DIABETES PREDICTION

Abstract:- Diabetes is an illness caused because of high glucose levels in the human body. Diabetes recently affects around 442 million people. It is also the major cause for heart stroke, kidney failure, lower limb amputation and blindness.

Diabetes should not be ignored because if it is untreated then Diabetes may cause some major issues in a person like: heart related problems, kidney problem, blood pressure, eye damage and it can also affect other organs of human body. One third go undetected in the early stage. Diabetes can be controlled if it is predicted earlier.

To achieve this goal this project work we will do early prediction of Diabetes in a human body or a patient for a higher accuracy through applying various Machine Learning Techniques. Machine learning techniques provide better results for prediction by constructing models from data sets collected from patients. In this work we will use Machine Learning Classification and ensemble techniques on a dataset to predict diabetes, which include K-Nearest Neighbor (KNN), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Naive bayes(NB) and Random Forest (RF). The accuracy is different for every model when compared to other models.

The Project work gives the accurate or higher accuracy model which shows that the model is capable of predicting diabetes effectively.

Our Result shows that the Support vector machine achieved higher accuracy compared to other machine learning techniques.

INTRODUCTION

Diabetes is a noxious disease in the world. Diabetes caused because of obesity or high blood glucose level, and so forth. It affects the hormone insulin, resulting in abnormal metabolism of crabs and improves the level of sugar in the blood. Diabetes occurs when the body does not make enough insulin.

According to (WHO) World Health Organisation about 422 million people suffer from diabetes particularly from low or idle income countries. And this could be increased to 490 billion up to the year of 2030. However prevalence of diabetes is found among various Countries like Canada, China, and India etc. Population of India is now more than 100 million so the actual number of diabetics in India is 40 million.

Diabetes is the major cause of death in the world. Early prediction of disease like diabetes can be controlled and save human life. To accomplish this, this work explores prediction of diabetes by taking various attributes related to diabetes disease. For this purpose we use the Pima Indian Diabetes Dataset, we apply various Machine Learning classification and ensemble techniques to predict diabetes.

Machine Learning is a method that is used to train computers or machines explicitly. Various Machine Learning Techniques provide efficient results to collect knowledge by building various classification and ensemble models from collected dataset. Such collected data can be useful to predict diabetes. Various techniques of Machine Learning are capable of prediction, however it's tough to choose the best technique. Thus for this purpose we apply popular classification and ensemble methods on the dataset for prediction.

LITERATURE REVIEW:

K.VijayaKumar et al. proposed a 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 Nnamoko et al. presented predicting diabetes onset: an ensemble supervised learning approach they used five widely used classifiers that 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.

Tejas N. Joshi et al. presented diabetes prediction using machine learning techniques aimed to predict diabetes via three different supervised machine learning methods including: SVM, Logistic regression, ANN. This project proposes an effective technique for earlier detection of diabetes disease.

Deeraj Shetty et al. proposed diabetes disease prediction using data mining assemble Intelligent diabetes disease prediction system that gives analysis of diabetes malady utilising diabetes patients database. In this system, they propose the use of algorithms like Bayesian and KNN (K-Nearest Neighbor) to apply on diabetes patients' databases and analyse them by taking various attributes of diabetes for prediction of diabetes disease.

Muhammad Azeem Sarwar et al. proposed 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. Comparison of the different machine learning techniques used in this study reveals which algorithm is best suited for prediction of diabetes. Diabetes Prediction is becoming the area of interest for researchers in order 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.

PROPOSED METHODOLOGY:

Goal of the paper is to investigate for model to predict dia- betes with better accuracy. We experimented with different classification and ensemble algorithms to predict diabetes. In the following, we briefly discuss the phase.

- Dataset Description- the data is gathered from UCI repository which is named as Pima Indian Diabetes Dataset. The dataset has many attributes of 768 patients.
  Table 1: Dataset Description

Sl. 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 the class variable of each data point. This class variable shows the outcome 0 and 1 for diabetics which indicates positive or negative for diabetics.
Distribution of Diabetic patients- We made a model to predict diabetes however the dataset was slightly imbalanced having around 500 classes labelled as 0 means negative means no diabetes and 268 labelled as 1 means positive means diabetic.

- Data Preprocessing- Data preprocessing is the most important process.
  Most healthcare related data contains missing value and other impurities that can cause effectiveness of data. To improve quality and effectiveness obtained after the mining process, Data preprocessing is done.
  To use Machine Learning Techniques on the dataset effectively, this process is essential for accurate results and successful prediction. For the Pima Indian diabetes dataset we need to perform pre-processing in two steps.

  Missing Values removal- Remove all the instances that have zero (0) as worth. Having zero as worth is not possible. Therefore this instance is eliminated. Through eliminating irrelevant features/instances we make feature subset and this process is called features subset selection, which reduces dimensionality of data and helps to work faster.
  Splitting of data- After cleaning the data, data is normalised in training and testing the model. When data is spitted then we train algorithms on the training data set and keep test data set aside. This training process will produce the training model based on logic and algorithms and values of the feature in training data. Basically the aim of normalisation is to bring all the attributes under the same scale.

- Apply Machine Learning- When data has been ready we apply Machine Learning Technique. We use different classification and ensemble techniques to predict diabetes. These methods are applied on the Pima Indians diabetes dataset. Main objective to apply Machine Learning Techniques is to analyse the performance of these methods and find accuracy of them, and also be able to figure out the responsible/important features which play a major role in prediction.
  The Techniques are follows-

  Support Vector Machine- Support Vector Machine also known as svm is a supervised machine learning algorithm. Svm is the most popular classification technique. Svm creates a hyperplane that separates two classes. It can create a hyperplane or set of hyperplanes in high dimensional space. This hyperplane can be used for classification or regression also. Svm differentiates instances in specific classes and can also classify the entities which are not supported by data. Separation is done by a hyperplane that performs the separation to the closest training point of any class.
  Algorithm-

  - Select the hyper plane which divides the class better.
  - To find the better hyper plane you have to calculate 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 highest margin.
    Margin = distance to positive point + Distance to negative point.

  K-Nearest Neighbor – KNN is also a supervised machine learning algorithm. KNN helps to solve both the classification and regression problems. KNN is a lazy prediction technique. KNN assumes that similar things are near to each other. Many times data points which are similar are very near to each other. KNN helps to group new work based on similarity measures. KNN algorithms record all the records and classify them according to their similarity measure. For finding the distance between the points, it uses a tree-like structure. To make a prediction for a new data point, the algorithm finds the closest data points in the training data set its nearest neighbors. Here K= Number of nearby neighbors. It is always a positive integer. The Neighbor's value is chosen from the set of classes. Closeness is mainly de-

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:-
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 formula.
- Then, Decide a random value of K which 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 the same, then the patient is diabetic, otherwise not.

Decision Tree- Decision tree is a basic classification method. It is a supervised learning method. Decision tree is used when the response variable is categorical. Decision tree has a tree-like structure based model which describes the classification process based on input features. Input variables are any types like graph, text, discrete, continuous etc. Steps for *Decision Tree Algorithm*-

- Construct a tree with nodes as input features.
- Select feature to predict the output from input feature whose information gain is highest.
- The highest information gain is calculated for each attribute in each node of the tree.
- Repeat step 2 to form a subtree using the feature which is not used in the above node.

Logistic Regression- Logistic regression is also a supervised learning classification algorithm. It is used to estimate the probability of a binary response based on one or more predictors. They can be continuous or discrete. Logistic regression used when we want to classify or distinguish some data items into categories.
It classifies the data in binary form only in 0 and 1 which refer to the case to classify a patient that is positive or negative for diabetes.
Main aim of logistic regression is to best fit which is responsible for describing the relationship between target and predictor variables. Logistic regression is based on a Linear regression model. Logistic regression model uses sigmoid function to predict probability of positive and negative class.
Sigmoid function P = 1/1+e – (a+bx) Here P = probability, a and b = parameter of Model.
Ensembling- Ensembling is a machine learning technique Ensemble means using multiple learning algorithms together for some task. It provides better prediction than any other individual model that's why it is used. The main cause of error is noise bias and variance, ensemble methods help to reduce or minimise these errors. There are two popular ensemble methods such as Bagging, Boosting, ada-boosting, Gradient boosting, voting, averaging etc. Here in these work we have used Bagging (Random forest) and Gradient boosting ensemble methods for predicting diabetes.

Random Forest- It is a type of ensemble learning method and also used for classification and regression tasks. The accuracy it gives is greater than compared to other models. This method can easily handle large datasets. Ran- dom Forest is developed by Leo Bremen. It is a popular en- semble Learning Method. Random Forest improves Performance of Decision Tree by reducing variance. It operates by constructing a multitude of decision trees at training time and outputs the class that is the mode of the classes or classification or mean prediction (regression) of the 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 uses the best split point.
- Split the node into sub nodes using the best split.
- Repeat a to c steps until the number of nodes has been reached.
- Built forest by repeating steps a to d for a num- ber of times to create a number of trees.
  The random forest finds the best split using the Gin-Index Cost Function which is given by:
  The first step is to take a glance at choices and use the foundations of each indiscriminately created decision tree

to predict the result and store the anticipated outcome at intervals at the target place. Secondly, calculate the votes for each predicted target and ultimately, admit the high voted predicted target as a result of the ultimate prediction from the random forest formula. Some of the options of Random Forest do correct predictions results for a spread of applications are offered.

MODEL BUILDING:

This is most important phase which includes model building for prediction of diabetes. In this we have implemented various machine learning algorithms which are discussed above for diabetes prediction.

Procedure of Proposed Methodology-

Step1: Import required libraries, Import diabetes dataset.

Step2: Pre-process data to remove missing data.

Step3: Perform a percentage split of 80% to divide the dataset as Training set and 20% to Test set.

Step4: Select the machine learning algorithm i.e. K- Nearest Neighbor, Support Vector Machine, Decision Tree, Logistic regression, and Random Forest algorithm.

Step5: Build the classifier model for the mentioned machine learning algorithm based on the training set.

Step6: Test the Classifier model for the mentioned machine learning algorithm based on the test set.

Step7: Perform Comparison Evaluation of the experimental performance results obtained for each classifier.

Step8: After analysing based on various measures, conclude the best performing algorithm.

EXPERIMENTAL RESULTS:

In this work different steps were taken. The proposed approach uses different classification and ensemble methods and is implemented using python. These methods are standard Machine Learning methods used to obtain the best accuracy from data.

In this work, we see that support vector machine classifiers achieve better compared to others. Overall we have used the best Machine Learning techniques for prediction and to achieve high performance accuracy.

Here, features that play an important role in prediction are presented to support vector machine algorithms. The sum of the importance of each feature playing a major role for diabetes have been plotted, where X-axis represents the importance of each feature and Y-Axis the names of the features.

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 methods in SVM, KNN, Random Forest, Decision Tree, and Logistic Regression. classifiers are used. And 77% classification accuracy has been achieved. The Experimental results can assist health care to take early prediction and make early decision to cure diabetes and save humans life.