DIABETES PREDICTION USING MECHINE LEARNING

Varghese Thomas[1], Tintu Varghese[2], Adarsh Santhosh[3], Adithya Suresh[4], Sreya James[5]
[1,3,4,5]PG- Master of Computer Application, Kristu Jyoti College of Management and Technology, Changanacherry, Kerala
[2]Assistant professor- Kristu Jyoti College of Management and Technology, Changanacherry, Kerala

ABSTRACT

Diabetes, sometimes known as diabetes mellitus or just diabetes, is a condition brought on by elevated blood glucose levels. The diagnosis of diabetes can be made using a variety of conventional techniques based on physical and chemical examinations. However, it can be difficult for medical professionals to predict diabetes early.due to the complicated interconnectedness of numerous factors, including how diabetes affects people,kidney, eye, heart, nerves, foot, and so forth. Methods of data science have the potential toaid other scientific disciplines by providing fresh insights into established issues. Among them is toassist in forecasting using medical data. Datascience has an emerging topic called machine learning that studies how machines learn from experience. The purpose of this research is to create a system that can detect diabetes in a patient with a higher risk.by merging the outcomes of many machine learning methods. Goals of this projectusing three different supervised machine learning techniques, such as SVM, ANN and logistic regression (Artificial neural network). This also seeks to suggest a successful a method for diabetic illness early diagnosis.

Keywords: Diabetes mellitus, K-nearest neighbour, Machine learning, Euclidean distance, Gradient Boosting,

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 is unable to produce sufficient insulin The World Health Organization estimates that 422 million individuals worldwide. suffering from diabetes, especially in low-income or developing countries. Also, this might be Up until the year 2030, the number climbed to 490 billion. However, diabetes prevalence is reported among numerous nations, including Canada, China, and India, etc. India today has a population of more than 100.The actual number of people with diabetes in India is 40 million. Diabetes is major cause of death in the world. 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 various Diabetes prediction using machine learning classification and ensemble techniques. Machine Learning is a technique used to explicitly instruct robots or computers. Different Machine Learning methods produce effective knowledge acquisition by developing a variety of from the obtained dataset, classification and ensemble models. Such information gathered may be helpful for diagnose diabetes. Various techniques of Machine Learning can capable to do prediction, however its tough to choose best technique. Thus for this purpose we apply popular classification and ensemble methods on dataset for prediction.

LITERATURE REVIEW

[7] The Hybrid Prediction Model presented by B.M. Patil, R.C. Joshi, and Durga Toshniwal (2010) combines Simple K-means clustering. Algorithm, then applying a classification algorithm on the output of that clustering methodology. Using the decision tree approach, classifiers are built. [3] Nawaz in order to forecast diabetes, Mohamudallyl and Dost Muhammad (2011) used the decision tree algorithm, neural networks, the Kmeans clustering algorithm, and visualisation. [5] The process

of selecting a Feature matching of the data to be learned using existing data is a component of machine learning algorithms approaches. The categorization technique was utilised by K. Rajesh and V. Sangeetha (2012). For effective classification, they employed the decision tree approach to extract hidden patterns from the dataset. [8] Amisha Lyer (2015) investigated latent patterns in the diabetes dataset using classification techniques. Simple Bayes in this model, Decision Trees were employed. Performance comparisons were done for both. As a result, the usefulness of both algorithms was demonstrated.

WORKING MODEL

The aim of this research is to find a model that can more accurately predict diabetes. To forecast diabetes, we tested various classification and ensemble methods. The period is briefly covered in the sections that follow.

➢ **Phases of prediction**

- **Dataset Description**

Datasets are groups of data. In the case of tabular data, a data set relates to one or more database tables, where each row refers to a specific record in the corresponding data set and each column to a specific variable. Each value in the data collection is listed of the factors for each individual in the data collection, such as an object's height and weight. Each A datum is a measure of value. A group of papers or files can also make up a data set.

- **Data Preprocessing**

Data preprocessing is a procedure that is used to transform the raw data into a clean data set. It is the process of transforming or encoding data into a form that a machine can easily parse. Data preprocessing's primary function in the learning process is to delete the unnecessary information and fill in the missing value. To enable machine assistance easily trained the most crucial phase is data preprocessing. primarily health-related information contains missing values and other contaminants that may affect the usefulness of the data. To enhance data preparation is done after the mining process to ensure quality and effectiveness. The approach is crucial for accurate results and good prediction when using machine learning techniques to the dataset.

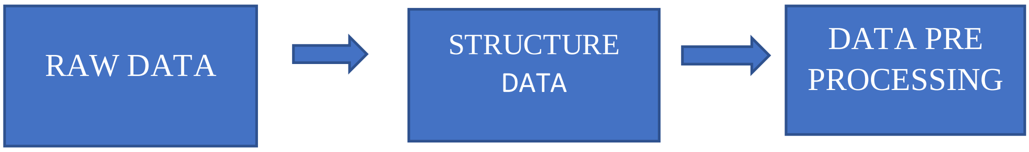
For Pima Indian diabetes data set we need to perform preprocessing in two steps:

➢ Missing Values removal

Eliminate all occurrences with a value of 0 (zero). A value of zero is not conceivable. As a result, this instance is stopped. The method of selecting feature subsets is known as features subset selection, and it decreases the number of irrelevant features and instances. Dimensionality of the data and speed up work.

➢ Splitting of data

Data is standardised for the model's training and testing once it has been cleaned. After the data is spilt, we train the algorithm on the training data set while putting the test data aside. As a result of this training process, a training model based on logic, algorithms, and values of the characteristic in training data. The primary goal of normalisation is to group all attributes similar scale.



Apply Machine Learning

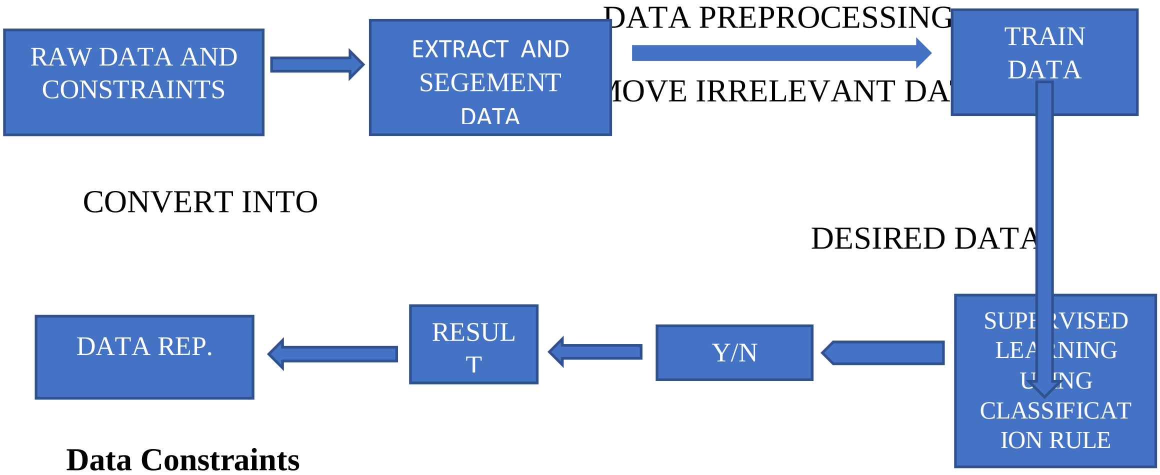
When the data is ready, machine learning techniques are used. To predict diabetes, we employ a variety of classification and ensemble algorithms. The procedures used on the diabetes dataset for Pima Indians. primarily to analyse the data using machine learning techniques performance of various techniques, confirm their accuracy, and determine the relevant or significant characteristics that are crucial to prediction.

Various Machine Learning Techniques:

- Support Vector Machine
- K-Nearest Neighbour
- Decision Tree
- Random Forest
- Gradient Boosting

PROPOSED METHODOLOGY

The Proposed method use KNN algorithm for classification and prediction of diabetes using trained data.



Data Constraints

Data is a collection global dataset. IN this system use Pima Indian data set is used for training a model. Data set contain 21 parameters and around 1000 dataset.

The dataset feature/parameters are:

- Age
- Gender
- Relation
- DOB
- Sugar tested value
- Symptoms
- Family history etc.

Train Dataset and Test Dataset

The initial collection of data utilised to understand the software is the training data. In order to set the feature and use the system-available data, we must first train the model in this case. This information is used to train the machine to perform certain tasks. It is the data that the model can learn from using an algorithm to train it and do tasks automatically. A software receives the test data as input. It demonstrates how the data is impacted by the action when this module, which specifies, is primarily used for testing.

ML Algorithm: K- NEAREST NEIGHBOUR

Thomas Cover's non-parametric method, called the k-nearest neighbour algorithm, is used in machine learning for classification and regression. This method is primarily used in the industry to categorise difficulties. An example of instance-based learning is the KNN algorithm method. This algorithm uses distance to classify objects and learn from training data. Its accuracy is greatly improved by normalising. the set of neighbours is the source of things whose values for class or object property are known. It can be viewed as training even though no explicit training steps are necessary, a set for the method 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. All records are captured by the KNN algorithm, which classifies them based on according to how similar they are. Uses a tree-like structure to determine the distance between the spots.

The technique locates the closest data points in the distribution to a new data point to produce a prediction. The nearest neighbours of the training data set. K stands for the number of immediate neighbours, which is always a positive integer. Value for neighbours is selected from a list of classes. Most definitions of proximity use the Euclidean concept of distance.

The Euclidean distance between two points X and Y i.e. X (xl,x2, . xn) and Y (y 1, y2,..yn) is defined by the following equation:

Euclidean distance = N(YI Xl)2 + (x2)2 (yn xn)2

_Algorithm

- Take a test dataset of attributes and rows.
- Find the Euclidean distance between the points.
- Arrange the calculated n Euclidean distances in non-decreasing order.
- Let k be a +ve integer, take the first k distances from this sorted list.
- Find those k-points corresponding to these k-distances.
- Let ki denotes the number of points belonging to the ith class among k points i.e. k>=0.
- If ki> kj,Vi not equal j, then put x in class i

Choosing the right value for k

We run the KNN algorithm numerous times with various values of K and choose the K that minimises the number of errors we experience while preserving the quality of the results to find the K that's appropriate for your data. the algorithm's capacity to correctly forecast outcomes when presented with novel data.

- As we decrease the value of K to 1, our predictions become less stable.
- Inversely, as we increase the value of K, our predictions become more stable due to majority voting and thus more likely to make more accurate predictions.
- In cases where we are taking a majority vote among labels, we usually make K an odd number to have a tiebreaker

MODEL BUILDING

This is the most crucial stage, during which a model for diabetes prediction is built. For predicting diabetes, we have used a variety of machine learning methods that were presented above.

Procedure

Step1: Import required libraries, Import diabetes dataset.

Step2: Pre-process data to remove missing data.

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

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

Step5: Build the classifier model for the mentioned ma- chine learning algorithm based on training set.

Step6: Test the Classifier model for the mentioned ma- chine learning algorithm based on 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

Different actions were taken in this work. These techniques are common machine learning techniques used to get the highest level of accuracy from data. Overall, to make predictions and achieve high performance accuracy, we applied the best machine learning approaches.

The main advantages of using this is the early prediction of diabetes and receive treatment. No human intervention is needed that means, its all about machine and machine learning, so no human intervention is needed during the entire process.

This system also havesome drawbacks they are: Data acquisition, Time and resources, Interpretation of result.Data acquisition: mechine learning requires massive data set to train on, and this should be inclusive and good quality.Time and resources: mechine learning needs enough time to let the algorithm learns and develop enough to fullfill with a considerable amound of accuracy. It also need massive resources to function. Interpretation of results: another challenge is the ability to accurately Interpret the result generated by the alogorithm

CONCLUSION

In light of its serious implications, diabetes prognosis is crucial in the current environment. Diabetes is one of the leading causes of death worldwide. The primary goal of this class project is to identify diabetes using a machine learning system, specifically nearest neighbour in K. Physicians can predict diabetes early on thanks to the system. So,that patients can receive traditional treatments and remedies. This system assists in obtain more accurate results. Numerous studies have been done on the diabetic imprint.For hospitals and doctors, developing a diabetes illness prediction system is useful. KNN aids in early disease prediction so that patients can receive better care from their doctors. The suggested model is the real-time application that is designed for various hospitals and less accurately forecasts disease time. We will get more accurate as we employ machine learning algorithms to anticipate diseases and effective outcomes.

REFERENCES

1. Iqbal H. Sarker, 'Performance Analysis of Machine Learning Techniques". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019
2. Rahul Joshi, "Analysis and prediction of diabetes diseases using machine learmng algorithm"•lnternational Research Journal of Engineering and Technology Volume: 04 Issue: 10 Oct -2017.
3. 3. Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabetes Disease Prediction Using Data Mining ". lnternational Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017.