Partners Universal International Innovation Journal (PUIIJ)

Volume: 01 Issue: 01 | January-February 2023 | www.puiij.com

A Survey Study on Big Data Analytics to Predict Diabetes Diseases Using Supervised Classification Methods

A.S. Hovan George[1], Aakifa Shahul[2], A. Shaji George[3] , T. Baskar[4], A. Shahul Hameed[5]

[1]Student, Tbilisi State Medical University, Tbilisi, Georgia.
[2]Student, SRM Medical College, Kattankulathur, Tamil Nadu, India.
[3]Director, Masters IT Solutions, Chennai, Tamil Nadu, India.
[4]Professor, Department of Physics, Shree Sathyam College of Engineering and Technology, Sankari Taluk, Tamil Nadu, India.
[5]General Manager, Department of Telecommunication, Consolidated Techniques Co. Ltd, Riyadh, Kingdom of Saudi Arabia.

Abstract - The complexity will significantly increase as the healthcare industry moves toward processing massive amounts of health records and accessing those records for analysis and implementation. Big Data from the health sector is becoming increasingly unstructured, so it is necessary to structure it, emphasise its size, and provide potential solutions. The healthcare sector faces numerous difficulties, which highlights the significance of developing data analytics. A prediction based on disease data is one of the missions. Currently, one of the leading causes of death worldwide is diabetes diseases (DD). One of the most common non-communicable diseases that affects people today is diabetes mellitus. Big data analytics can be used on this data to discover patterns and connections between the various factors that influence diabetes. The study of various supervised classification techniques is the main focus of this paper, and we also show the accuracy of each combined algorithm to give readers a clear idea of the most effective algorithm for predicting the development of diabetes.

Keywords: Big Data, Diabetes, Supervised Classification.

1. INTRODUCTION

Today, there are countless people suffering from different illnesses. The human services sector has generated a sizable amount of diabetes. A vast network of healthcare data is being used by researchers, hospitals, and doctors to better understand clinical context, avert future health problems, and even discover new treatment options. Big data is tackling some of the biggest challenges in the healthcare industry, even though there are many ways that data is used to influence healthcare. Researchers in the public and private sectors are committed to looking for a cure as well as more efficient treatment options.

One of the Non Communicable Diseases (NCD) that is spreading throughout the world is diabetic mellitus (DM). The condition known as diabetes mellitus, or simply diabetes, is one in which the body doesn't produce enough insulin. A number of factors, including age, insulin, blood pressure, skin thickness, and others, can affect diabetes.

There is now a lot of information available about diabetes that can be broken down to determine the relationships between various components in order to enhance the social insurance system and provide better treatment facilities.

A. BIG DATA AND ITS CHARACTERISTICS

Big Data is a term used to describe enormous amounts of information. This data establishes both structured and unstructured data that is quickly and incrementally developing. Because conventional database frameworks can't manage large data sets, associations are having trouble managing and investigating this data to produce some significant results.

Numerous disciplines, including science, engineering, finance, business, social work, and healthcare, can benefit from the processing and analysis of big data. [1].

Characteristics

The five different characteristics of data, including volume, velocity, variety, veracity, and value, are what the big data technology is based on. [2]

Volume: The amount of data produced by an organization or any entity.

Velocity: It represents the speed of data production and distribution.

Variety: It denotes the varying data formats of the data.

Veracity: It indicates the data uncertainty

Value: The value stresses on the data being formed out of certain business processes.

B. ADVANTAGES OF BIG DATA

- Accurately forecasts a disease.
- Simple electronic health record monitoring.
- Helps the doctor make the right decision.
- Reducing hospital visits is possible.
- Through the use of SMART technologies, the doctor can monitor the patient.
- keep people in good health.

2. LITERATURE SURVEY

The author of this paper [3] has discussed the challenges associated with diabetes data. This research paper examines all variables, including sexual orientation and age groups, that may affect Type 2 diabetes. This essay also demonstrates how large-scale data analysis outperforms all conventional methods for the management of diabetes data and the production of reliable results.

The diabetes diseases dataset from Loannis et almachine .'s learning algorithms is one of many medical data sets that can be predicted (DDD). Support vector machines (SVM), logistic regression, and naive bayes are used in this study to predict various medical datasets, including the diabetes dataset (DD). Based on their findings, the researchers compared the accuracy and performance of the algorithms, and they came to the conclusion that the SVM (Support Vector Machine) algorithm offers the best accuracy compared to the other algorithms mentioned above. On a small sample of data, the researchers used

© 2023, PUIIJ |DOI:10.5281/zenodo.7644341

these machine learning algorithms. Data origin, kind, and dimensionality were identified as accuracy-related factors in this study.

Using the CPCSSN dataset and three machine learning methods, Sajida et al. [5] were able to predict the early stages of diabetes diseases (DD) and prevent early death in humans. Bagging, Adaboost, and decision tree (J48) were used in this study to predict diabetes, and the researcher compared the results of those methods to come to the conclusion that the Adaboost method provided the most effectiveness and accuracy of all the methods in the Weka data mining tools.

In this study by Pradeep and Dr. Naveen [6], the effectiveness of machine learning techniques was compared and evaluated based on their accuracy. According to this study, there are differences in the technique's accuracy before and after pre-processing. This shows that the performance and accuracy of disease prediction are both impacted by the pre-processing of the data set.

In this study by Xue-Hui Meng et al. [7], the researchers used various data mining techniques to forecast the occurrence of diabetic diseases using real-world data sets and information gathered by distributed questioners. Weka and SPSS tools were used in this study for data analysis and prediction, respectively. Three techniques—ANN, logistic regression, and J48—are compared in this study. In the end, it was determined that the J48 machine learning technique offers effective and improved accuracy.

The Diabetes Data Prediction Using Spark and Analysis in Hue Over Big Data is discussed in this paper [8]. This paper uses big data analytics to investigate information on diabetes. The creator obtained the data from the database of Pima Indians. For this object, Hue and Hadoop MapReduce were used as a strategy. Flash was used to prepare quickly. The prescient examination calculation is then used in the Java code written in Hadoop to break down the clinical dataset. Inquiries were created in Hive, and Hue was used to create indexes.

Extrapolative analytics in healthcare are described as an exhausting sorting algorithm by Gopala Krishna Palam[9]. Today, lingering illness is the primary cause of death worldwide. As a result, the effectiveness of a health campaign system has significantly increased. On the other hand, based on statistical analysis and decision-making, the age range and generation of long-lasting disease prospect models represent a significant change in health proficiency.

In this document, effective methods for preventing chronic illness are used by data mining historical health proceedings. In this case, the diagnosis of diabetes is reached using the CNN class, End Tree, Carrier Support vector (SVM), and Artificial Neural Network (JST).

3. CLASSIFICATION ALGORITHM

Numerous statistical and machine learning techniques can be used to forecast diabetes diseases.

1. K-Nearest Neighbor algorithm (K-NN)

One of the simplest machine learning algorithms, it uses the technique of supervised learning. A new data point is classified using the K-NN algorithm based on similarity after all the existing data has been stored. Although the K-NN algorithm is most frequently used for classification problems, it can also be used for regression. Figure 1 depicts the K-NN algorithm in action.

© 2023, PUIIJ |DOI:10.5281/zenodo.7644341

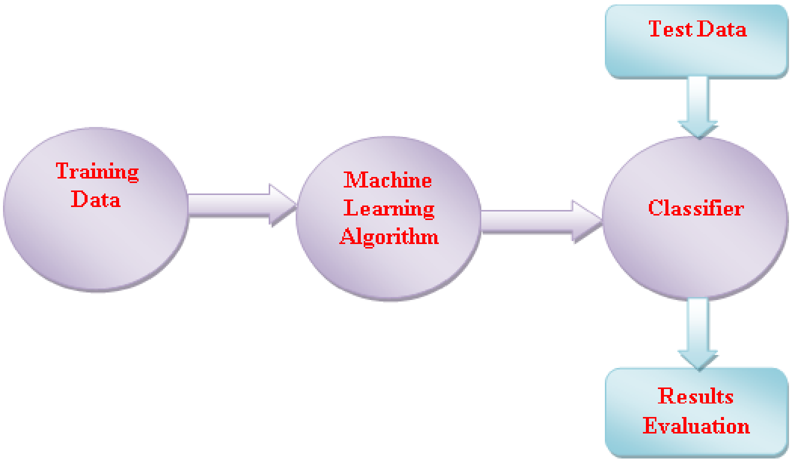

Fig -1: Working of K-NN Algorithm

Working of K-NN Algorithm

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 nearest neighbors as per the calculated Euclidean distance.

Step-4: Among these k neighbors, count the number of the data points in each category.

Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.

2. Random Forest

Decision Tree Forest is another name for the Random Forest. It is one of the widely used ensemble models based on decision trees. These models are more accurate than other decision trees. Both classification and regression applications use this algorithm. Figure 2 depicts the Random Forest Algorithm in action.

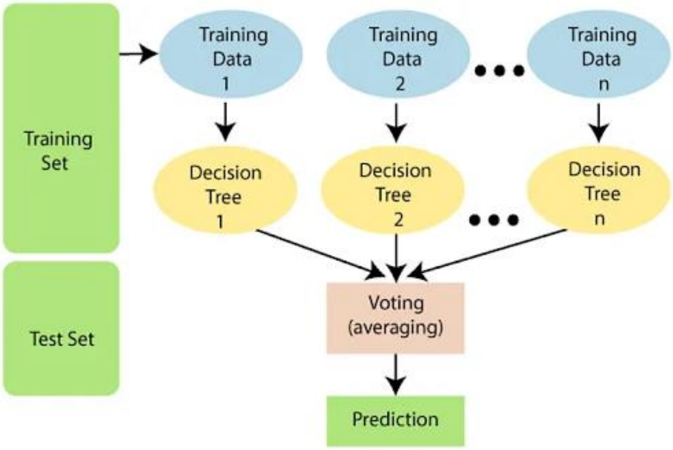

Fig -2: Working of Random Forest Algorithm

Working of Random Forest Algorithm

Step 1: Select random samples from a given data or training set.

Step 2: This algorithm will construct a decision tree for every training data.

Step 3: Voting will take place by averaging the decision tree.

Step 4: Finally, select the most voted prediction result as the final prediction result.

3. Naïve Bayes (NB)

The Nave Bayes algorithm is a supervised learning method for classification problems that is based on the Bayes theorem. It is primarily employed in text classification with a large training set. The Naive Bayes Classifier is one of the most straightforward and efficient classification algorithms available today. It aids in the development of quick machine learning models capable of making accurate predictions. Being a probabilistic classifier, it makes predictions based on the likelihood that an object will occur.

Working of Naïve Bayes' Classifier

Step 1: Convert the given dataset into frequency tables.

Step 2: Generate Likelihood table by finding the probabilities of given features.

Step 3: Now, use Bayes theorem to calculate the posterior probability.

4. Support Vector Machine (SVM)

One of the most well-liked supervised learning algorithms, Support Vector Machine, or SVM, is used to solve Classification and Regression problems. However, it is primarily employed in Machine Learning Classification issues.

The SVM algorithm's objective is to establish the best line or decision boundary that can divide n-dimensional space into classes, allowing us to quickly classify new data points in the future. A hyperplane is the name given to this optimal decision boundary.

SVM selects the extreme vectors and points that aid in the creation of the hyperplane. Support vectors, which are used to represent these extreme cases, are the basis for the SVM algorithm. Consider the diagram below, where a decision boundary or hyperplane is used to categorise two distinct categories:

4. PREDICTIVE ANALYTICS

Predictive analytics is a sophisticated form of analytics that has become well-known in the big data pertaining to health care. It uses the data mining method to predict what will happen in the future and gives necessary advice based on the predictions. It goes beyond data mining and is a radical branch of data engineering that predicts existences or probabilities according to the availability of the data. The two key concepts of predictive analytics, which is made up of numerous statistical and analytical techniques, are classification and regression.

It has the ability to manage both continuous and discontinuous changes. The predictive analysis, which combines the words predict and analysis, first analyses the given data and forecasts the meaning contained therein, as shown in the flow diagram in the figure. The process of the prediction is shown in Figure 3 below. They are the processes for gathering raw data, preprocessing it, and turning it into data that is manageable. This procedure is typically carried out using machine learning techniques.

Although there is a palpable relationship between statistics and data mining, data mining associated methods have em. The models for predictive analytics are constructed using the data mining tools and techniques, that identify the hidden patterns or the predictive information's that from the enormous volume of data, extracting the valuables available in the data and processing them applying the latest algorithms to detect the hidden information's in them[10].

Predictive analytics determines the likely future course of an event or the likelihood of specific current events; it is entirely focused on forecasting future probabilities and trends. It automatically analyses a sizable set of data with various variables, using techniques like text mining, neural networks, decision analytics, decision trees, genetic algorithms, regression modelling, and hypothesis testing, among others.

The variable known as the predictor, which is typically measured for a single person or the entire organisation in order to predict the future occurrences, the risks, and the opportunities hidden therein, is the fundamental component of predictive analytics[11-15].

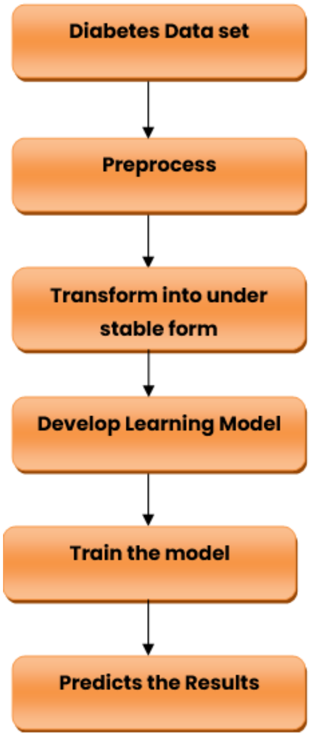

Fig -3: Prediction Model of Diabetes

5. BIG DATA IN PREDICTIVE ANALYTICS IN HEALTH CARE

Big data, which is able to handle a sizable amount of data gathered from the health care sectors and ensures the solutions to the major problems that arise in the medical industry, plays a significant role in predictive analytics, particularly in the medical domain.

The collection of large amounts of health care data gathered from patients and the general population using big data analytics in the medical field is emerging as a promising technology that can be used to advance prediction, performance, inventions, and comparative effectiveness. By incorporating big data and next-generation analytics into clinical research, the available information sources can become knowledgeable sources that power the healthcare system.

6. DATA SETS

The number of pregnancies, blood sugar levels, systolic and diastolic blood pressure, skin thickness, insulin, body mass index (BMI), plasma, Diabetes Pedigree Function (DPF), age, and class are the factors that were taken into consideration in this study to predict diabetes. These characteristics are useful in predicting diabetes. As a result, the modified dataset has twelve important attributes for predicting diabetes.

7. DISCUSSION

The use of big data in the healthcare system is essential. It helps with accurate disease prediction. Diabetes awareness is raised through the use of social media. Datasets gathered from clinical reports, doctor prescriptions, diagnostic reports, medical images, pharmacy information, insurance data, and social media data are used to forecast the type of diabetes and its risk in the future.

We gather data from WhatsApp and tweets on social media in cases and features like Name, DOB, Occupation, Food Habits, Food Causes Diabetes, Food Control Diabetes, Symptoms, Plasma, Glucose Density, Serum Insulin, Blood Pressure, History of Diabetes, BMI, Age, and Number of Pregnancies. These characteristics are employed to forecast diabetes.

The right disease prediction is made using machine learning algorithms. As compared to other algorithms, the SVM and Random forest algorithms yield better results, according to the study. Lack of exercise, poor diet, and lifestyle changes all contribute to an increase in diabetes. The country's finances will be impacted by the rise in diabetic patients.

The elderly and younger generations are moderately aware of diabetes, while the middle-aged group is less so. More people who are employed than farmers and stay-at-home moms are aware of diabetes. People are less aware of diabetes based on location. The best way to manage diabetes is through exercise and an the accuracy of each techniques is shown in table 1.

Table -1: Accuracy level for different dataset

S.NO	TECHNIQUES	DATASET	ACCURACY
1	Naïve Bayes, SVM & K-NN	50000 instances and 12 features	SVM Executes better than any other algorithm
2	SVM, Naïve Bayes & Random Forest	Electronic Health Record	Final output is the feature selection and classification algorithm achieved high accuracy, sensitivity, and specific and minimum error rate.
3	Random Forest, K-NN & SVM	650 records	Random Forest algorithm provides data more correctly and accurately.

8. CONCLUSION

People have very little awareness of diabetes in their minds. Big data analysis aids in providing patients with affordable treatment and care. By performing proactive diagnosis in order to construct the nation in economy mode with less risk, we can avoid the effects of diabetes in the future. This paper examined the diabetes prediction model and diabetes awareness.

REFERENCES

[1] Zheng W, Qin Y, Bugingo E, Zhang D, Chen J. Cost optimization for deadline-aware scheduling of big-data processing jobs on clouds. Future Gener Comput Syst. 2018; 82:244–55.

[2] Agrawal A, Choudhary A. Health services data: big data analytics for deriving predictive healthcare insights. Health Serv Eval. 2019 doi: 10.1007/978-1-4899-7673-4_2-1. [CrossRef] [Google Scholar]

[3] Wang, Lidong, and Cheryl Ann Alexander. "Big data analytics as applied to diabetes management." European Journal of Clinical and Biomedical Sciences 2.5 (2016): 29-38.

[4] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal.

[5] Perveen, S., Shahbaz, M., Guergachi, A., & Keshavjee, K. (2016). Performance analysis of data mining classification techniques to predict diabetes. Procedia Computer Science, 82, 115-121.

[6] Pradeep, K. R., & Naveen, N. C. (2016, December). Predictive analysis of diabetes using J48 algorithm of classification techniques. In Contemporary Computing and Informatics (IC3I), 2016 2nd International Conference on (pp. 347-352),IEEE.

[7] Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q., & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. The Kaohsiung journal of medical sciences, 29(2), 93-99.

[8] Guttikonda, Geetha, Madhavi Katamaneni, and MadhaviLatha Pandala. "Diabetes Data Prediction Using Spark and Analysis in Hue Over Big Data." 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC). IEEE, 2019.

[9] Gopala Krishna Palam, "The Practice of predictive analytics in healthcare", September 2015.

[10] Rumsfeld, John S., Karen E. Joynt, and Thomas M. Maddox. "Big data analytics to improve cardiovascular care: promise and challenges." Nature Reviews Cardiology 13, no. 6 (2016): 350.

[11] Jayanthi, N., B. Vijaya Babu, and N. Sambasiva Rao. "Survey on clinical prediction models for diabetes prediction." Journal of Big Data 4, no. 1 (2017): 26.

[12] Alharthi, Hana. "Healthcare predictive analytics: An overview with a focus on Saudi Arabia." Journal of infection and public health 11, no. 6 (2018): 749-756.

[13] Shyni, S., R. Shantha Mary Joshitta, and L. Arockiam. "Applications of big data analytics for diagnosing diabetic mellitus: issues and challenges." International Journal of Recent Trends in Engineering & Research 2, no. 06 (2016): 454-461.

[14] Tekieh, Mohammad Hossein, and Bijan Raahemi. "Importance of data mining in healthcare: a survey."In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, pp. 1057-1062. 2015.

[15] Raghupathi, Wullianallur, and Viju Raghupathi. "Big data analytics in healthcare: promise and potential." Health information science and systems 2, no. 1 (2014): 3.