Analyzing Diabetic Data Using Naive-Bayes Classifier

[1]A. Sharmila Agnal , [2]E. Saraswathi
Assistant Professor, Department of Computer Science and Engineering,
SRM Institute of Science and Technology, Chennai, India
sharmils3@srmist.edu.in, saraswae@srmist.edu.in

ABSTRACT--*Approximately 422 million people across the world have diabetes, particularly in countries where the average income is in the middle and lower end of the economic spectrum. Statistics reveal that every year, about 1.6 million deaths are recorded which can be directly attributed to diabetes. The graph suggests that number of cases as well as the prevalence of diabetes have been steadily incrementing over the past few decades. Through this new implementation of the Bayesian Classifier, raw medical data is analyzed and the risk of diabetes diagnosis based on each patient's medical information can be calculated. The raw data is converted into class labels and the likelihood of a positive potential diabetes case is derived, as a probability (≤1). This can not only be used by healthcare professionals but also by common users, and can be useful in detecting the risk and preventing it in time without taking any medical tests. This classifier uses very basic information that would be known to each patient or can easily be obtained.*

KEYWORDS: Diabetes, Naive-Bayes Classifier, Prediction, Healthcare, Decision Tree, Confusion Matrix.

1. INTRODUCTION

Diabetes mellitus or as it is simply called, diabetes is a disease which perpetuates in the metabolic method. It causes high blood sugar levels in an individual. The pancreas produce on of the most essential hormone of the human body, the insulin. The insulin extracts blood sugar and transports it for storage or to be used as cellular energy. For a diabetic patient, the body either produces insufficient insulin or is incapable of using the insulin that has been developed developed. Diabetes can be broadly categorized into four types: Type 1 , Type 2 , Pre-diabetes and Gestational Diabetes.

Type 1 Diabetes which is also called as the insulin-dependent diabetes or juvenile-onset diabetes [5] is endured by almost ten percent of people. Type 2 Diabetes is also named as insulin resistance diabetes or adult-onset diabetes[5]. When the level of blood sugar levels are higher than normal then the diabetes is called as Pre-diabetes. Type 3 Diabetes or Gestational diabetes occurs only during pregnancy when hormones are blocked by the insulin.[5]

Diabetes can be predicted or diagnosed by some of the common symptoms like increase in hunger, increased thirst, loss of weight, blurred vision, extreme tiredness. Also, other than these common symptoms these can be categorized differently in males and females. In males, erectile dysfunction, and poor muscle strength. In females, yeast infection, itchy and dry skin. These are some common symptoms of whatever type of diabetes it may be.

Diabetes ailment can be prevented on the basis of what type of diabetes it is. Type 1 Diabetes can be prevented by insulin treatment. The treatment replaces hormones that the body isn’t able to produce. Insulin treatment will be given depending on how long it lasts and quickly it starts working. Type 2 Diabetes can be prevented by changing diet and exercising. If this change in diet doesn’t work then the body needs medication that is some sort of drugs that will work on lowering the blood sugar level. Gestational diabetes can be prevented by taking care of diet and blood sugar levels.

The exponential rise in medical information has resulted into advanced, long procedure for processing and interpretation of the processed information for each patients and physicians. DM interference, diagnosis, risk reduction, and timely intervention don't seem to be trivial tasks since it's terribly difficult to process the noninheritable information into helpful knowledge.[3]

In this implementation, the data that is taken is categorized into Training data and Test data. Training data is the one where details of number of patients are present with the predicted result that either patient has a risk of diabetes or not. Test data is where implementation works and hence predicts that a person has a risk of diabetes or not. Test data is the data through which the implementation of the project is checked.[8] This training and test data are again categorized on the basis of attributes that are considered for the prediction and implementation purposes of the project. Attributes mean the ones that will be considered for predicting the results of a person for risk of diabetes.

Classifier means the categorization of the details of a patient on the basis of attributes that are considered for prediction of risk of diabetes. The data collected from people is then classified accordingly in an organized way for better results regarding the risk level of diabetes. In this work, the Naive Bayes Classifier is used for the results of the risk of diabetes. This classifies the data according to the attributes specified and hence implements using probabilistic characters. The result is in the form of probability whether the person has diabetes or not.

The implementation inputs details refer to the attributes that are categorized for the prediction of the risk level of Diabetes. The Features are:

1. Glucose
2. Body Mass Index
3. Diabetic Pedigree Function
4. Age
5. Blood Pressure
6. Insulin
7. Skin Thickness
8. Pregnancies

The Algorithm for implementation of this project is Naive Bayes Classifier. Naive Bayes is a classifier which gives results in a probabilistic character that is the probability of occurrence of a particular thing. In this work, the role of Naive Bayes is when the details of a patient input the values in all respective attributes that are age, blood pressure and so on it calculate the individual probability of each attribute and hence sets a range of each attribute. After this, it classifies and compares each attribute range hence after all calculation provides the result in probability form, of whether a person has a risk of Diabetes or not.

Naive Bayes plays an important role in classifying the attributes and hence calculating the probability of each attribute individually. Hence the Naive Bayes is used in implementation for prediction purposes. Based on probability results, this classifier predicts whether a person is suffering from diabetes or not. Prediction of diabetes depends on the range of each attribute. The prediction will result according to the cut-offs specified by default before all the calculations. Hence the prediction will predict the level of risk for diabetes. Prediction works on a probability basis. The results come out whether a person has a risk of diabetes or not. This prediction is with the help of the Naive Bayes classifier by classifying the attributes according to the details provided.

2. BACKGROUND AND RELATED WORK

Apache Hadoop is open supply and provides economical storage and less time computation platform. Map Reduce, Apache Hive, Spark SQL, Apache Pig and HBase square measure elements of the Hadoop system[4] within the existing system of the paper by Sangavi et al. R is employed for sleuthing polygenic disease and its stage of occurrence.

Hadoop is a clustering algorithm which works on cluster formation of unstructured and unspecified data. This algorithm comes under unsupervised learning and hence is less secure. Decision tree algorithm is used for processing the dataset but the implementation uses MapReduce for analysis[1]. However, due to map and reduce stages the process slows down and hence isn’t that precise.

2.1 EXISTING SYSTEM ALGORITHM

This algorithm has a Hadoop Distributed File System (HDFS) which is used for handling large volumes of data. By this the data is partitioned and stored across different cluster nodes.[1] Hence, according to this paper for predicting diabetes using Hadoop clearly states that this algorithm won’t be able to work precisely [1]. And also this algorithm doesn’t work on small files. Therefore, as compared to other algorithms, Hadoop won’t give results accurate and precise for detection of Diabetes.

3. PROPOSED SYSTEM

The proposed system mainly focuses on how much percentage of patients are prone to or have diabetes, as a set of positive and negative information using the MapReduce theorem, Naive Bayes classifier, and J48 decision tree.

3.1 NAIVE BAYES CLASSIFIER

Naive Bayes could be a method to construct classification models which assign category labels to drawback labels, that are diagrammatical as vectors of feature values, wherever category labels are derived from some finite set. Such classification isn't associate in Nursing algorithmic rule for coaching, however a group of algorithms supported a general principle.

All Naive Bayes classifiers assume that the worth of 1 explicit attribute is freelance of the worth of another attribute, given the category variable. In several sensible applications, parameter estimation for the naive model uses the principle of most chance. Alternatively, we are able to work with Naive Bayes models while not acceptive theorem chance or victimization any theorem ways. It examines all the symptoms from a given knowledge set and uses the contingent probability to work out the chance of polygenic disease.

There are several deciding factors like glucose level, BMI, weight, age, pressure level levels and hormone levels. This classifier needs an extremely low range of training information for parameter estimation.

By Bayes' theorem, the conditional probability is formulated as shown in equation 1.

$$p(C_a \mid Z) = p(C_a)\, p(C_a|Z) / p(Z) \qquad (1)$$

where,
$p(C_a \mid Z)$=posterior probability
$p(C_a|Z)$= Likelihood
$p(C_a)$= Class Prior Probability
$p(Z)$=Predictor Prior Probability

With the help of Bayesian probability terminology, we can rewrite the equation 1 as:

POSTERIOR=PRIOR*LIKELIHOOD/EVIDENCE ***(2)***

3.2 CONFUSION MATRIX

The performance of a classifier in a test data set with known true values is described using the confusion matrix which looks like a table.Thereby, the operation of an algorithm can be visualized with the help of confusion matrix.

This helps in fast and smooth identification of confusion between classes. For an example, one class is often mistaken to be another. Many performance measures are calculated from the confusion matrix. The amount of the right and wrong guess is summarized by the numerical values and reduced by each category. This plays a vital role in the confusion matrix. This matrix depicts the ways in which the split model is confused during prediction. It provides insight not only on the mistakes made by the classification model but also the types of mistakes made.

3.2.1 Definition of The Terms

- **i.** Positive or Po: Positive observation.
- **ii.** Negative or Ne: Negative observation
- **iii.** True Positive or tPo: Observation is positive as well as it has been predicted to be positive.
- **iv.** False Negative or fNe: Observation is positive, but it has been predicted to be negative.
- **v.** True Negative or tNe: Observation is negative as well as it has been predicted to be negative.
- **vi.** False Positive or fPo: Observation is negative, but it has been predicted to be positive.

3.3 CLASSIFICATION RATE/ACCURACY:

The following relation gives the Classification Rate/Accuracy:
Accuracy(X)

$$X= tPo + tNe / tPo + tNe + fPo + fNe \qquad (3)$$

3.3.1. Recall:

$$Recall=TruePositive/TruePositive+FalseNegative \qquad (4)$$

Recall is defined as the ratio between the total number of positive examples which have been correctly classified and the total number of positive examples. A Recall with high value will indicate that the class has been correctly recognized or in other words there are a small number of False Negative.

3.3.2 Precision:

$$Precision=TruePositive/TruePositive+FalsePositive \qquad (5)$$

To precision value is obtained by dividing the total number of positive examples which have been correctly classified by the total number of positive examples which have been predicted. Higher Precision values indicate that an example which has been labeled as positive is indeed positive.

i. **High recall, low precision:**

   This shows that most of the positive examples predicted are in fact correctly identified which means there is low False Negatives but lot of false positives.

ii. **Low recall, high precision:**

   This means that there is missing of many positive examples i.e, high False Negative but those we predict to be positive are in fact actually positive i.e, low False Positive.

3.4 CLASS LABELS:

As shown in [Fig.1], the dataset is classified as:

i. Glucose: Normal - Pre-Diabetic – High
ii. Blood Pressure: Normal – High
iii. Body Mass Index: Ideal – Overweight
iv. Diabetes Pedigree Function: Likely – Unlikely
v. Age: Adult - Middle Aged – Elderly

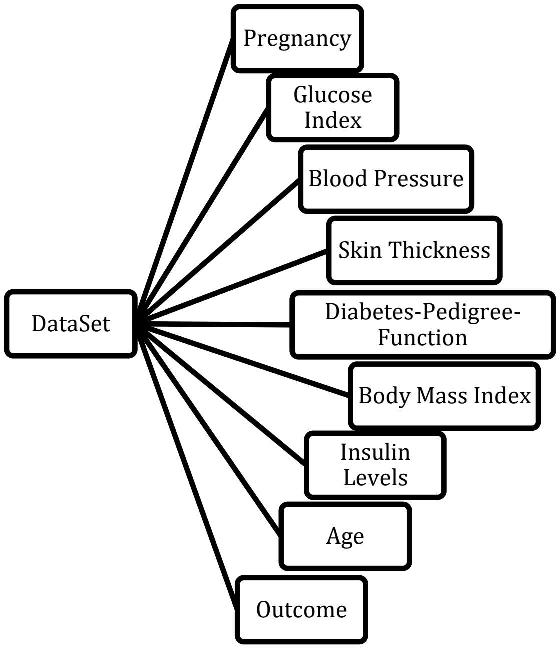

FIG. 1 DATASET CLASSIFICATION

4. SYSTEM ARCHITECTURE

The basic architecture of the system is illustrated by the flow diagram given in [Fig 2]. The diagram basically explains how the process goes. The process starts with the data collection and formation of a dataset for analysis. After the formation of the dataset, it goes through the process of data cleaning and dimension aggregation. MapReduce Theorem is used after the removal of not useful data. Hence after all the cleaning and reduction classifier analysis according to the data presented using the confusion matrix to analyze deviation from expected values.

4.1 Dataset:

A collection of sets of data which are related to one another and is composed of separate elements which can be manipulated on requirement basis is called a dataset. The Dataset used in this work has been obtained from UCI Machine Learning Repository.

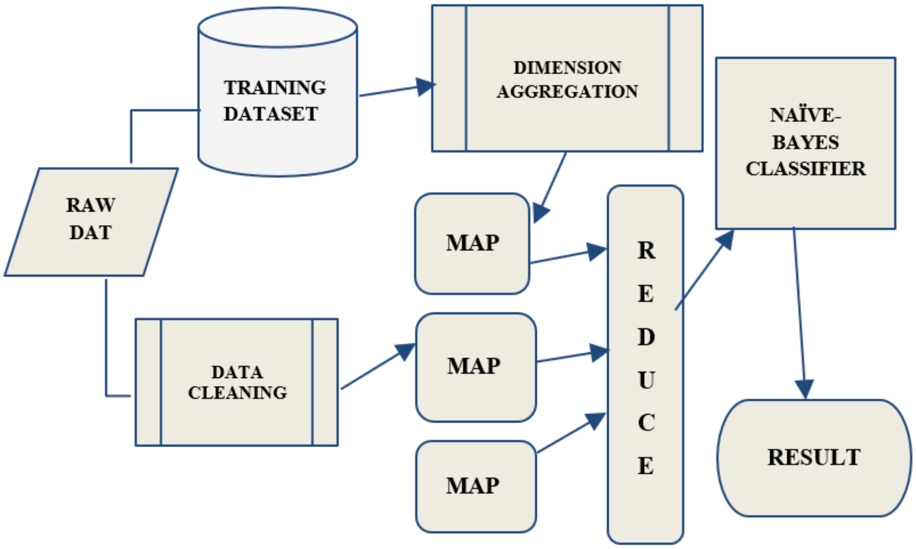

FIG. 2 DIABETES DATA ANALYSIS

4.2 Data Cleaning:

Data Cleaning is a process which helps to detect and correct corrupt or inaccurate data from a dataset. It also involves identification of incomplete, inaccurate, or irrelevant parts of the data followed by their replacement, modification, or deletion of the coarse data.

4.3 Data Aggregation:

In data aggregation, all the obtained information is gathered or aggregated as the name suggests and then expressed in the form of a summary. It serves several purposes such as statistical analysis. In this system, data aggregation is done using more relevant information about diabetic parameters based on specific attributes namely age, BMI, medical history (Diabetes Pedigree Function), glucose levels, skin thickness, Insulin, Blood Pressure of the patient.

4.4 MapReduce:

MapReduce is framework used for distributed processing of large amount of data and helps in computing clusters. This framework is used to schedule tasks of processing data, monitor the data, and re-execute any failed tasks.

4.5 Naive Bayes Classifier:

Naive Bayes classifier is an approach based on probability which works on each class label independently. It basically works on the Bayesian model of probability. For the estimation of any attribute it uses a method of maximum likelihood which works independently and hence gives a probability of occurrence of data. This classifier assumes the value independent of any feature or variable. It works very efficiently in supervised learning.

5. ANALYSIS & IMPLEMENTATION

5.1 NAÏVE-BAYES ALGORITHM

$$p(C_a \mid Z) = p(C_a)\, p(C_a|Z) / p(Z)$$

The result gives the likelihood of Diabetes Diagnosis as a Probability (≤1). The Outcome 0 means that the patient is Non-Diabetic. If the outcome is 1, the result is positive for Diabetes.

5.2 ANALYSIS AND STUDY

The data has been analyzed and presented categorically as various comparisons:

- **Bar Graph**:

X-axis: Level; *Y-axis*: No. of Patients [Fig.3]

- **Histogram**:

X-axis: Attribute data Points; *Y-axis*: No. of Patients (as shown in [Fig.4], [Fig. 5], and [Fig. 6])

- **Time Plot**:

X-axis: Time; *Y-axis*: Attribute data Points (as shown in [Fig. 7] and [Fig. 8])

- **Box Plot**: *X-axis*: Outcome; *Y-axis*: Age (as shown in [Fig. 9])

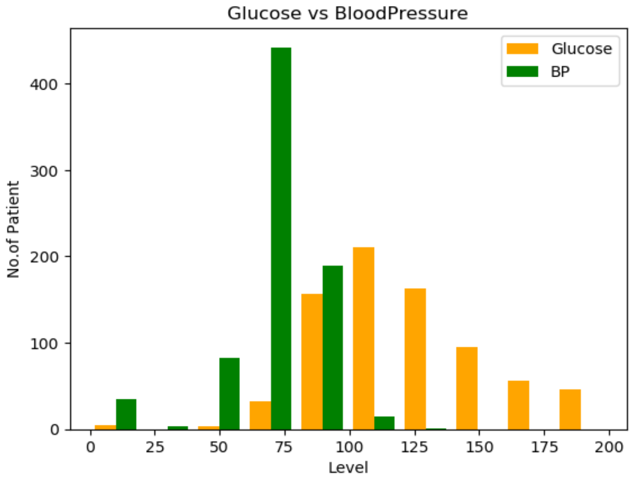

FIG. 3 BAR CHART OF GLUCOSE AND BLOOD PRESSURE

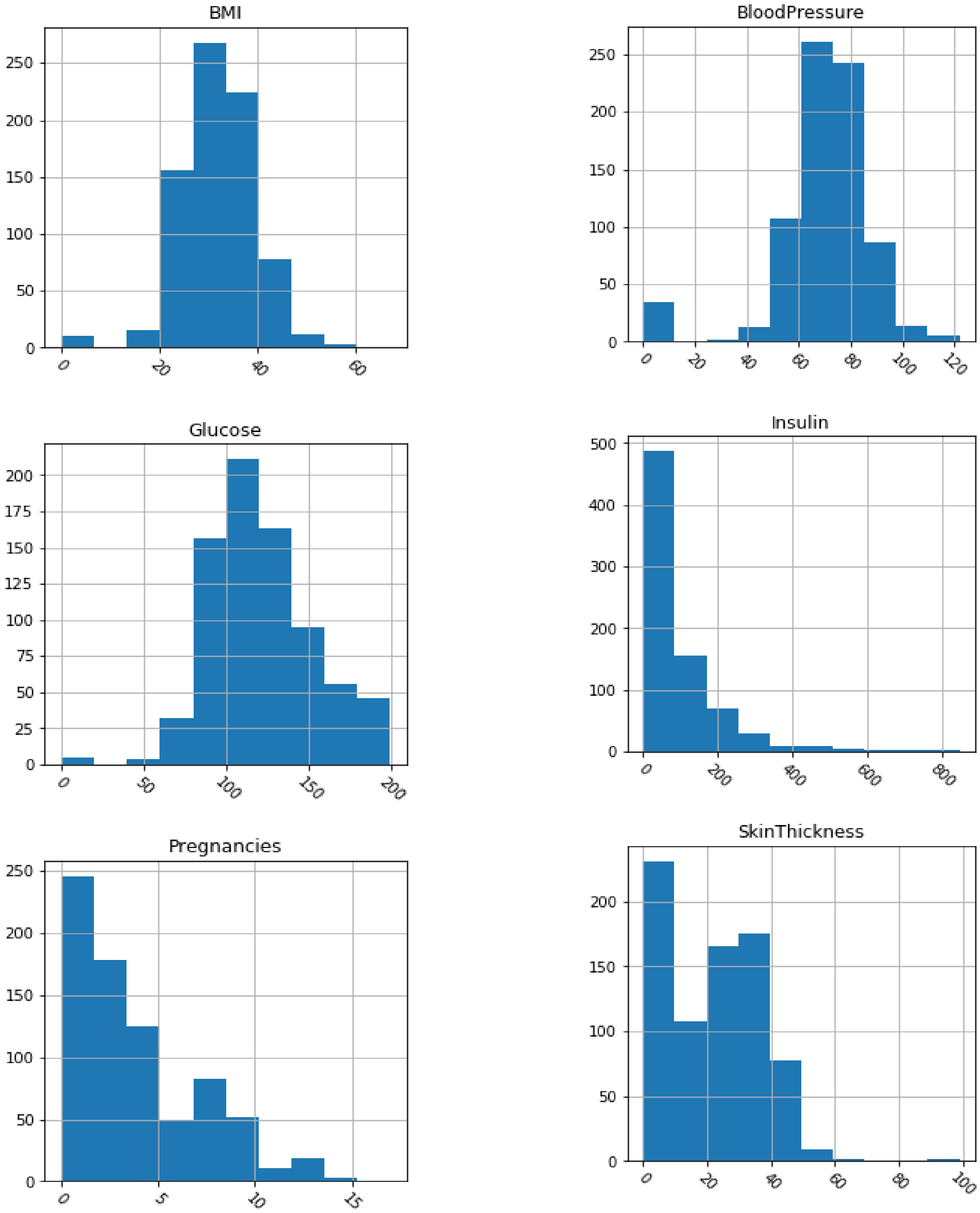

FIG 4. HISTOGRAM OF BMI, GLUCOSE AND PREGNANCIES

FIG 5. HISTOGRAM OF BLOOD PRESSURE, GLUCOSE LEVELS AND SKIN THICKNESS

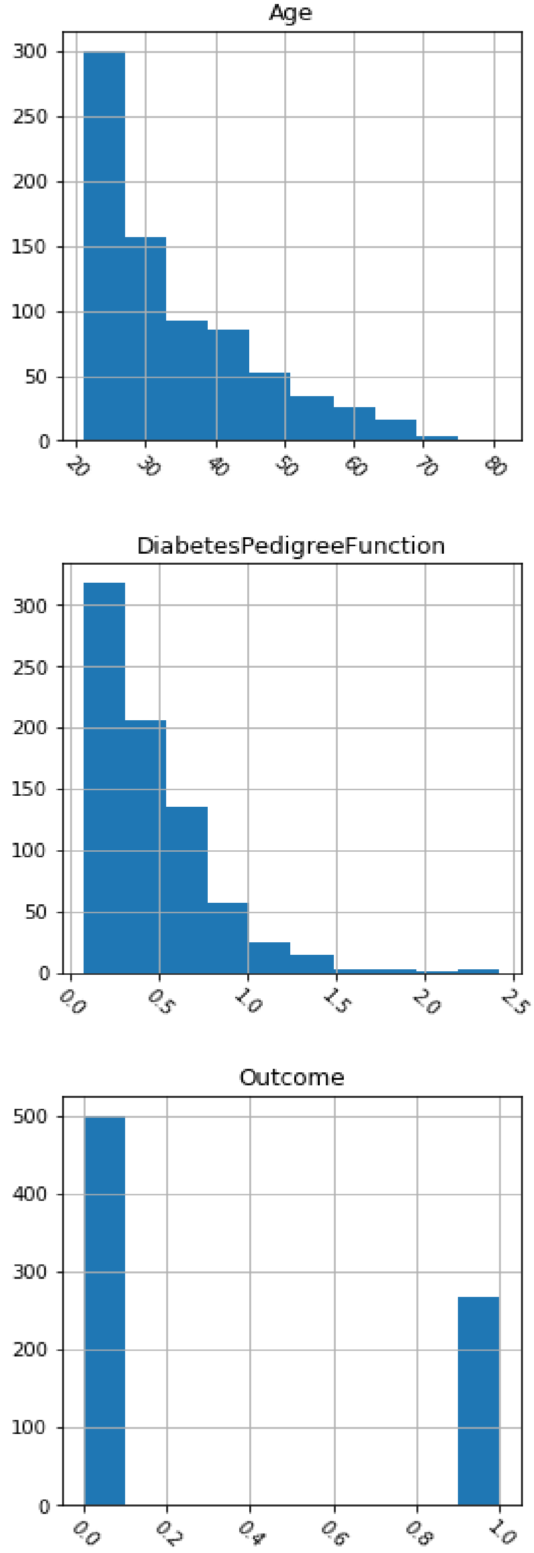

FIG 6. HISTOGRAM OF AGE, DIABETES PEDIGREE FUNCTION AND THE FINAL OUTCOME

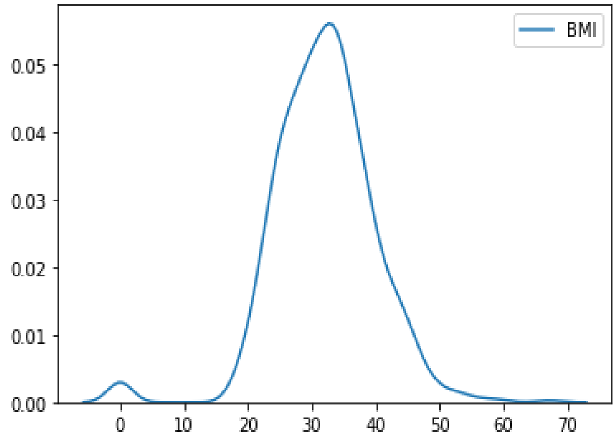

FIG. 7 TIME PLOT OF BMI

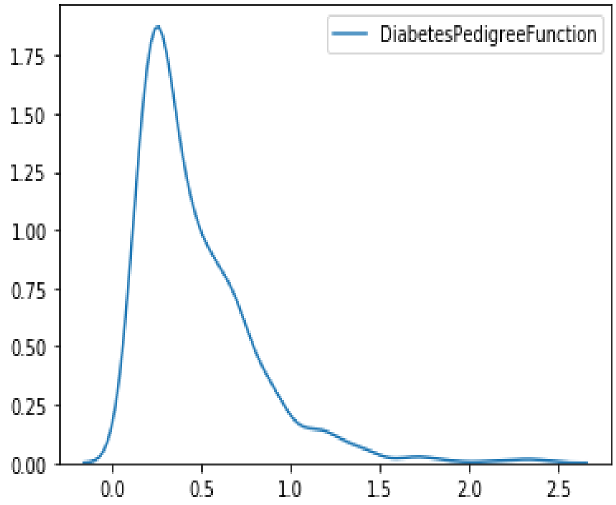

FIG 8. TIME PLOT OF DIABETIC PEDIGREE FUNCTION

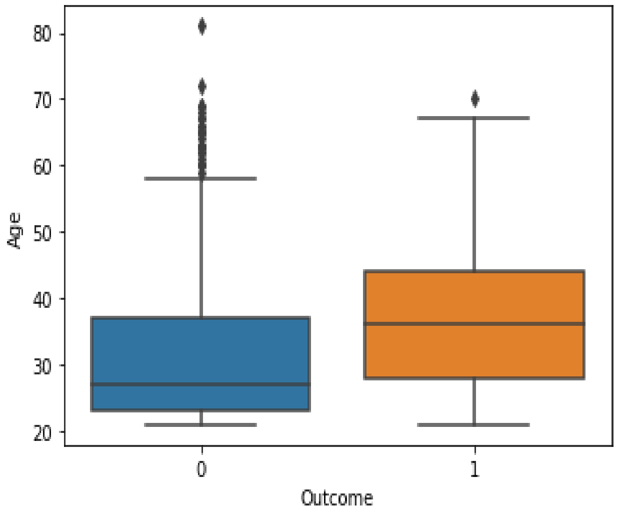

FIG.7 BOX PLOT OF AGE AND OUTCOME

6. CONCLUSION AND FUTURE WORK

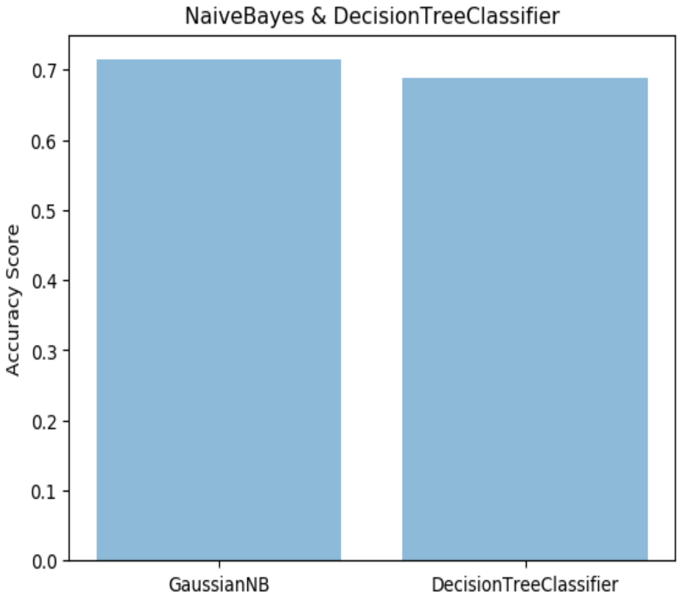

FIG.8 HISTOGRAM OF NAIVE BAYES AND DECISION TREE CLASSIFIER

Fig. 8 displays the accuracy level recorded based on the performance of the algorithms on the database that we have considered. This research work focuses widely on the methodological approach of Naive Bayes Classifier to detect diabetes accurately and precisely. The proposed approach can handle both large and small amounts of data in a better way than Hadoop Technology. This process being a classification algorithm is fast, secure, easy to adapt and provides accurate results. By making this more accurate, it can be useful in medical research and a mobile application can also be built based on these findings to make it more accessible to common users and users who have a risk of developing diabetes.

REFERENCES

[1]S. T. Prasad, S. Sangavi, A. Deepa, F. Sairabanu and R. Ragasudha, "Diabetic data analysis in big data with predictive method," International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies,2017.

[2]Sossi Alaoui, Safae & Farhaoui, Yousef & Aksasse, B, “Classification algorithms in Data Mining”. International Journal of Tomography and Simulation,2018.

[3] Geshwaree Huzooree, Kavi Kumar Khedo, Noorjehan Joonas “Glucose Prediction Data Analytics for Diabetic Patients Monitoring,2017.

[4] P.Bhardwaj and N. Baliyan, "Hadoop based Analysis and Visualization of Diabetes Data through Tableau," Twelfth International Conference on Contemporary Computing (IC3),2019

[5] S. Rani and S. Kautish, "Association Clustering and Time Series Based Data Mining in Continuous Data for Diabetes Prediction," Second International Conference on Intelligent Computing and Control Systems (ICICCS),2018

[6] Sahnius Usman, Mamun Bin Ibne Reaz, Mohd Alauddin Mohd Al “Risk Prediction Of Having Increased Arterial Stiffness Among Diabetic Patients Using Logistic Regression”,2018

[7] Peng Zhao, Illhoi Yoo “A Self-adaptive 30-day Diabetic Readmission Prediction Model based on Incremental Learning”,2017

[8] Raja, S. Kanaga Suba, and T. Jebarajan. "Reliable and secured data transmission in wireless body area networks (WBAN)." European Journal of Scientific Research 82, no. 2 (2012): 173-184.

[9] S.Ananthi , V.Bhuvaneswari “Prediction of heart and kidney risks in Diabetic Prone Population using Fuzzy Classification”, 2017.

[10] Jeffrey Dean and Sanjay Ghemawat “MapReduce: Simplified Data Processing on large clusters”, Communications of the ACM, 2008.

[11] M. Chen, J. Yang, J. Zhou, Y. Hao, J. Zhang and C. Youn, "5G-Smart Diabetes: Toward Personalized Diabetes Diagnosis with Healthcare Big Data Clouds," in IEEE Communications Magazine, vol. 56, no. 4, pp. 16-23, April 2018.

[12] Rajaram K., Usha Kiruthika S. (2010) Dynamic Contract Generation and Monitoring for B2B Applications with Composite Services. In: Das V.V., Vijaykumar R. (eds) Information and Communication Technologies. ICT 2010. Communications in Computer and Information Science, vol 101. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-15766-0_55.