International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

Real Time Diabetes Prediction using Naïve Bayes Classifier on Big Data of Healthcare

Krish Shah[1], Rajiv Punjabi[2], Priyanshi Shah[3], Dr Madhuri Rao[4]

[1,2,3]U.G. Students, Department of Information Technology, TSEC College, Mumbai, Maharashtra, India
[4]Professor and Head of Department , Examination Incharge, Department of Information Technology, TSEC College, Mumbai, Maharashtra, India

---

Abstract - *Diabetes is a chronic disease, with numerous cases enrolled annually. The number of deaths caused by diabetes has been expanding each year, and it is crucial to anticipate the factor so that they can be relieved at the soonest guaranteeing the patient's life is saved. This prediction is effectively acquired by utilizing Naïve Bayes Classifier. This algorithm classifies, based on the indications of whether an individual has diabetes or not. This model achieved an accuracy of around 81%. The proposed system also supports live streaming of data input, where the results are obtained in real-time for the entered patient.*

Key Words*: *Diabetes Prediction, Naïve Bayes Classifier, Classification, Apache Spark, Apache Kafka

1. INTRODUCTION

Diabetes is a widespread disease in the world. However, individuals may never realize how they contract the disease, what is going to happen to them, and what are the symptoms of the disease. Diabetes is a metabolic disorder which is distinguished by the high glucose level. Increase in blood glucose level harms the vital organs just as well as other organs of the human's body, causing other potential health ailments.

In 2017, 425 million individuals had diabetes around the world, up from an estimated 382 million individuals in 2013 and 108 million in 1980. Accounting for the shifting age structure of the worldwide population, the predominance of diabetes is 8.8% among adults, almost twice the pace of 4.7% in 1980. The WHO accounts that diabetes brought about 1.5 million deaths in 2012, making it the eighth leading reason for death. The worldwide number of diabetes cases may increase by 48% between 2017 and 2045.

Machine learning is a quickly developing trend in the health care industry, on account of the approach of wearable gadgets and sensors that can utilize data to evaluate a patient's wellbeing in real-time. The technology can likewise enable clinical specialists to analyse data to recognize patterns or warnings that may prompt improved conclusions and treatment.

In machine learning, classification comes under supervised learning approach in which the model classifies a new observation dependent on training data set collection of instances whose classification is known. Naïve Bayes Classifier is a classification technique based on Bayes' Theorem with a presumption of freedom among indicators. Naïve Bayes model is easy to construct and especially valuable for exceptionally large data sets. The paper focuses on the prediction of critical disease diabetes using Naïve Bayes Classifier.

The remainder of the paper is organized as follows: Section 2 is a Literature Survey describing the already existing work. Section 3 is about Methodology which highlights the dataset being worked on and the proposed methodology. Section 4 describes the Experimental Results that are obtained after building classifiers. Section 5 is about Results and Discussions which discusses the performance evaluation of all classifiers. Section 6 is about the Conclusion which concludes the overall results.

2. LITERATURE SURVEY

This section reviews the existing recent literature work and provides insights in understanding the challenges and tries to find the gaps in existing approaches.

The focus of the literature survey here is on the use of Machine Learning algorithms and Real-Time Big Data Analytics in the healthcare domain. The paper enumerates the primary areas where Big-Data has been applied thus far and points out areas within healthcare analytics, where an equivalent level of success are often gained through the adoption of this technology [1]. In this paper, data mining classification strategies, such as RIPPER classifier, Decision Tree, Artificial neural networks, and Support Vector Machine, are utilized on the cardiovascular malady. Performance is compared through sensitivity, specificity, accuracy, error rate, True Positive Rate and False Positive Rate [2]. This paper outlines and looks at different methods that are implemented for the classification of a medical diabetes diagnosis on different datasets. These techniques are examined and looked at based on their advantages, issues, classification accuracy [3]. This paper proposed Fuzzy Min-Max neural network, Regression Tree and Random Forest (FMM-CARTRF) combined hybrid classification method that achieved an accuracy of 78.39% for PIDD [4]. The study shows three machine learning classification algorithms SVM, Naïve Bayes and Decision Tree, which are utilized to recognize diabetes at an early stage [5].

Harry Zhang proposed a novel clarification on the classification performance of Naïve Bayes. It explains the dependence distribution of all nodes in a class. Therefore, Naïve Bayes is optimal regardless of the dependencies among the attributes [6]. In this paper, the performance assessment has been highlighted. The algorithm performance is measured on its ability to correctly classify instances of data using Naïve Bayes and J48 classification algorithm [7]. In this system, we propose the utilization of algorithms like Bayesian and KNN to apply on diabetes patient's database and investigate them by taking various traits of diabetes for forecasting of the disease [8]. The paper breaks down the effect of the entropy on the classification error, showing that low-entropy feature distributions yield excellent performance of Naïve Bayes [9]. In this paper, we review some variations of Naïve Bayes classifier and its use in information retrieval [10].

The system uses the Decision Tree, and K-Nearest Neighbor Algorithm calculates and compares the accuracy of C4.5 and KNN [11]. The paper features different Data Mining strategies such as classification, clustering, association and features related work to examine and anticipate human disease [12]. Identify and discuss the theoretical possibilities lying ahead for computational health informatics in this big data age [13]. This study intends to lead an expert analysis of the implementations and applications of machine learning, data mining strategies and tools utilized in the field of diabetes research [14].

The objective of this paper is to study the previous work done in the field of diabetes mellitus assessment and to examine them thoroughly and make conclusions which will provide the directions for future research in this sphere [15]. The inspiration for driving the examination was to look at the performance of the two algorithms and help health workers in better decision making. The level of accuracy, precision, sensitivity and specificity of the two algorithms will be the focus of this research [16]. The purpose of this paper was to study the Obesity and Body Mass Index (BMI) as a risk factor for American and Japanese population. The risk and consequences that obesity and BMI can have on the human being were mentioned in [17].

This paper canters around breaking down the performance of producer and consumer in Apache Kafka processing. The performance assessment of Kafka and its effect on the messaging system in real time big data pipeline architecture [18]. The study of Apache Kafka in Big Data Stream Processing is covered [19]. In this study, the system provides real-time disease diagnosis from ECG data using Logistic Regression. The findings obtained show the architecture, built with Apache Kafka and Apache Spark, design option in real time processing of ECG data [20].

3. METHODOLOGY

This section includes the methodology describing the approach used to research in order to perform analysis.

A. Apache Spark

Apache Spark is an analytic engine with in-memory and rapid processing for big data and machine learning. Apache spark engine assists in creating well-crafted and expressive APIs to permit data workers for effective execution of streaming, machine learning or SQL workloads that require quick repetitive access to datasets.

Apache Spark serves the purpose of model development. In our case, we have developed a Multinomial Naïve Bayes Classifier model using Apache Spark. The model processes the data, and the output is evaluated based on various measure conclude the performance of different types of Naïve Bayes classifier. It fulfils our purpose of distributed parallel computing.

B. Apache Kafka

Apache Kafka is a free public data stream handling software. It expects to offer high throughput, low latency and a platform for dealing with continuous real-time data records.

Apache Kafka is used for collecting input data from producers like clinics, laboratories, hospitals, survey records and other sources which are stored and then used as a dataset for the model. It also serves the purpose of sharing the data to the consumers who are research organizations or any entity who would like to access these records at a given offset timeline of up to past seven days.

C. Diabetes Dataset

Here is the description of the dataset that has used as an input to classifiers implemented using various algorithms. The name of the dataset is Annual Health Survey: Clinical, Anthropometric and Biochemical (CAB) Survey Database contributed by Ministry of Health and Family Welfare and Department of Health and Family Welfare available on data.gov.iv [21] which is an Open Government Data Platform India.

The total number of records is 121901. The total number of attributes are 14, including the target class attribute. The name of two target classes is Diabetic and Not Diabetic. The number of instances for Diabetic is 43049, and the number of instances from Nothing is 78852. We have derived the BMI attribute using height and weight attribute. Apache Spark and Scikit-learn library are used for generating model. Apache Kafka handles the live streaming of data.

Table -1: Diabetes Disease Dataset

Sr. No	Attribute Used	Attribute Type	Attribute Description
1	Sex	Numeric	0 for Male 1 for Female
2	Age	Numeric	Age in years

3	Weight	Numeric	Weight (Kg)
4	Length_height	Numeric	Height (cm)
5	Haemoglobin_lvl	Numeric	HaemoglobinLevel (g/dL)
6	BP_systolic	Numeric	Pressure in the arteries during contraction of heart (mm Hg)
7	BP_systolic_2	Numeric	Pressure in the arteries during contraction of heart (mm Hg)
8	BP_Diastolic	Numeric	Pressure in the arteries when heart rests between beats (mm Hg)
9	BP_Diastolic_2	Numeric	Pressure in the arteries when heart rests between beats (mm Hg)
10	Pulse_rate	Numeric	Number of beats per minute
11	Pulse_rate_2	Numeric	Number of beats per minute
12	Fasting_blood_glucose	Numeric	Sugar is in a blood sample after an overnight fast. (mg/dL)
13	BMI	Numeric	Body Mass Index is a measure of body fat
14	Outcome	Nominal	0 for not Diabetic 1 for Diabetic

D. Flowchart of Proposed Methodology

Here is a brief description of the flow of the proposed methodology.

The Proposed Classifier model considers input from several data sources. The input dataset is processed using two variations of machine learning algorithms of Naïve Bayes namely, Multinomial and Gaussian Naïve Bayes and for each algorithm respective classifier model is trained and tested, and the result is gathered. Based on the experimental results, the best performing algorithm can be determined, which will help in the accurate prediction of the disease.

The following figure Fig 1. depicts the approach that is applied to perform the comparative analysis in order to recommend the best algorithm for building a classification model in order to predict the diabetes disease.

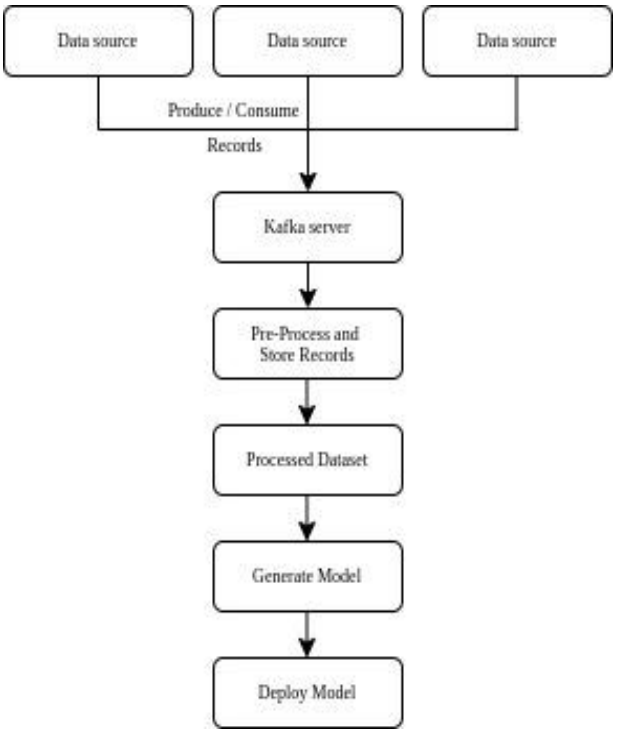

Fig- 1: Proposed Methodology Flowchart

The following describes the steps involved in the procedure of the Fig 1. Proposed Classifier Methodology

Stepwise Procedure of Proposed Methodology

- **Step 1:** - Accumulation of data from various data sources
- **Step 2:** - Preprocess the input data and store it.
- **Step 3:** - Perform percentage split of 75% to divide dataset as a Training set and Test set
- **Step 4:** - Use the training data set to train the machine algorithm i.e., Multinomial Naïve Bayes and Gaussian Naïve Bayes,
- **Step 5:** - Test each Naïve Bayes Classifier model for the mentioned based on the test data set
- **Step 6:** - Evaluate the performance result obtained for each classifier
- **Step 7:** - After analyzing, use the model for diabetes prediction

The proposed Gaussian Naïve Bayes Classifier is built using Scikit-learn library for Python programming language, and Multinomial Naïve Bayes classifier model is built in Apache Spark and the real time data streaming is handled by Apache Kafka. On successful execution of each step, we can evaluate the experimental results.

4. EXPERIMENTAL RESULTS

This section describes the experimental results obtained after training Multinomial and Gaussian Naïve Bayes Classifiers on the diabetes patient dataset. The purpose of these experimental results is for performance evaluation

of two classifier and to recommend the best algorithm suited for prediction.

A. Confusion Matrix

In machine learning, a Confusion Matrix is used to analyse the performance of the classification algorithm. The Confusion matrix is a tabular structure where the rows represent Actual class and columns represents Predicted class.

CONFUSION MATRIX STRUCTURE			
Total No. of Instances		**Predicted Class**	
		No (0)	**Yes (1)**
Actual Class	**No (0)**	True Negative	False Positive
	Yes (1)	False Negative	True Positive

Fig- 2: Confusion Matrix Structure

Specific terminology as that appears in the general Confusion Matrix Structure are described below. These terminologies will be further used for Performance Evaluation of each classifier.

- Actual Class: Class label representing the Actual class before building the classifier
- Predicted Class: Class label representing the Predicted Class after building the classifier
- True Positive: No. of occurrences anticipated positive and are actually positive
- False Positive: No. of occurrences anticipated negative and are actually negative
- True Negative: No. of occurrences anticipated positive but are actually negative
- False Negative: No. of occurrences anticipated negative but are actually positive

Multinomial Naïve Bayes and Gaussian Naïve Bayes algorithm are implemented, and their confusion matrix generated, are shown below:

MULTINOMIAL CONFUSION MATRIX			
Total No. of Instances		**Predicted Class**	
		No (0)	**Yes (1)**
Actual Class	**No (0)**	17642	4214
	Yes (1)	2101	6518

Fig- 3: Multinomial Confusion Matrix

As per Fig.3. According to Multinomial Naïve Bayes Confusion Matrix the values of True Negative=17642, False Negative=2101, False Positive=4214, True Positive=6518

GAUSSIAN CONFUSION MATRIX			
Total No. of Instances		**Predicted Class**	
		No (0)	**Yes (1)**
Actual Class	**No (0)**	18102	4192
	Yes (1)	1641	6540

Fig- 4: Gaussian Confusion Matrix

As per Fig.4. According to Gaussian Naïve Bayes Confusion Matrix the values of True Negative=18102, False Negative=1641, False Positive=4192, True Positive=6540

B. Classification Accuracy

The classification accuracy is one of the performance evaluation measures. Accuracy represents how well the classifier performs prediction of the instances based on the training data.

- Accuracy: It is the ratio of the no. of true predicted instance both positive and negative to the total no. of instances.

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ Number\ of\ Instances}$$

The following table, Table 2 represents the experimental classification accuracy results of Multinomial Naïve Bayes and Gaussian Naïve Bayes Classifier. The table displays the Accuracy value of each algorithm.

Table -2: Experimental Classifier Accuracy

Algorithm	Accuracy Value
Multinomial Naïve Bayes	0.793
Gaussian Naïve Bayes	0.808

C. Accuracy Measure

Values Following are the Classifier Accuracy Measure Values description:

- TP-Rate: It is the ratio of the no. of predicted positive instances to the actual total no. of positive instances

$$TP - Rate = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

- FP-Rate: It is the ratio of the no. of predicted negative instances to the actual total no. of negative instances

$$FP - Rate = \frac{FalsePositive}{FalsePositive + TrueNegative}$$

• Precision: It is the ratio of no. of predicted positives instances to the total of all predicted positive instances.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

• Recall: It is the ratio of the no. of predicted positive instances to the actual total no. of positive instances

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

• F-Measure: Used to represent overall performance. It is the weighted harmonic mean of the precision and recall

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

The following table Table III represents Accuracy Measure Value of two types of Naïve Bayes classifiers that are obtained after building two classifiers on diabetes dataset.

Table -3: Classifier Accuracy Measure Values

Algorithm	TP-Rate	FP-Rate	Precision	Recall	F-Measure
Multinomial	0.756	0.192	0.607	0.756	0.673
Gaussian	0.799	0.188	0.609	0.799	0.691

5. RESULTS AND DISCUSSIONS

This section discusses the overall experimental results obtained.

According to the Classification Accuracy Table II, the Accuracy of Gaussian Naïve Bayes is the highest, which is 0.808. The Accuracy of Multinomial Naïve Bayes is 0.793. Overall, according to the classification Accuracy of Gaussian Naïve Bayes Classifier was better than the other classifier.

According to the Classification Accuracy Measure depicted in Table III, the Gaussian Naïve Bayes Classifier has the highest F-measure value of 0.691. The Precision value of Gaussian Naïve Bayes classifier and Multinomial Naïve Bayes classifier has almost the same value of 0.609.

6. CONCLUSION

In this research work, two classifiers based on machine learning algorithm Naïve Bayes, Multinomial Naïve Bayes Classifier and Gaussian Naïve Bayes Classifier have been used for experimentation to predict Diabetes disease. The two classifiers thus built have been compared based on their accuracy value. Another performance evaluation method was classifier accuracy measure which included TP-rate, FP-rate, precision, recall, F-Measure.

The overall performance of Gaussian Naïve Bayes Classifier to predict the diabetes disease is better than Multinomial Naïve Bayes Classifier. Hence the viability of the proposed model is clearly portrayed throughout the experimental results.

REFERENCES

[1] Fuad Rahman, Marvin Slepian, "Application of Big-Data in Healthcare Analytics – Prospects and Challenges", IEEE 2017002E

[2] Milan Kumari, SunilaGodara, "Comparative Study of Data Mining Classification Methods in cardiovascular Disease Prediction", IJCST, Vol. 2, Issue 2, 2011, pp. 304-308

[3] Dilip Kumar Choubey, Sanchita Paul, "Classification techniques for diagnosis of diabetes: A review", International Journal of Biomedical Engineering and Technology, 2016

[4] M. Seera and C. P. Lim, "A hybrid intelligent system for medical data classification," Expert Syst. Appl., vol. 41, no. 5, pp. 2239–2249, 2014

[5] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," Procedia Comput. Sci., vol. 132, pp. 1578–1585, Jan. 2018

[6] Harry Zhang, "The Optimality of Naïve Bayes", Faculty of Computer Science at University of New Brunswick

[7] Tina R. Patil, S. S. Sherekar, "Performance Analysis of Naïve Bayes and J48 Classification Algorithm for Data Classification", International Journal of Computer Science and Applications, 2013

[8] Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabetes Disease Prediction Using Data Mining", International Conference on Innovations in Information, Embedded and Communication Systems, 2017

[9] I. Rish, "An empirical study of the Naïve Bayes classifier", T.J. Watson Research Center, 2001

[10] Davis D. Lewis, "Naïve Bayes at Forty – The Independence Assumption in Information retrieval." AT&T Labs

[11] Emrana Kabir Hashi, Md. Shahid Uz Zaman, Md. Rokibul Hasan, "An Expert Clinical Decision Support System to Predict Disease UsingClassification Techniques", International Conference on Electrical, Computer and Communication Engineering (ECCE), February 16-18, 2017, IEEE

[12] S. Patel, H. Patel, "Survey of data mining techniques used in healthcare domain", Int. J. of Inform. Sci. and Tech., vol. 6, pp. 53-60, March 2016

[13] R. Fang, S. Pouyanfar, Y. Yang, S. Chen, S. Iyengar, "Computational health informatics in the big data age: a survey", ACM Comput. Surv. New York, vol. 49, pp. 12-47, June 2016

[14] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, I. Chouvarda, "Machine learning and data mining methods in diabetes research", Computational and structural biotechnology journal, vol. 15, pp. 104-116, 2017

[15] Mohammad Atif, Jamshed Siddiqui, Faisal Talib, "An Overview of Diabetes Mellitus Prediction Through Machine Learning Approaches", 2019, 6th International Conference on Computing for Sustainable Global Development

[16] Dominikus BoliWatomakin, Andi Wahju Rahardjo Emanuel, "Comparison of Performance Support Vector Machine Algorithm and Naïve Bayes for Diabetes Diagnosis", 2019 5th International Conference on Science in Information Technology

[17] M. Kuwabara, R. Kuwabara, K. Niwa, I. Hisatome, G. Smits, C. A. Roncal-Jimenez, P. S. MacLean, J. M. Yracheta, M. Ohno, M. A. Lanaspa, R. J. Johnson, and D. I. Jalal, "Different risk for hypertension, diabetes, dyslipidemia, and hyperuricemia according to level of body mass index in Japanese and American subjects," Nutrients, vol. 10, no. 8, p. 1011, Aug. 2018

[18] Thandar Aung, Hla Yin Min, Aung Htein Maw, "Performance Evaluation for Real-Time Messaging System in Big Data Pipeline Architecture", 2018 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery

[19] Bhole Rahul Hiraman, ChapteViresh M., Karve Abhijeet C, "A Study of Apache Kafka in Big Data Stream Processing", 2018 International Conference on Information, Communication, Engineering and Technology

[20] Nur Banu Oğur, CelalÇeken, "Real Time Data Analytics Architecture for ECG", 2018 3rd International Conference on Computer Science and Engineering

[21] https://data.gov.in/catalog/annual-health-survey-clinical-anthropometric-bio-chemical-cab-survey