IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 3, March 2021

DOI: 10.17148/IARJSET.2021.8316

Analysis and Prediction of Diabetics Disease Using Machine Learning Algorithm

Dr. Prabha R[1], Akshay H.S[2]

Professor, Dept of CS&E, Dr.AIT, Bengaluru, India[1]

M.Tech Student, Dept of CS&E, Dr.AIT, Bengaluru, India[2]

Abstract: In today's world, Machine Learning Techniques (MLT) are used to predict the medical datasets at an early stage to save a human life. Medical datasets are accessible in different data which are used in real world applications. One of the missions is to predict the disease data and analysis. Currently Diabetic Disease (DD) is one among the leading cause of death in the world. Earlier to group and predict symptoms various data mining techniques were used by researchers in different time. The motive of this study is to design a model which can identify the diabetics in patients with maximum accuracy. In this system the most known predictive algorithm applied is random forest algorithm and using this algorithm make an ensemble hybrid model is designed combining individual methods, the performance of the algorithm are evaluated on various parameters precision, accuracy, F-measure and recall.

Keyword: Diabetics, Diseases Prediction, Machine Learning, Random forest Algorithm.

I. INTRODUCTION

The major goal of the classification technique is to forecast the target class accurately for each case in the data [1]. Classification Algorithms generally require that the classes be defined grounded on the data attribute values. Diabetes diseases commonly stated by health professionals or doctors as diabetes mellitus (DM), which describes a set of metabolic diseases of a person. The new proposed study follows the different machine learning techniques (MLTs) to predict the diabetes at an early stage to save human life. Diabetes is not a hereditary disorder in which could result an ultimately boom of glucose within the blood and lack of glucose in urine [2]. High Blood sugar can also increase the kidney diseases and heart illness. The excess of blood sugar can harm the tiny blood vessels. In diabetic person generally suffers from high blood sugar, Intensify thirst, Intensify hunger and frequent urination are some of the symptoms caused due to high blood sugar. In this work the parameters are used are pregnancies, glucose, blood pressure, skin thickness, insulin, BMI diabetes, pedigree function, age, outcome. Pima Indians Diabetics Dataset (PIDD) this dataset is taken from the national institute of diabetes. The objective of the dataset is to predict whether the patient has diabetes or not based on the constraints which were taken from the massive database.

II. LITERATURE SURVEY

The author in [3] proposed a Data Mining Techniques (DMT) that are helpful to predict the disease at an early stage of human life. Machine Learning is based on the prediction of disease data. To cluster and predict the symptoms in medical data, various data mining techniques were used by the researchers in different time and data set is taken from Pima Indian Diabetics Data set which can be accessed from online source and different types of algorithm are used. The work is based on the early prediction of the diabetes to save the life of the human.

Devi et.al., [4] proposed a data mining technique which approaches to help and diagnose patients disease. Diabetes Mellitus is a dangerous disease to affect the various organs of the human body and prediction can save human life and take control over the diseases. The dataset has taken 768 instances taken from PIMA Indian Dataset to determine the accuracy and prediction.

The authors in [5] focused on prediction of diabetes using classification algorithm, diabetes can be considered as the deadliest and chronic diseases which causes an increase in blood sugar. The tedious identifying process results in visiting of a patient to a diagnostic center and consulting doctor. The motive of this study is to design a model in prognosticate of diabetes with maximum accuracy. Three machine learning algorithms are used namely Decision Tree, SVM and Naïve Bayes. The performance of all the three algorithms are evaluated on various measures like Precision, Accuracy, F-Measure and Recall.

The author in [6] which tells about diabetes mellitus is a metabolic disorders and millions of people are affected. Many research studies have done on the diagnosis of diabetes are most research studies were carried out on Pima Indian diabetes data set. The most popular techniques like KNN algorithm used to identify the diabetes and pre-processing of data methods and find the accuracy.

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 3, March 2021

DOI: 10.17148/IARJSET.2021.8316

III. DESIGN ND METHODLOGY

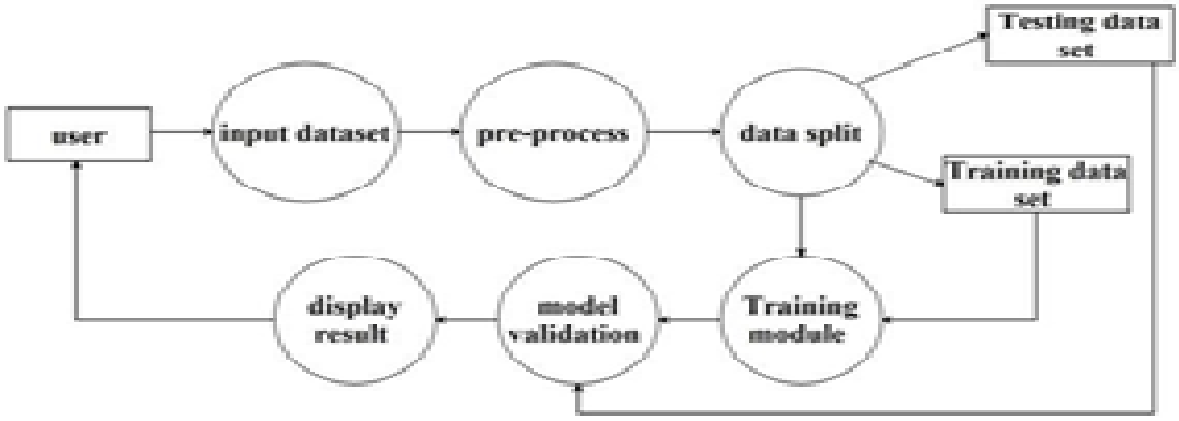

Figure.1 System Design

It is very important to complete all the tasks and meet the deadline given by the user. There are many project tools to help the manager and one of them is data flow diagram. In the data flow diagram mainly four steps involved namely Data acquisition and preprocessing, Feature selection and data preparation, Model construction and Training, Model validation and Result analysis. Data acquisition and preprocessing involves collecting of raw data from online sources in the form of statements and digits. A huge volume of raw data collected need to be separated individually for the prediction of the dataset. Data Preprocessing is a technique that is used to convert raw data into a clean data set. Feature selection and data preparation is a process of using domain knowledge of the data to create features that makes machine learning algorithms work. The dataset is divided into subsets will be used to train a model and testing will be done for the subset to train the model. The model construction is based on the feature engineering, it is a process of transforming the raw data into features and represent the model for the prediction, resulting in the improved accuracy of the data. Testing and Training will be done according to the dataset for the model. Model Validation and Result Analysis is the last step after testing and training the model, be validated to pass real time data for the prediction. Once prediction is done we will analyze the model and correct accuracy is found with the output to find the crucial information.

a) Python.

Python is an interpreted, general-purpose and high level programming language and it is dynamically typed and garbage collected. Python is easy to learn and powerful programming language and it is processed at run time by the interpreter. Python is interactive and it is easy to write a program and python is object oriented that supports the programming which encapsulates the code within objects. Python supports functional and structured programming methods as well as object oriented programming and can be easily integrated with C, C++ and java programming languages [7].

b) Machine Learning.

Machine Learning algorithms explore and influence different algorithms and multi-feature in the time series. The Machine Learning (ML) models have the frameworks which deals with programming and systems to construct like loops and recursion. The issues which are used to solve the algorithms and develop the knowledge by given specific data and past experience are used like, probability, mathematical optimization, logic and statistical science. ML are classified into two types 1) Grouping of algorithms by the use of learning style 2) Grouping of algorithms in the form of similarity function [8]. Machine learning algorithms broadly categorized as supervised and unsupervised Learning [1]. Supervised Learning is the machine learning technique that has a function which maps an input to output based on the pairs from the training examples. Unsupervised Learning is a machine learning technique which does not allow users to supervise the training model, it allows the model to work on its own to discover patterns and information which deals with unlabeled data.

c) Supervised Learning.

Supervised Machine Learning (SML) is a type of learning algorithm that reason from externally supplied instances to produce general hypotheses, which then make predictions about future instances. Supervised classification is one of the tasks most frequently carried out by the intelligent systems. Supervised Machine Learning (ML) classification techniques, compares various supervised learning algorithms as well as determines the most efficient classification algorithm based on the data set, the number of instances and variables (features).It is grouped into two types classification and regression

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 3, March 2021

DOI: 10.17148/IARJSET.2021.8316

d) **Unsupervised Learning**.

Unsupervised Machine Learning have the survey of methodologies which are used to learn complex, highly non-linear models with large amount of unlabeled data. This model have order structure and distribution of data. Unsupervised Learning is grouped into two types clustering and association problem [10].

e) Random Forest Algorithm.

Random forest algorithm was developed by Leo Bremen and is a popular machine learning algorithm that belongs to the supervised learning technique. There are two stages in Random Forest Algorithm, one is random forest creation and thus the choice is to make a prediction from the random forest algorithm and contains decision tree on various dataset and calculate the average and improves the predictive accuracy of the dataset Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

Random Forest Algorithm:

Step 1	Select random K data points from the training set.
Step 2	Build the decision trees associated with the selected data points (Subsets).
Step 3	Choose the number N for decision trees that you want to build.
Step 4	Repeat Step 1 & 2.
Step 5	For new data points, find the predictions of each decision tree, and assign the new data points to the category [11].

The training dataset is classified into training data1 training data 2 to number of training data after that the testing will be done for the given data set and like training set there is decision tree1, decision tree2 and n number of decision trees after that the average will be considered for the given dataset and prediction and accuracy of the algorithm will be done for the model based on this the percentage of the model can be calculated for the given dataset.

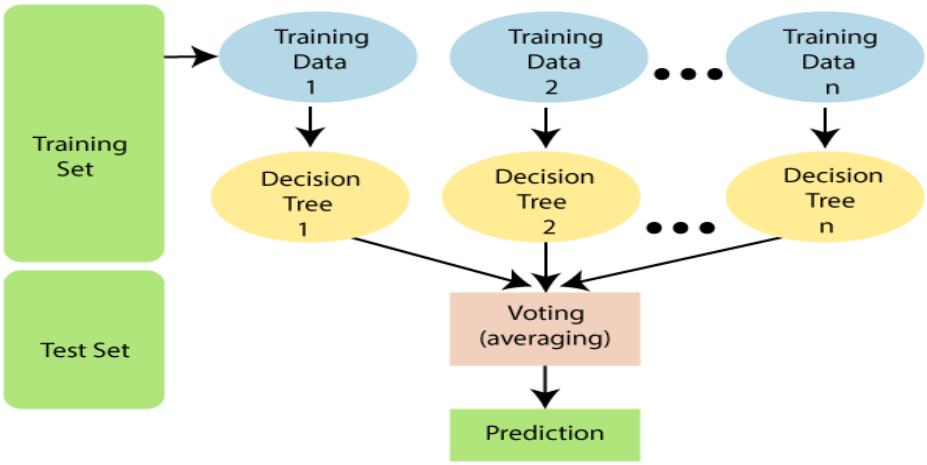

Fig.2 Random Forest

IV. RESULT

The Machine Learning algorithm has become very easy to find patterns and relation of various data. This project mainly revolves about predicting the diabetics of apatient and finding the accuracy of a given dataset using Random Forest Algorithm. The model is built using the training and test data which have the data preprocessing, data cleaning and model validation. The diabetic dataset is used to predict the disease. The dataset is publically available at UCI machine learning repository the dataset contain the diabetic disease information. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. After implementing several algorithm, random forest algorithm gave the best accuracy result compared to other algorithms. The Random forest algorithm gives 74.69% accuracy.

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 3, March 2021

DOI: 10.17148/IARJSET.2021.8316

Attribute No	Attribute	Description
1	Pregnancies	Number of times pregnant
2	Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3	BloodPressure	Diastolic blood pressure (mmHg)
4	SkinThickness	Triceps skin fold thickness (mm)
5	Insulin	2-Hour serum insulin (mu U/ml)
6	BMI	Body Mass Index (BMI)
7	DiabetesPedigreeFunction	Diabetes Pedigree function
8	Age	Age (in years)
9	Outcome	Class variable (0 or 1)

Fig.3 Attributes of the diabetics data

Histogram are one of the most common graphs used to display the numeric data. Distribution of the data whether the data is normally distributed or it's skewed to the left or right. pregnancies, glucose, age, BMI, blood pressure, insulin and many others features. The below diagram shows the different parameters like pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetics function, age and outcome for the prediction of diabetics for different scenarios.

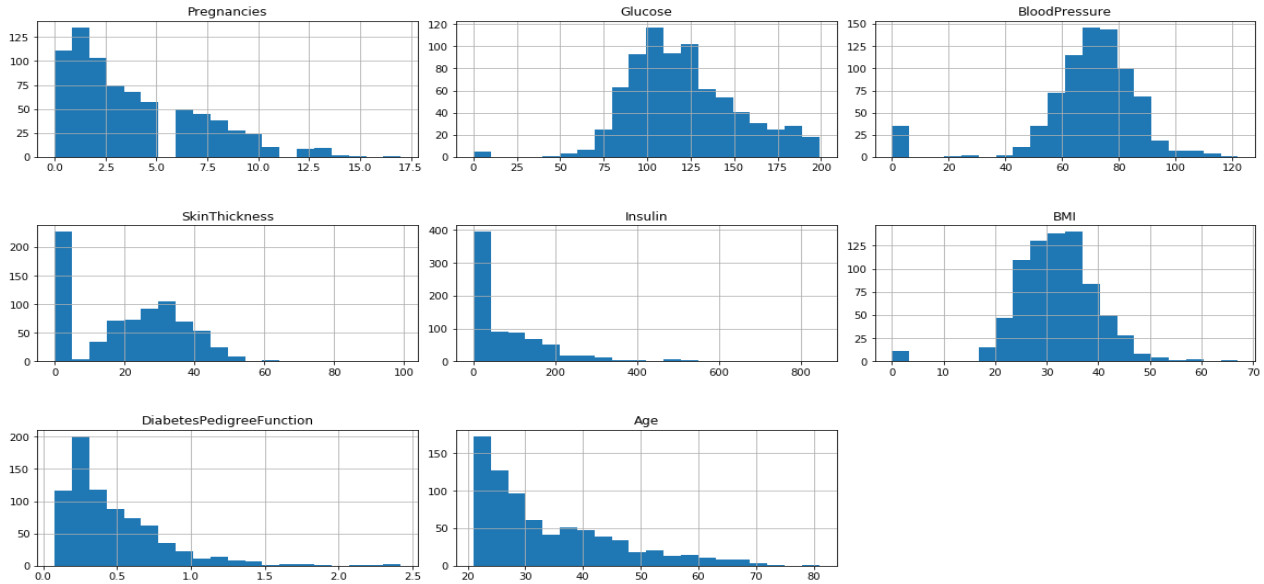

Fig.4 Graph

V. CONCLUSION

One of the important real-world medical problems is the detection of diabetes at its early stage**.** Diabetes is a heterogeneous group of diseases. The main motto of the diabetes association is "To prevent and cure diabetes and to improve the lives of all people affected by diabetes". To support the lives of the people all over the world, we are trying to detect and prevent the complications of diabetes at the early stage through predictive analysis by improving the classification techniques. The Random forest giving the highest specificity 74.691%, respectively holds best for the analysis of diabetic data.

REFERENCES

[1] S.Saru et al, Analysis and prediction of diabetes using machine learning Volume 5, Issue 4, April 2019 (ISSN: 2394 – 6598)
[2] Naveen Kishore G Prediction Of Diabetes Using Machine Learning Classification Algorithms VOLUME 9, ISSUE 01, JANUARY 2020 ISSN 2277-8616
[3] Minyechil Alehegn Rahul Joshi ,Analysis and Predection of Diabetes Mellitus using Machine Learning Algorithm Volume 118 No. 9 2018, 871-878 ISSN: 1311-8080
[4] Devi, Renuka & Shyla, J.. (2016). Analysis of various data mining techniques to predict diabetes mellitus. 11.727-730.
[5] Deepti Sisodia et al, Predection of Diabetes using Classification Algorithm, Procedia Computer Science 132(2018)1578-1585
[6] Bhavya M R et al, Diabetes Prediction using Machine Learning Vol. 9, Issue 7, July 2020 ISSN: 2278-1021
[7] G. van Rossum and F.L. Drake, Python Tutorial, http://docs.python.org/tut/tut.html.
[8] L. Li, Y. Wu, Y. Ou, Q. Li, Y. Zhou and D. Chen, "Research on machine learning algorithms and feature extraction for time series," *2017*

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 3, March 2021

DOI: 10.17148/IARJSET.2021.8316

Montreal, QC, Canada, 2017, pp.1-5,doi:10.1109/PIMRC.2017.8292668.

[9] Akinsola, J E T. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. International Journal of Computer Trends and Technology (IJCTT). 48. 128 – 138. 10.14445/22312803/IJCTT-V48P126.

[10] R. Raina, A.Madhava, and A. Y. Ng. "Large-scale Deep Unsupervised Learning using Graphics Processors. Internet: videolectures.net/site/normal_dl/tag=4836 8/ icml09_raina_lsd _01.pdf, 2009.

[11]VijiyaKumar, K., Lavanya, B., Nirmala, I., & Caroline, S. S. (2019). *Random Forest Algorithm for the Prediction of Diabetes. 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN).* Doi:10.1109/icscan.2019.88788 02

[12] Sneha, N., Gangil, T. Analysis of diabetes mellitus for early prediction using optimal features selection. *J Big Data* **6,** 13 (2019). https://doi.org/10.1186/s40537-019