International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Big Data Analysis for Diabetes Recognition using Classification Algorithms

S. Kamalakkannan, R. Thiagarajan, S. Mathivilasini, R. Thayammal

Abstract*: *In that paper, we've an inclination to project as checking the whole patient ill health victimization Naive Bayes classification and J48 decision tree. As a result of the information, enormous process comes from mu ltiple, heterogeneous, autonomous sources with sophisticated an d evolving relationships and continues to grow. So in that, we'll take results of what proportion share patients get ill health as a positive knowledge and negative knowledge. Huge info is difficult to work with victimization most database management systems and desktop statistics and internal representation packages. The projected shows a huge process model, from the data mining perspective. Victimization classifiers, we've an inclination to unit method congenital disease share and values unit showing as a confusion matrix. We've an inclination to projected a replacement classification theme which could effectively improve the classification performance inside the situation that employment dataset is out there. During this dataset, we have nearly 1000 patient details. We'll get all that details from there. Then we have a tendency to unit attending to sensible and unhealthy values square measure victimization naive Bayes classifier and J48 tree.

Index Terms*: *Big Data, Diabetes, J48 Tree, Naïve Bayes Classification.

I. INTRODUCTION

Diabetes is a chronic condition caused by awkwardness in insulin discharge, which has an unsettling effect on blood sugar levels. According to the American Diabetes Association, there was diabetes in 2012 among 29.1 million people in the United States (9.3 percent of the population). There were 8.1 million undiscovered people among this population. Diabetes occurrence continues to expand, with 1.7 million new analyzes being conducted in 2012. Moreover, 86 million people over the age of 20 had pre-diabetes and could be considered to be in a state in the midst of health and chronic disease. Diabetes remains the seventh leading cause of death in the United States in 2010, with 69,071 deaths directly related to it, and some more (an aggregate of 234,05) as the leading or contributing cause[3]. Critical efforts are therefore the compelling administration of health, pre-diabetes, and diabetes.The term heart disease is a chronic condition that in cludes the different heart-influencing diseases. Heart diseas e in the distinctive nations, including India, was the real reas on for losses. It kills one person in the United States at reg ular intervals.

The World Health Organization has assessed that around the world there are 12 million deaths, consistently dueto heart d isease. A large portion of the deaths from cardio-vascular dis eases occur in the United States and other created nations [4]. It is also the main cause of death in different creating nations. All in all, it is regarded as the essential explanation for grown-up deaths. In this paper, we focus on the algorithm for the decision tree, Naïve Bayes to analyze the datasets for diabetes. Using UCI machine learning datasets, experiments were conducted. The performance of all the methodologies was also evaluated.

II. BIG DATA AND DATA MINING

Huge information open source advances have picked up a lot of footing because of the showed capacity to parallel process a lot of information. Big Data distress large-volume, mounting files collections that are multipart and have several sovereign cradles. [8] Both parallel handling and method of conveying calculation to information has made it conceivable to process extensive datasets at fast. These key highlights and capacity to process huge information has been an extraordinary inspiration to investigate the engineering of the business driving enormous information preparing system by Apache, Hadoop. See how this huge information stockpiling and investigation is accomplished and exploring different avenues regarding RDBMS versus Hadoop condition has demonstrated to give an extraordinary knowledge into much discussed innovation. The measure of information created each day on the planet is detonating. The expanding volume of computerized and web based life and web of things is energizing it much further. The rate of information development is astounding and this information comes at a speed, with assortment (not really organized) and contains abundance of data that can be a key for picking up an edge in contending organizations. Capacity to dissect this colossal measure of information is bringing another period of profitability development, advancement and purchaser surplus [1]. "Enormous information is the term for an accumulation of informational indexes so substantial and complex that it ends up hard processing it utilizing conventional database administration devices or information preparing applications.

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author
Dr.S.Kamalakkannan*, Associate Professor, Department of Information Technology, Vels Institute of Science, Technology & Advanced Studies, Chennai.
R.Thiagarajan, Assistant Professor, Department of Computer Science, St.Joseph's College (Arts & Science), Chennai.
Dr.S.Mathivilasini, Department of Computer Science, Ethiraj College for Women, Chennai.
R.Thayammal, Department of Computer Science, Chennai.

The difficulties incorporate the zones of catch, curation, stockpiling, look, sharing, exchange, investigation, and perception of this information" [2]. Presently the term Data Mining, Finding for the correct valuable data or information from the gathered information, for future activities, is only the information mining [6].

III. DATASET DESCRIPTION & PREPROCESSING

The paper investigates the part of Choice Tree and Gullible Bayes Classifier as Information Mining procedures in deciding diabetes in ladies. The primary target is to estimate if the patient has been influenced by diabetes utilizing the information mining apparatuses by utilizing the medicinal information accessible.

Pima_diabetes-weka.filters.supervised.instance.SMOTE-C0-K5-P100.0-S1-weka.filters.unsupervised.instance.Randomize-S42.

Table 6.1 demonstrates a concise depiction of the dataset that is being considered.

Table I: .Dataset Description

Dataset	No. of attributes	No. of instances
Pima_diabetes-weka.filters.supervised.instance.SMOTE-C0-K5-P100.0-S1-weka.filters.unsupervised.instance.Randomize-S42	9	1036

The attributes descriptions are shown in Table II below.

Table II: Attributes Description

Attributes	Relabeled values
1. Number of times pregnant	preg
2. Plasma glucose concentration	plas
3. Diastolic blood pressure (mm Hg)	pres
4. Triceps skin fold thickness (mm)	skin
5. 2-Hour serum insulin	insu
6. Body mass index (kg/m2)	mass
7. Diabetes pedigree function	pedi
8. Age (years)	age
9. Class Variable (0 or 1)	class

IV. PROPOSED DATA MODEL

Load past datasets to the framework.

•Data pre-handling has done utilizing coordinating WEKA device.

•Following activities are performed on the dataset after that.

a. Supplant Missing qualities.

b. Standardization of qualities.

The last makes it simple to utilize the dataset, as the scope of the factors is confined from 0 to 1. Highlight choice has been utilized utilizing the CfsSubsetEval calculation, and the characteristics acquired after execution are as per the following:

Property choice on all info information

1. Number of times pregnant
2. Plasma glucose focus
5. 2-Hour serum insulin
6. Weight record (kg/m2)
7. Diabetes family work
8. Age (years)

The graphic insights of the dataset are introduced in Table III. Since the parameters are standardized the scope of all are in the range 0 to 1.

Table III: Descriptive Statistics of Transformed Dataset

Attribute	Minimum	Maximum	Mean	Standard Deviation
preg	0	1	0.24	0.197
plas	0	1	0.634	0.163
Insu	0	1	0.1	0.143
mass	0	1	0.488	0.112
pedi	0	1	0.177	0.142
age	0	1	0.22	0.191

- User input information to the framework keeping in mind the end goal to analyze whether he has the sickness or not.
- Build two models utilizing J48 Choice Tree and Credulous Bayes Calculations and prepare the informational index.
- Test the dataset utilizing two models.
- Get the assessment comes about.
- Get the anticipated voting from all classifiers and gives the analytic outcome.

V. PERFORMANCE EVALUATION CRITERIA FOR MODEL

To investigate and look at the execution of the information mining techniques introduced in our examination, we apply different insights

1) Mean Absolute Error (MAE)

$$MAE = 1/n\sum_{i=1}^{n} |Predicted - Actual| \quad (2)$$

MAE is the quantity used to measure how close predictions are to the actual outcomes.

2) Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\sum_{i=1}^{n} (Predicted - Actual)^2 / n} \quad (3)$$

Where n is the number of observations for corresponding predicted and observed values of the model [9].

A. Confusion Matrix

The data about genuine and anticipated characterization framework is hold by the Perplexity network. It shows the exactness of the answer for an arrangement issue. The sections in the perplexity grid have the accompanying significance with regards to our examination.

tp is the quantity of right forecasts that an occasion is certain.

fn is the quantity of erroneous expectations that an example is negative.

fp is the quantity of erroneous expectations that an example is certain and

tn is the quantity of right forecasts that a case is negative.

Table IV: Confusion Matrix

		Predicted	
		Positive	Negative
Actual	Positive	t_p	f_n

	Negative	f_p	t_n

1) Recall /True Positive Rate /Sensitivity:
True positive rate (TP) is the proportion of positive cases that were correctly identified, as calculated using the equation:

$$R = \frac{tp}{tp+fn} \quad (4)$$

2) Precision :
Precision (P) is the proportion of the predicted positive cases that were correct, as calculated using the equation:

$$P = \frac{tp}{tp+fp} \quad (5)$$

3) Accuracy :
The proportion of the total number of predictions that were correct is known to be as Accuracy (AC). It shows overall effectiveness of classifier. It is determined using the equation [7]:

$$AC = \frac{tp+tn}{tp+fn+fp+tn} \quad (6)$$

VI. RESULTS ANALYSIS

From the outcomes no inheritable , each the methods have an almost very little distinction in blunder rate, but the speed split of 70:30 for Guileless Bayes procedure provides the slightest mistake rate and time taken to assemble demonstrate is least once contrasted with alternative to J48 executions. Each the models area unit skilled within the conclusion of polygenic disease utilizing the speed split of 70:30 of the informational index. A created demonstrates for determination of polygenic disease would force all the additional making ready data for creation and testing.

Table V: Summary of Prediction

Algorithms	CC	IC	MAE	RMSE	RAE	RRSE	Time Taken
J48	74.28	25.72	0.3172	0.4571	63.5	91.51	0.05 seconds
Naïve bayes	72.67	27.33	0.3121	0.4403	62.49	88.15	0.03 seconds

Table VI: Results of various measures

Algorithms	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area	Class
J48	0.734	0.192	0.781	0.734	0.757	0.801	tested_negative
	0.808	0.266	0.765	0.808	0.786	0.801	tested_positive
	0.772	0.23	0.773	0.772	0.772	0.801	Weighted Average
Naïve Bayes	0.77	0.295	0.709	0.77	0.738	0.819	tested_negative
	0.705	0.23	0.767	0.705	0.735	0.819	tested_positive
	0.736	0.261	0.739	0.736	0.736	0.819	Weighted Average

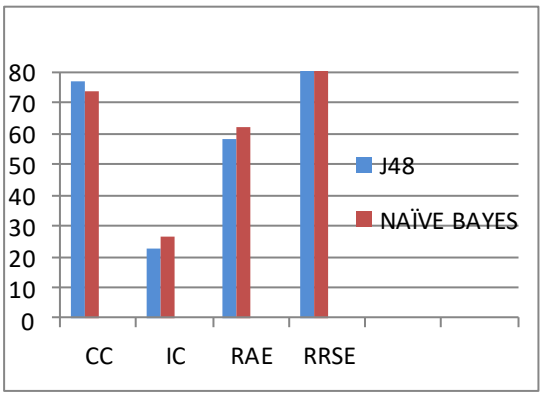

Fig 1: Comparative analyses of algorithms in terms of CC, IC, RAE, RRSE

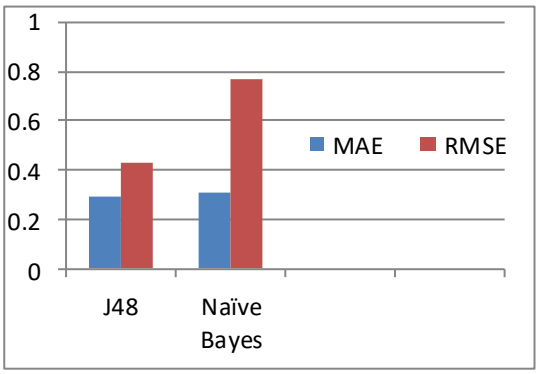

Fig 2: Comparative analysis of algorithms in terms of MAE, RMSE.

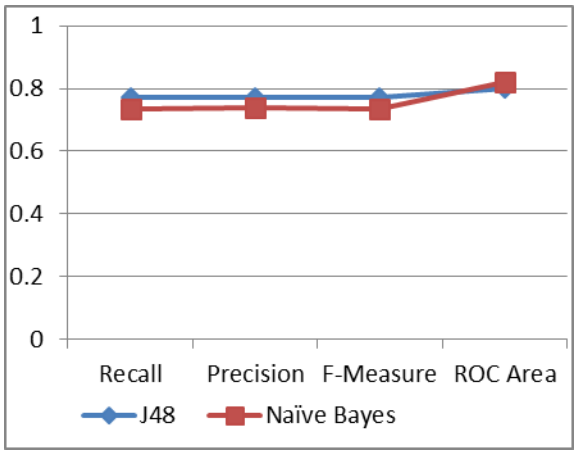

Fig 3: Comparative Analysis of algorithms in terms of Recall, Precision, Accuracy and ROC Area.

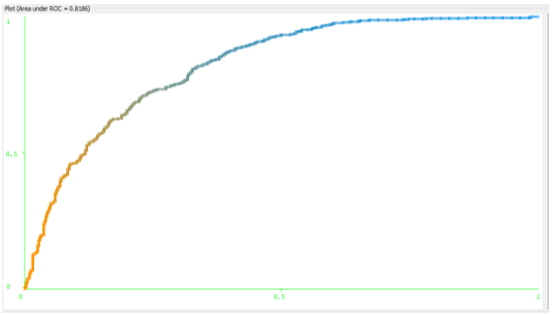

Fig 4: ROC Plot for Naïve Bayes

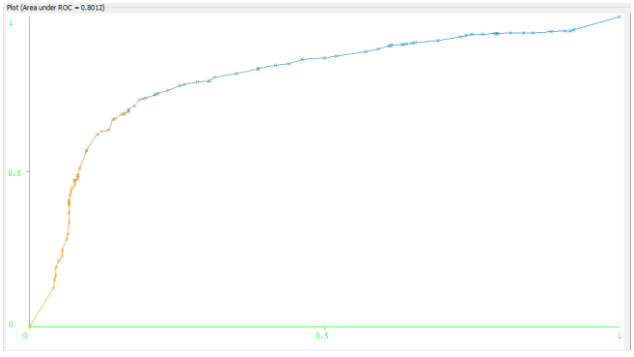

Fig 5: ROC Plot for J48

Table V indicates Naïve Thomas Bayes classifier being the smallest amount tough classifier have performed well with a exactitude of seventy two.67%,

whereas having relative supreme mistake sixty two.49%. The execution of J48 in ordering occurrences accurately is more than Naïve Thomas Bayes. .Fig five demonstrates similar examination of calculation as so much as properly Classified Instances, Incorrectly Classified Instances, Relative Absolute Error and Root Relative square Error. Fig half-dozen demonstrates similar investigation of calculation as so much as Mean Absolute Error, Root Mean sq. Error. Fig seven appearance at Recall, Precision, mythical monster space determined utilizing confusion framework. Likewise, the territory below the mythical monster is in addition thought of for all the four connected calculations above all Naïve Thomas Bayes,J48. . The mythical monster territory of Naïve Thomas Bayes is most elevated among J48. a lot of the zone secured higher is that the classifier. Further, the outcomes could be increased by applying large size reinvigorated informational indexes of sensible setting. Anyway we've to use different machine learning calculations utilizing reinvigorated informational indexes before summed up the outcomes.

VII. CONCLUSION

In this analysis work, the oftentimes used classification techniques J48, Naïve mathematician Algorithms area unit analyzed, on the medical dataset to search out the best answer for polygenic disorder. The performance indicators accuracy, precision, error rate area unit calculated for the given dataset. Accusation beside with a correct knowledge preprocessing technique will get well the accuracy of the classifier. The perform of information standardization had noticeable impact on categorization performance and significantly increased the performance of J48. Supported the parameters taken for analysis, the performances of the 2 algorithms area unit analyzed. The results show that the performance of J48 technique is considerably superior to the opposite technique for the classification of polygenic disorder knowledge. Through the tumor of suggestion machinery and its sustained dawn into the healing and aid sector, the belongings of polygenic disorder and their cyphers area unit well acknowledged. Call tree and Naïve mathematician rule learning at discovery elucidations to identify the malady by using the patterns found within the knowledge through language analysis. This letter of inquiry applies to advanced instrumentation within the computer science society. These tools area unit terribly helpful for analyzing these snags and additionally provide the likelihood of higher resolution-constructing processes. during this knowledge, we've used “Naive mathematician and call Tree Algorithm” for analysis as a result of this classification predicts correct results for several issues.

VIII. FUTURE WORK

Our future work can involve the integration of the varied nominative algorithms to enhance the accuracy so the designation will grow to be a lot of correct just in case of unnoticeably known knowledge sets. The Work is extended which will be analyzed with alternative massive knowledge tool. Hence, the project will meet the stress of future additionally. To enhance the general accuracy, it's necessary to use a lot of knowledge set with sizable amount of attributes and use the simplest feature choice methodology in future. Future works may additionally embrace hybrid classification models by combining a number of the information mining techniques.

REFERENCES

1. Big Data Analysis – Hadoop Performance Analysis by KetakiSubhashRaste Master of Science in Computer Science ,San Diego State University, 2014.
2. Wikipedia. Big data, 2014. http://en.wikipedia.org/wiki/Big_data, accessed April 2014.
   [3]. Ravi S. Behra, Pranitha Pulumati, Ankur Agarwal, Ritesh Jain, M.D Vinaya Rao, "Predictive Modeling for Wellness and Chronic Conditions", 2014 IEEE 14th International Conference on Bioinformatics and Bioengineering.
3. JyotiSoni, Uzma Ansari, Dipesh Sharma, SunitaSoni,” Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers”, International Journal on Computer Science and Engineering (IJCSE), ISSN : 0975- 3397 Vol. 3 No. 6June2011.
4. 5 Introduction to WEKA – A Toolkit for Machine Learning.
5. 6. A Survey Paper on Data Mining With Big Data, RohitPitre ,ComputerEngg. Vijay Kolekar ,ComputerEngg. K.J.C.O.E. & M. R. Pune K.J.C.O.E. & M. R. Pune, International Journal of Innovative Research in Advanced Engineering (IJIRAE) Volume 1 Issue 1 (April 2014).
6. 7 Naïve Bayes Classifier: A MapReduce Approach, SONGTAO ZHENG, North Dakota State University of Agriculture and Applied Science, Fargo, North Dakota,October 2014.
7. 8 S.Kamalakkannan “A Survey on Big Data Analytics in Diabetes Disease Using Classification Algorithm”, Journal of Applied Science and Computations (JASE) ISSN NO: 1076-5131. Volume V, Issue XII, December 2018.
8. 9 Pradeep Kumar, Abdul Wahid . “Performance Evaluation of Data Mining Techniques for Predicting Software Reliability”, World Academy of Science, Engineering and Technology, 2015.