Bulletin of Environment, Pharmacology and Life Sciences
Bull. Env. Pharmacol. Life Sci., Vol 10 [12]November 2021 : 44-49
©2021 Academy for Environment and Life Sciences, India
Online ISSN 2277-1808
Journal's URL:http://www.bepls.com
CODEN: BEPLAD
ORIGINAL ARTICLE **OPEN ACCESS**

Big Data Analytics Machine Learning Techniques on Diabetes Mellitus Dataset

Rajkumar N[1] and Chithra S[2]
[1]Department of Computer Applications, Krupanidhi Group of Institutions, Bengaluru, Karnataka
[2]Department of Computer Applications, Krupanidhi Degree College, Bengaluru, Karnataka
[1]Email: bcakric@krupanidhi.edu.in

ABSTRACT

Big Data and the Cloud are critical in assisting healthcare professionals in identifying solutions to their difficulties. Healthcare expenses are increasing at a tremendous speed, demanding the establishment of a sustainable, cost-effective, and strategically planned approach for mortality reduction. Diabetes is one of the most significant long-term health problems that a person can face. Diabetes patients who have a poor treatment outcome such as death may sustain long-term damage to their eyes, hearts, kidneys, and nerves. The dataset for this study was taken from UCI Machine Learning, Pima Indians Diabetes Database, with nine attributes. The purpose of this study is to investigate and analyze numerous machine learning Techniques methods in order to identify a viable set of forecasting rules based on a variety of measures, including kappa, accuracy, precision, recall, and sensitivity. A comprehensive study of a diabetes dataset is achievable using SVM, Random Forest, CART, k-NN, and LDA methods. According to the testing results, SVM predictions are more accurate than those produced by other algorithms.

Keywords*: diabetes dataset, SVM, Random Forest, CART, k-NN, and LDA methods.*

Received 14.09.2021 Revised 21.10.2021 Accepted 13.11.2021

INTRODUCTION

Health care records have exploded in popularity in recent years, exhibiting distinct features such as dependency, semi-established, and unstructured. The contemporary era is becoming increasingly influential in the provision of healthcare services. The issues associated with e-healthcare can be overcome by employing various technologies, including cloud computing, big data, and sensor data, as an entire internet [1]several chronic diseases troubling individuals worldwide, including Diabetes, asthma, and thyroid disease. Diabetic Mellitus (D.M.) is a metabolic disorder in which blood glucose levels remain elevated for an extended time [2]. Diabetes is a severe and chronic condition affecting anybody. Glucose is the typical breakdown result of carbohydrate intake, and it provides energy to the cells. Insulin, another hormone, or both lack the frame and the body [3].

As of 2014, an estimated 387 million people [4] worldwide were diagnosed with Diabetes, around 90% of those diagnosed with type 2 diabetes [5, 6]. Between 2012 and 2014, Diabetes claimed over 2-5 million lives per year. By 2035, the number of diabetics is expected to reach 592 million.

Health care records are massive, vital, and complex in various ways. A tremendous amount of data must be analyzed and interpreted to draw meaningful conclusions. Additionally, high-quality healthcare solutions [7] must be supplied to save expenses and keep people alive by cutting-edge technology to avert life-threatening diseases.

D.M. is a chronic disease that affects almost 350 million people globally, according to World Health Organization (WHO) [2]. By 2030, it is expected to be the world's sixth most significant cause of death. Individuals in low- and middle-income nations are more prone to this disease. The majority of people with Diabetes are entirely ignorant that they have the condition. There is a great deal of study and inquiry going on in D.M., with a particular emphasis on the core causes of Diabetes, such as inactivity, obesity, and bad eating habits. Diabetes, pneumonia, and cancer are just a handful of chronic diseases researchers search for improved therapies.

Big data for Diabetes.

The author [8] used k-NN, LDA, ID3, and C4.5 to outperform the other four algorithms in detecting Diabetes in a dataset. To discover and diagnose diabetic problems, suggested in [9], classified using Artificial Intelligence. Clustering and classification techniques generated a selection tree and then provided predictions [10]. The J48 algorithm was used to identify Type-2 diabetes [11]. This study estimated the model's accuracy to be 78.6% [12] and created a system for forecasting Diabetes and determining the type of Diabetes based on the CART principles.

Additionally, the authors differentiate between those at low risk and those at high risk. Diabetics can be categorized according to type using Joseph L. Breault's predictive technique [13]. In this work, the classification algorithms C4.Five and Naive Bayes were used. Diabetes was defined [14], and separate comments were made to group the diabetes facts.

Healthcare Predictive analytics

Hadoop and Map Reduce researchers produced a predictive version of their approach [15]. The authors proposed an efficient and effective remedy based on their data analysis.

Nagarajan et al. [16] suggested an expert medical gadget for determining diabetes type and risk stage. This method can predict diabetes types and treatment options. Simple k-means criteria and particular classification methods were employed to categorize the data. Researchers [17] developed an extensive data analytics system for diabetic patients. This study looks at Diabetes and coronary heart disease. To examine and show the decision tree algorithm's outcomes [18].

An SVM, C4.5-based model for predicting the recurrence of renal sickness in chronic kidney disease patients was developed by [19] suggested an adaptive rule-based classifier for identifying biological data. This method uses biological records to overcome problems like the poor fit and excessive noise. K-Nearest-Neighbor and decision tree used for the analysis of data. This study's classifier can detect and characterize DNA variants with high accuracy.

Anurag et al. [20] built an Indian diabetes prediction model using Hive, and R. K-NN rules were employed to build a prediction model. It was measured more accurately in chronic disease prediction models[21]. The researchers created CNN-MDRP, an algorithm dealing with structured and unstructured health facility data.

This paper's primary goal is to study the diabetes dataset and establish the optimum system version for accurate diabetes forecasts. This study used SVM, R.F., CART, k-NN, and LDA approaches to anticipate illness. We chose these semi-random algorithms for their variety in depiction and manner of learning.

MATERIAL AND METHODS

Proposed Methods

Diabetes is a chronic, life-threatening disease that serves as the basis for this paper's analysis. Numerous machine learning algorithms are applied in this proposed method to predict Diabetes, and the learning technique's accuracy is evaluated. This enabled us to determine which algorithm appropriately forecasts data compared to others.

System Model

The proposed system model is depicted in Figure 1. The following are some of the components of the proposed technique and their anticipated contributions:

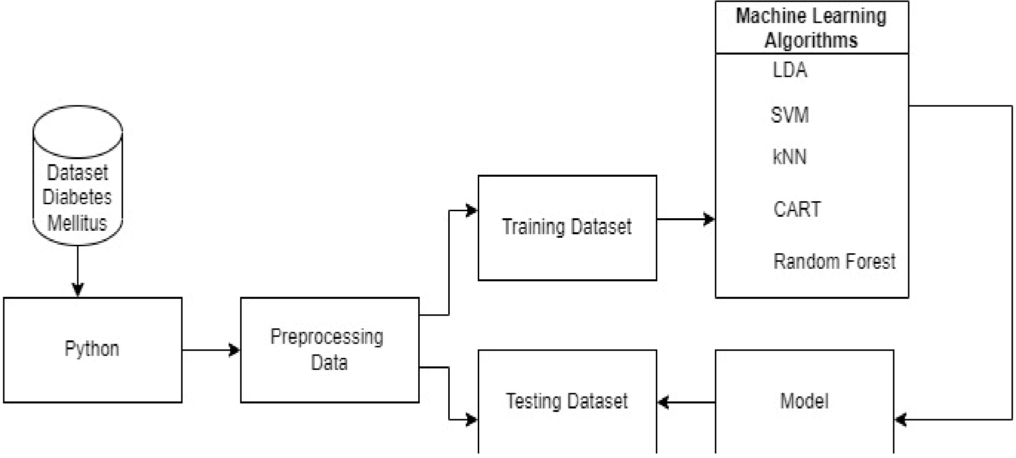

Figure 1. The Proposed System Technique Model.

B. Techniques used

Three steps comprise the technique proposed in this study. The diabetes dataset is loaded into Python as the first step of preprocessing. Further, the loaded dataset is subjected to the go validation process with 10 folds repeated three times. This is a typical structure or strategy for examining exclusive models. Finally, the preprocessed data is randomly divided into education and check sets, each with an 80/20 ratio. Analyze and train the data using various machine learning techniques, including CART, LDA, and k-

NN. Examine the dataset to determine whether the forecasts are accurate. Following then, the inquiry is mainly focused on accuracy and kappa measures.

Implementation

The following subsections detail the context, dataset, and measures used to analyze the outcomes.

Environment

The dataset is then stored in HDFS. Data cleaning and preprocessing queries are executed on the corresponding tables in HIVE. All testing and analysis are performed using Python. These techniques, including LDA, SVM, CART, R.F., and k-NN, are being explored. The diabetes dataset is separated into 80:20 proportions of training and testing data, as is usual, to assure data accuracy. The overall performance may be determined by examining its accuracy and precision and sensitivity, specificity, and Kappa values.

Diabetes Dataset

The dataset for this study was taken from UCI Machine Learning, Pima Indians Diabetes Database [22]. It comprised 768 diabetic patients of 21 to 81 years of age. The properties used in this project are listed in Table 1.

TABLE 1: The attributes of this dataset [22].

Variable	Description	Type(Normal(No) / Numeric(N) / Class(C))
Age	Age (years)	N
Pregnancies	Number of times pregnant	N
BloodPressure	Diastolic blood pressure (mm Hg)	N
BMI	Body mass index (weight in kg/(height in m)^2)	N
Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	N
Insulin	2-Hour serum insulin (mu U/ml)	N
DiabetesPedigree Function	Diabetes pedigree function	N
SkinThickness	Triceps skin fold thickness (mm)	N
Outcome	Class variable (0 or 1) 268 of 768 are 1, the others are 0	C

C. Research Objectives and Metrics

Test evaluations are guided by three primary criteria: precision, recall, and accuracy.

$$\text{Precision} = \frac{TP}{(TP + FP)}$$

$$\text{Accuracy} = \frac{CP}{TP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TP}{FP + TN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$\text{Kappa} = \frac{(Accuracy - Expected\ Accuracy)}{1 - Expected\ Accuracy}$$

RESULTS AND DISCUSSIONS

Python is used initially to import the diabetes dataset's CSV file. A variety of machine learning approaches were utilized to monitor and evaluate a diabetes dataset following statistical preprocessing to define the type of Diabetes that exists.

Table 2. Diabetes Attributes Mean and Standard Deviation [22]

Variable	Mean	Standard Deviation
Pregnancies	3.85	3.37
Glucose	121	32
Blood Pressure	69.1	19.3
Skin Thickness	20.5	15.9
Insulin	79.8	11.5
BMI	32	7.88
Diabetes Pedigree Function	0.47	0.33
Age	33.2	11.8
Outcome	0.35	0.48

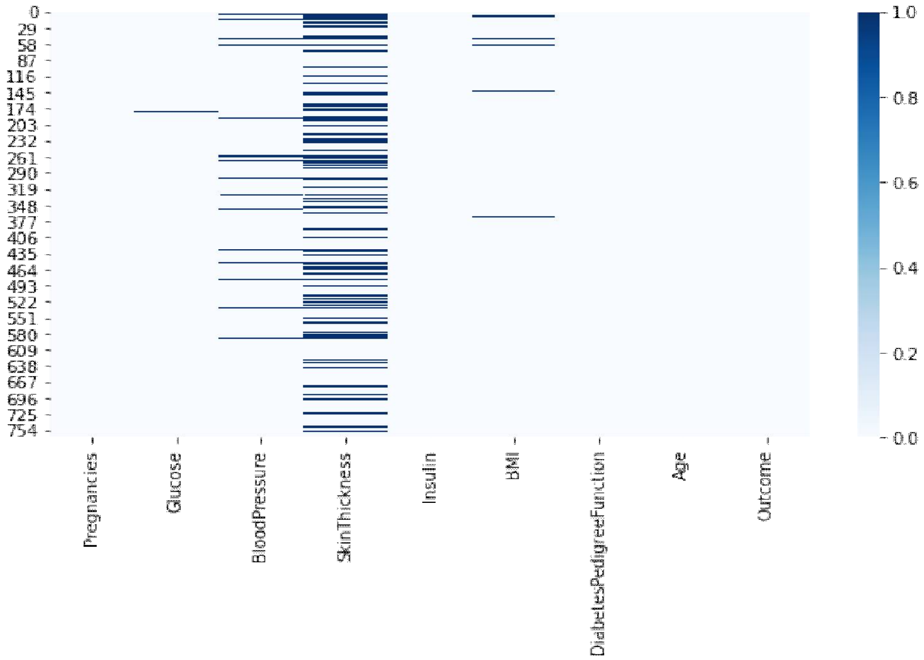

Fig. 2. Heatmap of Diabetes mellitus Attributes

A comparison was conducted by comparing all the attributes of the Diabetes dataset is depicted in Figure 2. The pre Process technique additionally preprocesses the supplied dataset using a scaling strategy. Preprocessing datasets enables us to evaluate the efficiency of innovative machine learning models. We repeated each approach three times to check for fact re-sampling, using the 10-fold pass validation methodology each time. Dot plots compare the evaluated outcomes to the accuracy and kappa values. Figure 3 examines multiple techniques for each model based on their accuracy and kappa coefficients. As a result, the SVM set of rules surpasses LDA, R.F., CART, and CART in terms of accuracy. The accuracy cost of the SVM method is 0.98, which is the highest quality among all algorithms. Algorithms with a kappa value of 0.98 are the most accurate in general. This rule set has a precision of 1 and a take into account value of 1. The accuracy and error rate of each set of rules is shown in Figure 4, and the SVM set of rules has the lowest error rate. The findings of this study demonstrate that when applied to a range of algorithms, the SVM set of rules works well—the confusion matrices for each outfit. The experiment's outcome demonstrates that the SVM rules accurately predict effects and effectively classify records.

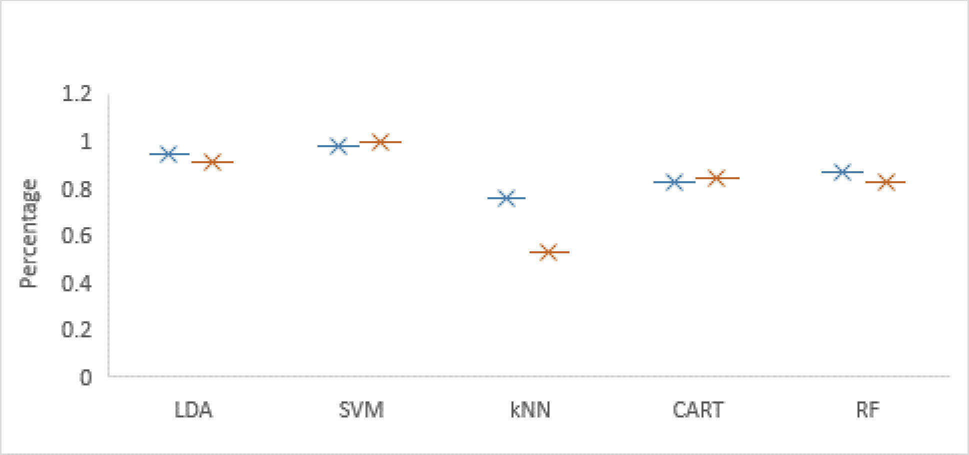

Fig. 3. Comparing Different Machine Learning Techniques with Accuracy and Kappa valueson the Diabetes Dataset

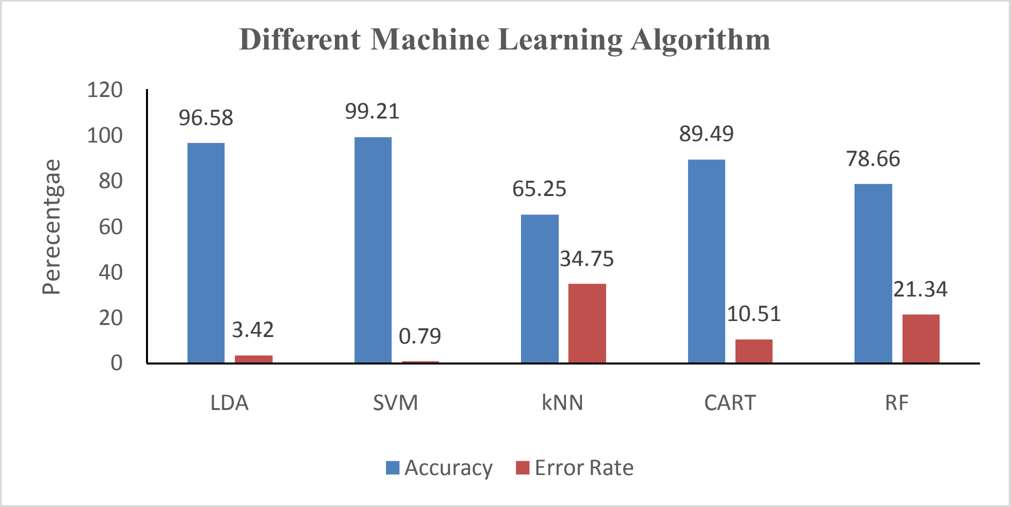

Fig. 4 illustrates various machine learning approaches applied to the Diabetes dataset.

CONCLUSION

Today's society is more concerned with their hectic schedules than their health, leading to long-term health problems such as Diabetes. The objective of this study is to present a complete comparison of different machine learning techniques. This comparison employs various measures, including Kappa, Sensitivity, specificity, accuracy, precision, and Recall. The empirical evidence demonstrates that the SVM techniques are more accurate and efficient in predicting facts.

ACKNOWLEDGMENT

The authors express gratitude towards the assistance provided by The Management, Krupanidhi Group of Institutions (KGI) and Krupanidhi Research Incubation Centre, KGI in completing the research. We also thank our Research Mentors who guided us throughout the research and helped us in achieving the desired results.

REFERENCES

1. U.Varshney, (2009).Pervasive Healthcare Computing: EMR/EHR, Wireless and Health Monitoring pp98..
2. Diabetes Fact sheet N°312". WHO.(2013). Retrieved25March2014.
3. Shoback, edited by David G. Gardner, Dolores, (2011). Chapter 17".Greenspan's basic & clinical endocrinology (9thed.).New York: McGraw-Hill Medical.ISBN0-07-162243-8.
4. IDF. (2014). International Diabetes Federation.100-145.
5. Williams's textbook of endocrinology (12th ed.). (2011). Philadelphia:Elsevier/Saunders.pp.1371–1435.ISBN978-1-4377-0324-5.
6. Shi,Yuankai; Hu,FrankB.(2014). "The global implications of diabetes and cancer". The Lancet 383 (9933): 1947–8. Doi: 10.1016/S0140-6736(14)60886-2.PMID2491022.
7. Ioannis M. Stephanakis, George K. Anastassopoulos and Aggelos D. Tsalkidis, (2002). Multiresolution Autoregressive Filtering for Pneumonia Detection Iin Medical Images, Pp. 1157-1159.

8. Rajesh K, Sangeetha V. (2012). Application of data mining methods and techniques for diabetes diagnosis. International Journal of Engineering and Innovative Technology (IJEIT). 2(3):224–9.
9. AfrandP, Yazdani NM, Moetamedzadeh H, Naderi F, Panahi MS. (2012). Design and implementation of an expert clinical system for diabetes diagnosis. Global J. of Sci, Engg and Tech .p.23–31.
10. Adidela DR, Lavanya DG, Jaya SG, Allam AR. (2012). Application offuzzy ID3 to predict Diabetes. Intr J Adv Comp Math Sci. 3(4):541–5.
11. Aljarullah AA. (2011). Decision tree discovery for the diagnosis of type II diabetes. ICIIT.p.303–7.
12. Kavitha K, Sarojamma RM. (2012). Monitoring of Diabetes with data mining via CART Method. Intr Jrnl of Emr Techy and Adv Engg., 2(11):157–62.
13. Joseph L.Breault.,(2010). Data Mining Diabetic Databases: Are Rough Setsa Useful Addition? Jamia Hamdard University, New Delhi, Proceedings of the 4thNational Conference, INDIACom-2010 Computing for Nation Development, February25-26,2010.
14. P. Padmaja, (2008). Characteristic evaluation of diabetes data using clustering techniques", IJCSNS,VOL.8No.11,20-27.
15. Saravanakumar, Eswari T, Sampath P &Lavanya S (2015). Predictive Methodology for Diabetic Data Analysis in Big Data, Sciencedirect,Vol.50 pp203-208
16. SrideivanaiNagarajan and R.M. Chandrasekaran, (2015). Design and implementation of Expert clinical system for diagnosing diabetes using data mining techniques", Indian J SCi and Tech,.Vol8(8),771-776.
17. K. Sharmila and S.A. Vetha Manickam, (2015). Efficient prediction and classification of diabetic patients frombig data using R. International Journal of Advanced Engg Research and Science,Vol2(9),56-58.
18. Basam Boukenze, Hajar Mousannif and Abdelkrim Haqiq, (2016). Performance of data mining techniques to predict in health care case study: Chronic kidney failure disease. Int. Journal of Database Managment systems,2016,Vol8 (30,1-9.
19. Dewan Md. Farid, Mohammad Abdullah Al-Mamun, Bernard Manderick and Ann nowe, (2016). An adaptive rule-based classifier for mining big biological data. International journal of Expert systems with applications, vol 64, 305-316.
20. A K srivastava, C Kumar and Neha Mangla, (2016). "Analysis of diabetic dataset and developing prediction model by using HIVE and R", Indian Journal of Science and Technology, Vol 9(47), 1-5.
21. Min Chen, YixueHao, Kai Hwang, Lu Wang and Lig Wang, (2017). Disease prediction by machine learning over big data from Healthcare communities", IEEE Access, pp 1-9.
22. Pima- Indians- diabetes- database (2015).https://www.kaggle.com/uciml/pima-indians-diabetes-database.

CITATION OF THIS ARTICLE

Rajkumar N and Chithra S. Big Data Analytics Machine Learning Techniques on Diabetes Mellitus Dataset. Bull. Env. Pharmacol. Life Sci., Vol 10[12] November 2021 : 44-49.