JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 461-472
https://publishoa.com
ISSN: 1309-3452

Design and Analysis of Improved Machine Learning Model for Heart Disease Prediction

Dr. Krishan Kumar Goyal[1], Dr. Sandeep Kumar Jain[2]
[1]Professor, Faculty of Computer Application, R.B.S. Management Technical Campus, Agra-282002
[2]Associate Professor, Department of Computer Science, Dr. Bhimrao Ambedkar University, Khandari, Agra-282002
kkgoyal@gmail.com[1], sandeepzen@yahoo.co.in[2]

Received: 2022 March 15; **Revised:** 2022 April 20; **Accepted:** 2022 May 10.

ABSTRACT:
An early diagnosis of such a disease is a key responsibility for many health care professionals in order to safeguard their patients from contracting such an illness. Cardiovascular disease is one of the universal ailments that are prevalent in today's society. Due to the fact that even a minor error can result in serious health issues or even the individual's death, the diagnosis and treatment of heart-related diseases require an increased level of precision, perfection, and correctness. This is on account of there are a variety of death cases that are associated with the heart, and the number of people checking for these conditions is growing dramatically and gradually. In order to effectively handle the situation, there is an essential requirement for an expectation framework for practicing mindfulness towards disorders. With the application of Machine Learning techniques, it is possible to evaluate the data and determine the reasons that contribute to cardiac disorders such as coronary heart disease, arrhythmia, and dilated cardiomyopathy. Machine learning is playing an important role in the medical sector. This article provides a description of a preprocessing method and an analysis of the accuracy of the prediction of heart disease after the data including noise has been preprocessed. It has also been seen that the accuracy has improved after the preprocessing step has been taken.

Keywords: Cardiovascular Disease, Machine Learning, Data Mining, Data preprocessing, Classification technique

1. INTRODUCTION

The detection of heart illness in its earliest stages is an important test since heart diseases claim the lives of 17.3 million people per year, and each error in the diagnosis of cardiac disease contributes to this staggering number. The analysis of heart diseases has made use of Data Mining (DM) characterization strategies, despite the fact that these strategies are constrained by certain challenges associated with the quality of the data, such as irregularities, noise, missing data, outliers, high dimensionality, and imbalanced data. Data

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 461-472
https://publishoa.com
ISSN: 1309-3452

planning was accomplished through the application of data pre-processing (DP) strategies with the intention of enhancing the presentation of disease modelling (DM)-based forecasting systems.

Data mining transforms the enormous assortment of crude healthcare data into data that can assist with settling on informed decision and prediction. There are some current investigations that are applied in data mining for prediction of heart disease. In any case, contemplates that have given consideration towards the critical highlights that assume a crucial part in anticipating cardiovascular disease are restricted. This examination means to distinguish huge highlights and data mining methods that can improve the precision of foreseeing heart disease. A methodology was utilized to direct how the model can be utilized to improve the precision of expectation of Heart Attack in any person.

2. RELATED WORK

Machine learning techniques such as Random Forest, XGBoost, Neural Network, and Support Vector Machine (SVM) were highlighted by ZubairHasan et al. as being able to accurately forecast the risk of heart disease. The results showed that random forest had an accuracy of 84.9 percent, XGBoost had an accuracy of 86.99 percent, neural networks had an accuracy of 85.4 percent, and support vector machines had an accuracy of 84.8 percent.

On the Framingham HD dataset, Nour El Houda CHAOUI and colleagues evaluated and contrasted the performance of three machine learning methods, namely Hybrid-SVM, Support Vector Machine, and Neural Network, with the goal of better predicting cardiovascular illnesses. The accuracy that was attained was 84.7 percent for neural networks, 86.03 percent for support vector machines, and 94 percent for hybrid SVMs, and [10].

Logistic regression was used by OzkanKilic et al. on the HD dataset in order to make predictions about heart disorders. The results of this method showed an improved accuracy of 85.2 percent than ever before. In order to forecast a cardiac illness, RifkiWijaya et al. utilised the same logistic regression methodology as in a prior study, along with the same dataset from that study. The accuracy was improved by 87.6 percentage points using this strategy.

The heart disease diagnosis that Chu-Hsing et al. presented using a standard dataset was evaluated using a group of classifiers that included multilayer perception (MLP), Support Vector Machine (SVM), and J48 Logistic Regression (LR). The results produced with the J48 classifiers demonstrate a higher degree of accuracy across a variety of performance parameters.

In spite of this and numerous other explores, the field is as yet open for scientists to direct their tests to work on the exactness of the AI procedures for anticipating infections that represent a danger to human existence, including heart related diseases

3. PROPOSED RESEARCH METHOD

The data related to healthcare may be accessed in enormous volumes and include a wealth of information, but it is not possible to mine the information for information based on its historical records. It is possible to use data mining techniques in order to differentiate between respected knowledge and clinical frameworks. The analysts were given a presentation on the many different ways of information mining that may be used to diagnose cardiac problems. Data mining is the process of analysing vast amounts of information in order to uncover previously hidden instances, information,

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 461-472
https://publishoa.com
ISSN: 1309-3452

and correlations. This research uses the Cleveland dataset from the UCI archive to support the subsequent model since it has highlights for the majority of the putative risk variables for coronary sickness. The paper's objective is to validate the subsequent model.

3.1 Dataset

The dataset was retrieved from the UCI Machine Learning repository using information about Cleveland Heart Disease. The dataset contains 303 individual pieces of data with a total of 76 attributes; however, we have chosen to focus on only 14 of those attributes for our research (13 predictors and 1 class), which include fasting blood sugar, resting blood pressure, resting electrocardiographic, chest pain, exercise-induced angina, depression, cholesterol, maximum heart rate, gender, age, major vessels, and slope. Three Boolean types, five Continuous types, and six Categorical kinds are included in our dataset. Table 1 displays the many aspects of the dataset for your perusal.

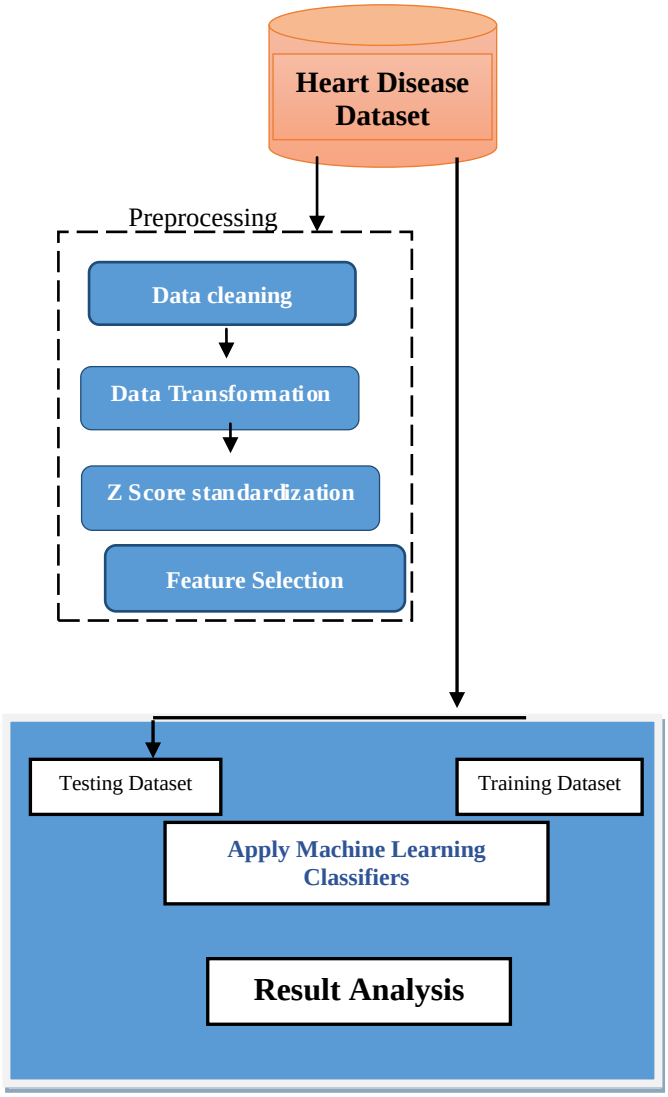

Figure 1. Proposed System Overflow

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 461-472
https://publishoa.com
ISSN: 1309-3452

3.2 Data Preprocessing

The term "preprocessing" refers to a collection of operations that are carried out on the data in order to work on the nature of the data [10]. These procedures include dealing with missing characteristics, converting the sort of element, and a variety of other processes. The Cleveland dataset has a smaller number of missing values for its characteristics. During the step of preprocessing, six rows of data were found to have missing values. These rows were then computed with the help of the Impute Missing Values function. By using the KNN Equation (1), a total of 303 observations were gathered for the dataset.

$$Dist(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \qquad (1)$$

Where Xi ,Yi are some known values and predicted values.

ii) Data transformation is carried using Min Max Normalization Equ (2) where it each numerical feature value into new value depending on the maximum and minimum values of the features [7] by applying (2).

X new = (X – X min) / (X max – X min) ---------- (2)

Where X is a set of the observed values present in X. X min and X max are the minimum and maximum values in X.

iii) Z-Score standardization is used to handle outliers in a dataset Equ (3). It scales the data ranges from 0 to 1.

$$\bar{X} = \frac{X - \mu}{\sigma} \qquad (3)$$

Where $\bar{X}$ is a new select value after applying standardization, X is a selected value from a numerical feature.

iv) Integer Encoding: It splits the Boolean feature and gives 0 for absence and 1 for presence in each new feature.

v) One Hot Encoding: It splits the categorical feature into a separate number of features and gives 0 for absence and 1 to 4 for presence in each new feature.

vi) Feature selection is carried out using Chi-squared test statistic between the feature and the target class by applying Equ (4).

$$X^2 = \sum \left((\text{observed} - \text{expected})^2 / (\text{expected}) \right) \qquad (4)$$

During the course of our research, we made use of the Chi-Squared Attribute evaluator in conjunction with the Ranker search technique. The Ranker search technique assigns points to the characteristics after ranking them based on the Chi-squared test statistic associated with those features.

Most Significant Feature	Ranker scores	Least Significant Feature	Ranker scores
cp	214.6	resrecg	4.9
exang	151.5	fbs	1.6
oldpeak	139.8	trestbps	< 0.001

Table 1.Chi squared Attribute Test evaluator

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 461-472
https://publishoa.com
ISSN: 1309-3452

After removing the three elements that we deemed to be of the least importance, we focused on the remaining 10 aspects of the dataset. It was for the purpose of model training, and the characteristics that were chosen include cp, exang, oldpeak, chol, thalach, thal, sex, age, and slope. In the context of our research, the data are categorised and divided as follows: eighty percent of the data are used for training, and twenty percent are used for testing the model. It makes it easier for the learning models to be trained effectively, which enables them to deliver accurate classifications with a higher level of precision.

3.3. Classification Technique

Classification is a type of supervised machine learning model that takes the output of a label as its input and uses that output in conjunction with other labels or categorical data to determine the conclusion of the model [2]. The classification model was developed for the purpose of training based on a large number of known labelled and categorical aspects of the data input [5]. In the succeeding step, the model made an attempt to address the aim of the model by making use of the test set to differentiate the amount of the known objective for the models and to try to address the objective of the model. [9]

i) SVM: When using a support-vector machine, each data point is plotted in an n-dimensional space, with the value of each variable being the value of particular coordinates. Classification is then carried out based on the hyper plane that distinguishes between the two data classes. After this, the features of new instances might be utilised in order to make a prediction regarding the class that a new instance ought to belong to.

ii) Naive Bayes: The statistical classification technique known as Naive Bayes is based on Bayes' theory and uses it to make its determinations. It calculates the likelihood of a relationship between each feature in the test data and each target by applying the prior probability of Bayes's theorem. The target with the highest probability is chosen as the result of the model [38]. Bayes's theorem uses prior probability to determine likelihood of a relationship. The probability can be found using (10):

$$P(C_i|F_j) = \frac{P(F_j|C_i)P(C_i)}{P(F_j)} \quad \text{--------(5)}$$

Where P(Ci) probability of specific class, P(Fj |Ci) probability of specific feature (Fj) appear with specific class (Ci), p(Fj) probability of specific feature (Fj)

iii) Random Forest Classifier: It is a characterization calculation works by making numerous trees from the dataset [5]. During the structure of every tree, the randomization is applied to discover the worth the split hub [12]. Predictions (average) from all individual regression trees onx'

$$\hat{f}_{=B}\frac{1}{B}\sum_{b=1}^{B} f_b(x') \text{ -------- (6)}$$

Estimate of uncertainty can be made as the standard deviation of the prediction of all individual regression trees onx'

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b(x')-\hat{f})^2}{B-1}} \text{ --------- (7)}$$

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 461-472
https://publishoa.com
ISSN: 1309-3452

iv) Decision Tree Classifier: In the decision tree, all of the provisions are placed as the root hub. After that, the elements are segregated by finding the entropy, which is used to determine the proportion of concordance in the information. The weights that are computed on the features while the process of learning is taking place are utilised to categorise testing data, after which the classes are allocated based on the weights that were computed on the features. By utilising Equ (8), one can determine both the data gained and the entropy, and the node on the tree that has the most significant value in terms of the data acquired is selected to serve as the root hub [9].

$$Entropy(F) = \sum_{i=1}^{C}(-P_i \log_2 P_i)$$

$$Gain(F, A) = Entropy(F) - \sum_{i=1}^{K}(\frac{|F_i|}{|F|} Entropy(F_i)) \quad \text{----} \quad (8)$$

Where C is number of outputs, Pi is probability of occurrences each output from all output, K number of spilt data, F feature with some data, F i spilt data from feature F.

3.4. Tool

Python Jupyter is a core supported programming language that provides an integratedenvironment to create and share documents that contains live code, equations and visualizations. [13]. Python Jupyter can be utilized for include preparing, dataset division, model preparing, model testing, network research, information representation and execution assessment [5].

4. RESULTS AND DISCUSSION

In this paper, four machine learning classification techniques are used to predict the accuracy.

4.1 Dataset

S.no	Feature Names	Type	Description
1	#3 (age)	Continuous	Patient age in years
2	#4 (sex)	Boolean	1 = male. 0 = female
3	#9 (cp)	Continuous	Chest pain type. Values range from 1 to 4. Value 1: typical angina. Value 2: atypical angina. Value 3: non-anginal pain. Value 4: asymptomatic.
4	#10 (trestbps)	Continuous	Resting blood pressure measured in mm Hg on admission to the hospital

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 461-472
https://publishoa.com
ISSN: 1309-3452

5	#12 (chol)	Continuous	cholesterol of the patient measured in mg/dl
6	#16 (fbs)	Categorical	Fasting blood sugar of the patient. If > 120 mg/dl, Value 1 = true. Value 0 = false
7	#19 (restecg)	Categorical	Resting electrocardiographic results.0-Normal, 1-Having ST wave abnormality 2-Showing probable or left ventricular hypertrophy
8	#32(thalach)	Continuous	Maximum heart rate achieved
9	#38 (exang)	Categorical	Exercise included angina 1-Yes 0-No
10	#40 (oldpeak)	Continuous	Depression induced by exercise relative to rest
11	#41 (slope)	Categorical	Slope of the peak.1-up sloping , 2-flat, 3-down sloping
12	#44 (ca)	Categorical	Number of major vessels (0-3) colored by fluoroscopy. Values can range 0 to 3
13	#51 (thal)	Categorical	Represents heart rate of the patient. It can take values 3, 6, or 7.Value 3 = normal. Value 6 = fixed defect. Value 7 = reversible defect
14	#58 (Target)	Boolean	Diagnosis classes.0-healthy , 1 – have heart disease

Table 2: Heart Disease Data Set with 14 Attributes

4.2. Algorithms for Error Matrix

Algorithm	TP	FN	FP	TN
SVM	145	30	26	102
Naive Bayes	141	18	19	125
Random Forest	155	13	23	112
Decision Tree	172	11	21	99

Table 3: Error matrix for classifiers without Preprocessing

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 461-472
https://publishoa.com
ISSN: 1309-3452

Above Table 3 shows the error matrix without data preprocessing. It represents number of correct classification (True Positive and True Negative) and misclassification (False Positive and False Negative) for each algorithm.

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F-score (%)
SVM	81.51	84.79	82.86	83.814
Naïve Bayes	87.78	88.12	88.68	88.399
Random Forest	88.11	87.07	92.26	89.589
Decision Tree	89.43	89.12	93.99	91.490

Table 4: Evaluation result without Data preprocessing

The above table 4 represents various performance measures for SVM, Naive Bayes, and Random Forest and Decision tree without data preprocessing.

4.3. Evaluation result with Data preprocessing

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F-score (%)
SVM	91.39	93.36	95.35	97.43
Naive Bayes	90.56	91.79	97.48	96.06
Random Forest	92.8	93.57	98.85	96.13
Decision Tree	92.64	93.36	98.92	96.06

Table 5: Algorithm Evaluation report after data preprocessing

The table 5 demonstrates the performance measures of various algorithms with data preprocessing techniques. The result shows the performance matrix was high when compared to table 3.

4.4 Performance metrics result

The predicted accuracy, precision, and recall were the measures that were utilised in order to evaluate the performance of each model. The confusion matrix is a two-by-two matrix that compares the model's predicted class values to the actual class values. These measurements are based on the confusion matrix.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 461-472
https://publishoa.com
ISSN: 1309-3452

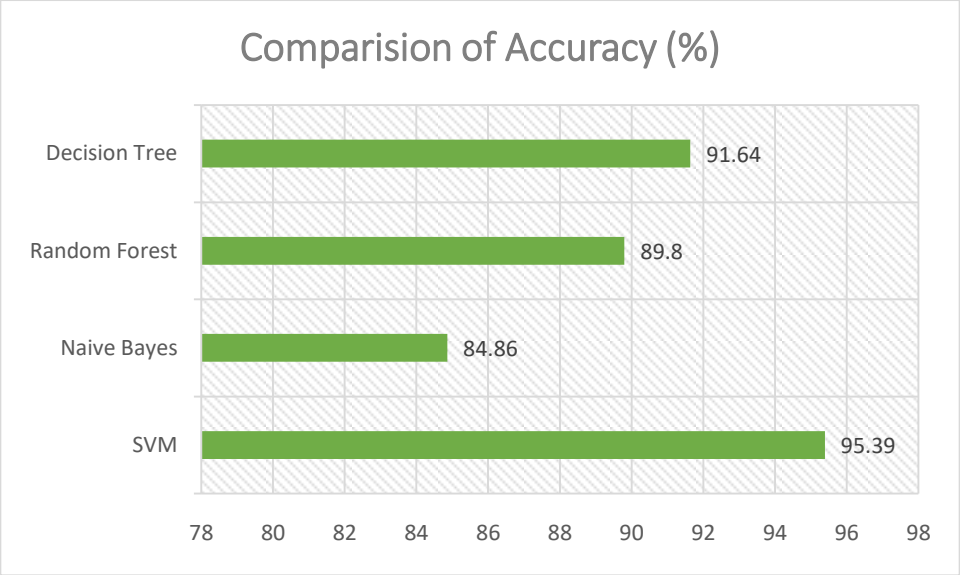

Fig 2 Accuracy metrics comparison with and without preprocessing

In Fig 2, the accuracy metrics is compared for four classifiers and the result shows that Random forest algorithm performs 89.8% after processing the data. Other algorithms have an accuracy of 95.39% of SVM which is highest among all, 90.56% of NB, 84.8%ofFscore.

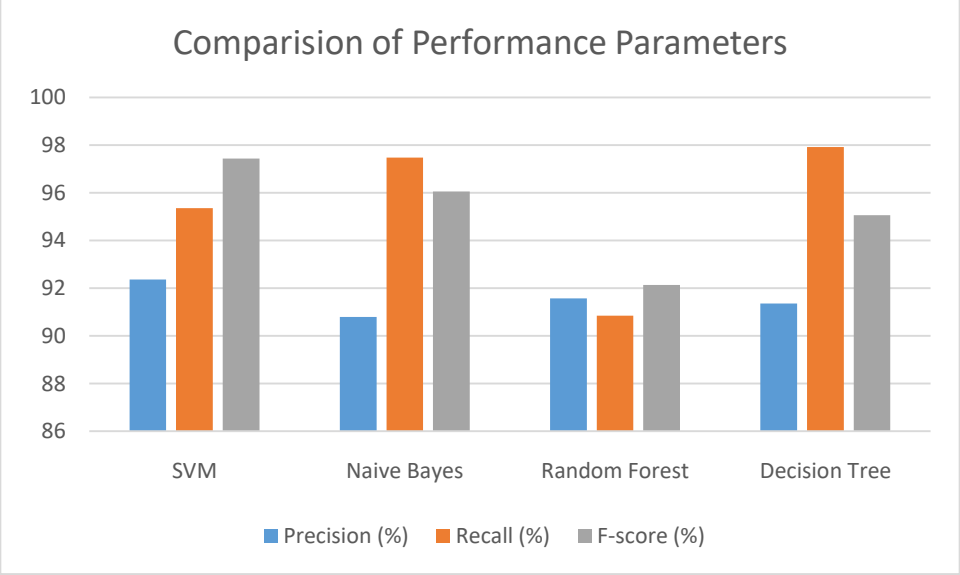

Fig 3. Comparison of Performance Parameter

In Fig 3, the performance metrics of precision, recall and F-score is compared with and without preprocessing data. The metrics are analyzed with four algorithm

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 461-472
https://publishoa.com
ISSN: 1309-3452

namely SVM, Naive Bayes, RF, DT. The result obtained shows that after preprocessing the metrics values are high.

The model that was constructed by utilising the primary dataset for Heart Disease. This model obtained an accuracy of 92.8 percent, a precision of 93.57 percent, a recall of 98.85 percent, and an F-score as high as 96.13 percent. Random forest technique had the best precision of any model that we developed.

4.4Performance measures and an Error Matrix

In order to evaluate how well the model can distinguish between the various classes included in the data set, an error matrix is utilised [2]. A correct classification is denoted by the abbreviations TP and TN, whereas an incorrect classification is denoted by the abbreviations FP and FN. When it comes to the precise categorization, TP and TN are categorised higher than FN and FP [2], as can be seen in the table that can be seen below.

		Actual Values: Positive (1)	Actual Values: Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Table 6. Error Matrix

i) TP: person is healthy and predict as healthy
ii) TN: person is ill as well predict as ill
iii) FP: person is healthy and predict as ill
iv) FN: person is ill and predict as healthy

5. CONCLUSION AND FUTURE WORK

The vast quantity of unstructured healthcare data is transformed through the process of data mining into information that can facilitate the making of informed decisions and predictions. The overarching goal here is to specify several machine learning approaches that can contribute to accurate prediction of heart disease. The strategies that were implemented to enhance the accuracy of machine learning classifiers in the process of diagnosing heart disease have been validated as effective, resulting in findings that are superior to those obtained by earlier research. As a direct consequence of applying preprocessing methods to the dataset, the precision of the machine learning algorithm known as the Random Forest algorithm is significantly higher than that of other algorithms. Our goal is to have more accurate and reliable expectations while also reducing the number of attributes and testing.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 461-472
https://publishoa.com
ISSN: 1309-3452

REFERENCES

[1] Katarya, Polipireddy Srinivas (2020), "Predicting Heart Disease at Early Stages using Machine Learning", DOI: 10.1109/ICESC48915.2020.9155586, Electronic ISBN: 978-1-7281-4108-4, IEEE.

[2] Dr.B.Azhagusundari, Ms. C. Keerthana (2021) "Enhanced Weighted Quadratic Random Forest Algorithm For Heart Disease Prediction", Design Engineering, pp. 4119- 4133.

[3] Noor Basha, Ashok Kumar P S, Gopal Krishna C, Venkatesh P (2019),"Early Detection of Heart Syndrome Using Machine Learning Technique", DOI: 10.1109/ICEECCOT46775.2019.9114651, ISBN: 978-1-7281-3261-7, IEEE.

[4] B.KeerthiSamhitha, SarikaPriya.M.R, Sanjana.C, SujaCherukullapurathMana and Jithina Jose (2020), "Improving the Accuracy in Prediction of Heart Disease using Machine Learning Algorithms ", DOI:10.1109/ICCSP48568.2020.9182303 , ISBN: 978-1-7281-4988-2, IEEE.

[5] Chu-Hsing Lin, Po-Kai Yang, Yu-ChiaoLin, Pin-Kuei Fu (2020), "On Machine Learning Models for Heart Disease Diagnosis", DOI: 10.1109/ECBIOS50299.2020.9203614 , ISBN: 978-1-7281-8712-9, IEEE.

[6] RahmaAtallah, Amjed Al-Mousa (2019)," Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method",DOI: 10.1109/ICTCS.2019.8923053, ISBN: 978-1-7281-2882-5, IEEE.

[7] Senthilkumarmohan,ChandrasegarThirumalai, and GautamSrivastava (2019), "Effective Heart Disease Prediction using Hybrid Machine Learning Techniques",

[8] DOI: 10.1109/ACCESS.2019.2923707,ISBN: 2169-3536, IEEE.

[9] Ms. C. Keerthana, Dr. B. Azhagusundari (2020) "Heart Disease Data Pre-Processing Using Enhanced Data Mining Techniques", Vol 29(08), pp.6274-6282. Available at: http://sersc.org/journals/index.php/IJAST/article/view/36064, IJAST

[10] Amin UlHaq, Jianping Li, Muhammad HammadMemon, et al.(2019), "Heart Disease Prediction System Using Model Of Machine Learning and Sequential Backward Selection Algorithm for Features Selection",DOI: 10.1109/I2CT45611.2019.9033683,ISBN:978-1-5386-8075-9, IEEE.

[11] M.Ganesan, Dr.N.Sivakumar (2019), "IoT based heart disease prediction and diagnosis model for healthcare using machine learning models",DOI: 10.1109/ICSCAN.2019.8878850, ISBN: 978-1-7281-1525-2, IEEE.

[12] Xu S, Zhang Z, Wang D, Hu J, Duan X, Zhu T. Cardiovascular risk prediction method based on CFS subset evaluation and random forest classification framework. In: 2017 IEEE 2nd international conference on big data analysis (ICBDA). IEEE. p. 228–32.

[13] Otoom AF, Abdallah EE, Kilani Y, Kefaye A, Ashour M. Effective diagnosis and monitoring of heart disease. Int J Softw Eng Appl. 2015;9(1):143–56.

[14] Vembandasamy K, Sasipriya R, Deepa E. Heart diseases detection using Naive Bayes algorithm. Int J Innov Sci Eng Technol. 2015;2(9):441–4.

[15] Chaurasia V, Pal S. Data mining approach to detect heart diseases. Int J Adv Comput Sci Inf Technol (IJACSIT). 2014;2:56–66.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 461-472
https://publishoa.com
ISSN: 1309-3452

[16] Parthiban G, Srivatsa SK. Applying machine learning methods in diagnosing heart disease for diabetic patients. Int J Appl Inf Syst (IJAIS). 2012;3(7):25–30.

[17] Deepika K, Seema S. Predictive analytics to prevent and control chronic diseases. In: 2016 2nd international conference on applied and theoretical computing and communication technology (iCATccT). IEEE. p. 381–86.

[18] Dwivedi AK. Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Comput Appl. 2018;29(10):685–693.