International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 02 | Feb 2022 www.irjet.net p-ISSN: 2395-0072

DIABETES PROGNOSTICATION UTILIZING MACHINE LEARNING

Suyog Nemade

Student Dept. of Information Technology Engineering, Shah & Anchor Kutchhi Engineering College, Mumbai, India

---

Abstract - *Diabetes is a disease caused by an increase in glucose level in the human body. Diabetes should not be ignored Left untreated, diabetes can cause serious problems for a person such as: heart problems, kidney problems, high blood pressure, eye damage can also affect other parts of the human body. Diabetes can be controlled if it is predicted early. In order to achieve this project goal we will make early diagnosis of Diabetes in the human body or patient with high accuracy by applying, Various Machine Learning Strategies. Machine learning strategies Provide a better predictive effect by creating models from data sets collected from patients. In this activity we will use the Learning Machine Planning and integrate databases to predict diabetes. Neighboring K-Nearest (KNN), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Gradient Boosting (GB) and Random Forest (RF). Accuracy is different for all models compared to other models. Project Work provides a more accurate or accurate model showing that the model is able to accurately predict diabetes. Our result shows that Random Forest has achieved higher accuracy compared to other machine learning strategies.*

Key Words: Diabetes, Machine, Learning, Prediction, Dataset, Ensemble

1.INTRODUCTION

Diabetes is one of the most dangerous diseases in the world. Diabetes caused by obesity or high blood glucose levels, and so on. It affects the hormone insulin, which leads to altered crab metabolism and improves blood sugar levels blood. Diabetes occurs when the body does not produce enough insulin. According to the World Health Organization (WHO), about 422 million people have diabetes, mainly from low-income or low-income countries. And this could increase to $ 490 billion by 2030. However the spread of diabetes is found in various countries such as Canada, China, and India etc. The population of India is now over 100 million so the actual number of diabetics in India is 40 million. Diabetes is the leading cause of death in the world. Early predictors of diseases such as diabetes can be controlled and save a person's life. To achieve this, this work assesses diabetes prognosis by taking into account a variety of diabetes-related traits. For this purpose we use the Pima Indian Diabetes Dataset, using various machine classifications and compiling Diabetes Prediction Strategies. Machine Learning The method used to train computers or equipment explicitly. Variety Machine Learning Strategies provide an effective Information Collection effect by creating a variety of classifications and integrating models from collected databases. Such data can be helpful in predicting diabetes. Various machine learning strategies can predict, yet it is difficult to choose the best method. So for this purpose we use popular classifications and databases to predict.

2. LITERATURE REVIEW

K.VijiyaKumar et al.[1] The proposed Forest Forest Algorithm for Prediction of Diabetes creates a system that can predict diabetes in a patient with high accuracy using the Random Forest algorithm in machine learning techniques. The proposed model provides the best possible diabetes forecast results and the result showed that the predictive system is able to predict diabetes effectively, efficiently and most importantly, instantly. Nonso Nnamoko et al. [5] presented to predict the onset of diabetes: an integrated targeted learning method they used to break down into five most commonly used categories in ensembles and a meta-classifier used to compile their results. The results were presented and compared with similar studies that used the same database within the textbooks. It is shown that by using the proposed method, predicting the onset of diabetes can be done with high accuracy. Tejas N. Joshi et al. [4] introduces Diabetes Prediction Using Machine Learning Strategies aims to predict diabetes with three different machine learning modes include: SVM, retrieval, ANN. This project proposes an effective way to diagnose diabetes early. Deeraj Shetty et al. [6] Proposed prognosis of diabetes using data mines includes the Intelligent Diabetes Disease Prediction System which analyzes diabetes using a database of diabetic patients.In this program, they propose the use of algorithms such as Bayesian and KNN (K-Nearest Neighbor) for use on a diabetic patient's website and analyzed by taking various diabetic traits to predict diabetes. Muhammad Azeem Sarwar et al. [7] The proposed study of predicting diabetes using machine learning algorithms in health care using six different machine learning algorithms The effectiveness and accuracy of the algorithms used are discussed and compared. A comparison of the different machine learning techniques used in this study reveals which algorithm is most suitable for predicting diabetes. Diabetes Prediction is becoming an area of interest for researchers to train their diabetic patient through the appropriate section of the database. Based on previous research work, it has been noted that the classification process is not significantly improved. So a system is needed as a Diabetes Prediction is an important component of computers, handling issues identified based on previous research.

3. PROPOSED METHODOLOGY

The purpose of this paper is to investigate the model to predict diabetes with better accuracy. We experimented with different categories and compiled algorithms to predict diabetes. Next, we will briefly discuss the paragraph.

A. Dataset Description- data are collected at a UCI site called the Pima Indian Diabetes Dataset. The database has many patient characteristics of 768.

Table 1: Dataset Description

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

Attribute 9 is a class variation of each data point. This variability of class indicates a result of 0 and 1 in diabetics indicating whether it is good or bad for diabetics.

Distribution of Diabetic patient- We did the diabetes modeling model but the data set was inconsistent as it had about 500 classes labeled 0 mean you do not have diabetes and 268 labeled 1 means good means you have diabetes.

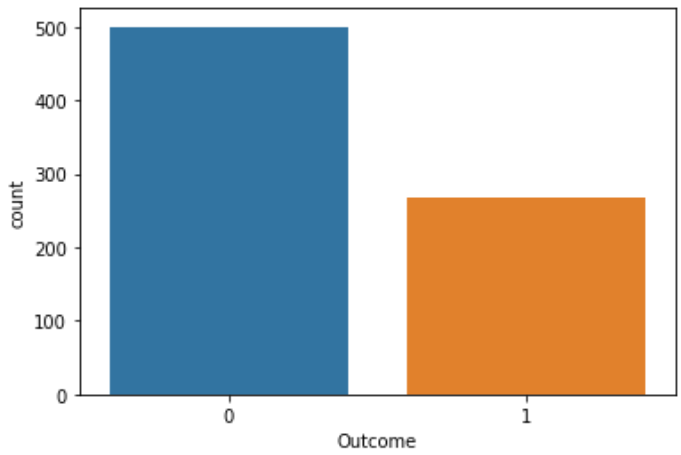

Figure 1: Ratio of Diabetic and Non Diabetic Patient

B. Data Preprocessing-Data processing is a very important process. In particular data related to health care contains a missing vale and other contaminants that may affect the performance of the data. In order to improve the quality and performance obtained after the excavation process, data analysis is performed. Using Machine Learning Techniques on Database Successfully This process is essential for achieving accurate and successful results forecast. In the Pima Indian diabetic data database we need to do a preliminary analysis in two steps.

1). Missing Values removal- Subtract all zero conditions (0) as a value. Having a zero number is not possible. This event is therefore deleted. By removing unnecessary features / conditions we create a feature set and this process is called small selection selection features, which reduces data dialing and helps to work faster.

2). Splitting of data-After data cleaning, the data is usually trained and tested model. When data is spit out then we train the algorithm on the training data set again keep test data set aside. This training process will produce a training model based on concepts and algorithms and feature values in training data. Basically the purpose of the practice is to bring all the elements under the same scale.

C. Apply Machine Learning- Once the data is ready we use the Machine Learning Technique. We use a different one classification and integration strategies, predicting diabetes. Methods used in the Pima Indians diabetes database. The main goal is to use Machine Learning Strategies to analyze the effectiveness of these methods and to determine their accuracy, and to be able to identify the responsible / important factor that plays a major role in prediction.

The following strategies

1) Support Vector Machine-Vector Support Machine also known as svm is a supervised machine learning algorithm. Svm is the most popular method of partitioning. Svm creates a hyperplane that divides the two classes. It can form a hyperplane or a set of hyperplane in an area of high magnitude. This hyper plane can be used for splitting or retreating. Svm separates the conditions directly classes and can also classify non-data-based businesses. Separation is done using a hyperplane that makes the separation to the nearest training area of any class.

Algorithm-

• Choose the top flight that separates the class best.

• To get the best flight you must calculate the distance between the planes and the data called Margin.

• If the distance between classes is low then the chances of missing a pregnancy are high and vice versa. So we need to

• Choose a class with a higher limit.

Margin = distance to positive point + Distance to negative point

2) K-Nearest Neighbor -KNN is also an algorithm for machine learning learning. KNN helps solve both planning

and retrospective problems. KNN is a way of predicting laziness. KNN thinks the same things are near. Many times the same data points are very close. KNN helps consolidate new work based on the similarity scale. The KNN algorithm records all records and classifies them according to their similarity. To find the distance between points uses a tree like a building. To predict a new data point, the algorithm finds nearby data points in a training data set - nearby neighbors. Here K = The number of the nearest neighbors, is always the positive number. Neighbor value is selected from the class set. Intimacy is best described in terms of the Euclidean range. 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 numbers: -

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

Algorithm-

• Take a sample database of columns and rows called Pima Indian Diabetes data set.

• Take a set of test data for attributes and lines.

• Find the Euclidean distance with the help of a formula-

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

• Then, Determine the random value of K. number. to nearby neighbors

• Then with the help of these small distances and the Euclidean distance find the column for each nth.

• Find the same output values. If the values are the same, the patient has diabetes, otherwise it is not.

3) Decision Tree-The decision tree is the basic method of differentiation. It is a supervised learning method. Decision tree used where response variability is category. Decision tree has a tree-based model that describes the classification process based on the input factor. Variables include input of any type such as graph, text, obscure, continuous etc. Steps to the Decision Tree.

Algorithm-

• Build a tree with notes as an input element.

• Select a feature to predict output from the input benefit feature with advanced information.

• The maximum information gain is calculated for each attribute in each tree area.

• Repeat step 2 to create a thin thread using a feature that can be applied to the surface

4) Logistic Regression-Moving backwards is also an algorithm for dividing supervised learning. It is used to measure the probability of a binary response based on one or more predictions. They can be continuous or separate. Annotation is used when we want to classify or classify other data objects into categories. Separating data in a binary way only means 0 and 1 when referring to the condition of distinguishing a good or bad patient from sugar.

The main purpose of the retrospective is a positive balance that is responsible for defining the relationship between the target and the volatility of the forecast. The reversal of objects is based on the line reversal of the line. The retrospective model uses the sigmoid function to predict the chances of a positive and negative class.

Sigmoid function P = 1/1 + e - (a + bx) Here P = probability, a and b = Model parameter.

Ensembling - Ensembling is a machine learning method Integration means using multiple learning algorithms together to do a specific task. It offers better predictions than any other model which is why it is used. The main cause of error is bias and noise variation, the combination methods help to reduce or minimize these errors. There are two popular ways to combine - Wrap, Encourage, ada-boosting, Gradient boosting, voting, rating etc. Here In this activity we have used Bagging (random forest) and Gradient integration techniques for predicting diabetes.

5) Random Forest – It is a form of integrated learning and is used for classification and retrospective activities. The accuracy it offers is on a grater and compared to other models. This method can easily handle large data sets. The random forest was built by Leo Bremen. It is an integrated learning approach. The Random Forest Improves the Function of the Decision Tree by minimizing variability. It works by building a pile of decision trees during training and issuing a phase which is a mode of division or division or direct prediction (reversal) of each tree.

Algorithm-

• The first step is to select the "R" elements from all the "m" features when R << M.

• Among the "R" features, the area using the best separating point.

• Divide a node into sub nodes using the best partitions.

• Repeat a to c until "l" is the number of nodes reached.

• Build a forest by repeating steps from a to d to get the "a" number of times to build an "n" number of trees.

The Random Forest finds the best classification using the Gin-Index

Cost Work provided by:

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

The first step is to carefully consider the selection and use of the bases for each randomly generated decision tree to predict the outcome and maintain the expected outcome from time to time in the targeted area. Second, count the votes for each predicted target and finally, accept the highly voted target as the result of a complete prediction from a random forest formula. Some of the Random Forest options make appropriate predictions as to the outcome of a given application.

6) Gradient Boosting -Gradient Boosting is a powerful integration method used for forecasting and is a split method. It brings the church reader together in order make strong student models for prediction. It uses the Decision Tree model. separates complex data sets and is a very efficient and popular method. In the gradient improvement model the performance is improved by repetition.

Algorithm-

• Consider a sample of target values such as P

• Measure the error in target values.

• Review and adjust weights to reduce M-error.

• P [x] = p [x] + alpha M [x]

• Model students are analyzed and calculated for the loss function F

• Repeat the steps until you wish and the target result is P.

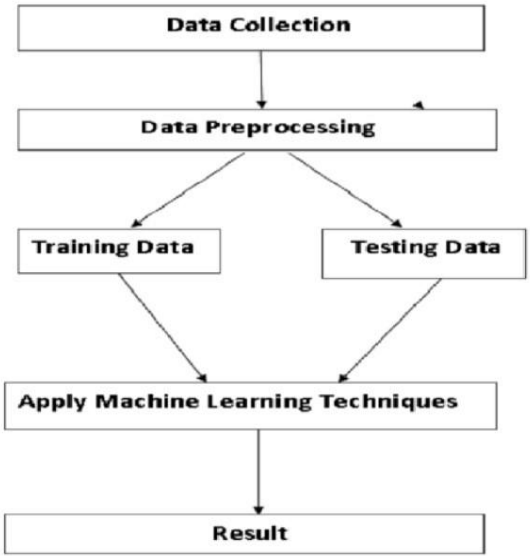

Figure 2: Overview of the Process

4. MODEL BUILDING

This is the most important stage in the development of a diabetic prediction model. In this case we have used the machine learning algorithms discussed above to predict diabetes.

Procedure of Proposed Methodology-

Step 1: Enter the required libraries, Enter the diabetic database.

Step 2: Pre-delete data to delete missing data.

Step 3: Perform an 80% split to split the database as a Training set and 20% create a test set.

Step 4: Choose a machine learning algorithm namely K Close Neighbor, Vector Support Machine, Decision Tree, Backbone, Random Forest and Gradient development algorithm.

Step 5: Create a differentiated model of machine learning algorithm based on the training set.

Step 6: Test the Classifier model of machine learning algorithm based on the test set.

Step 7: Perform an Assessment Comparison of the psychological performance results obtained by each student.

Step 8: After analyzing based on the various steps combine an excellent algorithm

5. EXPERIMENTAL RESULTS

In this process various steps were taken. The proposed method uses different methods of partitioning and merging

and is implemented using python. These methods are the most common mechanical learning methods used to obtain the best accuracy of data. In this work we see that the informal forest class benefits best compared to others. Overall we have used the best Machine Learning methods to predict and achieve the highest accuracy of performance. The diagram shows the effect of these Machine Learning methods.

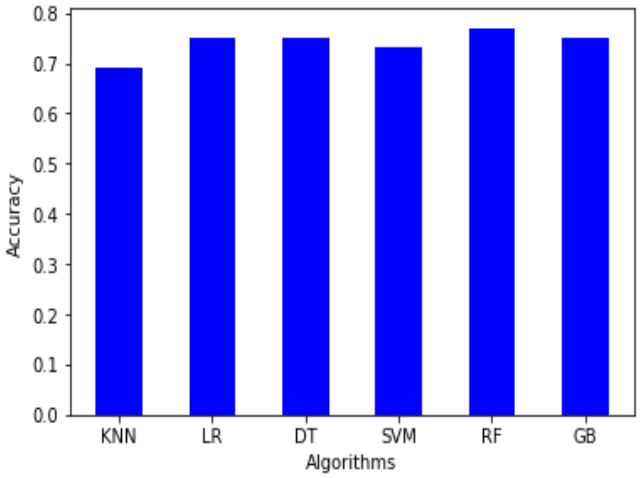

Figure3: Accuracy Result of Machine learning methods

Here is a feature that has played an important role in predicting the introduction of a random forest algorithm. The total value of each factor that plays a major role in diabetes has been determined, where the X-axis represents the value of each factor and feature and Y-Axis feature names.

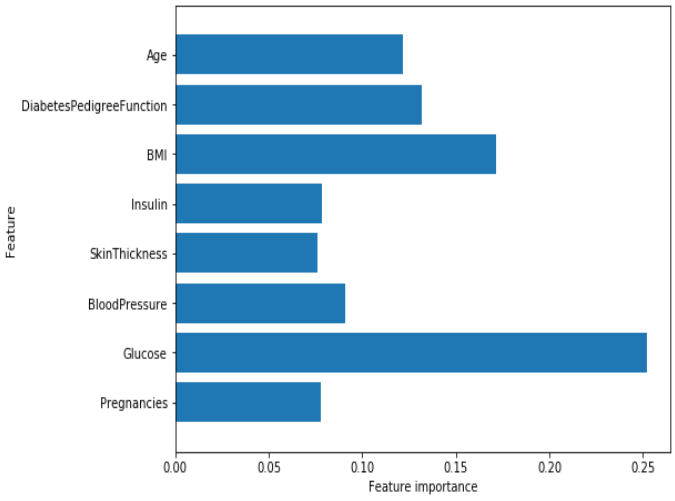

Figure 4: Feature Importance Plot for Random Forest

6. CONCLUSION

The main objective of this project was to design and implement Diabetes Predictability using Mechanical Learning Methods and Performance Analysis of those methods and successfully achieved. The proposed method uses a variety of classification and integration learning methods where SVM, Knn, Random Forest, Decision Tree, Logistic Regression and Gradient Boosting divisions are used. And 77% stage accuracy was achieved. Test Results can be astronomical health care to take early predictions and make early decisions to treat diabetes and save lives.

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 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.