International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 06 | June 2022 www.irjet.net p-ISSN: 2395-0072

Diabetes Disease Prediction Using Machine Learning Algorithms

Nitin M B[1], Anil Venkatraju N[2], Preethi B[3]

[1]*Nitin M B, Student, Dept. of CSE, BIET*
[2]*Anil Venkatraju N, Student, Dept. of CSE, BIET*
[3]*Preethi B, Asst. Professor, Dept. of CSE, BIET*

Abstract - *Diabetes mellitus is a very common and serious disease in many American Indian nations, as well as many other citizens of the world. Few known risk factors such as parental diabetes, genetic predisposition, overeating are considered to be the main risk factors for diabetes while the exact nature of genes or genes is unknown. The findings of this study may be used by health professionals, participants, students and researchers who are involved in research and development of diabetes predictors. We used the Random Forest Classifier algorithm to predict diabetes and later obtained a accuracy of 0.903.*

Key Words: American Indian Tribes, Random Forest.

1.INTRODUCTION

Diabetes is a condition that occurs when your blood sugar, also called sugar, is too high. Blood sugar is your main source of energy and it comes from the food you eat. Insulin, a hormone produced by pancreas, helps glucose in food to enter your cells for energy. Sometimes our bodies do not produce enough insulin. Glucose is in your blood and does not reach your cells. As we predict diabetes in pregnant women and this type of diabetes is called Gestational diabetes. Most of the time this type of diabetes disappears once the baby is born. However, if you have ever had gestational diabetes, you are more likely to have type 2 diabetes later in life. Sometimes diabetes during pregnancy is type 2 diabetes.

Early predictions of diabetes can be controlled to save lives. To get the best results, we test for diabetes prognosis by taking a variety of traits related to diabetes. We use the *Pima Indian Diabetes Dataset and use classification strategies and use these methods to predict diabetes. The purpose of this report is to look at a good diabetes predictor model with better accuracy.

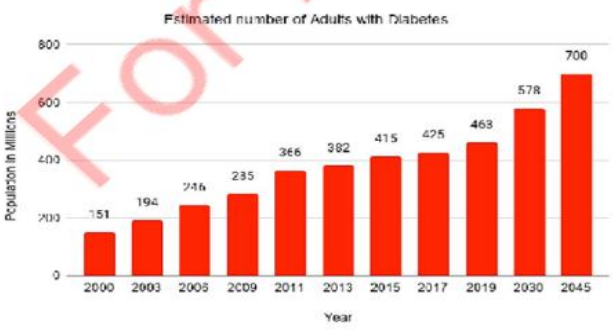

Fig -1: Drastic increase in number of Diabetes cases

1.1 MACHINE LEARNING

Machine Learning is a program of computer algorithms that can be learned from the model by self-improvement without being explicitly coded by the programmer. Machine learning is part of Artificial Intelligence that integrates data and mathematical tools to predict output that can be used to create data.

Success comes with the idea that the machine can read individually from the data (i.e., for example) in order to produce accurate results. Machine learning is closely related to data mining and speculative Bayesian model. The machine receives data as input and uses an algorithm to generate responses.

Typical machine learning activity is to provide recommendation. Google gives recommendations based on the persons browsing history and all these recommendations are based on the historical data. The data collected about past events and these events (browsing history). The data can be generated manually or automatically.

Machine learning is also used for a variety of tasks such as fraud detection, speculative care, portfolio optimization, automated work and more.

1.2 MACHINE LEARNING V/S TRADITIONAL PROGRAMMING

Traditional systems are very different from machine learning. In traditional editing, the editor code is all rules in consultation with an expert in the software industry. Each rule is based on a sound foundation; the machine will output the output following the logical statement. As the system grows more complex, more rules should be written. It can quickly become uncontrollable to care.

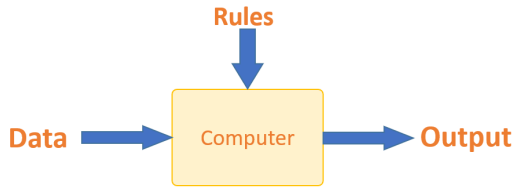

Fig -2: Traditional Programming

Fig -3: Machine Learning

Machine learning should overcome this issue. The machine learns how inputs and outputs are related and writes the rule. Program planners do not need to write new rules every time there is new data. Algorithms adapt and respond to new data and information to improve efficiency over time.

1.3 MACHINE LEARNING ALGORITHMS

There are many machine learning algorithms. The choice of algorithm is based on purpose. In the machine learning example below, the task is to predict the type of flower among the three species. Predictions are based on the length and width of the petal. The diagram shows the results of ten different algorithms. The image at the top left is a database. The data is divided into three categories: red, light blue and dark blue. There are certain collections. For example, from the second picture, everything on the left is red, in the middle, there is a mixture of uncertainty and light blue while the bottom is in the dark. Some images show different algorithms and how they try to separate data.

Machine learning can be combined into two broad learning activities: Supervised and Supervised.

1.3.1 SUPERVISED LEARNING

An approach to create artificial intelligence , here the computer algorithm is trained on input data which has been labelled for particular output.

The model is trained till it detects the underlying patterns and relationships between input and output labels which leads to obtaining accurate labelling results when interpreted with never-before-seen data

Supervised learning performs good at classification and regression problems for predicting the volume of sales for a given future date.

A supervised learning is based on training. During training the system is fed with labelled datasets which in turn instructs the system what output related to each specific value.

The trained model is presented with test data. The data which has been labelled and these labels are not disclosed to the algorithm.

The aim of testing the data is to measure how accurately the algorithm performs well on unlabelled data.

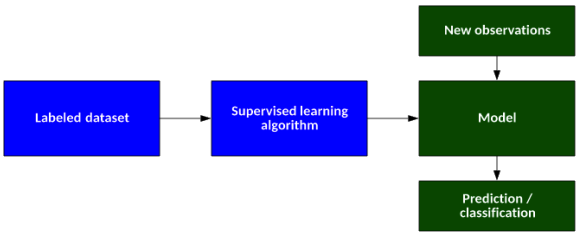

Fig -4: Supervised learning.

A) SUPERVISED LEARNING CATEGORY

Supervised techniques uses a model to reproduce outputs known from a training set. The system receives input as well as output data and its task is to create appropriate rules that maps input to the output. This training process should continue until the performance is high. After the training the system must be able to assign an output which has not been observed during the training phase. This process is fast and accurate. There are two types of supervised learning techniques i) classification and ii) regression

i. CLASSIFICATION

The technique aims to reproduce class assignments. It predicts the response value and the data is separated into classes. Example:- recognition of spam messages in the mail box.

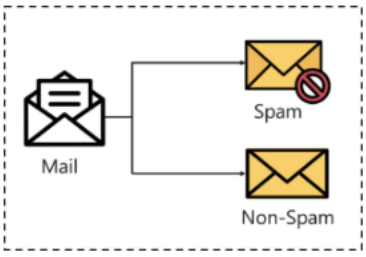

Fig -5: Classifying mail as spam or not spam.

ii. REGRESSION

A fundamental concept in machine learning in supervised learning here the algorithm is trained with both input features and output labels. It helps in establishing relationship among the variables by estimating how one variable affects the other. Example:- to predict the price of a product or price of house in a city and predicting the value of the stock in the market.

1.3.2 UNSUPERVISED LEARNING

In unsupervised learning the AI system groups unsorted information according to similarities and differences but there are no categories provided.

The algorithms are now allowed to classify , label or group the data points which are present in the dataset without having any external guidance in performing the task.

Performs more complex processing tasks as compared to supervised learning and subjecting the system to unsupervised learning is one way of testing AI.

These algorithms develop specific output from unstructured input and looks for relationship between each sample or input object.

Analyses underlying structure of datasets by extracting useful information or features from them

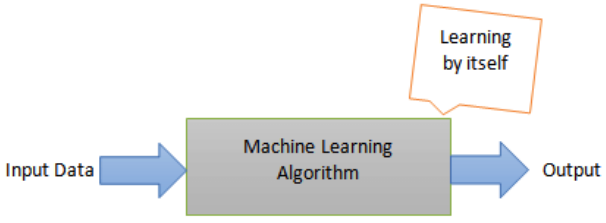

Fig -6: Unsupervised Learning

A) UNSUPERVISED LEARNING CATEGORY

i. CLUSTERING

A method of unsupervised learning and a common technique for statistical data analysis. Clustering involves grouping of data points. For a given set of datapoints we use clustering to classify each datapoint into specific group.

The datapoints which are in same group should have similar properties or features and datapoints in different groups will have dissimilar properties or features.

We use clustering to gain valuable insights from our data to notice what groups the datapoints fall into when clustering is applied.

ii. ANOMALY DETECTION

Anomaly detection is a process of identifying unexpected items or events in the dataset which differs from the norm.

Anomaly detection is applied on unlabelled data which is called unsupervised anomaly detection.

The anomaly detection has two basic components

a. Anomalies occur very rare in a data
b. Their features differ from normal instances

1.4 CHALLENGES AND LIMITATIONS OF MACHINE LEARNING

- **Challenges of machine learning are**

a. Not enough training data

A machine takes lots of data for most of the algorithms to function. The tasks like speech and image recognition the machine needs lakhs of examples. Now we can clearly justify that training data is insufficient.

b. Poor quality of data

The training data will have lots of errors outliers and noise and it is impossible for a machine learning model to detect proper underlying pattern and the model does not perform well. We may end up cleaning the data for getting accurate machine learning model.

c. Irrelevant features

Our training data must contain more relevant and less number of irrelevant features. For a successful machine learning to come up it should have good set of features including feature selection , extraction and creating new features in the machine learning model.

d. Nonrepresentative training data

The training data should be representative of new cases that we need to generalize. If a model trained using nonrepresentative training set as the model won't be accurate in predictions as it will be biased against one class or group. Our model works well if we used representative data during training and model won't be biased among one or two classes when works on testing data.

e. Overfitting and underfitting

Overfitting:-A machine learning model is said to be overfitted when trained with lot of data. If a model is trained with so much of data it starts learning from noise and inaccurate data entries in our dataset. The model will not classify data correctly because of too many details and noise.

Underfitting:- A machine learning algorithm is said to have underfitting when it cannot capture the underlying trends of the data. Underfitting means our model doesn't fit the data enough well. The underfitting destroys the accuracy of machine learning model.

- **Limitations of Machine Learning**

In recent times AI and ML developers have made AI and ML think more like humans performing complex tasks and making decisions based on in-depth analysis. Sometimes machine learning implementation is not necessary and it is not well thought out and also can cause more problems than it solves therefore machine learning is not an appropriate solution.

Since machine learning has profoundly impacted the world we are slowly evolving to the philosophy called "dataism" meaning the people trust data more than their personal beliefs.

There are five key limitations of Machine Learning algorithms

1. Ethical Concern:- We have been benefited from relying on computer algorithms to automate processes , analyse large amounts of data and also make complex decisions. Algorithms come in for bias at any level development because the algorithms are developed and trained by humans so it is impossible to eliminate bias. For example if self-driving car met with an accident who is responsible the driver, car manufacturer or the developer. Hence machine learning cannot make ethical or moral decisions on its own.

2. Deterministic problems:- A powerful technology suited for many domains which includes weather forecasting , climate and atmospheric research. The ML models can be used to calibrate and correct sensors which allows us to adjust the operation of sensors that measure environmental indicators like temperature pressure and humidity. The limitation is depending on the amount of data and complexity of model and this can be intensive and takes up to a month.

3. Lack of Data:- Neural Networks re complex architectures and it needs enormous amounts of training data to produce viable result. If the neural network architecture grows and also the requirement of the data will grow and if we may try to reuse the data but reusing the data doesn't bring good results.

4. Lack of Interpretability:- Let us consider we are working for the financial firm and we are assigned a task to build a model which detects fraudulent transactions and our model should be able to justify how it classifies transactions.

5. Lack of reproducibility:- The lack of reproducibility is due to complex and growing issue which is lack of code transparency and model testing methodologies. If the models are developed to take into account the latest research advances the model will not work in real time environment.

1.5 RANDOM FOREST

Random forest is an ensemble learning method for classification and regression and other tasks which operates by constructing multitude of decision trees in training time. For the classification tasks the output of random forest is class selected by most trees. Similarly for regression the mean or average prediction of an individual tree is returned.

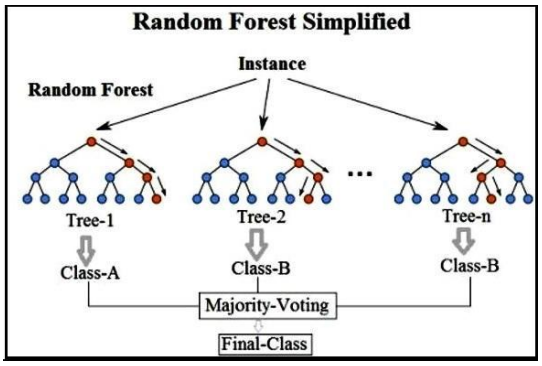

Fig -7: Random Forest

2.LITERATURE SURVEY SUMMARY

It is the comparison of various authors who have worked on the Diabetes Prediction and also have tried various machine learning algorithms for the same.

[1] P. Saeedi, I. Petersohn, P. Salpea, B. Malanda, S. Karuranga, N. Unwin, S. Colagiuri, L. Guariguata, A. A. Motala, K. Ogurtsova, J. E. Shaw, D. Bright, and R.Williams are used Logistic Regression to generate smoothed age-specific diabetes prevalence estimates (including previously undiagnosed diabetes) in adults aged 20-79 years.

[2] A. Mir and S. N. Dhage are used WEKA tool to predict diabetes disease by employing Naive Bayes, Support Vector Machine, Random Forest and Simple CART algorithm by taking an approach in Big Data Analytics which is emerging approach in healthcare.

[3] D. Sisodia and D. S. Sisodia are used Decision Tree, SVM and Naive Bayes are used in this experiment to detect diabetes at an early stage.

[4] J. Smith, J. Everhart, W. Dickson, W. Knowler, and R. Johannes are used ADAP an early neural network model to forecast Diabetes in high risk population of PIMA Indians. The algorithm's performance was analysed using standard measures for clinical tests: sensitivity, specificity, and a receiver operating characteristic curve. The crossover point for sensitivity and specificity is 0.76. They have further examined these methods by comparing the ADAP results with those obtained from logistic regression and linear perceptron models using precisely the same training and forecasting sets.

[5] Mitushi Soni and Dr. Sunita Varma are Machine Learning Classification and ensemble techniques on a dataset to predict diabetes. They are K-Nearest Neighbor (KNN), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Gradient Boosting (GB) 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 shows that the model is capable of predicting diabetes effectively. Our Result shows that Random Forest achieved higher accuracy compared to other machine learning techniques.

3. EXISTING SYSTEM

It is evident from the literature survey that the incidence of diabetes mellitus is increasing and that although there is evidence that the complications of diabetes can be prevented, there are still patients who lack the required knowledge and skills to manage and control their condition. It is generally accepted that diabetic patients must take responsibility for their own care and treatment. Patients therefore have to acquire the relevant knowledge, for their diabetes condition and we also need to educate family members of the patient This study is an attempt to determine patients and family members knowledge and views on diabetes mellitus, to make recommendations towards improved diabetic education which might lead to improved adherence to the diabetic treatment regimen.

4.PROPOSED SYSTEM

Classification is one of the most important decision-making techniques in many real-world problems. In this project, the main objective is to classify the data as diabetic or non-diabetic and improve the classification accuracy. For many classifications problem, the higher number of samples chosen but it doesn't lead to higher classification accuracy. In many cases, the performance of algorithm is high in the context of speed but the accuracy of data classification is low. The main objective of our model is to achieve high accuracy. Classification accuracy can be increased if we use much of the data set for training and few data sets for testing. This study analysed ways to categorise diabetic and non-diabetic data. Therefore, it is noted that Random Forest has achieved higher when compared with other machine learning models

5.OBJECTIVES

1. To determine the good accuracy score of the patient and predict that patient is diabetic or non-diabetic

2. To develop a model which acts as health assistant to the patient to take early prediction and make early decision to cure diabetes.

6.METHODOLOGY DIAGRAM

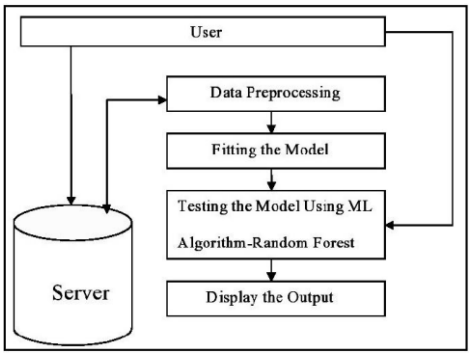

Fig -8: User uploads his medical data related to Diabetes to the server

The system is implemented in four phases. It includes collection and pre-processing of dataset. The training is done with train data and validation with test data. *This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. 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. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

In the PIMA Indian Diabetes Dataset there are 9 attributes they are pregnancies glucose blood pressure skin thickness insulin bmi and diabetes pedigree function age and outcome

a. Pregnancies = npreg = number of times women is pregnant
b. Glucose = glu = plasma glucose concentration
c. BloodPressure = bp = Diastolic Blood Pressure (mm Hg)
d. SkinThickness = skin thickness fold thickness
e. BMI = bmi = body mass index (kg/m^2)
f. Diabetes Pedigree Function = ped
g. Age = age (years).
h. Outcome = women has diabetes or not.

The dataset consists of 768 patients; 268 patients are diabetic and the rest of them are non-diabetic. The output variable takes '0 or '1' values, where '0' and '1' are depict the non- diabetic instance and diabetic instance respectively.

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
10	129	62	36	0	41.2	0.441	38	1(positive)
0	134	58	20	291	26.4	0.352	21	0(negative)
3	102	74	0	0	29.5	0.121	32	0(negative)
7	187	50	33	392	33.9	0.826	34	1(positive)
3	173	78	39	185	33.8	0.97	31	1(positive)
10	94	72	18	0	23.1	0.595	56	0(negative)
1	108	60	46	178	35.5	0.415	24	0(negative)
5	97	76	27	0	35.6	0.378	52	1(positive)

Table -1: The table showing women have been predicted diabetes or not

Feature label	Variable type	Range
Number of times pregnant	Integer	0–17
Plasma glucose concentration in a 2 h oral glucose tolerance test	Real	0–199
Diastolic blood pressure	Real	0–122
Triceps skin fold thickness	Real	0–99
2 h serum insulin	Real	0–846
Body mass index	Real	0–67.1
Diabetes pedigree function	Real	0.078–2.42
Age	Integer	21–81
Class	Binary	Tested positive for diabetes = 1

Table -2: The table showing the range of labels present in the diabetes dataset

7.CONCLUSION

The main aim of our project was to design and implement Diabetes Prediction Using Machine Learning Methods and performance analysis of methods and it has been achieved. The proposed approach uses various classification and ensemble learning method in which Support Vector Machine, K-Nearest Neighbour ,Random forest, Decision Tree, Logistic Regression and Gradient Boosting classifiers are used. Accuracy achieved is 77%. The experimental results can be assistant health care to take early prediction and make early decision to cure diabetes.

8.REFERENCES

[1] Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045

[2] Diabetes disease prediction using machine learning on big data of healthcare

[3] Prediction of diabetes using classification algorithms

[4] Using the adap learning algorithm to forecast the onset of diabetes mellitus

[5] Diabetes Prediction using Machine Learning Technique