International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

Diabetes Prediction using Machine Learning Techniques

[1]*Dr. J Rajaram Professor, drrajram81@gmail.com,*

[2]*D Navya Assistant Professor, dubbaka.navya@gmail.com,*

[3]*Bandam Naresh Assistant Professor, nareshbandam4@gmail.com,*

[4]*Banothu Usha Assistant Professor, banothuusha@gmail.com.*

Department of CSE Engineering,
Nagole , **Institute of Engineering and Technology collage** *in Hyderabad.*

Abstract:

Diabetes is a disease that develops as a result of a high glucose level in the bloodstream of a person. A person's diabetes should not be disregarded; if left untreated, diabetes may lead to serious health complications in the long run. Such as heart disease, renal disease, high blood pressure, and so on it may cause eye damage and can also have an impact on other organs in the human body. Diabetes may be managed if it is identified and treated early on. In order to accomplish this is the objective during this project's effort; we will look at early diabetes detection. In a human body or on a patient in order to gets more precision Different Machine Learning Techniques are being used. Machine gaining knowledge of methods by constructing models using data gathered from patients, it is possible to get better results for prediction. This is the case in this effort that we will put to use Classification and ensemble learning with machine learning Using statistical methods on a dataset, diabetes may be predicted. Which of the following are K-Nearest? KNN (Kindest Neighbour), Logistic Regression (LR), and Decision Tree (DT), Support Vector Machine (SVM), Gradient Boosting (GB), and Support Vector Machine (SVM) The Forest of Chance (RF). Every model has a different level of accuracy than the others. Whenever they are contrasted with other models. The project work provides the opportunity to the model's ability to forecast diabetes with high accuracy or greater accuracy demonstrates that the model is capable of doing so. As a result of our research, we have discovered that when compared to other methods, Random Forest produced greater accuracy. Techniques using machine learning.

Keywords: Diabetes, Machine, Learning, Prediction, Dataset, Ensemble

1) . INTRODUCTION

Diabetes is one of the most dangerous illnesses in the world. Diabetes is caused by obesity, excessive blood glucose levels, and other factors, among others. It has an effect on the insulin hormone, resulting in aberrant glucose levels. Crabs' metabolism is improved, as is the amount of sugar in their blood. Blood. Diabetes develops when the body does not produce enough insulin. Insulin. In accordance with the World Health Organization (WHO), Diabetes affects about 422 million people worldwide, with the majority of them living in low- or middle-income nations. And this may be the case. Up to the year 2030, the total amount of money will have risen to 490 billion. However Diabetes is reported to be prevalent in a number of different countries. Such as Canada, China, and India, among others. India has a population of 1.2 billion people. As the population of India has grown to more than 100 million, the real number of diabetics in the country is 40 million. Diabetes is a leading cause of mortality in the United States. Throughout the whole globe early detection of diseases such as diabetes may save lives. Maintain control while saving a person's life In order to achieve this, this research investigates diabetes prediction by examining a variety of variables. Diabetes-related characteristics are listed below. In order to do this, we using the Pima Indian Diabetes Dataset, we run a number of tests. Techniques for machine learning classification and ensemble learning to be able to anticipate diabetes Machine learning is a technique that is used to learn new things. This method is used to explicitly teach computers or machines. Various Machine Learning Techniques are effective in delivering results. Gather knowledge by creating different classifications and categorizations ensemble models derived from a dataset collection such information was gathered Diabetes may be predicted with the use of statistics. Various methods are used. Machine Learning is capable of making predictions; however this is not always the case. It's difficult to decide on the ideal method. As a result, for this reason On the basis of popular classification and ensemble techniques, we develop a dataset for the purpose of prediction

2) REVIEW OF THE LITERATURE

K.VijiyaKumar et al. [11] presented a random Forest algorithm for the prediction of diabetes and developed a system that may be used to diagnose the disease. Can make an early diagnosis of diabetes in a patient who has a

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

genetic predisposition The Random Forest method, which is used in machine learning techniques, provides more accuracy. The suggested model provides the following:

Findings for diabetes prediction were the best, and the outcome indicated showed the diabetes disease prediction system is capable of accurately, efficiently, and most importantly, accurately forecasting the diabetes disease immediately. Nonstop Nnamoko and colleagues [13] proposed a method for predicting An method based on ensemble supervised learning for the detection of diabetes onset They used five commonly used classifiers for the classification process. Ensembles are created, and a meta-classifier is utilised to group them together. Outputs. The findings are given and contrasted with those of similar research that utilised the same dataset that has been published before. It has been shown that diabetes can be prevented by using the suggested approach. It is possible to anticipate the beginning of a storm with greater precision. Tejas Researchers N. Joshi and colleagues [12] presented Diabetes Prediction Using Artificial Intelligence. The goal of Machine Learning Techniques is to forecast diabetes via machine learning. SVM, Logistic regression, and ANN are three distinct supervised machine learning techniques that may be used together. This research proposes a method for the early identification of the flu that is successful. Diabetes is a condition that affects the body's glucose levels. Deeraj Shetty et al. [15] suggested diabetes illness prediction using data mining to build an Intelligent Diabetic Disease Prediction System that provides analysis of diabetes malady using a diabetes patient's database, which they call the Intelligent Diabetes Disease Prediction System.

This system proposes the usage of algorithms such as genetic algorithms. In order to use Bayesian and KNN (K-Nearest Neighbor) methods and analyse diabetic patients' databases by taking different measures, the characteristics of diabetes may be used to predict the development of diabetic illness. Muhammad Azeem Sarwar and colleagues [10] suggested a research project on in this study; machine learning techniques were used to predict diabetes. They used six distinct machine learning algorithms in the healthcare field. The performance and accuracy of the algorithms that were used were evaluated. Is addressed, as well as contrasted a comparison of the many options the use of machine learning methods in this research shows In terms of diabetes prediction, which algorithm is the most effective? Diabetes Prediction is becoming a popular topic of research for scientists. The researchers in order to train the computer algorithm in order to recognise Patients may be classified as diabetic or not by using the appropriate classifier. It is a collection of data it has been determined based on prior study work. It has been noted that the categorization procedure is not very complex.

3) PROPOSED METHODOLOGY

Goal of the study is to explore a model that can more accurately predict diabetes than the current one. In order to forecast diabetes, we experimented with a variety of different classification and ensemble methods. The following section provides a high-level overview of the period. A. Dataset Description: The information has been collected from the University of California; Irvine.Pima Indian Diabetes Dataset is the name of the repository where the data is stored. There are numerous characteristics of 768 patients in the dataset.

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

Each data point has a class variable, which is the ninth attribute. This class variable displays the result for diabetics as a number between 0 and 1, indicating whether the outcome is positive or negative for diabetes.

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

We created a model to represent the distribution of diabetic patients. predict diabetes, however the dataset was somewhat skewed, with about 500 classes classified as 0 means negative, meaning no diabetes, and 268 classes labelled as 1 means positive, meaning diabetes. Indicates that you are diabetes

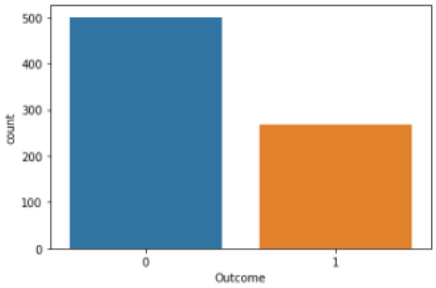


Figure 1: Ratio of Diabetic and Non Diabetic Patient

B. Data Pre-processing-

The most important process is data pre-processing. The majority of healthcare-related data includes missing values and other contaminants that may impair the effectiveness of the information provided. Data pre-processing is carried out in order to enhance the quality and efficacy of the results produced following the mining process. To make use of Machine Learning Techniques are used successfully on the dataset. This procedure is critical for obtaining an accurate result and achieving success. Prediction. We need to do this with the Pima Indian diabetes dataset. Pre-processing should be done in two stages. 1st, remove all occurrences of missing values by selecting them and pressing delete. Have a monetary value of zero (0) It is not possible to have a value equal to 0. As a result, this instance has been eliminated. We create a feature subset by removing non-essential characteristics and occurrences from the dataset. This procedure is referred to as features subset selection, and it is carried out by lowers the dimensionality of data and assists in working more quickly

2) Data segmentation- After cleaning the data, the data is normalised and divided into two groups: training and testing the model. When information is collected when we have spitted, we train the algorithm on the training data set. Keep the results of the tests separate. The results of this training procedure will be the training model is built on logic and methods, as well as the values of the feature in the training data. The primary goal of normalisation is to put all of the characteristics into the same range of values.

C. Implement Machine Learning-

Once the data has been collected and analysed. We make use of the Machine Learning Technique. We use a variety of methods. Techniques like as classification and ensemble learning are being used to predict diabetes. The techniques were tested on a diabetic dataset including Pima Indians. The main goal is to apply Machine Learning Techniques to a variety of problems. In addition to analysing the performance of various techniques and determining their accuracy, researchers have been able to identify the responsible/important characteristic that plays a significant part in prediction.

The procedures are as follows:

Support Vector Machine (SVM):

Support Vector Machine is a kind of vector machine. Supervised machine learning (also known as svm) is a machine learning algorithm that is supervised. The most often used categorization method is Svm. Svm

The hyper plane that separates two classes is generated. It has the potential to generate hyper plane or group of hyper planes in three dimensions or higher space. This hyper plane may be used for categorization or for a variety of other purposes. Regression is also possible. Svm distinguishes between occurrences in certain situations. Classes and it may even categorise things that are not supported by data in the database. Separation is

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

accomplished via the use of hyper planes. It is responsible for the separation to the nearest training point of any class.

Algorithm-

• Choose the hyper plane that divides the class the most effectively.

• To determine which hyper plane is the best fit for the data, you must first calculate the distance between the planes and the data. This is referred to as Margin.

• If the distance between the classes is short, the distance between the classes is the likelihood of miscarriage is great, as is the likelihood of fertilisation. As a result, we must

• Choose the class that has the highest profit margin. Margin equals the distance between the positive and negative points. Point in the negative

K-Nearest Neighbour (KNN) –

KNN is a supervised machine learning method that is also used in image recognition. KNN contributes to the solution of both problems. Problems involving categorization and regression KNN are a method for making lazy predictions. KNN makes the assumption that comparable objects are nearby. Each's company many times, data pieces that are similar to one another are combined.

They are extremely close to one another.KNN assists in the organisation of new work. Based on the measure of similarity The KNN algorithm records all of the information. Keep track of all of your documents and categorise them according to how similar they are measure. The method of calculating the distance between two locations is called the structure is similar to a tree. To create a forecast based on fresh information when a data point in the training data set is reached, the algorithm searches for the data points that are the closest to it – its nearest neighbours. Here, K denotes the number of it's always a positive integer when it comes to close neighbours. Neighbour's the value is selected from a collection of classes. The degree of closeness is mostly denoted by fined in terms of Euclidean distance. The Euclidean distance between two points P and Q i.e. P (p1, p2, Pn) and Q (q1, q2,qn) is defined by the following equation:-

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

lgorithm-

• Create an example dataset consisting of columns and rows with the following names: As an example, the Pima Indian Diabetes data set

• Create a test dataset with a variety of characteristics and rows.

• Calculate the Euclidean distance using the following formula:

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

Then, choose a random value for K, where K is the number of closest neighbours.

• Then, using the lowest distance and the maximum distance, calculates the distance between the two neighbours. Using the Euclidean distance, get the nth column of each.

• Determine the same output values as before.

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

If the readings are the same, the patient has diabetes; otherwise, the patient does not. Three-dimensional decision tree is a fundamental categorization system. Method. It is a technique of supervised learning. Diagram of a decision tree when the response variable is categorical, this term is utilised. Diagram of a decision tree is built on a tree-like structure-based model that explains the classification process depending on the input feature The variables that are used as inputs include any kinds, such as graphs, text, discrete, continuous, and so on. Pertaining to the Decision Tree Algorithm-

• Construct a tree with nodes as the input feature and display it.

• Choose the feature to use to forecast the output from the input feature that has the greatest information gain.

• The greatest amount of information is gained is determined for each and every property in each and every node of the tree

• Repeat step 2 to create a sub tree that makes use of the feature. Which is not utilised in the preceding node?

Logistic Regression-

Logistic regression is a classification method that may be learned via supervised learning. In order to estimate the likelihood of a binary answer depending on one or more variables, it is employed. A greater number of predictors they may be either continuous or discrete in nature. When we wish to categorise or distinguish certain data objects into categories, we utilise logistic regression to do this. It categorises the data in binary form, which implies solely in the numbers 0 and 1.which refers to categorise patients as positive or negative for diabetes, according on their results.

The primary goal of logistic regression is to find the best fit, which is the person in charge of explaining the connection between targets as well as the predictor variable Logistic regression is a technique that is based on the linear regression model is used. The logistic regression model is used. The sigmoid function is used to forecast the likelihood of being in the positive or negative class. The sigmoid function P = 1/1+e - (abs) is defined as P denotes probability, and

Model parameters are denoted by the letters a and b. Assembling is a machine learning method that is used to put together puzzles. When several learning algorithms are used together for a task, this is referred to as an ensemble. It offers more accurate forecasting than any other method. It is utilised since it is different from any other unique model. The most important thing to remember is that the source of error is noise bias and variation; ensemble techniques are used to correct for this. Assist in reducing or minimising the occurrence of these mistakes there are two of them. Bagging, boosting, and other common ensemble techniques are examples of this. ado-boosting, Gradient boosting, voting, averaging, and other similar techniques Bagging has been used in this piece of art (Random forest) as well as ensemble techniques based on gradient boosting for forecasting diabetes.

5) Random Forest - This is a kind of ensemble learning method that may be used for classification and regression problems, as well as for machine learning tasks. If you compare the accuracy it provides to other methods, it is much better models. This technique is capable of dealing with big datasets with ease. Leo Bremen is the creator of the Random Forest algorithm. It is a well-known method of ensemble learning. By decreasing variance, Random Forests help to improve the performance of Decision Trees. It is in operation. By building a large number of decision trees during the training phase time and returns the class that corresponds to the mode of the classes, respectively Individual trees are classified or predicted using a mean prediction (regression) method. Algorithm-

• The first step is to choose the "R" features from the drop-down menu.

Where R is more than M, the total features is "m."

• Among the "R" characteristics, the node with the best performance a severance point

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

• Decide on the optimum split to use when dividing the node into sub nodes.

• Repeat steps a through c until the "l" number of nodes has been reached. A conclusion has been reached

• To construct the forest, go through Steps A through D as many times as necessary to generate "n" number of trees. Using the Gin-Index, the random forest determines which split is the best. A cost function that is provided by the expression:

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

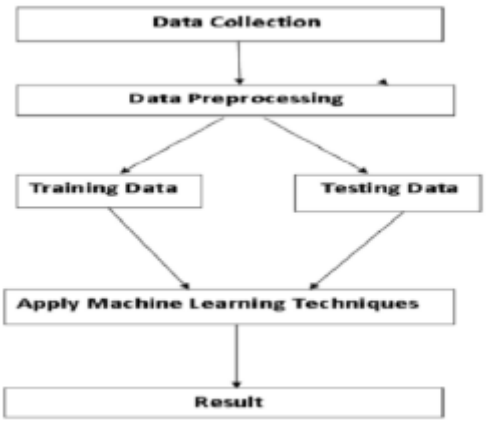


Figure 2: Overview of the Process

This is the most critical step, which involves the development of a model for the prediction of diabetes. Here, we have applied different machine learning methods, such as those described above, for the purpose of predicting diabetic complications.

The Proposed Methodology's Implementation Procedure

In the first step, import the necessary libraries and the diabetes dataset.

Step2: Pre-process the data in order to eliminate any missing information.

Step3: Perform an 80 percent split on the dataset to separate it into two groups.

Training set accounts for 80 percent of the total, with the remaining 20 percent going to Test set.

Step4: Choose a machine learning algorithm, such as Nearest Neighbour, Support Vector Machine, Decision Tree, or any other method you choose.

Logistic regression, Random Forest, and Gradient Boosting are all methods of predicting the future. Algorithm.

Step5: Using the training data, create a classifier model using the machine learning method that was described before.

Step6: Using the test set, evaluate the Classifier model for the previously stated machine learning method.

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

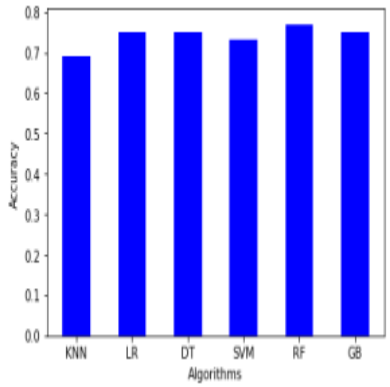
Step 7: Carry out a comparison Evaluation of the experimental performance results achieved for each classifier based on the outcomes of the experiment.

Step8: After doing an analysis based on different metrics, choose the method that performs the best.

4) RESULTS OF EXPERIMENTAL STUDIES

In this study, a number of alternative approaches were used. The suggested approach makes use of a variety of classification and ensemble techniques.

Python was used to create the programme. These techniques are standard Machine Learning methods that are used to achieve the highest level of accuracy from data collection. We can observe in this study that the random forest is effective. When compared to other classifiers, this one performs better. In general, we Predictions have been made using the finest Machine Learning methods available. As well as to attain high levels of precision in performance. The illustration depicts These Machine Learning techniques produced the following results.



Here feature played important role in prediction is presented 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 Y-Axis the names of the features.

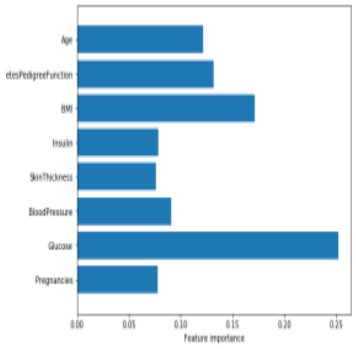


Figure 4: Feature Importance Plot for Random Fores

5) CONCLUSION

The main aim of this project was to design and implement Diabetes Prediction Using Machine Learning Methods and Performance Analysis of that methods and it has been achieved successfully. The proposed approach uses various classification and ensemble learning method in which SVM, Knn, Random Forest, Decision Tree, Logistic Regression and Gradient Boosting classifiers are used. And 77% classification accuracy has been achieved. The Experimental results can be asst health care to take early prediction and make early decision to cure diabetes and save humans life.

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 Computation Automation and Networking, 2019.

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

[4] Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineering Research and Application, Vol. 8, Issue 1, (Part -II) January 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, "Diabetes Disease Prediction Using Data Mining ".International Conference on Innovations in Information, Embedded and Communication 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, and P., Agrawal, "Classification of Diabetes Mellitus Using Machine Learning Techniques," International Journal of Engineering and Applied Sciences, vol. 2, 2015.