various classification and ensemble methods. The period is briefly covered in the sections that follow.

A. Dataset Description- the data is gathered from UCI repository which is named as Pima Indian Diabetes Da- taset. The dataset have many attributes of 768 patients.

Table 1: Dataset Description

S No.	Attributes
1	Pregnancy
2	Glucose
3	Blood Pressure
4	Skin thickness
5	Insulin

6	BMI(Body Mass Index)
7	Diabetes Pedigree Function
8	Age

The 9th attribute is class variable of each data points. This class variable shows the outcome 0 and 1 for diabetics which indicates positive or negative for diabetics.

Distribution of Diabetic patient- We made a model to predict diabetes however the dataset was slightly imbal- anced having around 500 classes labeled as 0 means nega- tive means no diabetes and 268 labeled as 1 means positive means diabetic.

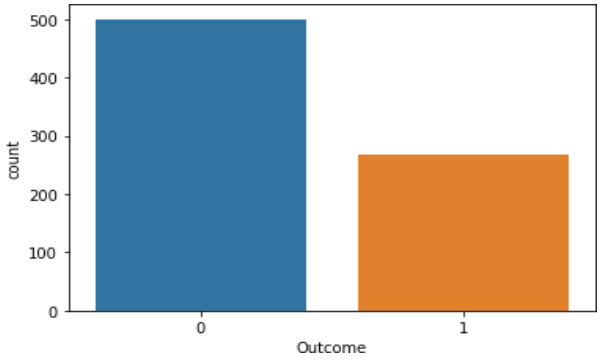


Figure 1: Ratio of Diabetic and Non-Diabetic Patient .

B. Data Preprocessing - The most crucial process is data preprocessing. Most data pertaining to healthcare has missing values and other contaminants that can affect how useful the data is. Data preparation is done to enhance the quality and effectiveness of the results produced through mining. This procedure is crucial for accurate results and good prediction when applying machine learning techniques to the dataset. For the Pima Indian diabetes dataset, pre processing must be done twice.

1. **Missing Values Removal**: Remove all instances with a value of zero (zero). Zero as a value is not conceivable. As a result, this instance is stopped. Feature subset selection is the process of creating a feature subset by the elimination of extraneous characteristics and instances. This procedure minimises the dimensionality of the data and helps to more quickly.

2. **Data splitting**:Data is separated into training and testing the model after it has been cleaned. After the data is spilt, we train the algorithm on the training data set while putting the test data aside. Based on the logic, methods, and values of the feature in the training data, this training process will create the training model. The primary goal of normalisation is to put all qualities on the same scale.

C. **Use machine learning**:Once the data is ready, we use machine learning techniques. To predict diabetes, we employ a variety of classification and ensemble algorithms. The procedures used on the diabetes dataset for Pima Indians. The primary goal is to use machine learning techniques to examine how well these methods function and determine accuracy.

1) **Support Vector Machine**: The SVM, also referred to as a support vector machine, is a supervised machine learning algorithm. The most used classification method is SVM. A hyperplane is made by SVM to divide two classes. In high-dimensional space, it can produce a hyperplane or collection of hyperplanes. This hyperplane can also be utilised for regression or classification. SVM distinguishes examples within particular classes and has the ability to categorise entities that lack data support. The nearest training point for any class is used for separation, which is carried out using a hyperplane.

Algorithm-

- Select the hyper plane which divides the class bet- ter.
- To find the better hyper plane you have to calcu- late the distance between the planes and the data which is called Margin.
- If the distance between the classes is low then the chance of miss

conception is high and vice versa. So we need to

- Select the class which has the high margin. Margin = distance to positive point + Distance to negative point.

2) K-Nearest Neighbor - Another supervised machine learning technique is KNN. KNN aids in the resolution of the classification and regression issues. KNN is a slack prediction method. KNN assumes that related things are located close to one another. Similar data points are frequently found close together. KNN aids in classifying fresh work using a similarity metric. The KNN algorithm records every record and categorises them based on how similar they are. Uses a tree-like structure to determine the distance between the spots. The algorithm locates the nearest data points in the training data set, or the new data point's nearest neighbours, to create a forecast. Here, K is always a positive integer and stands for "number of close neighbours." Value of the neighbour is picked from

a list

$$d(P,Q) = \sum_{i=1}^{n} (P_i - Q_i)^2$$

Algorithm-

- Take a sample dataset of columns and rows named as Pima Indian Diabetes data set.
- Take a test dataset of attributes and rows.
- Find the Euclidean distance by the help of formu- la-

$$EculideanDistance = \sqrt{\sum_{i=1}^{y} \sum_{j=1}^{m} \sum_{l=1}^{n-1} (R_{(j,l)} - P_{(i,l)})^2}$$

- Then, Decide a random value of K. is the no. of nearest neighbors
- Then with the help of these minimum distance and Euclidean distance find out the nth column of each.
- Find out the same output values.

If the values are same, then the patient is diabetic, other- wise not.

3) **Decision Tree** – Decision trees are a fundamental categorization technique. This type of learning is supervised. For categorical response variables, a decision tree is employed. A decision tree is a structure-based model with a tree-like structure that represents the classification process depending on input data. Input variables can be of any type, including text, discrete, continuous, and graph. Using the Decision Tree Algorithm: Steps-

- Create a tree using nodes as the input feature.
- Pick the input characteristic with the biggest information gain while predicting the output.
- For each attribute in each tree node, the highest information gain is determined.
- Repeat step 2 to create a subtree utilising the feature not present in the node above.

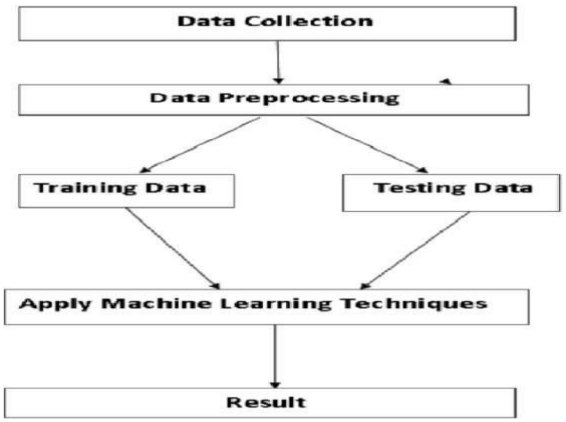
4) Logistic Regression: Another advised learning classification approach is logistic regression. It is employed to estimate, depending on one or more predictors, the likelihood of a binary response. Both continuous and discrete ones are possible. When we wish to categorise or separate some data items into categories, we utilise logistic regression.

It categorises the data in binary form, which entails only the digits 0 and 1, which relate to patients who are positive or negative for diabetes. The main goal of logistic regression is to find the optimal fit that best describes the connection between the target and predictor variables. The linear regression model is the foundation of logistic regression. The sigmoid function is used in the logistic regression model to forecast the likelihood of both positive and negative outcomes. class. sigmoid process P = 1/1+e - (a+bx) P stands for

probability, and a and b represent model parameters.

Ensembling is a machine learning strategy that refers to combining several learning algorithms for a certain goal. It is employed because it offers greater prediction than any other individual model. Noise bias and variation are the primary causes of error; ensemble methods assist in minimising or reducing these errors. There are two widely used ensemble methods, including voting, averaging, ada-boosting, bagging, and gradient boosting. Here For forecasting diabetes, we have employed Bagging (Random forest) and Gradient Boosting Ensemble approaches in this paper.

5) Random Forest: This ensemble learning methodology, which is employed for both classification and regression applications, is a kind. It provides more accuracy as compared to other models. Large datasets can be handled by this strategy with ease. Leo Bremen has created Random Forest. It is a well-known en- semble learning technique. By lowering variance, Random Forest enhances the performance of Decision Tree. It works by building a large



number of decision trees during training period, and it outputs the class that represents the mean of all classes, or mean classification, or mean regression, of all individual trees.

Algorithm-

- The first step is to select the "R" features from the total features "m" where R<<M.
- Among the "R" features, the node using the best split point.
- Split the node into sub nodes using the best split.
- Repeat a to c steps until "l" number of nodes has been reached.
- Built forest by repeating steps a to d for "a" num- ber of times to create "n" number of trees.

The random forest finds the best split using the GinIndex Cost Function which is given by:

$$Gini = \sum_{k=1}^{n} p_k * (1 - p_k) \; Where \; k = Each \; class \; and \; p = proption \; of \; training \; instances$$

The first stage entails looking at options, using the bases of each randomly formed decision tree to forecast the conclusion, and storing the predicted outcome at intervals around the desired location. Calculate the votes for each predicted target, and then, as a consequence of the final prediction made using the random forest algorithm, accept the predicted target with the highest number of votes.

For a variety of applications, Random Forest offers some solutions that produce accurate

6) Gradient Boosting - The most effective ensemble technique for prediction, gradient boosting is a classification technique. To create powerful learning models for prediction, it combines weak learners. It employs judgement.

Algorithm-

- Consider a sample of target values as P
- Estimate the error in target values.
- Update and adjust the weights to reduce error M.
- P[x] =p[x] +alpha M[x]
- Model Learners are analyzed and calculated by loss function F
- Repeat steps till desired & target result P.

4.MODEL BUILDING

This step, which includes model building for diabetes prediction, is the most crucial one. The numerous machine learning techniques for diabetes prediction that were outlined above have been used in this.

The proposed methodology's process

Step 1: Import diabetic dataset and necessary libraries.

Pre-process data to omit missing data in step two.

Step 3: Perform an 80/20 split of the dataset to create the training set and the test set.

Step4: Choose the machine learning algorithm from the list, which includes K-Nearest Neighbor, Support Vector Machine, Decision Tree, Logistic Regression, Random Forest, and Gradient Boosting.

step 5.Build the classifier model based on the training set for the aforementioned machine learning method .

Step 6: Use the test set to evaluate the Classifier model for the given machine learning algorithm.

Step 7: Conduct a comparison evaluation of each classifier's experimental performance outcomes.

Step 8: Select the highest performing algorithm after analysis based on various metrics.

5.RESULTS OF EXPERIMENTS

Different actions were taken in this work. The proposed approach is built in Python and makes use of various classification and ensemble algorithms. These techniques are common Machine Learning techniques used to get the maximum accuracy out of the data. We can see from this work that the random forest classifier performs better than the competition. Overall, to make predictions and achieve high performance accuracy, we applied the best machine learning approaches. The outcome of these machine learning techniques is shown in Figure.

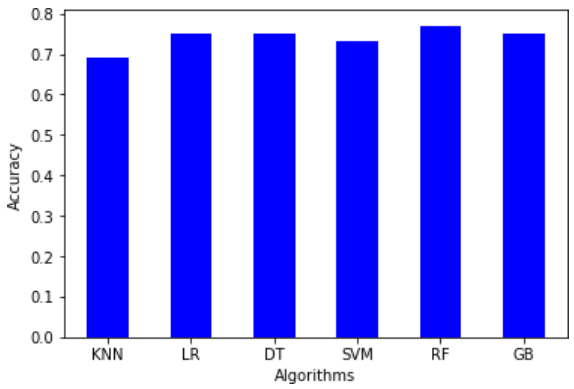


Figure3: Accuracy Result of Machine learning methods

Here feature played important role in prediction is present- ed for random forest algorithm. The sum of the importance of each feature playing major role for diabetes have been plotted, where X-axis represents the importance of each feature and YAxis the names of the features.

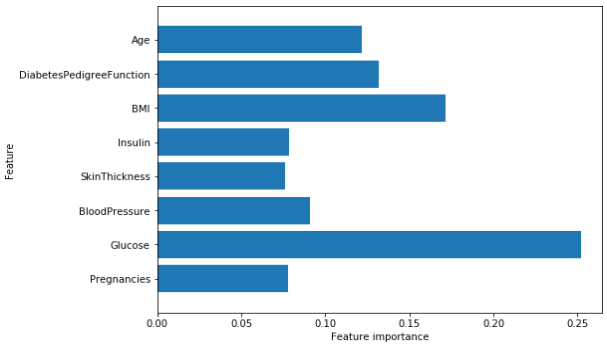


Figure 4: Feature Importance Plot for Random Forest

6.CONCLUSION

This project's primary goal was to build and implement methods for predicting diabetes using machine learning, and to assess the effectiveness of those methods. The suggested strategy makes use of a variety of classification and ensemble learning techniques, including classifiers from SVM, Knn, Random Forest, Decision Tree, Logistic Regression, and Gradient Boosting. Additionally, classification accuracy of 77%

was attained. The experimental results can help medical professionals make early predictions and decisions to treat diabetes and save a patient's life.

7.REFERENCES

[1] Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance's for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.

[2] K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ". Proceeding of International Conference on Systems Compu- tation Automation and Networking, 2019.

[3] Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Perfor- mance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 Feb- ruary, 2019.

[4] Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineer- ing Research and Application, Vol. 8, Issue 1, (Part -II) Janu- ary 2018, pp.-09-13.

[5] Nonso Nnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC), 2018.

[6] Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabe- tes Disease Prediction Using Data Mining ". International Con- ference on Innovations in Information, Embedded and Com- munication Systems (ICIIECS), 2017.

[7] Nahla B., Andrew et al,"Intelligible support vector machines for diagnosis of diabetes mellitus. Information Technology in Biomedicine", IEEE Transactions. 14, (July. 2010), 1114-20.

[8] A.K., Dewangan, andP., Agrawal, "Classification of Diabetes Mellitus Using Machine Learning Techniques," International Journal of Engineering and Applied Sciences, vol. 2, 2015.

Author's Profile

A. HEMANTHA KUMAR has received his M.Tech degree in CSE from Sathyabama Deemed University in 2006, Chennai. He is dedicated to teaching field from 2001. He has guided P.G and U.G students. His research areas included Computer Networks, Network Security and Machine Learning. At present he is working as Associate Professor in Audisankara College of Engineering and Technology, Gudur, Tirupati (Dt), Andhra Pradesh, India.

RANGINENI SWETHA has Pursuing her MCA from Audisankara College of Engineering and Technology (AUTONOMOUS),

Gudur, affiliated to JNTUA in 2022. Andhra Pradesh, India.