International Research Journal of Education and Technology
ISSN 2581-7795

# PREDECTING DIABETES MELLITUS ACCURETELY USING MACHINE LEARNING ALGORITHMS IN BIGDATA

[1]N.V.Poornima, [2]Dr.B.Srinivasan
[1]Research scholar
[2]*Associate Professor*
[1,2]*Department of computer science*
[1,2]*Gobi arts &Science College*
*Gobichettipalayam.*

## ABSTRACT

The medical industry contains huge amount of data. It is a large in size and very difficult to predict a disease in traditional methods. Diabetics Mellitus is non-communicable diseases. It targets more in middle income countries. As shown by the International Diabetes Federation (IDF), 463 million people have diabetes in the world and 88 million people in the Southeast Asia region. Of these 88 million people, 77 million have a spot with India. The power of diabetes in the general population is 8.9%, as demonstrated by the IDF. As indicated by the IDF gauges, India has the second most noteworthy number of youngsters with type 1 diabetes after the United States. It additionally adds to the biggest extent of occurrence instances of type 1 diabetes in youngsters in the SEA district. Per the World Health Organization, 2% of all passing in India are because of diabetes. As per the developing dreariness as of late, in 2040, the world's diabetic patients will arrive at 642 million, which implies that one of the ten grown-ups later on is experiencing diabetes. There is no uncertainty that this disturbing figure needs incredible consideration. With the quick improvement of Machine learning, Machine Learning has been applied to numerous parts of clinical wellbeing. Diabetes is one of the common and growing diseases in several countries and all of them are working to prevent this disease at early stage by predicting the symptoms of diabetes using several methods we are using Big data analytics to predict the diabetes data accurately. Big data analytics creates awareness about the diabetes among the patients. It helps a patient to resolve and care the diseases with Electronic Health Records (EHR). Based on dataset, big data predict upcoming risk in diabetes and provide a treatment accordingly.in this paper machine learning concepts used to prevent and diagnose the diabetes were discussed and presented an overview of Big Data, Machine learning tools and models.

Keywords: Big Data, Machine Learning, Diabetes.

## INTRODUCTION TO MACHINE LEARNING AND BIG DATA

### BIGDATA

Anyway as of late, Nature of Data is changed. Furthermore, Systems or Organizations or Applications are creating colossal measure of Data in assortment of arrangements at exceptionally quick rate.

That implies Data isn't basic Structured Data (Not as basic Rows and Columns). It doesn't have any appropriate configuration, only Raw Data with no arrangement. It is "extremely troublesome or impractical" to utilize Old Technologies, Traditional Relational Databases and Tools to store, oversee interaction and

report this Data. Customary Databases can't Store, Process and Analysis this sort of Data.

At that point how to tackle this issue? Here Big Data Solutions come into picture.

Enormous Data Solutions tackle every one of these issues without any problem.

Allow us to begin with understanding what Big Data is and how significant it is a major part of our life.

We don't have a clear definition to Big Data. Notwithstanding, we will attempt to address this inquiry in an unexpected way.

In Simple Words, Big Data is a strategy to tackle information issues that are not resolvable utilizing Traditional Databases and Tools.

In alternate manner, Big Data implies not simply immense measure of Data. Big Data implies gigantic measure of information creating at quick rate in various organizations.

Enormous Data is a Technique to "Store, Process, Manage, Analysis and Report" a colossal measure of assortment information, at the necessary speed, and inside the necessary opportunity to permit Real-time Analysis and Reaction.

Big Data is Data with has the accompanying three qualities:

- Extremely Large Volumes of Data
- Extremely High Velocity of Data
- Extremely Wide Variety of Data

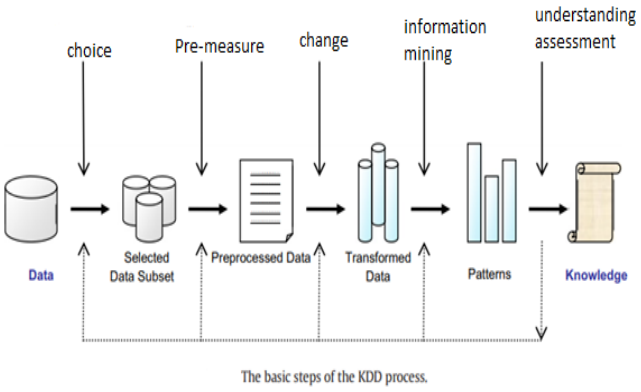
Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers to learn automatically without human intervention or assistance and adjust actions accordingly.

**MACHINE LEARNING**

Machine learns from experience. For many scientists, the term “machine learning” is identical to the term “artificial intelligence”, the purpose of machine learning is the construction of computer systems that can adapt and learn from their experience.

Knowledge discovery in databases (KDD) is a field enveloping speculations, strategies and methods, attempting to sort out information and concentrate valuable information from them. It is viewed as a multistep interaction (choice, pre-measure, change, information mining, and understanding assessment) portrayed in Fig The main advance in the whole KDD measure is information mining, embodying the utilization of AI calculations in breaking down information. A total meaning of KDD is given as; KDD is the nontrivial cycle recognizing legitimate, novel, conceivably helpful, and eventually justifiable examples in information.

International Research Journal of Education and Technology
ISSN 2581-7795




The basic steps of the KDD process.

Machine learning typically classified into three broad categories. They are:
a) Supervised learning,
In which the system infers a function from labelled training data
b) Unsupervised learning,
In which the learning system tries to infer the structure of unlabelled data, and
C) Reinforcement learning
In which the system interacts with a dynamic environment.

## SUPERVISED MACHINE LEARNING ALGORITHMS

It can apply what has been realized in the past to new information utilizing named guides to foresee future occasions. Beginning from the examination of a known preparing dataset, the learning calculation creates a gathered capacity to make expectations about the yield esteems.

In supervised learning, there are two kinds of learning tasks:

Classification and regression. Classification models try to predict distinct classes, such as e.g. blood groups,

While regression models predict numerical values.

Regression is used when we are trying to predict an output variable that is continuous. Whereas, classification is used when we are trying to predict the class that a set of features should fall into. Some of the most common techniques are Decision Trees (DT), Rule Learning, and Instance Based Learning (IBL), such as k-Nearest Neighbours (k-NN), Genetic Algorithms (GA), Artificial Neural Networks (ANN), and Support Vector Machines (SVM).

## UNSUPERVISED MACHINE LEARNING ALGORITHMS

In contrast, unsupervised machine learning algorithms are used when the information used to train is neither classified nor labelled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabelled data.

## SEMI-SUPERVISED MACHINE LEARNING ALGORITHMS

Semi-supervised machine learning algorithms fall some place in the middle of supervised and unsupervised learning since they utilize both marked and unlabelled information for preparing. Normally a limited quantity of named information and a lot of unlabelled information. The frameworks that utilization this technique can extensively improve learning precision.

## REINFORCEMENT MACHINE LEARNING ALGORITHMS

Reinforcement machine learning algorithm is a learning strategy that interfaces with its current circumstance by creating activities and finds blunders or rewards. Experimentation search and postponed reward are the most significant qualities of support learning. This strategy permits machines and programming specialists to consequently decide the ideal conduct inside a particular setting to boost its presentation.

ADVANTAGES OF BIG DATA IN HEALTHCARE

- Predicts an diseases precisely
- Easy checking Electronic Health Records
- Helps to take choice accurately by specialist
- Hospital visits can be decreased

- Doctor can monitor patients utilizing brilliant advancements
- Keep a group sound

**6 COMPONENTS OF MACHINE LEARNING**

1) Feature Extraction + Domain

Knowledge

As a matter of first importance we truly need to comprehend what kind of information we are managing and what ultimately we need to receive in return. Basically we need to see how and what highlights should be separated from the information. For example accept we need to construct programming that recognizes male and female names. Every one of the names in text can be considered as our crude information while our highlights could be number of vowels in the name, length, first and last character, and so on of the name.

2) Feature Selection

In many scenarios we end up with a lot of features at our disposal. We might want to select a subset of those based on the resources and computation power we have. In this step we select a few of those influential features and separate them from the not-so-influential features. There are many ways to do this, information gain, gain ratio, correlation etc.

2) Choice of Algorithm

There are wide scopes of calculations from which we can pick dependent on whether we are attempting to do expectation, grouping or bunching. We can likewise pick among direct and non-straight calculations. Naive Bayes, Support Vector Machines, Decision Trees, k-Means grouping are some normal calculations utilized.

4) Training

In this progression we tune our calculation dependent on the information we as of now have. This information is considered preparing set as it is utilized to prepare our calculation. This is the part where our machine or programming learn and improve with experience.

5) Choice of Metrics/Evaluation Criteria

Here we choose our assessment measures for our calculation. Basically we think of measurements to assess our outcomes. Usually utilized proportions of execution are precision, recall, f1-measure, robustness, specificity-sensitivity, error rate etc and so

6) Testing

Finally, we test how our machine learning algorithm performs on an inconspicuous arrangement of experiments. One approach to do this is to parcel the information into preparing and testing set. The preparation set is utilized in sync for while the test set is then utilized in this progression. Procedures, for example, cross-approval and leave-one-out can be utilized to manage situations where we need more information containers certainly isn't thorough and can't do finish equity to a wide field like Machine Learning. And still, at the end of the day, the vast majority of the occasions a Machine Learning venture would include a large portion of the previously mentioned containers, if not all

**FAMOUS MACHINE LEARNING ALGORITHMS**

**LINEAR REGRESSION**

Linear Regression is one of the simplest and most important supervised Machine Learning algorithms. Basically It is a statistical technique that is used for predictive analysis.

Linear regression makes predictions for continuous/discrete or numeric variables such as sales, salary, age, product price, etc.

International Research Journal of Education and Technology
ISSN 2581-7795

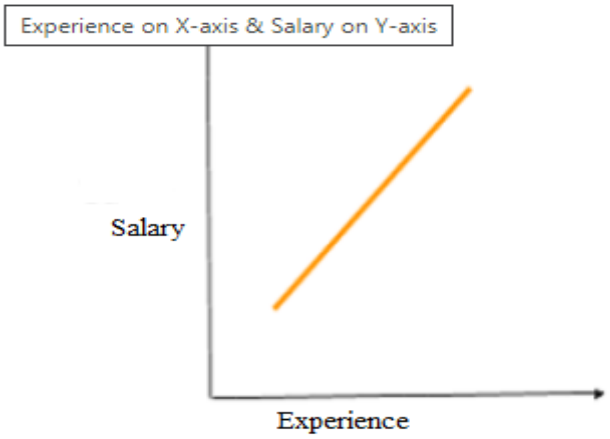
Linear regression shows a direct connection between a dependent variables and one or more independent variable, consequently called as linear regression. Since linear regression shows the linear relationship, which implies it discovers how the worth of the dependent variable is changing as per the worth of the independent variable.

Linear regression is used to produce a slope between the single input and the expected outcome. The model is referred to as a simple linear regression when there is a single input variable,

In a simple linear regression the coefficients required by the model is estimated and make the predictions on new data logically. That is, the line for a simple linear regression model can be written as:

$$y = b0 + b1 * x + \varepsilon$$

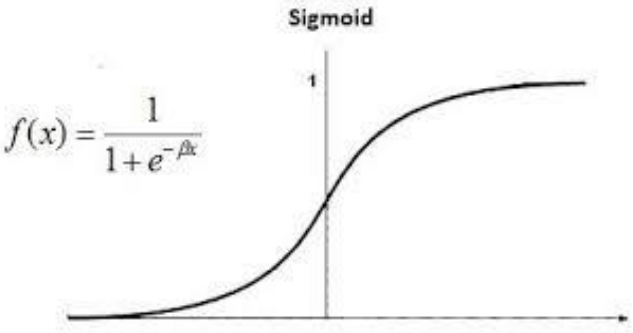
Where B0 and B1 are the coefficients we must estimate from the training data and ∈ is an error term.




The above diagram displays the linear Regression, it takes one input and produce the output based on the input outcome will be produced. Employee salary will be determined based on our experience.

## LOGISTIC REGRESSION

Logistic Regression is a classification algorithm used to find the probability of event success and event failure. It is used when the dependent variable is binary (0/1, True/False, Yes/No) in nature. It supports categorizing data into discrete classes by studying the relationship from a given set of labelled data. It learns a linear relationship from the given dataset and then introduces a non-linearity in the form of the sigmoid function.

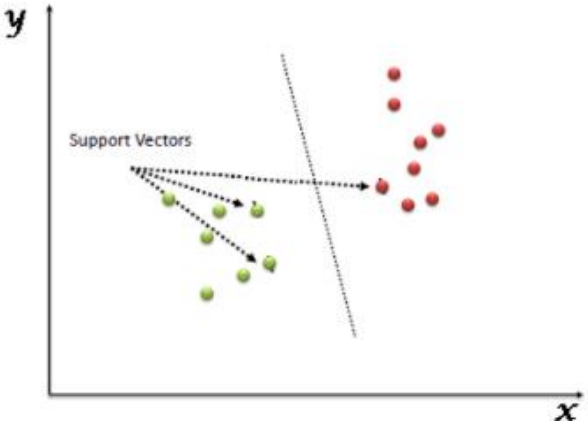



Logistic regression is also known as Binomial logistics regression. It is based on sigmoid function where output is probability and input can be from -infinity to +infinity.

## SUPPORT VECTOR MACHINE

Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well (look at the below snapshot).

International Research Journal of Education and Technology
ISSN 2581-7795




Support Vectors are simply the co-ordinates of individual observation. The SVM classifier is a frontier which best segregates the two classes (hyper-plane/ line).

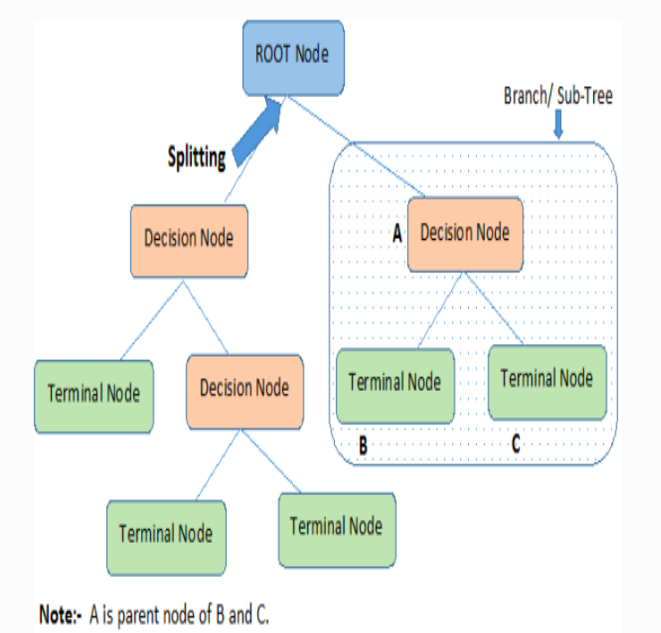
Classification is a two-step process, learning step and prediction step, in machine learning. In the learning step, the model is developed based on given training data. In the prediction step, the model is used to predict the response for given data. Decision Tree is one of the easiest and popular classification algorithms to understand and interpret.

## DECISION TREE ALGORITHM

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too.

The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data (training data).

In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node




### How do Decision Trees work?

The decision of making strategic splits heavily affects a tree's accuracy. The decision criteria are different for classification and regression trees.

Decision trees use multiple algorithms to decide to split a node into two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words, we can say that the purity of the node increases with respect to the target variable. The decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes.

## INSTANCE BASED LEARNING

Storing and using specific instances improves the performance of several supervised learning algorithms. These include algorithms that learn decision trees, classification rules, and distributed networks

## K-NEAREST NEIGHBOUR ALGORITHM

International Research Journal of Education and Technology
ISSN 2581-7795

. KNN: K Nearest Neighbour is one of the essential calculations in AI. AI models utilize a bunch of information esteems to foresee yield esteems. KNN is perhaps the least difficult type of AI calculations for the most part utilized for order. It arranges the information point on how its neighbour is ordered. K-NN calculation can be utilized for Regression just as for Classification however for the most part it is utilized for the Classification issues.

K-NN is a non-parametric calculation, which implies it doesn't make any suspicion on fundamental information.

It is likewise called a lazy learner algorithm since it doesn't gain from the preparation set quickly rather it stores the dataset and at the hour of characterization, it's anything but an activity on the dataset.

KNN orders the new information focuses dependent on the closeness proportion of the prior put away information focuses. For instance, on the off chance that we have a dataset of tomatoes and bananas. KNN will store comparative estimates like shape and shading. At the point when another item comes it will check its closeness with the shading (red or yellow) and shape.

K in KNN addresses the quantity of the closest neighbours we used to characterize new information focuses.

**Why do we need a K-NN Algorithm?**

Assume there are two classes, i.e., Category A and Category B, and we have another information point x1, so this information point will lie in which of these classifications. To tackle this sort of issue, we need a K-NN calculation. With the assistance of K-NN, we can undoubtedly recognize the classification or class of a specific dataset.

**How does K-NN work?**

Step-1: Select the number K of the neighbors

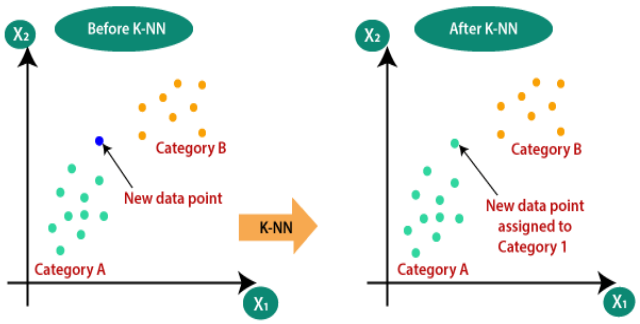
Step-2: Calculate the Euclidean distance of K number of neighbors

Step-3: Take the K closest neighbors according to the determined Euclidean distance.

Step-4: Among these k neighbors, check the quantity of the information focuses in every class.

Step-5: Assign the new information focuses to that class for which the quantity of the neighbor is greatest.

Step-6: Our model is prepared.




**How to select the value of K in the K-NN Algorithm?**

There is no specific method to decide the best incentive for "K", so we need to attempt a few qualities to track down the best out of them. The most favoured incentive for K is 5.

A low incentive for K like K=1 or K=2, can be loud and lead with the impacts of anomalies in the model.

Huge qualities for K are acceptable, yet it might discover a few troubles.

**CONCLUSION**

This paper gives the clear explanation of machine learning concepts in the big data environment.by using these algorithms the artificial neural network has been created .via the ANN the

human work is eliminated and the accuracy and computational time is reduced.

## REFERENCES


1. Cichosz, S. L., Johansen, M. D., & Hejlesen, O. (2016). Toward big data analytics: review of predictive models in management of diabetes and its complications. *Journal of diabetes science and technology*, *10*(1), 27-34.
2. Rallapalli, S., & Suryakanthi, T. (2016, November). Predicting the risk of diabetes in big data electronic health Records by using scalable random forest classification algorithm. In *2016 International Conference on Advances in Computing and Communication Engineering (ICACCE)* (pp. 281-284). IEEE.
3. Muni Kumar, N. (2016). Survey on map reduces based apriori algorithms in medical field for the prediction of diabetes mellitus. *RESEARCH JOURNAL OF FISHERIES AND HYDROBIOLOGY*, *11*(4), 13-18.
4. Shetty, S. P., & Joshi, S. (2016). A tool for diabetes prediction and monitoring using data mining technique. *International Journal of Information Technology and Computer Science (IJITCS)*, *8*(11), 26-32.
5. Mishra, S., Chaudhury, P., Mishra, B. K., & Tripathy, H. K. (2016, March). An implementation of Feature ranking using Machine learning techniques for Diabetes disease prediction. In *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies* (pp. 1-3).
6. Joshi, R., & Alehegn, M. (2017). Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach. *International Research Journal of Engineering and Technology*, *4*(10), 426-435.
7. Komi, M., Li, J., Zhai, Y., & Zhang, X. (2017, June). Application of data mining methods in diabetes prediction. In *2017 2nd International Conference on Image, Vision and Computing (ICIVC)* (pp. 1006-1010). IEEE.
8. Kumar, P. S., & Pranavi, S. (2017, December). Performance analysis of machine learning algorithms on diabetes dataset using big data analytics. In *2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions)(ICTUS)* (pp. 508-513). IEEE.
9. Jayanthi, N., Babu, B. V., & Rao, N. S. (2017). Survey on clinical prediction models for diabetes prediction. *Journal of Big Data*, *4*(1), 26.
10. Chen, W., Chen, S., Zhang, H., & Wu, T. (2017, November). A hybrid prediction model for type 2 diabetes using K-means and decision tree. In *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)* (pp. 386-390). IEEE.
11. Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. *Procedia computer science*, *132*, 1578-1585.
12. Sarwar, M. A., Kamal, N., Hamid, W., & Shah, M. A. (2018, September). Prediction of diabetes using machine learning algorithms in healthcare. In *2018 24th International Conference on Automation and Computing (ICAC)* (pp. 1-6). IEEE.
13. Mir, A., & Dhage, S. N. (2018, August). Diabetes disease prediction using machine learning on big data of healthcare. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)* (pp. 1-6). IEEE.
14. Liu, B., Li, Y., Sun, Z., Ghosh, S., & Ng, K. (2018, February). Early Prediction of Diabetes Complications from Electronic Health Records: A Multi-Task Survival Analysis Approach. In *AAAI* (pp. 101-108).
15. Chen, M., Yang, J., Zhou, J., Hao, Y., Zhang, J., & Youn, C. H. (2018). 5G-smart diabetes: Toward personalized diabetes diagnosis with healthcare big data clouds. *IEEE Communications Magazine*, *56*(4), 16-23.