Diabetes Prediction Using Machine Learning Algorithms

Melbin Varghese
Department of Computer Applications
Amal Jyothi College of Engineering, koovapally
Kottayam, Kerala.
Melbinvarghese2021@mca.ajce.in

Mrs. Nimmy Francis
Asst. Professor in Computer Science
Amal Jyothi College of Engineering, koovapally
Kottayam, Kerala
nimmyfrancis@amaljyothi.ac.in

***Abstract*—Diabetes mellitus also known as diabetes is a metabolic disorder that is characterized by a high level of blood sugar over a long period of time. It is a disease that occurs when the blood glucose, also known as blood sugar, content is too high in the blood. Blood glucose is the source of energy and it is made from the food we eat. Pancreas make insulin which helps in breaking down the protein and carbohydrates from the food and convert into glucose and energy for the body. This insulin is used to break down this glucose and convert into energy for the body. Sometimes the body doesn't make enough insulin to work on the glucose. In such situations the glucose doesn't get processed and it just stays in the blood stream. Over time, the presence of a lot of glucose in the blood stream can cause chronic health issues. Generally diabetes has no actual cure but certain steps can be taken to keep diabetes in check and stay healthy. This paper discusses about the applications of Machine Learning in prediction and understanding the diabetes in people.**

***Keywords* — machine learning, KNN Algorithm, logistic regression, Diabetes prediction analysis, report**

I. INTRODUCTION

Diabetes is a silent killer. It is one of the most feared modern disease that is killing more people in the current times only second to the heart related diseases. And the number of the diabetic people are on the rise every day. More people are prone to be a diabetic patient than any other diseases. The number of the diabetic patients are on the rise. And it is rising at an alarming rate. The bad thing about diabetes is that it is not found out early into the development of the condition. The main reason for diabetes is the hereditary diabetes and life style related diabetes. Hereditary diabetes is passed on from parents to children. It is inherited. Life style diabetes is caused by the unhealthy lifestyle the people are leading. In todays fast paced world, the life style choices are not helping people with their health standard. The food which is fast food nowadays is a main factor in the rise of diabetes in young people. The diabetics is a causing factor for many of other diseases like heart disease, kidney issues, nerve damage, stroke, thyroid etc. Type 1 & 2 diabetes along with gestational diabetes are the main types of diabetes. With the type 1 diabetes the body doesn't actually make the insulin that is actually required. The cells in the pancreas that actually generates the insulin that is required by the body is actually attacked and destroyed by the immune system of the body. Children and young adults are usually diagnosed with Type 1 diabetes, although it can appear at any age. The insulin needs to be taken by patients with Type 1 diabetes every day in order to stay alive..[1] With type 2 diabetes your body does not make or use insulin well. Type 2 diabetes can be developed at any age, even during childhood. The most common type of diabetes that is seen in most people is the Type 2 diabetes[2].Gestational diabetes is the diabetes that is developed in some women when they are pregnant. This diabetes usually goes away after the baby is delivered[3]. But the problem is that this diabetes can develop type 2 diabetes later in life.

II. MACHINE LEARNING ALGORITHMS

Machine learning is the technique of analyzing the data that automates analytical model building. It is related to artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with very minimum human interaction. With the advent of new computing technologies, the present day machine learning is totally different from how it was in its inception. Machine learning was made from the pattern recognition methodology and from the theory that machine learning is possible without being programmed for specific tasks. The most important aspect of machine learning is the iterative aspect as the data models are exposed to new data they adapt independently. Previous computations are learned to make reliable, reputable results. Machine learning has gained a fresh pace as of lately. Machine learning is a study of artificial intelligence with computer science focusing on the data and algorithms to follow the ways humans learn improving its accuracy. The medical field is using the advanced machine learning algorithms to make predictions on the historic dataset. With the advancements made in machine learning methodology

DOI: 10.5281/zenodo.5109314

detection and prediction of diabetes can be done with ease and with less effort from human intervention. Machine learning will be a necessity in efficiently diagnosing not just diabetes but also other medical conditions. Because of the availability of many dataset related to different diabetes types are available the machine learning algorithms could be used efficiently to reduce the strain on humans. There is basically two types of machine learning algorithms that can be employed to predict diabetes through our research -

1. Supervised learning
2. Unsupervised Learning

In Supervised Learning, an algorithm is made to learn to map an input to a certain output. That is done on labelled datasets that have been collected over the course of time. The algorithm is learned successfully if the mapping is done correct. If the mapping is not done correct then required alterations can be done to the algorithm in order to correctly learn. Trained data models made with Supervised Learning algorithms can be used for predictions on new data that can be collected in the future. Supervised Learning gives experience to the algorithm which can be used to predict the outputs for new unseen data. Experience helps in optimizing the performance of the algorithm. Supervised Learning is classified into 2 types – Regression and Classification. Regression is when the algorithm learns from the labelled datasets and then a continuous-valued output is predicted for the newly given dataset. It is used whenever a number is required as on output. Linear regression and Logistic regression are examples of regression algorithm.

Classification is the type of learning where the algorithm maps new data to any one of the two classes in our dataset. The classes can be 1 or 0 and 'Yes' or 'NO'. Decision Tree, Naïve Bayes Classifier, Support Vector Machines etc are examples of Classification algorithm.

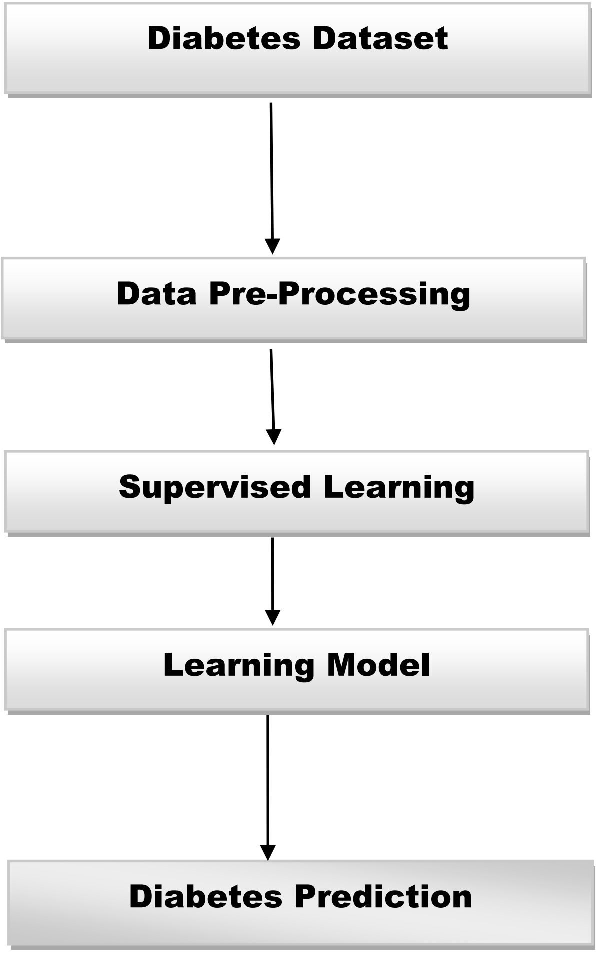
In unsupervised learning the data consist of values without any labels and the output is not pre-determined. The model predicts on the basis of self-learning. The main purpose of these models is to predict, Classify, detect, segmentation and The most common use of machine learning is analysis, recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

III. MOTIVATION

The alarming amount of increase in the number of the diabetes patients around the world should be a concern for everyone. The technology especially smart computing should play its role in predicting, identifying and classifying diabetes among people. This can drastically people in the medical field and medical research on understanding and identifying diabetes among people. With such a system the healthcare professionals can have the time required to analyze and identify the diabetes in the people be drastically reduced. This system can also be used to provide verification to the medically collected data of the patients.

IV. PROPOSED METHODOLOGY

The purpose of this paper is find a Machine Learning model which can predict whether the patient has diabetes with accuracy from the given dataset. The model should be able to classify correctly the dataset into Diabetic and Non-Diabetic groups. The given datasets will be bifurcated into training dataset and testing data sets. A dataset with limited data cannot provide accuracy with learning so we need to train the model with more data in the dataset. The results obtained from this algorithm models can be analyzed to create a diabetes classification. If the classification model can detect a diabetic entry the value will be set to 1 else the value will be 0.The models like Logistic Regression, KNN, Random Forest and gradient boosting classifier will be used to learn to detect diabetes. A trained dataset can be used on these models to predict whether a person is diabetic or not. The structure of this study will be as follows:



DOI: 10.5281/zenodo.5109314

Figure 1 : Diabetes prediction methodology

The steps we need to follow are:

1. Data collection
2. Defining data
3. Pre-processing
4. Building model
5. Analysis
6. Results
7.

With the algorithm we follow the below steps :

1. Import the libraries
2. Import the dataset
3. Define the dataset
4. Dataset training and testing
5. Algorithm execution
6. Results comparison and evaluation.

The dataset we use in this study is the Pima Indian Diabetes Dataset from UCI repository. The data in the dataset are as follows.

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

Figure 2: Description of the dataset

The next attribute is the class variable for each of the attribute in the dataset. This attribute represents the values 0 and 1 which indicates the presence of diabetes in the patients.The dataset that we currently have is imbalanced because of all the entries available in the dataset around 500 out of the total entries in dataset have classes with 0 values implying no diabetes with just 268 labelled as 1 with positive diabetic entries.

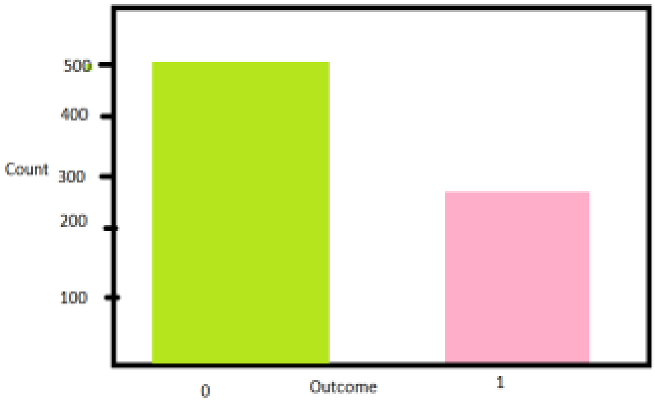


Figure 3: Count of diabetic patients

A. Data Processing: This is the most vital process in the whole machine learning methodology. The missing data and the impurities in the dataset can reduce the quality of the output that can be generated from the dataset. Data preprocessing can be performed in order to increase the effectiveness of the data that is received after the data processing technique. On our dataset we can perform data preprocessing by following the below methods

1).Removing the missing values – All the entries that have 0 as an entry needs to be removed. Having) is not a valid instance. So all the entries with 0 needs to be removed. We create a feature subset by removing all the irrelevant entries.

2).Data Splitting – Once the data is cleaned, it needs to normalized for both the training and test models. Once the split data is available, the training data set is used to train the algorithm. A training model based on the features of the training is created after the training process.

B. Machine Learning: Once the data is ready we exercise Machine Learning process. We apply various classification and algorithms to prognosticate diabetes prone patients. The performance of these methods are probed to find their accuracy and identify the major features which can help us in our augury of diabetes. The following techniques can be used –

1).Logistic Regression - Logistic regression is the quotidian sort of machine learning algorithm used. It has got a high level of accuracy and is usually very definitive. 0s and 1s are the results that are usually noted with logistic regression. It is normally used when data needs to be categorized. Sigmoid function is used in Logistic regression to vaticinate the probability of valid and invalid class.

Sigmoid function P = 1/1+e-(a+bx)

Here P = probability, a and b are model parameters.

2). Support Vector Machine - It is also called as Supervised machine learning algorithm. It creates a hyperplane that separates different classes. This hyperplane is also used for categorizing and also regression. It can specify entries in particular classes and also categorize the instances which are not supported by the data.

Algorithm:

- The hyper plane which divides the class the best is selected.
- The distance between the planes and the data called the Margin is calculated. This is used to find the better plane.
- Higher the distance between the classes, lower the chance of miss conception.
- We need to select the class with high margin.

  Margin = Distance to negative point + Distance to positive point.

3). K – Nearest Neighbor - It is also known as the KNN algorithm. It is a supervised machine learning algorithm. It works on the principle that items of same attributes stay near to each other. The idea of similarity measure is used to group a new work. It notes the records and categorize them on the basis of their similarity measure. The algorithm finds the nearest data points in the training dataset to prognosticate a new data point. K is the positive number of nearest neighbors. The distance between the neighbors is defined in Euclidean distance. The Euclidean distance between 2 points P and Q is defined as:

$$d(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

Algorithm:

1. A sample dataset from the Pima Indian Diabetes data set is taken.
2. A test data set of attributes and rows is taken.
3. The Euclidean distance is found with the following formula

$$d(p,q) = d(q,p) = \sqrt{(q_{1-}p_1)^2 + (q_3 - p_3)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n}(q_i - p_i\ )^2}$$

4. A random value of K is selected.
5. The nth column of each neighbor is found using the Euclidean distance and the minimum distance.
6. The output values is found out.

If the values are the same ,then diabetes is detected if not the person is not diabetic.

4). Random Forest – It is a machine learning algorithm used for classification and regression data. It has got the highest score of accuracy in comparison to other models. Any data set of any size can be easily handled by this method. The performance of the Decision Tree algorithm can be vastly improved by Random Forest tree algorithm by reducing the variance. It functions by creating a host of decision trees while training and outputting the mode of the classes or the regression of the discrete trees.

Algorithm:

1. Select ‘R’ features from ‘M’, total features where R<<M.
2. Find the best node from R features.
3. Split the node into sub nodes using the best split method.
4. Repeat the above steps until ‘l’ number of nodes is reached.
5. Repeat the above steps to built forest for “a” number of times to create “n” number of trees.

5).Decision Tree – It is a supervised learning method which is also a basic classification method. It is used to categorize the response variable. It has a tree like structure. It describes classification process based on the input features.

Algorithm:

1. With the nodes as input feature construct a tree
2. Select the input feature with the highest information gain to predict the output.
3. For each feature in each node of the tree the highest information gain is calculated.
4. Repeat step 2 to form a subtree using the feature which is not used in the above node.

V. BUILD THE MODEL

The model building is a cardinal stage in the prediction of diabetes. In this spell we implement the algorithms we have discussed above.

Procedure –

1. Necessary libraries are imported
2. The diabetes dataset is imported

3. Missing data is removed with the pre-process of the data
4. Split the dataset into training and test data sets in a 8:2 ratio
5. Select any of the algorithms, Random Forest, Logistic regression, Decision Tree, KNN, or the Support Vector Machine algorithm.
6. Based upon the selected algorithm using the training set build a classifier model
7. Evaluate the classifier model on the test set
8. Using the performance values received for each classifier conduct a comparison evaluation
9. Find the algorithm with the best performance based on the obtained performance values.

VI. Results

With our approach we use varied classification and models to implement our methodology using the Python language. By using this methods we are looking for a machine learning algorithm with the highest accuracy. By comparing each working model we can see that the Random Forest classifier has higher overall performance in comparison with other classifiers.

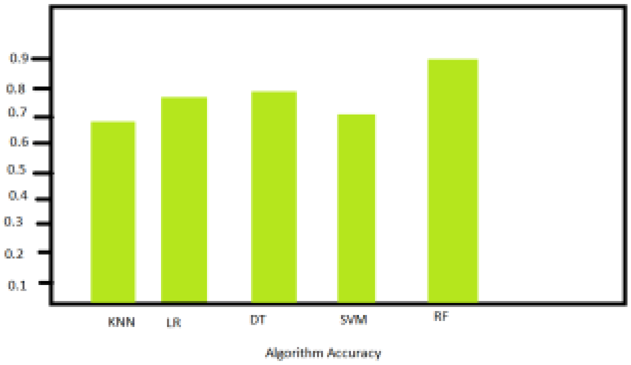


Figure 4: Results of the test

VII. CONCLUSION

With this research we have applied different machine learning algorithm classifiers in order to predict diabetes from the patients data set. We were able to create different classifier models using different machine learning algorithms. With this study we were able to identify the model with the highest accuracy. Of all the models, the Random Forest had the highest accuracy.

VII. References

[1] Shapiro AM, Lakey JR, Ryan EA,et al. Islet transplantation in seven patients with type 1 diabetes mellitus using a glucocorticoid-free immunosuppressive regimen. N.Engl. J. Med. 343, 230-238(2000).

[2] Cheng L, Hammond H, Ye Z, Zhan X, Dravid G. Human adult marrow cells support prolonged expansion of human embryonic stem cells in culture. Stem. Cells. 21, 131142(2003).

[3] Guan LX, Guan H, Li HB, et al. Therapeutic efficacy of umbilical cord-derived mesenchymal stem cells in patients with type 2 diabetes. Exp.Ther. Med.9(5), 1623-1630 (2015)

[4] Lucas Felipe Klein, Sandro José Rigo, Sílvio César Cazella, Ângela Jornada Ben: An Application for Mobile Devices Focused on Clinical Decision Support: Diabetes Mellitus Case. ISAmI 2016: 57-65