International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-11 Issue-6, May 2022

Health Care Data Analytics – Comparative Study of Supervised Model

Madhu H. K., D. Ramesh

Abstract: ***In the present pandemic situation, health care data is generated voluminously in an unstructured format posing challenge to technology in perspective of analysis, classification and prediction. The data generated is converted to structured format. Suitability of methodology keeping in mind low computational complexity and high accuracy is a major concern which has emerged as a problem in data science. In this research work real time heart disease data set is considered to evaluate the accuracy of six supervised methods –SVM (Support Vector Machine), KNN (K-Nearest Neighbor), GNB (Gaussian Naïve Bayes), LR (Logistic Regression), DT (Decision Tree) and RF (Random Forest). Analysis through ROC curve and confusion matrix predominantly justify RF classifier and LR gives efficient results compared to other methods. This is a preprocessing stage; every researcher has to perform before deciding the methodology to be considered for further processing.***

Keywords: SVM (Support Vector Machine), KNN (K-Nearest Neighbor), GNB (Gaussian Naïve Bayes), LR (Logistic Regression), DT (Decision Tree) and RF (Random Forest).

I. INTRODUCTION

The health care data collected from ERP models used in hospitals are dumped in data centers and are also structured by researchers for archiving, retrieval and processing. This has created lot of structured voluminous data on the background where segregation of data itself is challenging for backend intelligent system. Assessing these data through machine learning algorithms has become the need of the hour. Many procedures done by doctors/diagnosis by pathologists are automated. Initial diagnosis/identification/classification of data has been successfully conducted through various intelligent algorithms available. In this research work an attempt has been made to justify the efficiency of choosing appropriate intelligent algorithm looking at the nature of data considered. Healthcare data projects the history of the patients with respect to his/her present health situations. Most of the automated systems are trained models with apriori knowledge available to further predict/identity based on the requirement of the applications.

Manuscript received on 26 April 2022.
Revised Manuscript received on 29 April 2022.
Manuscript published on 30 May 2022.
* Correspondence Author
Mr. Madhu H. K.*, Research Scholar, Sri Siddhartha Institute of Technology, Tumkur (Karnataka), India.
Dr. D. Ramesh, Professor and HOD, Sri Siddhartha Academy of Higher Education, Tumkur (Karnataka), India.

As a supervised model healthcare data needs intelligent algorithms to auto train and successfully accomplish the required results. From this requirement suitability of intelligent algorithm is an important aspect to be considered by researcher/ product developers to develop models with appropriate data structures and intelligent algorithms with low complexity and high accuracy. To understand the suitability of data and algorithms in this research work six supervised intelligent algorithms are implemented and compared for UCI heart disease data set which has 76 attributes and 303 samples.

Every algorithm is evaluated through the time considered for training and running. Training time depends on the number of samples considered for training the model. During this phase, intelligent algorithm set range or threshold for attributes to define a class classify them into attributes are the dimensional perspective view of the sample. The complexity of many models depends on the number of attributes considered.

Running time is the time taken to execute and furnish the output, this depends on the complexity of mathematical model defined in the algorithm. The six algorithms considered are SVM, KNN, GNB, LR, DT and RF are supervised algorithms used here to classify 2-class classifiers. These are considered as they are mathematically simple, accommodative to attributes and data sets. The suitability of these algorithms is tested on UCI heart disease data set and training time and running time computation is tabulated in the coming sections. Every researcher has to perform such analysis before adapting any algorithm.

The rest of the paper is organized as follows, literature survey on the algorithms with applications in section-2, comparative analysis with results is discussed in section-3. Conclusion in section-4.

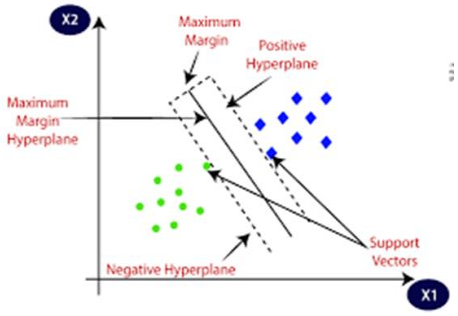
II. LITERATURE SURVEY

The supervised model like SVM [1,2,3,4,6,8,9], KNN [12,13,15,16,18], GNB [19,20,22,24,26], LR [28,29,30,31], DT [33,34,36,37] and RF [41,42,43,44] are most preferred in healthcare analytics to recognize, predict or classify data based on the application. The six supervised models with applications are listed to understand the appropriate decision of best choice. Each model is mathematically defined as per survey and a detailed literature about the applications using the model and accuracy are stated.

2.1 Support Vector Machine

SVM is a managed AI procedure that can be utilized to tackle arrangements or relapse issues [1,2,3,4,6,8,9].

The SVM classifier divide the two classes the most accurately. The SVM calculation's motivation is to track down the ideal line or choice limit for sorting n-layered space among classes with the goal that new information focuses can be promptly positioned in the suitable classification later on. A hyperplane is the name for the ideal decision limit.



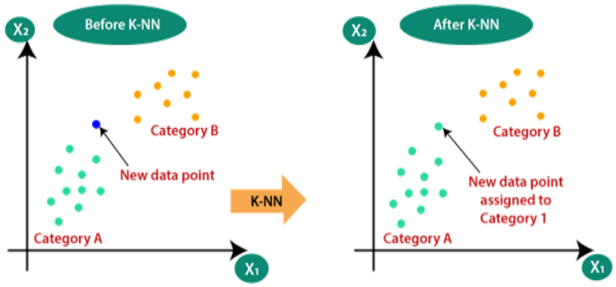
A hyperplane has the equation a.x + b = 0, where a is a vector for the hyperplane and b is an offset. It is a positive point if the value of a.x + b >0, else it is a negative point.

R. Vijayarajeswari et al. [1] discuss categorization and feature extraction methodologies. Hough transform is used to find features in mammography images, and SVM is used to classify them. The employment of an SVM classifier improves classification accuracy. Nico Surantha et al. [2] set out to create an accurate model for identifying sleep stages based on HRV parameters collected from electrocardiograms (ECG). The accuracy test results for the SVM are 82.1 percent for two classes. For breast cancer diagnosis, Chen et al. [3] suggested a RSSVM. On the Wisconsin Cancer Dataset, the effectiveness of the RS SVM is investigated B. Richhariya and M. Tanveer [4] offer a novel machine learning strategy for categorization based on the universal support vector machine (USVM). For both healthy and seizure EEG signals, the suggested USVM attained the greatest classification accuracy of 99 percent. Bissan Ghaddar and Joe Naoum-Sawayawe [5] look at the problem of featureselection in support vector machine classification, which is concerned with creating an accurate binary classifier with a little number of features to achieve high accuracy. In their study, Mingjing Wang and Huiling Chenin [6] used many various diagnosis problems of cancer and diabetes to conduct feature selection and parameter optimization simultaneously for SVM. FOA-SVM was proposed by Liming Shen et al., [7], where the FOA technique successfully and effectively addresses the set of parameters in SVM. Additionally, four important data sets are used to test the usefulness and efficiency of FOA-SVM. A SVM-based outfit learning framework for bosom disease determination is researched by Haifeng Wang et al., [8]. When contrasted with the best single SVM model on the SEER dataset, the proposed WAUCE model lessens change by 97.89% and further develops precision by 33.34%. Mustafa et al. [9] propose a hybrid technique that combines feature selection with classification using the artificial bee colony (ABC) algorithm. The goal of their research is to see how removing unnecessary and obsolete characteristics from datasets affects the classification results using the SVM classifier. The proposed technique by V. Kumari et al. [10] uses SVM, a ML-method, as the classifier for diagnosis of diabetes. The findings of the experiments demonstrate that the support vector machine can be used to successfully diagnose diabetes illness.

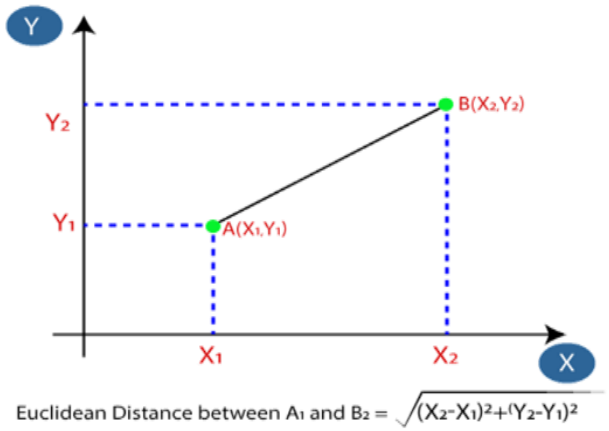
2.2 K-Nearest Neighbor

KNN [12,13,15,16,18] is a Supervised Learning-based Machine Learning method that is one of the simplest methods for classification. The K-NN technique stores all information accessible and adds another piece of information in light of its similitude to the current information. This implies that new information can be effortlessly arranged into an appropriate class utilizing the K-NN calculation when it shows up.

Consider the below diagram:



The new infomation is assigned to the category by computing the Euclidean distance between the data elements. The formula for calculating the Euclidean distance is,



Najat Ali et al., [11] objective is to explore the performance of k-NN on heterogeneous datasets, where information can be depicted as a combination of mathematical and straight-out highlights. In their work, a few similitude measures have been characterized in view of a mix between notable distances for both mathematical and parallel information, and to explore k-NN performance for arranging such heterogeneous informational collections. Krati Saxena et al., [12], present an approach for diagnosing Diabetes Mellitus based on the K-Nearest Neighbor Algorithm, which is among the most promising technologies in artificial intelligence. Iqbal H. Sarker et al. [13] provided a diabetic mellitus classification and analysis based on k-nearest neighbor learning for eHealth services.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

To create an effective classification model, they identified the ideal value of K by considering the low mistake rate. According to Rajendrani Mukherjee et al., [14], a machine learning classifier was used to provide predictive analytics on the disease. This study offered an improved KNN method that did not choose the value of k at random. M.

Akhil Jabbar et al., [15], look into using KNN with feature subset selection for heart disease diagnosis. The findings of the experiments suggest that using feature subset selection in KNN improves the accuracy of heart disease diagnosis in the Andhra Pradesh population. Annushree Bablani et al., [16], proposed a KNN approach for detecting dishonesty using EEG signals from the brain. Hjorth factors such as activity, mobility, and complexity are used in the proposed technique. After doing subject-by-subject analysis, the mobility parameter produces the greatest results, providing up to 96.7 percent. M.Akhil Jabbar et al. [17] proposed a new method for classifying cardiac disease. They evaluated the proposed strategy with a focus on heart illness on A.P as well as other machine learning data sets from the UCI library to validate it. The findings of seven data sets of experiments suggest that their method is a competitive classification method. Mai Showman et al. [18] look into using KNN to assist healthcare practitioners in the detection of cardiac disease. The results reveal that using KNN may reach a greater accuracy of 97.4%, which is higher than any other reported finding on a benchmarked data set.

2.3 Gaussian Naïve Bayes

GN-Bayes [19,20,22,24,26] calculation is a regulated learning calculation, which depends on Bayes hypothesis and utilized for classification of data. Credulous Bayes Classifier is one of the straightforward and best Classification calculations which helps in building the quick AI models that can make fast characterizations. It is a probabilistic classifier, and that implies it orders based on the likelihood of an article.

$$\mathrm{P(A|B)} = \frac{P(B|A)P(A)}{P(B)}$$

Abhilasha et al., [19], researches Bayes Net, Naive Bayes and their blend have been carried out utilizing WEKA. It has been inferred that the blend of Bayes Net and Naive Bayes gives the most extreme arrangement effectiveness out of these three classifiers. V. R. Balaji et al., [20], plan to apply a unique dynamic graph cut technique for skin lesion segmentation, followed by a Nave Bayes classifier for skin illness categorization. They tested their proposed strategy using the ISIC 2017 dataset and discovered that the findings outperformed many state-of-the-art methods. Rahma Fitria et al., [21], use data mining techniques to analysis a diabetic dataset from the UCI Repository. This dataset was subjected to three different classification algorithms: NB Classifier, Multilayer Perceptron's (MLP's), and DT. The results showed that the Nave Bayes Classifier had the highest accuracy, with 76.30 percent. Nazim Razali et al., [22] intend to classify whether a diabetes diagnostic result is positive or negative using numerous data mining approaches such as NB, Sequential Minimal Optimization (SMO), and Simple Logistic Regression. Majed Alwateer et al. [23] present an innovative technique to healthcare data processing. The proposed method employs a hybrid algorithm with two phases. The Whale Optimization Algorithm is used as a feature selection strategy in the initial step to reduce the number of features for huge data. The second step then uses the Nave Bayes Classifier to do real-time data classification. The classification was conducted out in the stacking ensemble learner by Yueling Xiong et al., [24]. To further evaluate the model's classification ability, the suggested CSNB stacking method was applied to nine cancer datasets. The experimental findings demonstrated the efficacy and robustness of the suggested Naive Bayes method in processing various types of cancer data when compared to previous classification methods. The goal of Shweta Kharya et alwork .'s is to create a Graphical User Interface for entering patient screening records and detecting the likelihood of breast cancer disease in future women using Naive Bayes Classifiers. The system was built on the Java platform and trained on benchmark data from Irvine's repository. The goal of S. Vijayarani et al., [26] research's is to forecast kidney disorders utilizing classification algorithms like NB and SVM. This study was primarily concerned with determining the optimal classification algorithm based on classification accuracy and execution time.

2.4 Logistic Regression

The Supervised Learning procedure incorporates one of the most noticeable Machine Learning calculations [28,29,30,31]. It's for 2-class classifiers, so it very well may be Yes or No, 0 or 1, valid or False, etc, however rather than giving accurate qualities like 0 and 1, it gives probabilistic qualities that fall somewhere in the range of 0 and 1. The order issues are addressed utilizing strategic relapse. For Logistic Regression, apply the equation below.

$$\log\left[\frac{\mathrm{y}}{1-\mathrm{y}}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

[27], Antipov et al. offer a CHAID-based method for detecting accuracy of classification of heterogeneity across segments of observations. The method was used to churn data from the Irvine Repository. For the microarray cancer diagnostic challenge, JI ZHU et al. [28] offer penalized logistic regression (PLR) as an alternative to SVM. They show that PLR and the SVM perform equally in cancer classification when utilizing the same collection of genes, but PLR has the advantage of also giving an estimate of the underlying likelihood. Danielle M. et al., [29], approves past discoveries that more drawn-out work hours increment the gamble of unfriendly occasions and mistakes in medical care, and furthermore tracked down the comparative connection with intentional additional time. This auxiliary examination of 11,516 enrolled medical caretakers took a gander at nurture qualities, work hours, and antagonistic occasions and mistakes utilizing bivariate and multivariate strategic relapse. Maren et al., [30], offer a set of principles and heuristics that doctors can use to create a logistic regression-based categorization model for binary outcomes that will help them make better clinical decisions.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

R. Geetha Ramani et al., [31] aim to present a survey of current strategies for knowledge discovery in databases utilizing data mining techniques that are currently in use for Parkinson Disease classification. The disease dataset was obtained from UCI, and it was found to be 100 percent accurate.

2.5 Decision Tree

Learning discrete valued target functions using decision tree learning [33,34,36,37] is a supervised learning method in which the learnt function is represented by a decision tree. DT is one of the most accurate classification algorithms available. Researchers use decision tree learning to determine which features to focus on during the decision-making process, as well as how each feature relates to the choice's potential outcomes and the past. An electronic framework joined with WeChat focused on imported jungle fever patients was proposed by Wang et al., [32] that would conceivably turn into an excitement to lighten the weight of imported intestinal sickness. A choice tree technique was taken advantage of to give significant understanding into the connection between imported intestinal sickness cases and medical services establishments. Manikandan et al., [33] proposed an IoT-based planning strategy, called the Hash Polynomial Two-factor Decision Tree (HP-TDT) to increment booking proficiency and decrease reaction time by ordering patients as being ordinary or in a basic state in negligible time. Yan-song et al., [34] present the SPSS and SAS programs that may be used to view tree structure and introduces commonly used strategies for developing decision trees. The J48 algorithm, which is used to generate Univariate Decision Trees, was studied by Neeraj Bhargava et al. [35]. Weka is a data mining application that offers a variety of methods for analyzing data sets. Umar Sidiq et al. [36] used a data set obtained from a famous laboratory. The complete study will be carried out on the open-source platform anaconda in a Windows 10 environment. A variety of classification approaches will be used in an experimental investigation. The maximum accuracy was achieved by the DT, which was 98.89 percent. Mai Showman et al., [37], study the use of a variety of strategies to improve the performance of different types of DT in the detection of heart disease. The sensitivity, specificity, and accuracy of the various Decision Trees are determined to assess their performance.

2.6 Random Forest

A Random Forest [41,42,43,44]is basically a mixture of Decision Trees. A choice tree is based on a whole dataset, utilizing every one of the highlights/factors of interest, though an irregular timberland haphazardly chooses perceptions/lines and explicit elements/factors to fabricate various choice trees from and afterward midpoints the outcomes.

$$RFfi_i = \frac{\sum_j \text{normfi}_{ij}}{\sum_{j\in \text{all features}, k\in \text{all trees}} \text{normfi}_{jk}}$$

When using machine learning to solve a problem, several iterative tests are used to determine the optimum solution for the problem by fine-tuning it. Given the numerous machine learning methods available, a researcher will choose the most promising model for the initial trial. According to R. Saravana Kumar et al., [38], big data is first partitioned into multiple clusters using the k-means algorithm based on some dimension. Then, using the random forest classifier algorithm, each cluster is classed, resulting in a decision tree that is classified according to the provided criteria. The study by Mohammed Senan et al. [39] gave insight into the diagnosis of CKD patients in order to combat their condition and obtain therapy at an early stage of the disease. A total of 400 patients contributed to the dataset, which included 24 characteristics. The random forest method beat all other applicable algorithms, achieving 100 percent accuracy across the board. Indu Yekkala, et al., [40] employed the Random Forest method to categorize healthy and non-healthy heart disease patients using a cardiac dataset from the UCI repository. The goal of Serkan Balli et al., [41] research is to detect human motions using data from smart watch sensors. The data comes from the smart watch's accelerometer, gyroscope, step counter, and heart rate sensors. Ahmad Taher Azar et al. [42] present an RFC technique for diagnosing lymph disorders. In their paper, they describe a hybrid technique for diagnosing lymph disorders based on GA and RFC. The results showed that RFC has a classification accuracy of 83.9 percent. Md Mursalin et al., [43] presents an original investigation technique for identifying epileptic seizure from EEG signal utilizing Improved Correlation-based Feature Selection strategy (ICFS) with Random Forest classifier (RF). The test results exhibit that the proposed strategy shows better execution contrasted with the ordinary Correlation-based technique and furthermore beats another cutting-edge technique for epileptic seizure recognition utilizing a similar benchmark EEG dataset. A detailed descriptions of the methods are covered in literature with usage of them by various researchers to encompass the strength and weakness of these methods in various applications. From the survey it is evident that many works have been reported in literature that the supervised models are used more in healthcare analytics for various decisions. To use appropriate models, in this research work, supervised models are implemented and tested on UCI real time data set to analyze the model in terms of efficiency and computational complexity.

III. COMPARATIVE ANALYSIS AND DISCUSSION OFRESULTS

To demonstrate the efficacy of the six supervised models considered are applied on UCI heart disease data sets and following metrices used are accuracy, precession, recall, f1-score and roc curve for comparison. In order to justify the classifiers' performance, four basic measurement metrics are used.

1. TP (True Positive) – correctly classified patients with the disease,
2. TN (True Negative) – correctly classified patients with no disease,
3. FP (False Positive) – misclassified patients with no disease,
4. FN (False Negative) – misclassified patients with the disease.

Based on these numbers the metrics defined are as follows:

- Accuracy
- Precision
- Recall
- F1-score

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

- Receive Operating Characteristic (ROC)-curve

In order to conduct experiment, UCI heart disease dataset, consisting of 303 individual data with 13 features and 2 label are considered.

The data set attributes are, Age, Sex, Chest-pain type, Resting Blood Pressure, Serum Cholesterol, Fasting Blood Sugar, Resting ECG, Max heart rate achieved, Exercise induced angina, ST depression induced by exercise relative to rest, Peak exercise ST segment, Number of major vessels (0–3) colored by fluoroscopy, Thal and Diagnosis of heart disease (target).

In the actual dataset, we had 76 features/attributes but for our study, by applying dimensionality reduction technique [45]14 features are selected. To determine the performance of the numerous supervised algorithms, the data set is divided into 25% testing data 75% training data. Below table lists the training time and running time of all the six supervised algorithms.

Classification Method	Training Time	Running Time	Description
Decision Tree	O(n*logn*d)	O(Max depth of tree)	Use DT for large data with low dimensions.
Support Vector Machine	$O(n^2)$	O(k*d)	SVM should avoid for large value of n
Random Forest	O(n*logn*k)	O(T*k)	**RF is faster than all algorithms.**
Gaussian Naïve Bayes	O(n*d)	O(c*d)	C is a feature of each class.
K-Nearest Neighbor	O(k*n*d)	Hear time is Linear for total instances(n) and dimensions(d).	
Logistic Regression	O(n*d)	Applied for low latency data sets.	

As seen in the below tables, random forest has shown highest accuracy for the given dataset.

KNN Classification Report **KNN Accuracy 77.09 %**

	Precision	Recall	F1-score	Support
0	**0.61**	**0.61**	**0.61**	**33**
1	**0.70**	**0.70**	**0.70**	**43**
Accuracy			**0.66**	**76**
Macro avg				
Weighted avg	**0.72**	**0.70**	**0.70**	**76**

Naive Bayes Classification Report **Naive Bayes 85.46 %**

	Precision	Recall	F1-score	Support
0	**0.84**	**0.82**	**0.83**	**33**
1	**0.86**	**0.88**	**0.87**	**43**
Accuracy			**0.86**	**76**
Macro avg	**0.85**	**0.85**	**0.85**	**76**
Weighted avg	**0.86**	**0.86**	**0.85**	**76**

Linear SVM classification report **Linear SVM accuracy 51.1 %**

	Precision	Recall	F1-score	Support
0	**0.44**	**1.00**	**0.61**	**33**
1	**1.00**	**0.02**	**0.05**	**43**
Accuracy			**0.45**	**76**
Macro avg	**0.72**	**0.51**	**0.33**	**76**
Weighted avg	**0.76**	**0.45**	**0.29**	**76**

Logistic regression classification report **Logistic Regression accuracy 87.67 %**

	Precision	Recall	F1-score	Support
0	**0.89**	**0.76**	**0.82**	**33**
1	**0.83**	**0.93**	**0.88**	**43**
Accuracy			**0.86**	**76**
Macro avg	**0.86**	**0.84**	**0.85**	**76**
Weighted avg	**0.86**	**0.86**	**0.85**	**76**

Decision tree classification report **Decision Tree accuracy 85.02 %**

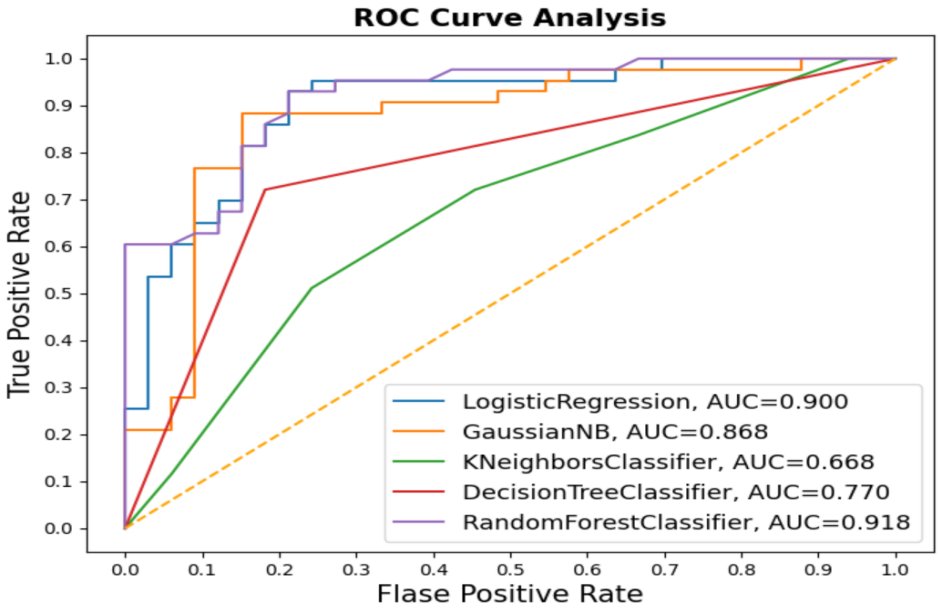
	Precision	Recall	F1-score	Support
0	**0.77**	**0.73**	**0.75**	**33**
1	**0.80**	**0.84**	**0.82**	**43**
Accuracy			**0.79**	**76**
Macro avg	**0.79**	**0.78**	**0.78**	**76**
Weighted avg	**0.79**	**0.79**	**0.79**	**76**

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Random Forest classification report **Random Forest 92.95 %**

	Precision	Recall	F1-score	Support
0	**0.78**	**0.85**	**0.81**	**33**
1	**0.88**	**0.81**	**0.84**	**43**
Accuracy			**0.83**	**76**
Macro avg	**0.83**	**0.83**	**0.83**	**76**
Weighted avg	**0.83**	**0.83**	**0.83**	**76**

ROC Curve:



In this research work, analyze the performance of SVM, KNN, GNB, LR, DT and RF algorithm for diagnosis of heart disease. We propose the use of random forest for more accurate result.

IV. CONCLUSION

Form this work, it is very much evident that the efficacy / output of any algorithm is based on the nature of data set considered. The research focus should be more on screening on data for any further choice of processing techniques. The algorithms should be low computational with high accuracy even with voluminous data. The size of data should be never be a constrained to an algorithm. The existing supervised / unsupervised algorithms have paved a way for scope to improve and restructure them to be computational simple and independent on nature of data as defined in big data.

REFERENCES

1. R. Vijayarajeswari, Parthasarathy, S. Vivekanandan and A. Alavudeen Basha "Classification of mammogram for early detection of breast cancer using SVM classifier and Hough transform". 2019 Elsevier. https://doi.org/10.1016/j.measurement.2019.05.0830263-2241/.
2. Nico Surantha, Tri Fennia Lesmana and Sani Muhamad. "Sleep stage classification using extreme learning machine and particle swarm optimization for healthcare big data". Bina Nusantara University, Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia.
3. Hui-Ling Chen, Bo Yang, Jie Liu and Da-You Liu. "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis". 0957-4174/$. 2011 Elsevier Ltd. doi:10.1016/j.eswa.2011.01.120.
4. B. Richhariya and M. Tanveer. "EEG signal classification using universum support vector machine". https://doi.org/10.1016/j.eswa.2018.03.053. 0957-4174/© 2018 Elsevier Ltd.
5. Bissan Ghaddar and Joe Naoum-Sawaya. "High dimensional data classification and feature selection using support vector machines". http://dx.doi.org/10.1016/j.ejor.2017.08.040. 0377-2217/© 2017 Elsevier.
6. Mingjing Wang and Huiling Chen. "Chaotic multi-swarm whale optimizer boosted support vector machine for medical diagnosis". https://doi.org/10.1016/j.asoc.2019.105946. 1568-4946/© 2019 Elsevier.
7. Liming Shena, Huiling Chena and Zhe Yu. "Evolving support vector machines using fruit fly optimization for medical data classification". http://dx.doi.org/10.1016/j.knosys.2016.01.002. 0950-7051/© 2016 Elsevier.
8. Haifeng Wanga and Bichen Zheng. "A support vector machine-based ensemble algorithm for breast cancer diagnosis". https://doi.org/10.1016/j.ejor.2017.12.001. 0377-2217/© 2017 Elsevier.
9. Mustafa Serter Uzer, Nihat Yilmaz and Onur Inan. "Feature Selection Method Based on Artificial Bee Colony Algorithm and Support Vector Machines for Medical Datasets Classification". Hindawi Publishing Corporation. The Scientific World Journal. Volume 2013, Article ID 419187, 10 pages http://dx.doi.org/10.1155/2013/419187.
10. V. Anuja Kumari and R.Chitra. "Classification Of Diabetes Disease Using Support Vector Machine". IJERA ISSN: 2248-9622 www.ijera.com. Vol. 3, Issue 2, March -April 2013, pp.1797-1801.
11. Najat Ali, Daniel Neagu and Paul Trundle. "Evaluation of k-nearest neighbor classifier performance for heterogeneous data sets". SN Applied Sciences (2019). https://doi.org/10.1007/s42452-019-1356-9.
12. Krati Saxena, Dr. Zubair Khan and Shefali Singh. "Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm". International Journal of Computer Science Trends and Technology (IJCST) – Volume 2 Issue 4, July-Aug 2014.
13. Iqbal H. Sarker, Md. Faisal Faruque, Hamed Alqahtani and Asra Kalim. "K-Nearest Neighbor Learning based Diabetes Mellitus Prediction and Analysis for eHealth Services". EAI Endorsed Transactions on Scalable Information Systems. 03 2020 - 05 2020 | Volume 7 | Issue 26 | e4.

14. Rajendrani Mukherjee, Aurghyadip Kundu, Indrajit Mukherjee, Deepak Gupta, Prayag Tiwari, Ashish Khanna and Mohammad Shorfuzzaman. "IoT-cloud based healthcare model for COVID-19 detection: an enhanced k-Nearest Neighbour classifier-based approach".
15. M. Akhil Jabbar, B. L. Deekshatulu and Priti Chandra. "Heart Disease Classification Using Nearest Neighbor Classifier With Feature Subset Selection". Annals. Computer Science Series. 11th Tome 1st Fasc. – 2013.
16. Annushree Bablania, Damodar Reddy Edlaa and Shubham Dodia. "Classification of EEG Data using k-Nearest Neighbor approach for Concealed Information Test". ICACC-2018. 1877-0509. 2018 The Authors. Published by Elsevier B.V.
17. M.Akhil jabbar, B.L Deekshatulua and Priti Chandra. "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm". International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013.
18. [18] Mai Shouman. "Applying k-Nearest Neighbor in Diagnosing Heart Disease Patients". ICKD 2012.
19. Abhilasha Nakra and Manoj duhan. "Comparative Analysis of Bayes Net Classifier, Naive Bayes Classifier and Combination of both Classifiers using WEKA". I.J. Information Technology and Computer Science, 2019, 3, 38-45.
20. V.R. Balaji, S.T. Suganthi, R. Rajadevi, V. Krishna Kumar, B. Saravana Balaji and Sanjeevi Pandiyan. "Skin disease detection and segmentation using dynamic graph cut algorithm and classification through Naive Bayes classifier". https://doi.org/10.1016/j.measurement.2020.107922. 0263-2241/2020 Elsevier.
21. Rahma Fitria, Desvina Yulisda and Mutammimul Ula. "Data Mining Classification Algorithms For Diabetes Dataset Using Weka Tool". Jurnal Sistem Informasi ISSN P : 2598-599X; E: 2599-0330.
22. Nazim Razali, Aida Mustapha, Syed Zulkarnain Syed Idrus, Mohd Helmy Abd Wahab and Siti Aida Fatimah Madon. "Analyzing Diabetic Data using Classification". JICETS 2019. doi:10.1088/1742-6596/1529/2/022105.
23. Majed Alwateer, Abdulqader M. Almars, Kareem N. Areed, Mostafa A. Elhosseini, Amira Y. Haikal and Mahmoud Badawy. "Ambient Healthcare Approach with Hybrid Whale Optimization Algorithm and Naïve Bayes Classifier". Sensors 2021, 21, 4579. https://doi.org/10.3390/s21134579.
24. Yueling Xiong, Mingquan Y and Changrong W. "Cancer Classification with a Cost-Sensitive Naive Bayes Stacking Ensemble". Hindawi Computational and Mathematical Methods in Medicine. Volume 2021, Article ID 5556992, https://doi.org/10.1155/2021/5556992.
25. Shweta Kharya, Shika Agrawal and Sunita Soni. "Naive Bayes Classifiers: A Probabilistic Detection Model for Breast Cancer". International Journal of Computer Applications (0975 – 8887). Volume 92 – No.10, April 2014.
26. Dr. S. Vijayarani and Mr.S.Dhayanand. "Data Mining Classification Algorithms For Kidney Disease Prediction". International Journal on Cybernetics & Informatics (IJCI) Vol. 4, No. 4, August 2015.
27. Evgeny Antipov and Elena Pokryshevskaya. "Applying CHAID for logistic regression diagnostics and classification accuracy improvement". Journal of Targeting, Measurement and Analysis for Marketing (2010) 18, 109 – 117. doi: 10.1057/jt.2010.3.
28. JI ZHU and TREVOR HASTIE. "Classification of gene microarrays by penalized logistic regression". Biostatistics (2004), 5, 3, pp. 427–443. Doi: 10.1093/biostatistics/kxg046.
29. Danielle M. Olds and Sean P. Clarke. "The effect of work hours on adverse events and errors in health care". 2010 National Safety Council and Elsevier Ltd. doi:10.1016/j.jsr.2010.02.002.
30. Maren E. Shipe, Stephen A. Deppen, Farhood Farjah and Eric L. Grogan. "Developing prediction models for clinical use using logistic regression: an overview". Jan 07, 2019. doi: 10.21037/jtd.2019.01.25. http://dx.doi.org/10.21037/jtd.2019.01.25.
31. Dr. R. Geetha Ramani and G. Sivagami. "Parkinson Disease Classification using Data Mining Algorithms". International Journal of Computer Applications (0975 – 8887). Volume 32– No.9, October 2011.
32. Xi-Liang Wang, Jie-Bin Cao, Dan-Dan L, Dong-Xiao Guo, Cheng-Da Zhang, Xiao Wang, Dan-Kang Li, Qing-Lin Zhao, Xiao-Wen Huang and Wei-Dong Zhang. "Management of imported malaria cases and healthcare institutions in central China, 2012–2017: application of decision tree analysis". Wang et al. Malar J (2019) 18:429. https://doi.org/10.1186/s12936-019-3065-7.
33. Ramachandran Manikandana, Rizwan Patanb, Amir H. Gandomi c,d, Perumal Sivanesana and Hariharan Kalyanaraman. "Hash polynomial two factor decision tree using IoT for smart health care scheduling". https://doi.org/10.1016/j.eswa.2019.112924. 0957-4174/© 2019 Elsevier Ltd.
34. Yan-yan SONG1 and Ying LU. "Decision tree methods: applications for classification and prediction". Shanghai Archives of Psychiatry, 2015, Vol. 27, No. 2.
35. Dr. Neeraj Bhargava, Girja Sharma, Dr. Ritu Bhargava and Manish Mathuria. "Decision Tree Analysis on J48 Algorithm for Data Mining". International Journal of Advanced Research in Computer Science and Software Engineering. Volume 3, Issue 6, June 2013 ISSN: 2277 128X.
36. Umar Sidiq, Dr. Syed Mutahar Aaqib and Dr. Rafi Ahmad Khan. "Diagnosis of Various Thyroid Ailments using Data Mining Classification Techniques". International Journal of Scientific Research in Computer Science Engineering and Information Technology · January 2019. DOI: 10.32628/CSEIT195119.
37. Mai Shouman, Tim Turner and Rob Stocker. "Using Decision Tree for Diagnosing Heart Disease Patients". Proceedings of the 9-th Australasian Data Mining Conference (AusDM'11), Ballarat, Australia.
38. R. Saravana kumar and P. Manikandan. "Medical Big Data Classification Using a Combination of Random Forest Classifier and K-Means Clustering". I.J. Intelligent Systems and Applications, 2018, 11, 11-19. DOI: 10.5815/ijisa.2018.11.02.
39. Ebrahime Mohammed Senan, Mosleh Hmoud Al-Adhaileh and Fawaz Waselallah Alsaade. "Diagnosis of Chronic Kidney Disease Using Effective Classification Algorithms and Recursive Feature Elimination Techniques". Hindawi Journal of Healthcare Engineering Volume 2021, Article ID 1004767, 10 pages. https://doi.org/10.1155/2021/1004767.
40. Flora Amato, Luigi Coppolino, Giovanni Cozzolino, Giovanni Mazzeo, Francesco Moscato and Roberto Nardone. "Enhancing random forest classification with NLP in DAMEH: A system for Data Management in eHealth Domain". https://doi.org/10.1016/j.neucom.2020.08.091. 0925-2312/2021 Elsevier.
41. Indu Yekkala and Sunanda Dixit. "Prediction of Heart Disease Using Random Forest and Rough Set Based Feature Selection". International Journal of Big Data and Analytics in Healthcare. Volume 3, Issue 1, January-June 2018.
42. Serkan Balli, Ensar Arif and Musa Peker. "Human activity recognition from smart watch sensor data using a hybrid of principal component analysis and random forest algorithm". sagepub.com/journals-permissions. DOI: 10.1177/0020294018813692. journals.sagepub.com/home/mac.
43. Ahmad Taher Azar, Hanaa Ismail Elshazly, Aboul Ella Hassanienb and Abeer Mohamed Elkorany. "A random forest classifier for lymph diseases". 2013 Elsevier Ireland Ltd. All rights reserved. http://dx.doi.org/10.1016/j.cmpb.2013.11.004.
44. Md Mursalina, Yuan Zhanga, Yuehui Chena and Nitesh V Chawla. "Automated epileptic seizure detection using improved correlation-based feature selection with random forest classifier". http://dx.doi.org/10.1016/j.neucom.2017.02.053. 0925-2312/© 2017 Elsevier.
45. Madhu H.K. and D. Ramesh. "Dimensionality Reduction of Healthcare Data through Niche Genetic Algorithm". International Journal of Computer Applications (0975 – 8887). Volume 183 – No. 53, February 2022.

AUTHORS PROFILE

Mr. Madhu H. K., is a research scholar at Sri Siddhartha Institute of Technology, Tumkur (SSAHE), having 21 years of Teaching experience at Department of M C A, in Bangalore Institute of Technology, Bengaluru. His research interest includes Data Mining and Big Data Analytics.

Dr. D. Ramesh, Professor and HOD from Sri Siddhartha Academy of Higher Education, Tumkur, India. His vision is to emerge as a center of excellence for imparting technical knowledge in the field of computer applications, nurturing technical competency and social responsibility among budding software professionals. He has around 30 years of teaching experience.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)