ISSN 2564-0127 (Print)
ISSN 2564-0135 (Online)

Canadian Journal of Medicine 3(2021)36-44

Canadian Journal of Medicine

WWW.CIKD.CA

Original Article

Analysis and Prediction of Heart Disease Using Machine Learning and Data Mining Techniques

Md. Murad Hossain[1*], Salman Khurshid[2], K. Fatema[3], M. Zahid Hasan[4], Mohammad Kamal Hossain[5]

[1, 2, 3, 4,5]Department of Statistics, Faculty of Science, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Gopalganj-8100, Bangladesh

Keywords:
Heart disease, Data mining, Machine learning, Random forest, Naïve Bayes, Logistic regression, Support Vector Machine (SVM), K-Nearest Neighbor, J48, Decision tree, WEKA

Received
09 January 2021
Received in revised form
18 February 2021
Accepted
24 March 2021

***Correspondence:**
murad.stat@bsmrstu.edu.bd

ABSTRACT

In clinical, sciences expectation of heart malady is one of the foremost troublesome undertakings. Nowadays, coronary illness may be a significant reason for bleakness and mortality in present-day society. Coronary illness could be a term that doles intent on countless ailments identified with the heart. Clinical determination is incredibly a big, however entangled errand that must be performed precisely, effectively, and unequivocally. Although huge advancement has been imagined within the finding and treatment of coronary illness, further examination is required. The accessibility of enormous measures of clinical information prompts the requirement for amazing information examination instruments to get rid of valuable information. Coronary illness determination is one in all the applications where information mining and AI instruments have demonstrated victories. This study used the machine learning algorithms KNN, Naïve Bayes, Random forest, Logistic regression, Support vector machine, J48, and Decision tree by WEKA software to spot which method provides maximum performance and accuracy. Using these algorithms with WEKA software, we made an ensemble (Vote) hybrid model by combining individual methods. Our research aims to access the effectiveness of various machine learning algorithms to diagnose the center disease and find the feasible algorithm, which is that the best for a heart condition.

Coronary sickness is one of the noteworthy afflictions among the propelled age people. The articulation "coronary disease" is much of the time used equally with the articulation of "cardiovascular malady" [1]. The latest review on Cardiovascular illnesses in Bangladesh

demonstrated that the inescapability of hypertension in adults was around 20-25%, followed by ischemic heart disease in adults (10%), rheumatic heart disease (1.2 per thousand), and congenital heart disease (8 per thousand new imagined youngsters) [2].

Physical inertness, tobacco use, sodium affirmation, hypertension, diabetes, heaviness, and air tainting are critical issues for the risk of cardiovascular disease (CVD), and these dangerous factors are rising in Bangladesh [3]. The top explanation behind mortality, inauspiciousness, and clinical facility affirmation in the country is cardiovascular disease, according to the National Health Bulletin [4]. Russia has the most critical pace of the coronary disease. In Russia, CVD is critical prosperity stress, with 57% of all going in the country being an outcome of CVD. Mississippi is the state with the most raised death rate from coronary disease at 233.1 per 100,000 people from the masses [5]. An investigation about the coronary disease is coordinated to see coronary peril components and illness regularity among Indians, Pakistanis, and Bangladeshis, and every South Asian and European. The pros analyzed data using SPSS/PC + version 6 [6]. The purpose of this assessment was to choose the penicillin consistency for rheumatic fever patients in an NCCRF/HD referral clinical facility in Dhaka, Bangladesh. An organized cross-sectional examination was driven among 160 patients from a picked NCCRF/HD facility in Dhaka. Data were accumulated by methods for a very close gathering using a standard sorted out study [7]. The Himachal Pradesh-Rheumatic Fever/Rheumatic Heart Disease (HP-RF/RHD) Registry database of 1918 patients was inspected. Atrial fibrillation (AF) was resolved to have a 12-lead ECG. The relationship of AF with nature and earnestness of valvular brokenness was inspected by using a multivariable determined backslide model, and the nature of connection was represented as chances extent (OR) with 95% sureness ranges (C. I.) [8]. The quantifiable assessments generally used various key backslide examination for independently organized case-control considers. The authentic group STATA (structure 5.0) was used. The dependent variable was the proximity of ischemic coronary disease, and the primary variable was demoralization before the finding or pseudo end date [9]. In this assessment, the pros used STATA/SE real programming transformation 11.2. The z test was used for joined techniques, covariates, and oddities, the t-test for minor examination impacts, and the chi-square test for heterogeneity [10].

Methods and Materials

We conducted our experiment in WEKA version 3.8.3 tool. In the present study, we used nine machine learning classifier algorithms that are given below:

Random Forest Algorithm

Random forest is a flexible and easy machine-learning algorithm to use that produces a great result most of the time. It is one of the most used algorithms; because of its simplicity and diversity, it can be used for both classification and regression methods. Random forest is also used in e-commerce to determine whether a customer will like the product or not.

K-Nearest Neighbors (KNN) Algorithm

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm which can be used to solve both classification and regression problems. KNN works by finding the distances between a query and all the data in the example, selecting the specified number (K) closest to the query. The value of k must be an odd number.

Naïve Bayes Algorithm

A Naive Bayes classifier is an algorithm, which is used Bayes theorem to classify objects. Naive Bayes classifier assumes strong, naïve and independence between attributes of data points. Popular uses of naive Bayes classifiers include spam filters, text analysis, and medical diagnosis. This classifier is widely used for machine learning analysis because it is simple to implement. The Bayesian classification is a supervised learning method as well as statistical classification. It can also solve diagnosis and predictive problems. The conditional probabilities for each feature value in the test data are obtained by getting the count of instances with that feature value in a particular class and dividing the value by the count of instances with the same class in the training set. Bayes theorem provides the process of calculating posterior probability P(c|x) from P(c), P(x), and P(x|c). Look at the equation below:

$$P(c|x) = \frac{p\left(\frac{x}{c}\right) * p(c)}{p(x)}$$

$$P(c/x) = P(x_1/c) * P(x_2/c) * P(x_3/c) * \ldots\ldots\ldots\ldots\ldots * P(x_n/c) * P(c)$$

Here P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).
P(c) is the prior probability of class.
P(x|c) is the likelihood which is the probability of predictor of class. P(x) is the prior probability of predictor.

Support Vector Machine (SVM) Algorithm

A support vector machine (SVM) is a supervised machine learning algorithm that can be employed for both classification and regression process. It can solve linear as well as non-linear problems and work effectively for many practical problems. The idea of support vector machine (SVM) is simple: The algorithm creates a line or a hyperactive plane, which separates the data into two classes. Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for both classification and regression analysis. It is one of the most popular and widely used machine learning techniques. This algorithm is also known as binary approach algorithm because it is used for binary classification like present or absence, either normal or abnormal, impactful or none impactful. In this study, it is used for the prediction of heart disease that is heart disease or non-heart disease. SVM uses kernel trick for performing non-linear classification. A kernel is used to transform low-dimensional space into high-dimensional space. There are three types of kernel such as linear, polynomial, and radial. In our method, we use a polynomial kernel.

J48 Algorithm

J48 decision tree classification is the process of building a model of classes from a set of records that contain class labels. J48 is an extension of ID3. Some additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. In WEKA, data mining tool J48 is an open-source Java implementation of the C4.5 algorithm. J48 inspect the standardized data growth that essentially the outcomes the dividing the data by selecting an element. To construct the conclusion, the element extreme regular data growth is utilized. The intense technique brings to a halt if a subset related to the similar category in all the instances. J48 creates a result node use the projected values of the class. J48 can select particular attributes, lost attribute values of the information, and contrary element values.

Simple Logistic Algorithm

Logistic regression is a statistical method used for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome can be measured with a dichotomous variable (in which there are only two possible outcomes). The dependent variable in the logistic regression is binary or dichotomous, i.e., it only contains data coded as one (TRUE, success, etc.) or 0 (FALSE, failure, etc.). The logistic regression aims to find out the optimum fitting (yet biologically reasonable) model to describe the relationship between the dichotomous characteristic of interest (dependent variable= response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula for predicting a logit transformation of the probability of the presence of the characteristic of interest:

$logit(p) = b_0+b_1X_1+b_2X_2+b_3X_3+\ldots\ldots+b_kX_k$

Where p is the probability of the presence of the dependent variable, the logit transformation can be defined as the logged odds:

$$Odds = \frac{p}{1-p} = \frac{Probability\ of\ presence\ of\ characteristic}{Probability\ of\ absence\ of\ characteristic}$$

$$logit(p) = ln\left(\frac{p}{1-p}\right)$$

One Rule Algorithm

One-R, which stands for "One Rule", is a simple yet accurate and more precise classification algorithm that generates one rule for each predictor in the data, then selects the rule with the smallest total error as its "one rule". To develop a rule for a predictor, we have to construct a frequency table for each predictor against the target. It has been shown that One-R produces rules, which are only slightly less accurate than state-of-the-art classification algorithms, producing rules that are simple for humans to interpret, and implementing the One Rule (OneR) Machine Learning classification algorithm with an enhancement for numeric data and missing values together with extensive diagnostic functions. It is useful as a baseline for machine learning models, and the rules are often helpful.

Zero Rule Algorithm

The Zero Rule Algorithm is a better baseline than the random algorithm. It uses more information about a given problem to create one rule to make predictions. This rule is different depending on the problem type. Zero Rule (ZeroR) is an effective procedure for classification algorithms whose output is simply the most frequently occurring classification in a data set. If 65% of data items have been classified rightly, ZeroR will presume that all data items have it and be right 65% of the time. Zero-R is the simplest classification method. It is that type of classification method that would lean on the target and ignore other attribute invasions. The baseline for both classification and regression problems is called the Zero Rule algorithm. For a regression predictive modeling problem where a numeric value is predicted, the Zero Rule algorithm predicts the mean of the training dataset—also called Zero-R or 0-R. Zero-R classifier predicts the majority category (class). Although there is no strength of prediction in Zero Rule, it is useful for determining a baseline performance as a benchmark for other classification methods. It is important to have a performance baseline on machine learning problems. It will give a point of reference to compare all other models that one can construct. For a classification predictive modeling problem where a categorical value is

predicted, the Zero Rule Algorithm predicts the class value with the most observations in the training dataset.

Vote Meta Classifier Algorithm

Meta level classifier is the combination of two or baser level classifier. Meta level classifier is generally better than a single base-level classifier because it has an over-fitting problem (Figure 1).

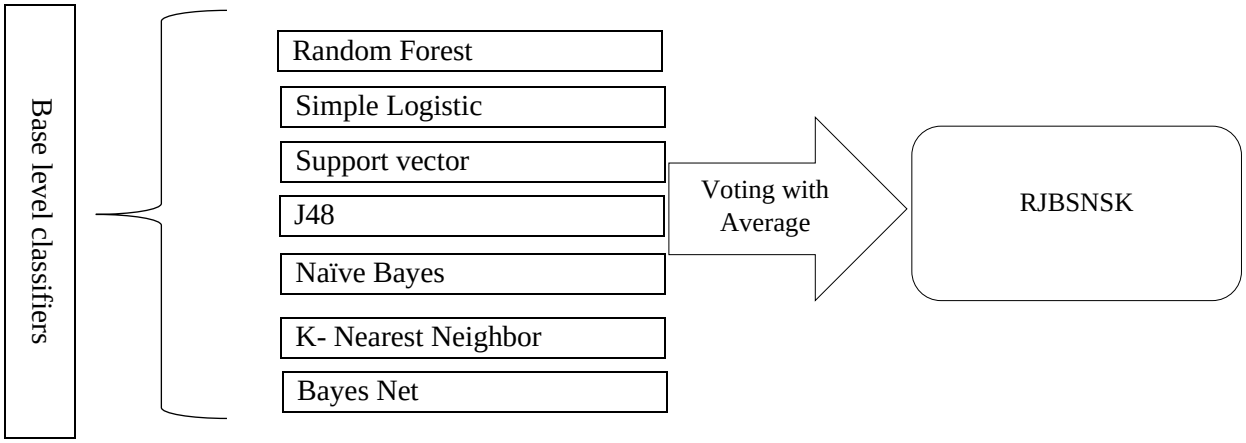

Figure 1. Meta classifier

Confusion Matrix

A confusion matrix refers to a table, which is often used to describe the performance of a classification model ("classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related method can be confusing (Figure 2).

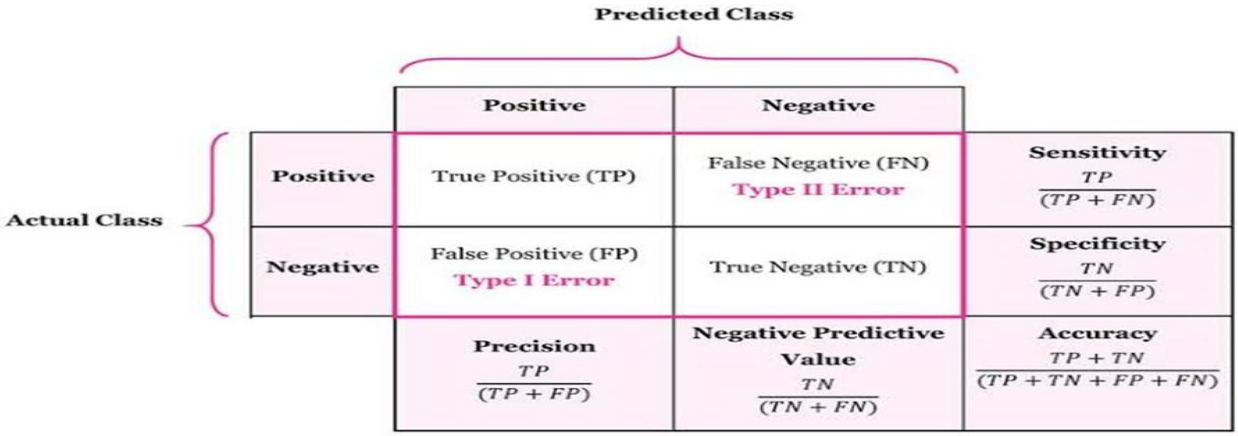

Figure 2. Confusion matrix

Data Sources

Data on heart disease is rarely available in our country. Some heart disease datasets are available in our country, which is not appropriate for using machine learning classifier algorithms. In this respect, we use the Kaggle Heart disease UCI dataset, which contains 899 observations with 14 attributes, as shown in Table 1.

Table 1
Dataset Contains 14 Attributes

Attribute	Representation	Description
1.Age	age	Age in years
2.Sex	sex	Gender instance (0 = Female, 1 = Male)
3.ChestPain	cp	Chest pain type (1: typical angina, 2: atypical angina, 3: non-anginal pain, 4:asymptomatic)
4.Rest Blood Pressure	trestbps	Resting blood pressure in mm Hg
5.SerumCholestoral	chol	Serum cholesterol in mg/dl
6.Fasting Blood Sugar	fbs	Fasting blood sugar > 120 mg/dl (0 = False, 1= True)
7.Res Electro-cardiographic	restecg	Resting ECG results (0: normal, 1: ST-T wave abnormality, 2: LV hypertrophy)
8.MaxHeartRate	thalach	Maximum heart rate achieved
9.Exercise Induced	exang	Exercise induced angina (0: No, 1: Yes)
10.Old peak	oldpeak	ST depression induced by exercise relative to rest
11.Slope	slope	Slope of the peak exercise ST segment (1: up-sloping, 2: flat, 3: down-sloping)
12.Major Vessels	ca	Number of major vessels colored by fluoroscopy (values 0 - 3)
13.Thal	thal	Defect types: value 3: normal, 6: fixed defect, 7: irreversible defect
14.Class	class	Diagnosis of heart disease (0: Non-heart disease, 1: Heart disease)

Results

Our dataset contains 899 observations with 14 attributes. We imply all the nine algorithms in our dataset. We use WEKA version 3.8.3 with stratified 10-fold cross-validation to conduct all the algorithms (Table 2).

a. In Random forest classifier, we set the number of Iteration as 68.
b. In K-Nearest Neighbors Algorithm we set the neighbor number (K) as 45.
c. In Naïve Bayes classifier, we use supervised discretization to convert numeric attributes to nominal ones.
d. In Support Vector Machine, we take the complexity parameter C as 6.
e. In J48 algorithm, we take the confidence factor as 0.15.
f. In the Simple Logistic Algorithm, the maximum number of boosting iterations is 500, and we set the number of boosting iteration as 50.
g. In One Rule algorithm, we set the minimum Bucket size as 2, which is used for discretizing (the process of transferring continuous functions, models, variables into discrete counterparts) numeric attributes.
h. In Zero Rule, we use default values for all the options.
i. In Meta classifier, among base-level classifiers, we set the number of Boosting Iteration as five in Simple Logistic classifier.

Table 2
The Results of each Machine-learning Algorithm

Classifier	TP rate	FP rate	Precision	Recall	F-measure	Roc	Accuracy	Class
Random Forest	0.767	0.147	0.809	0.767	0.788	0.855		Non-heart disease
	0.853	0.233	0.818	0.855	0.835	0.855		Heart disease
	0.814	0.194	0.814	0.814	0.814	0.885	81.4238%	Weighted Average
KNN	0.775	0.141	0.817	0.775	0.795	0.884		Non-heart disease
	0.859	0.225	0.824	0.859	0.841	0.884		Heart disease
	0.821	0.188	0.821	0.821	0.820	0.884	81.0912%	Weighted Average
Naïve Bayes	0.807	0.139	0.825	0.807	0.816	0.893		Non-heart disease
	0.861	0.193	0.845	0.861	0.853	0.893		Heart disease
	0.836	0.169	0.836	0.836	0.836	0.893	83.6485%	Weighted Average
Support Vector Machine (SVM)	0.767	0.152	0.805	0.767	0.786	0.808		Non-heart disease
	0.848	0.233	0.817	0.848	0.833	0.808		Heart disease
	0.812	0.196	0.812	0.812	0.812	0.808	82.6013%	Weighted Average
J48	0.740	0.137	0.815	0.740	0.776	0.815		Non-heart disease
	0.863	0.260	0.803	0.863	0.832	0.815		Heart disease
	0.808	0.205	0.808	0.808	0.806	0.815	80.7564%	Weighted Average
Simple Logistic regression	0.777	0.160	0.799	0.777	0.788	0.888		Non-heart disease
	0.840	0.223	0.822	0.840	0.831	0.888		Heart disease
	0.812	0.194	0.812	0.812	0.812	0.888	81.2013%	Weighted Average
One Rule	0.743	0.230	0.725	0.743	0.733	0.756		Non-heart disease
	0.770	0.257	0.786	0.770	0.778	0.756		Heart disease
	0.758	0.245	0.758	0.758	0.758	0.756	75.7508%	Weighted Average
Zero Rule	0.000	0.000	-	0.000	-	0.495		Non-heart disease
	1.000	1.000	0.551	1.000	0.710	0.495		Heart disease
	0.551	0.551	-	0.551	-	0.495	55.0612%	Weighted Average
Vote Meta	0.795	0.129	0.834	0.795	0.814	0.893		Non-heart disease
	0.871	0.205	0.839	0.871	0.854	0.893		Heart disease
	0.836	0.171	0.836	0.836	0.836	0.893	83.6485%	Weighted Average

Comparison of Different Machine Learning Classifier Algorithm

The paper centers around the Machine Learning calculation execution dependent on its actual positive rate, bogus positive rate, ROC territory, F-measure, review, exactness, supreme blunder rate, root mean square mistake rate, level of effective classifier, and level of erroneously characterized that implies precision. Therefore, our primary object is to discover the best Machine Learning calculation, which is the best accurately, characterized the coronary illness dataset agreeing to the predefined values. After individual portrayal, we look at nine Machine Learning calculations in a similar casing by graphically showing precision and ROC bend (Figure 3 and 4).

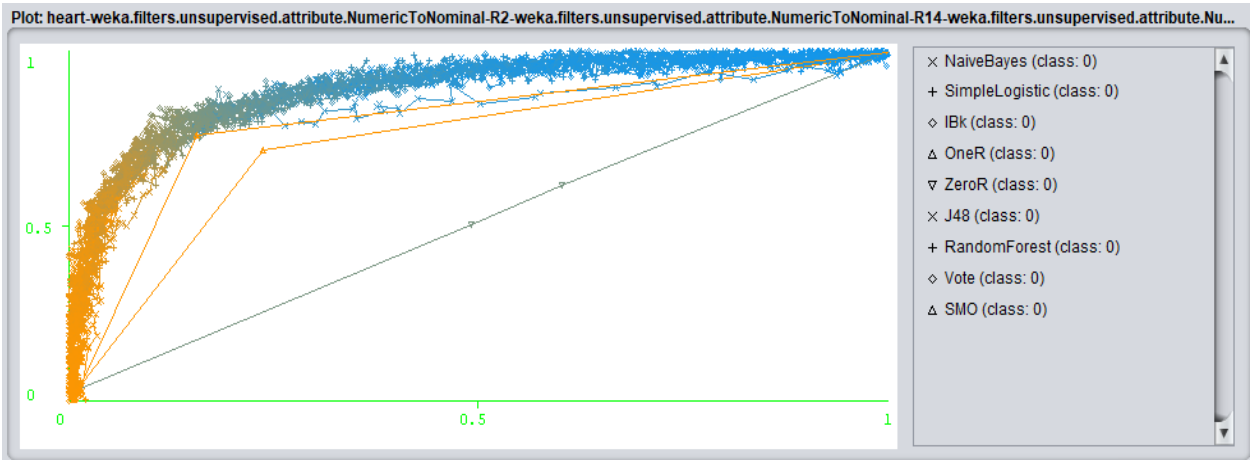

Figure 3. Multiple ROC curve based on Non-heart disease

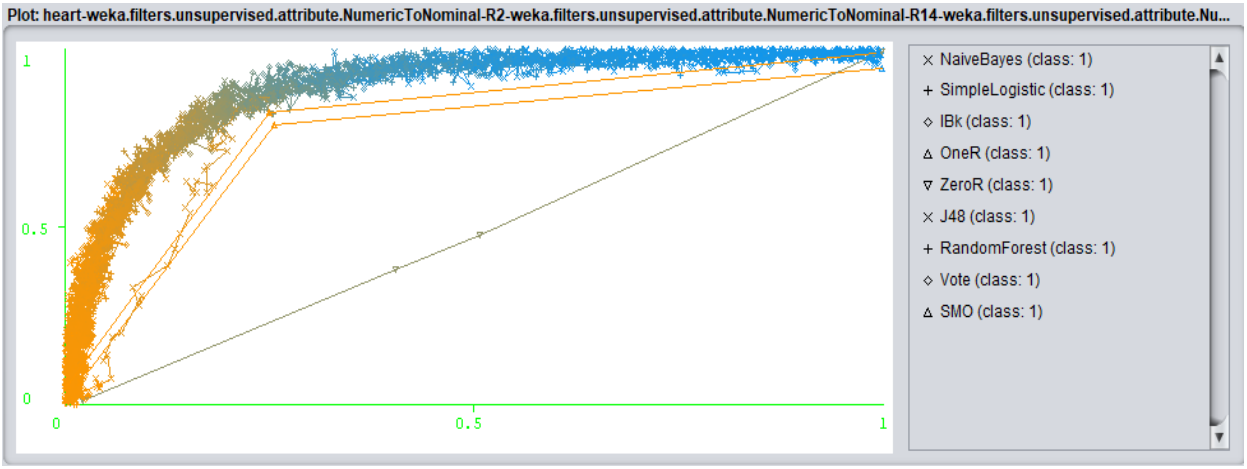

Figure 4. Multiple ROC curve based on heart disease

We speak to the individual ROC bend for every classifier dependent on the coronary illness and non-coronary illness in a singular investigation. Nevertheless, in numerous correlations, we utilize nine AI calculations. Since the ROC bend region of Random Forest (0.885, Naïve Bayes 0.893, and K-closest neighbors 0.884, Simple Logistic 0.888, and Vote Meta 0.893 calculation of the calculation is practically comparative. In the above diagram, we utilize nine AI calculations for both coronary illness and non-coronary illness. It relates ROC bend region is Support vector machine 0.808, J48 0.815, One Rule 0.756, Zero Rule 0.495.

At last, from the examination, we see that the Meta Vote and Naïve Bayes classifiers speak to the equivalent and best ROC bend zone for both coronary illness and non-coronary illness.

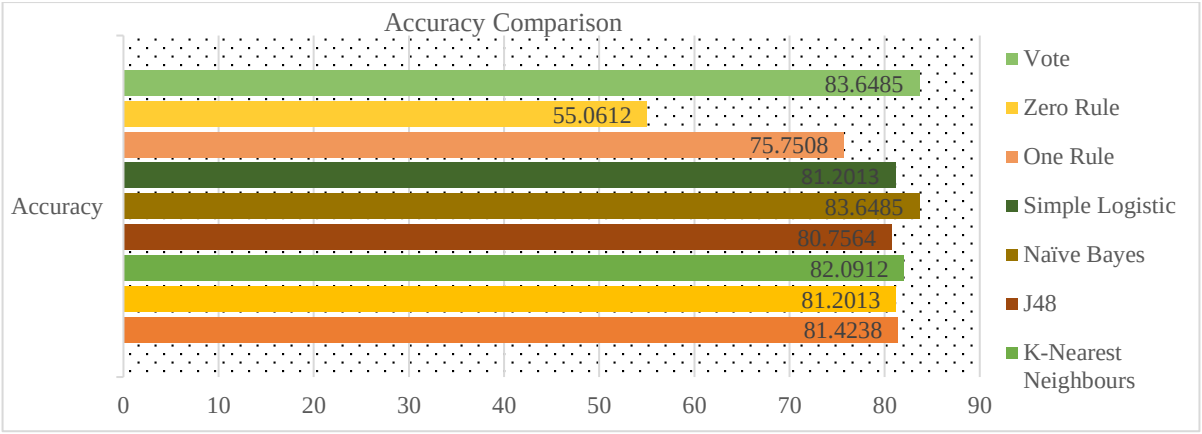

Figure 5. Comparison of Different Machine Learning algorithm's accuracy based on heart disease dataset.

Discussion

Coronary illness has become a reason for expanding worry for Bangladesh, with patients experiencing it finishing off the rundown of people with non-transmittable sicknesses. In reality, cardiovascular ailments, particularly coronary course infection, are developing by scourge extents systematically. Our examination is led by WEKA instrument. Among the distinctive AI calculations, we are able to presume that Meta Vote classifier, which is the mixture of Random Forest, Simple Logistic, Naïve Bayes, Bayes net, J48, K-closest neighbor, Support vector machine, and Naïve Bayes give the equivalent and best precision of 83.6485%. Therefore the other calculations speak to the exactness as Support vector machine (81.2013%), Simple Logistic (81.2013%), Random Forest (81.4238%), J48 (80.7564%), K-closest neighbors (82.0912%), One Rule (75.7508%), Zero Rule (55.0612%) for accurately arranging of the center malady dataset. Our examination aims to personality absolutely the best AI calculation, which provides more exactness for distinguishing coronary illness tolerance. We recommended that Naïve Bayes and Meta vote calculation be more precise to foresee the coronary illness from our exploration.

References

[1] Mayo clinic. Patient care and health information on heart disease. Updated May 06,2021 (https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118)

[2] Malik A. National Heart Foundation of Bangladesh gives information about heart disease in Bangladesh. Updated December,2020 (http://www.nhf.org.bd/founder.php)

[3] Hassan M. People with hypertension to rise globally, significantly higher in South Asia. Updated May2019 (https://www.dhakatribune.com/health/2018/10/02/3-out-of-4-bd-individuals-run-the-risk-of-developing-cardiac-diseases)

[4] Maswood MH. On August 19, 2019 The New age Bangladesh published an article "Cardiac patients rising in Bangladesh, 2.5 lakh die annually". Updated: Sep 29,2018

(http://www.newagebd.net/article/51904/cardiac-patients-rising-in-bangladesh-25- lakh-die-annually)

[5] Felman A. Medical News Today report on everything you need to know about heart disease. Updated on September 29, 2020 (https://www.medicalnewstoday.com/articles/237191.php)

[6] Raj B, Nigel U, Martin W. Heterogeneity of coronary heart disease risk factors in Indian, Pakistani, Bangladeshi, and European origin populations: cross sectional study. BMJ Clin Res. 1999;319(7204):215-220.

[7] Farzana K, Faruque MO, Zareen S, Choudhury K, Hossain, A. Factors Affecting Therapeutic Compliance among the Patients with Rheumatic Heart Disease in Bangladesh. Cardiovascular Journal. 2018;10(2):180-185.

[8] Negi PC, Sondhi S, Rana V, Rathoure S, Kumar R, Kolte N, et al. "Prevalence, risk determinants and consequences of atrial fibrillation in rheumatic heart disease: 6 years hospital based-Himachal Pradesh- Rheumatic Fever/Rheumatic Heart Disease (HP-RF/RHD) Registry". Indian Heart J. 2018;70: S68-S73

[9] Hippisley-Cox J, Fielding K, Pringle M. Depression as a risk factor for ischaemic heart disease in men: population based case-control study. BMJ. 1998; 316:1714–1719.

[10] Livesey G., Livesey H. Coronary Heart Disease and Dietary Carbohydrate, Glycemic Index, and Glycemic Load: Dose-Response Meta-analyses of Prospective Cohort Studies. Mayo Clin. Proc. Innov. Qual. Outcomes. 2019; 3:52–69.