Elimination and Backward Selection of Features (P-Value Technique) In Prediction of Heart Disease by Using Machine Learning Algorithms

Ritu Aggrawal[1], Saurabh Pal[2]
[1]Research Scholar, VBS Purcanchal University, Jaunpur
[2]Deapartment of Computer Applications, VBS Purcanchal University, Jaunpur
drsaurabhpal@yahoo.co.in

Abstract:

Background: Early speculation of cardiovascular disease can help determine the lifestyle change options of high-risk patients, thereby reducing difficulties. We propose a coronary heart disease data set analysis technique to predict people's risk of danger based on people's clinically determined history. The methods introduced may be integrated into multiple uses, such for developing decision support system, developing a risk management network, and help for experts and clinical staff.

Methods: We employed the Framingham Heart study dataset, which is publicly available Kaggle, to train several machine learning classifiers such as logistic regression (LR), K-nearest neighbor (KNN), Naïve Bayes (NB), decision tree (DT), random forest (RF) and gradient boosting classifier (GBC) for disease prediction. The p-value method has been used for feature elimination, and the selected features have been incorporated for further prediction. Various thresholds are used with different classifiers to make predictions. In order to estimating the precision of the classifiers, ROC curve, confusion matrix and AUC value are considered for model verification. The performance of the six classifiers is used for comparison to predict chronic heart disease (CHD).

Results: After applying the p-value backward elimination statistical method on the 10-year CHD data set, 6 significant features were selected from 14 features with p <0.5. In the performance of machine learning classifiers, GBC has the highest accuracy score, which is 87.61%.

Conclusions: Statistical methods, such as the combination of p-value backward elimination method and machine learning classifiers, thereby improving the accuracy of the classifier and shortening the running time of the machine.

Key Words: p-value technique, Statistical Method, Chronic heart disease, Confusion matrix, Machine learning, ROC, AUC.

1. Introduction

Identifying the evidence of risk factors that increase the incidence of cardiovascular illness is one of the significant achievements in the study of disease transmission in the 20th century (Einarson et al. 2018). In addition, analysts can choose to establish multivariate risk prediction calculations to help clinicians perform risk assessment. In the last 10 years, the author has proposed many risk scores (Sofi et al. 2014). These are all created for hazard assessment in a limited time of ten years or less. In order to meet this demand, some reports have introduced the whole life risks of CVD, CHD and stroke. Some experts work to calculate life span and long-opportunities in a class or class of hazardous variables (WHO 2012). Their findings emphasize the importance of the level of risk factors in early adulthood to the risk of CVD, just as CVD risk factors have a huge impact on all-cause mortality. They also pointed out that ten years of work might reduce the real dangers, especially among young people and ladies. These outcome highlight require for continuing models of CVD threat expectations that are very important for young adults and represent a competitive reason for non-CVD mortality (Singh et al. 2020). In any case, obviously, no calculation method has been proposed to measure the direct ability of 10-year CVD risk as a risk factor. The trouble of finding a long enough and thoroughly developed methodological complexity associated with integrating competing death risks into multivariate risk assessments for various reasons (Proust-Lima et al. 2016).

This exploratory article clarifies a procedure for assessing the 10-year risk of hard CVD function among people liberated from baseline conditions. Our risk scale will consider changes to the serious danger of non-CVD deaths, and will utilize standard danger factors that can be gathered during doctor visits. This process depends on the

Framingham CHD dataset, which adds some effective risk score calculations for thorough observation of CVD events. It is presented in a way that all analysis relies on utilization of 10-years of collected data (Den et al. 2012). From the feature elimination process, we select a factor that p-value is less than or equal to 0.5. Backward feature elimination is applied for this purpose.

Machine learning is very valuable for different problem arrangements. One of the uses of this method is to predict needed variables based on the estimation of autonomy factors. The medical field is an application field of information mining, because it has a large number of information assets. Realizing that it is valuable to include selection and feature reduction. Feature determination is concerned about distinguishing some relevant features that are sufficient to learn objective thoughts. The quantitative reduction of periodontal disease is a risk factor for cardiovascular infections, and the underlying basis for the results was studied in the author. Four scientists freely separated RR, CI, and p-value from each survey and assessed the level of confusing changes. Later on, periodontal contamination will build the danger of cardiovascular sickness by 19% (Janket et al. 2003). For observe whether it expects to be implemented in cognition, ability and behavior in the long run, and observe whether it expects endurance. Patients in the middle and middle stages maintain good execution ability in complex exercises of psychology (ADAScog and VSAT), worldwide (CDR-SB) and daily activity measurement (IADL) (P estesms <0.001, medium vs. fast; P estimate <0.003 to 0.03 the difference between transition and fast). To assess the infection period and determine the potential (pre-exercise) incidence of 597 Alzheimer's disease patients who were monitored for 15 years (Janket et al. 2010). In order to test on the Internet, calculations based on certification can distinguish the hazard indicators of Parkinson's disease among the British people. A total of 1323 members are selected for evaluation each year, and more than 79% of the evaluations are completed. The annual risk score corresponds to the moderate index of PD, while the pattern score is related to the moderate index during development (all p-values <0.001). The PD event analysis performed during the development process is completely related to the standard risk score (risk ratio 5 4.39, P 5.045). In 47 individuals were discovered GBA variations or G2019S LRRK2 variations (Noyce et al. 2017). Used the framework of coronary heart disease conclusions based on rough set-based quality reduction and stretch ranking 2 Fluffy Basic Principles Framework (IT2FLS). The rough set-based quality reduction using camouflage firefly calculation is explored to find the ideal reduction, which reduces the amount of calculation and improves the execution of IT2FLS. The results of the examination show the key advantages of the proposed framework, in contrast to other AI strategies (especially Naive Bayer, Support Vector Machines, and artificial neural network) (Long et al. 2015). Feature selection strategies can be used as an important method to reduce the cost of conclusions by selecting important attributions. Foresee Arrange the models and use the Cleveland and statlog risk heart data sets to understand that the selected focus plays a crucial role in the prediction of CAD (Aggrawal and Pal 2021). The arrangement accuracy of any decision calculation and the accuracy of the decision model have reached 90-95%, depending on three different ratio parts (Reddy et al. 2019). In order to forecast the patient's CAD, stochastic gradient boosting calculations and recursive feature elimination (RFE) are used to select the best features in the data. To select the most useful features, uses a loop called recursive feature elimination (RFE). The calculation accuracy of the stochastic gradient boosting increase is 95.45% (Kakulapati et al. 2017). This structure can reasonably locate the fundamentals to predict the risk of patients according to the given health status boundary. The main principle of this examination is to help unspecified experts make the right choice regarding the risk of coronary artery disease. The standards created by the proposed framework are organized by original rules, trimming rules, no duplication rules, classification rules, sorting rules, and Polish (Chaurasia and Pal 2020). Evaluating the implementation of the system is similar to the accuracy of the implementation process, and the results show that the structure has significant potential in predicting the risk of coronary artery disease more accurately. The productive heart disease prediction system achieved the most significant accuracy of 86.7% (Saxena and Sharma 2016). Feature selection based on fast correlation (FCBF) technology can guide redundant features, thereby improving the nature of coronary artery disease arrangements. Around then, it was orchestrated by different request estimations, for example, K nearest neighbors, support vector machines, naive Bayes, random forests and multi-layer perception just as artificial enhancement by particle swarm optimization (PSO) and ant colony optimization (ACO) techniques Neural Networks. Using the simplified model proposed by FCBF, PSO and ACO, the most extreme layout accuracy of 99.65% can be achieved (Khourdifi and Bahaj 2019). In order to make a artificial Lampyridae classifier, and further compare it with the Takagi Sugeno Kang fluffy classifier and the ANN classifier to predict the accuracy, susceptibility, particularity and connection coefficient of Mathew. Despite other execution measurements, the essence of MCC is to test the capabilities of AI classifiers. The use case is completed in Scilab, and it is inferred from the obtained results that the constructed ALC is better than the TSK fluffy classifier and the ANN classifier. The results are encouraging. It is speculated that the accuracy of male diabetic patients is 87.60% and the accuracy of female diabetic patients is 87.27% (Narasimhan and Malathi 2019).

2. Methods

Several techniques and methods were used in this experiment to assess the ten-year threat of CHD. The method section is divided into two parts, one part describes applied machine learning algorithms, and the other part describes experimental methods.

I. Applied Machine Learning Algorithms

In this section, we discuss machine learning algorithms, which will be used as methods throughout the research article.

Logistic Regression (LR)

Logistic or logit models are used to prove the possibility of a particular category or function. Some functional categories can be expanded to display. The probability of each article identified in the picture will be reduced to a value between 0 and 1, and the number will be 1 (Balu et al. 2019).

Consider a model with two indicators x_1 and x_2 and a parallel response variable Y, we mean p = P (Y = 1). We accept the direct link between the index factor and the log chance of the function Y = 1.

This direct relationship can be written in the accompanying digital structure.

Among them, l is the logarithmic chance, b is the base of the logarithm, and β_i is the boundary of the model:

$$l = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Random Forest (RF)

Random forest is an ensemble learning technique for ordered, recursive, and different tasks (Dudek 2015, Chaurasia and Pal 2021). It works by developing a large number of decision trees in preparation time and generating classes as a method for arranging or recursively predicting a single tree. Arbitrary decision trees are suitable for decision trees and tend to overfit their preparations. Random forest consists of large lateral decision trees, but its accuracy is lower than that of gradient boost trees. Nevertheless, the nature of the information will affect its display.

Decision Tree (DT)

The structure of the decision tree is similar to a flowchart, in which each internal center point tests the quality, each branch measures the test results, and each leaf center measures the cycle class label (Adebayo and Chaubey 2019). The way from root to leaf is related to representation rules.

In decision-making investigations, decision trees and closely related influence outlines are used as visual and scientific selection aid tools to determine the normal benefits of competing choices.

Naïve Bayes (NB)

The naive Bayes classifier is a set of basic probabilistic classifiers that rely on the application of Bayesian assumptions and reliable autonomy assumptions between features. They are one of the most direct Bayesian organization models. However, they can be used in conjunction with core thickness evaluation and achieve a higher level of accuracy.

The naive Bayes classifier is highly adaptable, and requires the boundaries of various directly influencing factors in learning problems (Krawczyk 2017). In contrast to the costly iterative guesswork of evaluating some different types of classifiers, the most extreme probability preparation should be made by evaluating the clarity of the closed structure that requires immediate time.

Utilizing Bayes' hypothesis, the contingent likelihood can be written as:

$$p(C_k|x) = \frac{p(C_k)p(x)|C_k)}{p(x)}$$

Among them, for each of the K potential results or Ck classes, x = (x1,...,xn) represents n features.

K-Nearest Neighbor (KNN)

K-nearest neighbor calculation is a non-parametric strategy for sequence and recurrence. In both cases, the information includes the k nearest preparation models in the composition space. k-NN is a kind of occasion-based learning or slow implementation, in which the ability is only approximated locally, and all calculations are retained until the work evaluation. Since this calculation depends on the separation of groups, normalizing preparation information can greatly improve its accuracy (Chaurasia and Pal 2018).

Whether it is characterization or recurrence, a useful method can be to distribute the load to neighbors' promises so that the closer neighbors provide more normal services than the more inaccessible neighbors.

Gradient Boosting Classifier (GBC)

Gradient boosting is an AI program for redundancy and change problems. It serves as a set of prior models and decision trees to form a hypothetical model. Like other advanced technologies, it develops the model in a phased, distinct style and summarizes the model by allowing enhancements on optional works (Stamate et al. 2018).

For now, let us consider a gradient boosting calculation with M stages. The slope of each stage is increased by m (1 <= m <= M), assuming that the model F_m is not perfect. In order to improve Fm, some new estimator's h_m(x) should be added to our calculations. Therefore,

$$F_{m+1}(x) = F_m(x) + h_m(x) = y$$

Or,

$$h_m(x) = y - F_m(x)$$

Logistic Regression and P-value Interpretation: Backward Elimination (Feature Selection)

Regression surveys create conditions for describing the measurable link between at least one indicator factor and the reply variable (Suguna et al. 2019).

The p value for each term tests the invalid hypothesis, that is, the coefficient is equivalent to zero. A low p-value (<0.05) demonstrates that it can disregard the invalid hypothesis. In the final analysis, indicators with low p-value may become an important extension of the display, because changes in indicator values can be identified by changes in response variables.

On the other hand, the adjustment of the larger p-value suggested index has nothing to do with the change in response.

Iteration Log

Iteration log is the release of log probability at each iteration. The main logarithm (iteration 0) is the log probability of an "invalid" model; that is, a model without any indicators (Harrell Jr 2015). In each iteration, the log probability will increase, and the purpose is to expand the log probability. When the contrast between progressive logging is small, it is said that the model is satisfied, the iteration is stopped, and the results are displayed.

Log likelihood

The estimation of log probability is not important in it. Rather, this number can be used to help consider established models (Smith and Levy 2013).

Number of observation

This is the amount of perception used in the survey. If lack quality in any of the factors used for strategic relapse, this number may be more moderate than the absolute number of perceptions in the information index. Statistics naturally uses list erasure, which means that if any factors are missing from the strategic relapse, the entire case will be rejected for review.

Pseudo R-squared

Logistic regression is different from R-squared in OLS recurrence. There are many types of pseudo R-squared measurements (Ye et al. 2019). This metric does not mean the R-squared method in OLS regression.

Dependent Variables

This is a relative variable in logistic regression.

Coef.

These are the calculated quality of the recurrence conditions and are used to predict the required variables based on the free factors (Gao et al. 2016). They are based on log chances. Like OLS relapse, the expectation condition is-

$$\begin{aligned}\mathrm{logit(p)} = \log(p/1-p) \\ &= b_0 + b_1 * Sexmale + b_2 * age + b_3 * cigsPerDay + b_4 * totChol + b_5 * sysBP + b_6 \\ &* glucose\end{aligned}$$

Where, p is the possibility of being in the structure. On the factors used in this model, the logistic regression conditions are-

$$\begin{aligned}\mathrm{log(p/1-p)} &= -9.1264 + 0.5815 * \mathrm{Sexmale} + 0.0655 * \mathrm{age} + 0.0197 * \mathrm{cigsPerDay} + 0.0023 * \mathrm{totChol} \\ &+ 0.0174 * \mathrm{sysBP} + 0.0076 * \mathrm{glucose}\end{aligned}$$

These assessments teach us about the links between independent factors and related variables, which require variables to be on a logarithmic scale. These evaluations tell us that the expected logarithmic ratio = 1 expansion measure, which will be expected by each additional unit in the indicator and keep all the different indicators stable.

Std. Err.

These are the standard mistakes related with the coefficients (Cole 2004). The standard blunder is utilized for testing whether the boundary is fundamentally unique in relation to 0; by separating the boundary gauge by the standard mistake we acquire a z-value. The standard mistakes can likewise be utilized to frame a certainty span for the boundary.

z and P>|z| Values

These subdivisions provide the z-value and 2 p-values to test invalid guesses with a coefficient of 0 (Miyamoto et al. 2018). In fact, the coefficient of p-value not completely equal to α is very large. For example, if we choose an alpha of 0.05, the coefficient of p-estimated value equal to or less than 0.05 is actually crucial, that is, we can ignore the invalid theory and point out that the coefficient is inherently unique with respect to 0.

Odds ratio (OR) and Logistic Regression (LR)

An odds extent is an extent of connection between a presentation and an outcome. The OR addresses the odds that an outcome will happen given a particular presentation, appeared differently in relation to the odds of the outcome occurring without that introduction.

Exactly when a LR is resolved, the LR coefficient (b) is the evaluated increase in the log odds of the outcome per unit increase in the assessment of the introduction (Park 2013).

The OR can similarly be used to choose if a particular introduction is a danger factor for a particular outcome, and to take a gander at the diverse danger factors for that outcome.

OR=1 Exposure doesn't impact chances of result
OR>1 Exposure related with higher chances of result
OR<1 Exposure related with lower chances of result

$$Odds\ Ratio = \frac{Number\ of\ exposed\ cases(a) \Big/ Number\ of\ unexposed\ cases(c)}{Number\ of\ exposed\ non-cases(b) \Big/ Number\ of\ unexposed\ non-cases(d)}$$

Or,

$$Odds\ Ratio = \frac{Number\ of\ exposed\ cases(a) * Number\ of\ unexposed\ non-cases(d)}{Number\ of\ exposed\ non-cases(b) * Number\ of\ unexposed\ cases(c)}$$

Confidence Intervals (CI)
The 95% certainty stretch is utilized to gauge the accuracy of the Odds proportion. A huge CI demonstrates a low degree of exactness of the OR, while a little CI shows a higher accuracy of the OR. It is essential to note in any case, that dissimilar to the p-value, the 95% CI doesn't report a measure's factual hugeness (Park et al. 2016). Practically speaking, the 95% CI is frequently utilized as an intermediary for the presence of factual criticalness in the event that it doesn't cover the invalid worth. By the by, it is improper to decipher an OR with 95% CI that traverses the invalid an incentive as demonstrating proof for absence of relationship between the presentation and result.

$$Upper\ 95\%\ CI = e\text{^}[\ln(OR) + 1.96\sqrt{1/_a + 1/_b + 1/_c + 1/_d}]$$
$$Lower\ 95\%\ CI = e\text{^}[\ln(OR) - 1.96\sqrt{1/_a + 1/_b + 1/_c + 1/_d}]$$

Model Validation
In AI, model approval is alluded to as the cycle where a prepared model is assessed with a testing informational index. The testing informational collection is a different segment of a similar informational collection from which the training set is determined. The fundamental motivation behind utilizing the testing informational collection is to test the speculation capacity of a prepared model.

Confusion Matrix
Confusion matrix is an arrangement of row and column that is frequently used to depict the exhibition of an order model on a bunch of test information for which the genuine qualities are known (Aggrawal and Pal 2020).
True positives (TP): These are cases in which we anticipated truly, and they do have.
True negatives (TN): We anticipated no, and they don't have.
False positives (FP): We anticipated indeed, however they don't really have. (Type I mistake)
False negatives (FN): We anticipated no, yet they really have. (Type II mistake)

Model Evaluation (Statistics)
Accuracy: In general, how frequently is the classifier right? Accuracy is defined as:

$$Accuracy\ of\ Model = \frac{TP + TN}{TP + TN + FP + FN}$$

Misclassification: Generally, how regularly is it wrong? The formula is:

$$Misclassification = 1 - Accuracy$$

True Negative Rate or Specificity: When it is actually no then how frequently does it anticipate no? The formula is:

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

Positive Predicted Value (PPV): It decides, out of the entirety of the positive discoveries, the number of are genuine positives.

$$PPV's = \frac{TP}{TP + FP}$$

Negative Predicted Value (MPV): It decides, out of the entirety of the negative discoveries, the number of are genuine negatives.

$$NPV's = \frac{TN}{TN + FN}$$

Positive Likelihood Ratio (LR+):
It is acquire when TPR divided by the FPR.

$$LR+= \frac{Sensitivity}{1 - specificity}$$

Negative Likelihood Ratio (LR-):
It is the likelihood of a patient testing negative that has an infection separated by the likelihood of a patient testing negative who doesn't have a sickness.

$$LR-= \frac{1 - Sensitivity}{specificity}$$

Threshold Values:
In order to describe logistic regression value for binary categories, we should describe a classification threshold (decision threshold) (Besse et al. 2013). A value exceeding this limit means "infection"; and the underneath specifies "no disease". The classification threshold should be always 0.5. The threshold is a subordinate problem, so it is the value we should adjust.

ROC Curve:
Receiver operating characteristic (ROC) curve is a graph demonstrating the introduction of a grouping model at all portrayal limits (Chaurasia and Pal 2020). This curve plots two limits:
True Positive Rate (TPR) and False Positive Rate (FPR)

Area under Curve:
Area under the curve (AUC) measures the entire two-dimensional area under the entire ROC twist from (0, 0) to (1, 1).

AUC gives an absolute extent of execution over all possible portrayal edges (Gao and Wang). One technique for interpreting AUC is as the probability that the model positions a sporadic positive model more significantly than a discretionary negative model.

II. Experimental Methodology

The ongoing cardiovascular data is focused on residents of Framingham, Massachusetts. The objective is to envision whether the patient has 10-year risk of future coronary ailment (CHD).The dataset gives the patient's information. It consolidates in excess of 4,000 records and 15 features.

After preprocessing the data set, logistic regression has been applied to obtain statistical results, such as standard error, z-value, p-value, and confidence interval (25-95%). In addition, these P values will be used to select features with P values <= 0.5. Six machine learning algorithms are applied to obtain accuracy. At the next level, all these results obtained from the classifier will enter the verification level, where ROC, AUC values and confusion matrix are checked. Figure 1 describes the steps used in this experiment.

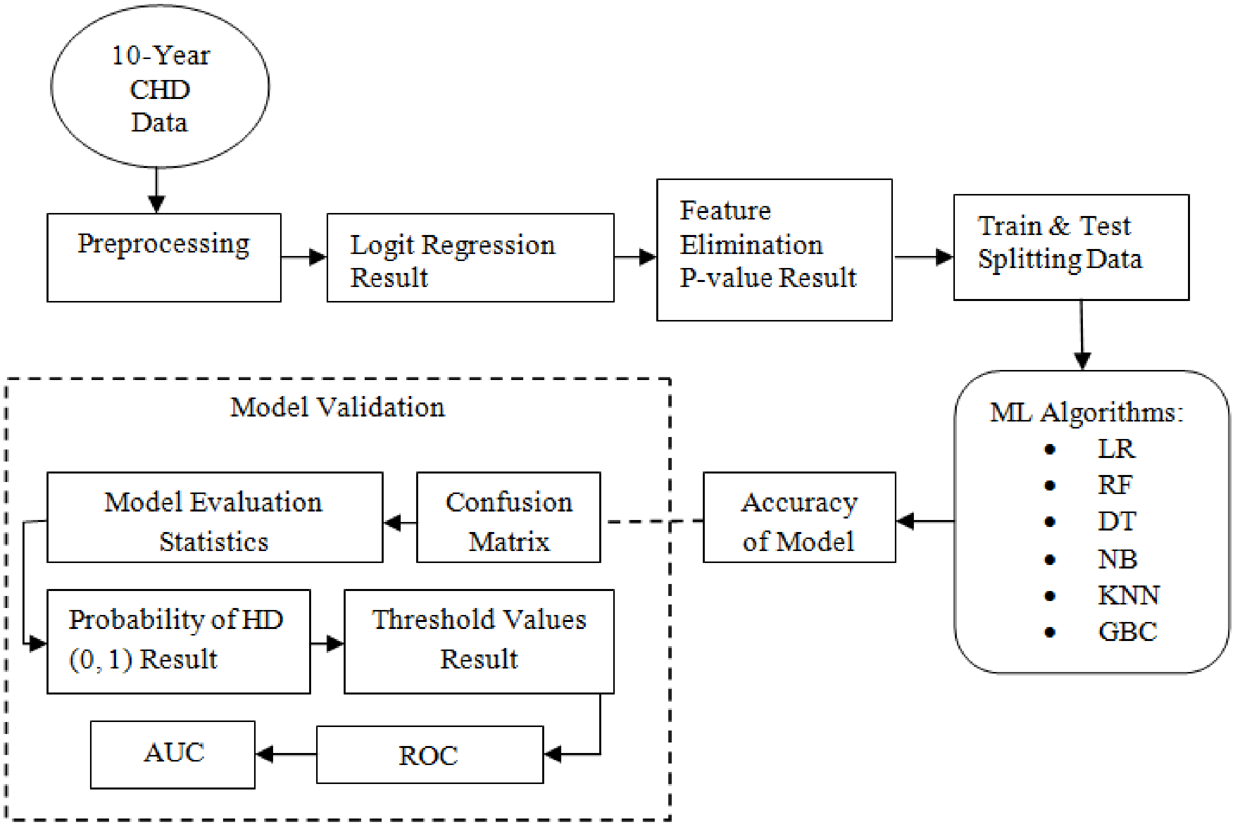


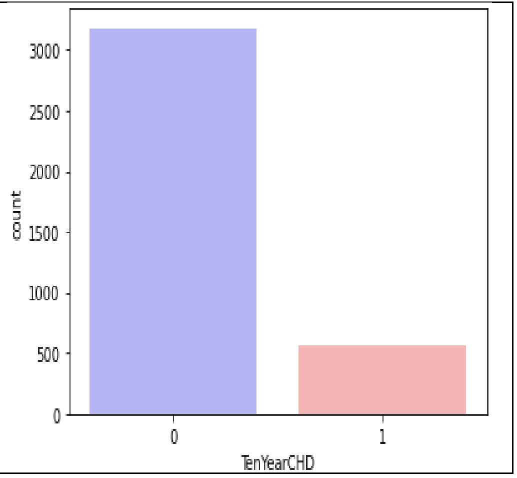
Figure 1. Step-by-step instructions for application methods

3. Experimental Setup

The experimental data is taken from Framingham Heart Research Data Set (Kannel et al. 1979). The data set contains 4240 records and 15 attributes. Variable information is provided in Table 1 below. In the following data set, some values are missing in the attributes, such as cigsPerDay, BPMeds, totChol, BMI, heartRate and glucose. The total number of missing values was 489, so these rows with missing values were excluded for further analysis. Now, among 3751 records, 3179 patients have no 10-year danger of coronary illness, and 572 patients are at risk after this time period.

Table 1. 10-year CHD dataset attributes Information

Risk Category	**Attributes**	**Attribute Type**	**Description**	**10-Year Risk Factor**
Demographic	sex	Nominal	male or female	Number of patients who have 10-year CHD Risk: No 0 3179 Yes 1 572
	age	Continuous	age of patient	
Behavioral	currentSmoker	Nominal	smoker or not	
	cigsPerDay	Continuous	per day cigarette consumption	
Medical(history)	BPMeds	Nominal	hypertension medication or not	
	prevalentStroke	Nominal	previous record of stroke or not	
	prevalentHyp	Nominal	hypertensive history or not	

	diabetes	Nominal	diabetes history or not	
Medical(current)	totChol	Continuous	cholesterol level	
	sysBP	Continuous	systolic blood pressure	
	diaBP	Continuous	diastolic blood pressure	
	BMI	Continuous	body Mass Index	
	heartRate	Continuous	heart rate	
	glucose	Continuous	glucose level	
Predict variable (desired target)	TenYearCHD	Binary	10-year risk of CHD (1-Yes, 0-No)	

4. Results

In this part, all the outcomes delivered by the model and their significance is investigated and clarified.

Logistic Regression

In following case, use the opposite approach, eliminate these features one by one with the most important P values, and then relapse again and again until all attributes have P values below 0.05. In Table 2 below, there are different P values with different attributes.

Table 2. Statistical significance of attributes

	Coef	std err	z	P>|z|	[0.025	0.975]
const	-8.6532	0.687	-12.589	0.000	-10.000	-7.306
Sex_male	0.5742	0.107	5.345	0.000	0.364	0.785
age	0.0641	0.007	9.799	0.000	0.051	0.077
currentSmoker	0.0739	0.155	0.478	0.633	-0.229	0.377
cigsPerDay	0.0184	0.006	3.000	0.003	0.006	0.030
BPMeds	0.1448	0.232	0.623	0.533	-0.310	0.600
prevalentStroke	0.7193	0.489	1.471	0.141	-0.239	1.678
prevalentHyp	0.2142	0.136	1.571	0.116	-0.053	0.481
diabetes	0.0022	0.312	0.007	0.994	-0.610	0.614
totChol	0.0023	0.001	2.081	0.037	0.000	0.004
sysBP	0.0154	0.004	4.082	0.000	0.008	0.023
diaBP	-0.0040	0.006	-0.623	0.533	-0.016	0.009
BMI	0.0103	0.013	0.827	0.408	-0.014	0.035
heartRate	-0.0023	0.004	-0.549	0.583	-0.010	0.006
glucose	0.0076	0.002	3.409	0.001	0.003	0.012

Backward elimination and Feature Selection

Table 3 below shows the P value and the corresponding statistics. Select only based on those important features with a P value less than 0.5, such as Sex_male, age, cigsPerDay, totChol, sysBP and glucose.

Table 3. Selected attributes based on P-value

	coef	std err	z	P>|z|	[0.025	0.975]
const	-9.1264	0.468	-19.504	0.000	-10.043	-8.209

Sex_male	0.5815	0.105	5.524	0.000	0.375	0.788
age	0.0655	0.006	10.343	0.000	0.053	0.078
cigsPerDay	0.0197	0.004	4.805	0.000	0.012	0.028
totChol	0.0023	0.001	2.106	0.035	0.000	0.004
sysBP	0.0174	0.002	8.162	0.000	0.013	0.022
glucose	0.0076	0.002	4.574	0.000	0.004	0.011

Odds Ratio, Confidence Intervals and P-values

In Table 4 below, the odds ratio, confidence interval and P value are calculated.

Table 4. Results of Odds Ratio, Confidence Intervals and Pvalues

	CI 95% (2.5%)	**CI 95% (97.5%)**	**Odds Ratio**	**P-value**
const	0.000043	0.000272	0.000109	0.000
Sex_male	1.455242	2.198536	1.788687	0.000
age	1.054483	1.080969	1.067644	0.000
cigsPerDay	1.011733	1.028128	1.019897	0.000
totChol	1.000158	1.004394	1.002273	0.000
sysBP	1.013292	1.021784	1.017529	0.000
glucose	1.004346	1.010898	1.007617	0.000

Model evaluation with corresponding Statistics

As shown in Table 5, calculation has been made for accuracy, misclassification, sensitivity, specificity, positive predictive value, negative predictive value, positive likelihood ratio and negative likelihood ratio of the classifier. The accuracy of the classifier GBC is higher, so there are fewer classification errors.

Table 5. Model Statistics

Model Statistics	**LR**	**RF**	**DT**	**NB**	**KNN**	**GBC**
Accuracy of the model = TP+TN/(TP+TN+FP+FN)	0.8748	0.8695	0.7603	0.8561	0.8695	0.8761
The Misclassification = 1-Accuracy	0.1251	0.1304	0.2396	0.1438	0.1304	0.1238
Sensitivity or True Positive Rate = TP/(TP+FN)	0.0543	0.0978	0.1956	0.1086	0.1521	0.0108
Specificity or True Negative Rate = TN/(TN+FP)	0.9893	0.9772	0.8391	0.9605	0.9696	0.9969
Positive Predictive value = TP/(TP+FP)	0.4166	0.375	0.1451	0.2777	0.4117	0.3333
Negative predictive Value = TN/(TN+FN)	0.8822	0.8858	0.8819	0.8853	0.8912	0.8783
Positive Likelihood Ratio = Sensitivity/(1-Specificity)	5.1164	4.2978	1.2163	2.7550	5.0141	3.5815
Negative likelihood Ratio = (1-Sensitivity)/Specificity	0.9558	0.9231	0.9585	0.9279	0.8743	0.9921

Figure 2 below is a graphical representation of Table 5. The accuracy of the classifier is in order (GBC <LR <RF = KNN <NB <DT), that is, the accuracy of GBC is higher, 87.61%, then the accuracy of LR is 87.48%, and the same accuracy of RF and KNN is 86.95%, the accuracy of NB is 85.61%, and the minimum accuracy of DT is 76.03%. Other indicators have their usual meanings.

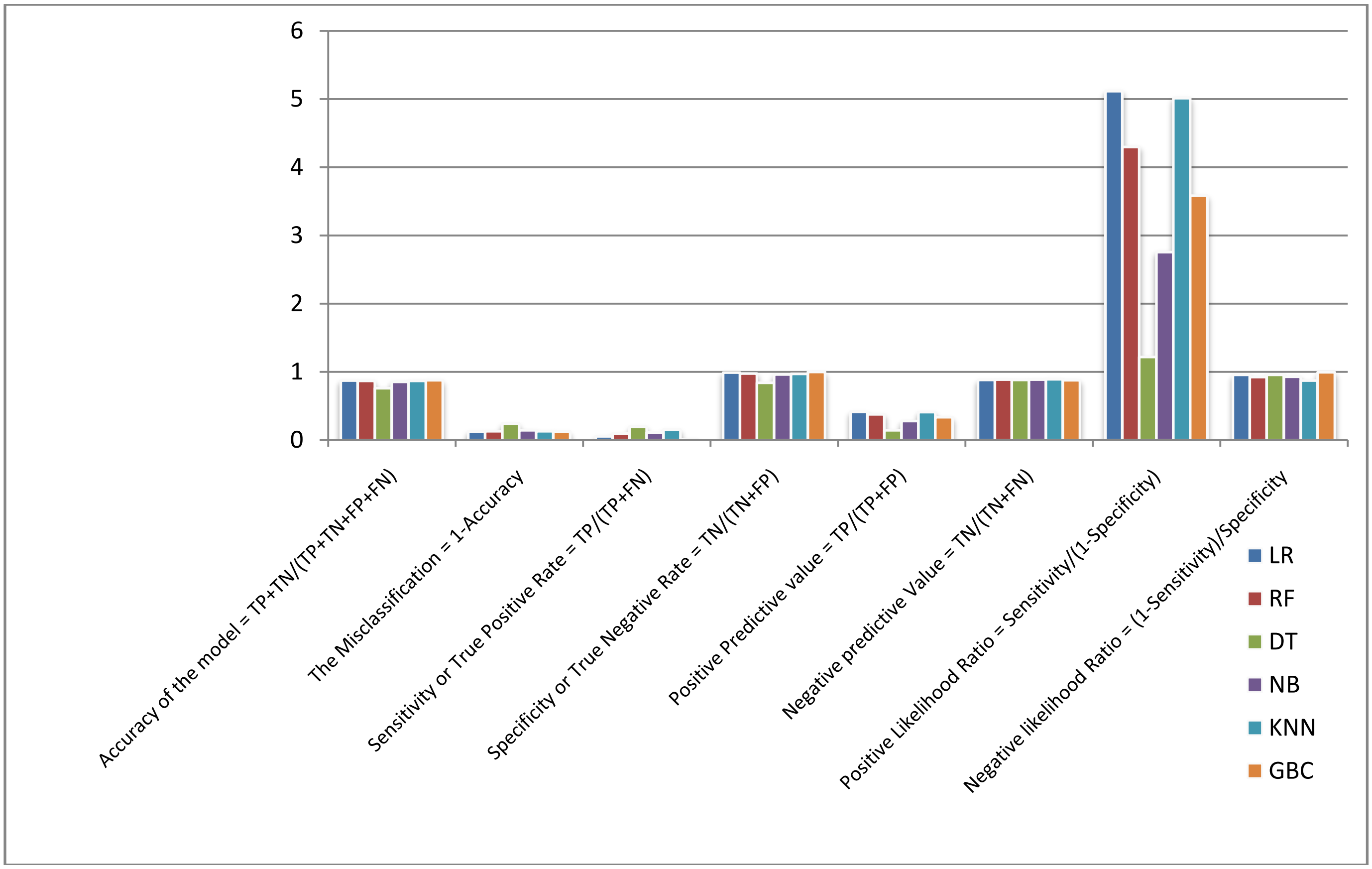


Figure 2. Graphical presentation of model statistics

Threshold Values Prediction (0.5)

In Table 6, the threshold value (0.5) calculated by the classifier to predict whether the patient has a heart disease.

Table 6. Prediction of probability of 10-year CHD at threshold value 0.5

	LR		**RF**		**DT**		**NB**		**KNN**		**GBC**	
	Prob of no heart disease (0)	**Prob of Heart Disease (1)**	**Prob of no heart disease (0)**	**Prob of Heart Disease (1)**	**Prob of no heart disease (0)**	**Prob of Heart Disease (1)**	**Prob of no heart disease (0)**	**Prob of Heart Disease (1)**	**Prob of no heart disease (0)**	**Prob of Heart Disease (1)**	**Prob of no heart disease (0)**	**Prob of Heart Disease (1)**
0	0.87	0.12	0.94	0.06	1.00	0.00	0.96	0.03	0.80	0.20	0.90	0.09
1	0.95	0.04	0.85	0.15	1.00	0.00	0.90	0.09	1.00	0.00	0.95	0.04
2	0.78	0.21	0.84	0.16	1.00	0.00	0.90	0.09	1.00	0.00	0.84	0.15
3	0.80	0.19	0.92	0.08	1.00	0.00	0.90	0.09	0.80	0.20	0.83	0.16
4	0.89	0.10	0.92	0.08	1.00	0.00	0.97	0.02	1.00	0.00	0.88	0.11

Figure 3 shows the confusion matrix of various classifiers. The values of TP, TN, FP and FN have usual meanings when predicting coronary heart disease in 10 years.

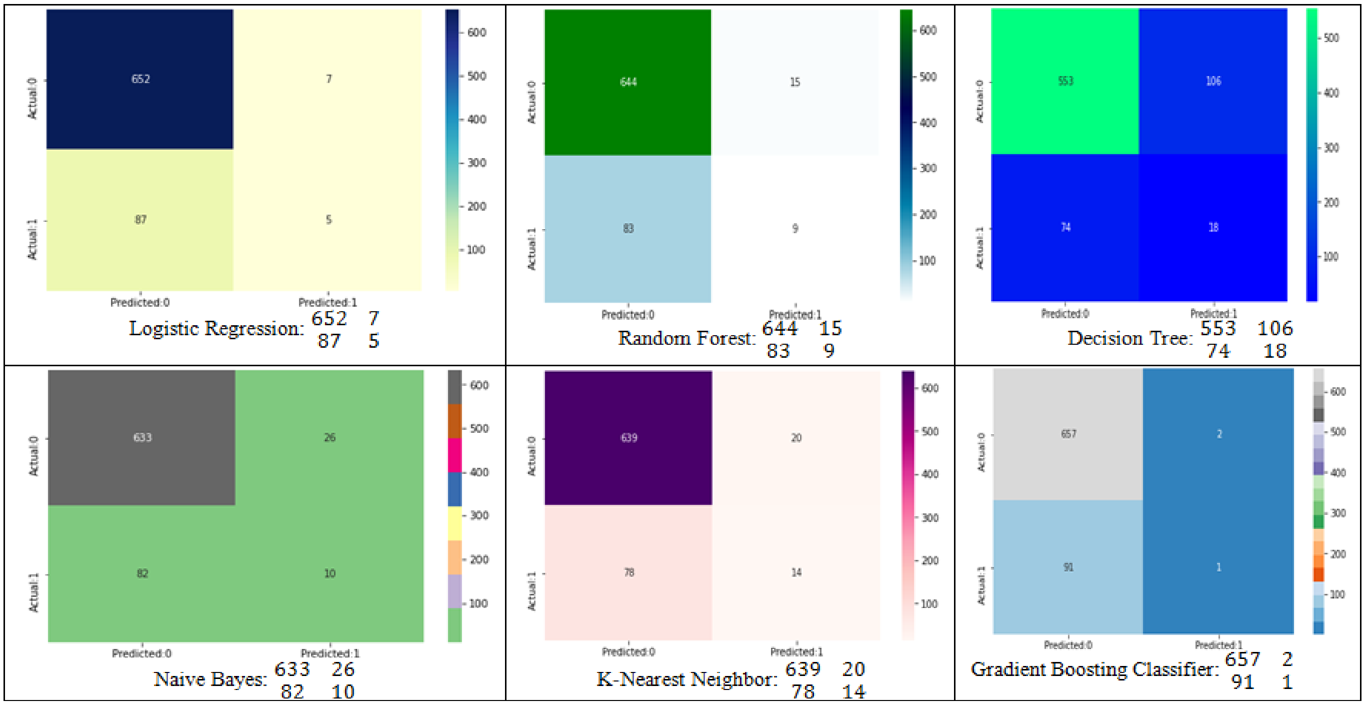


Figure 3. Confusion Matrix of Classifiers

In Figure 3, the classifiers have different AUC values and their corresponding ROCs.

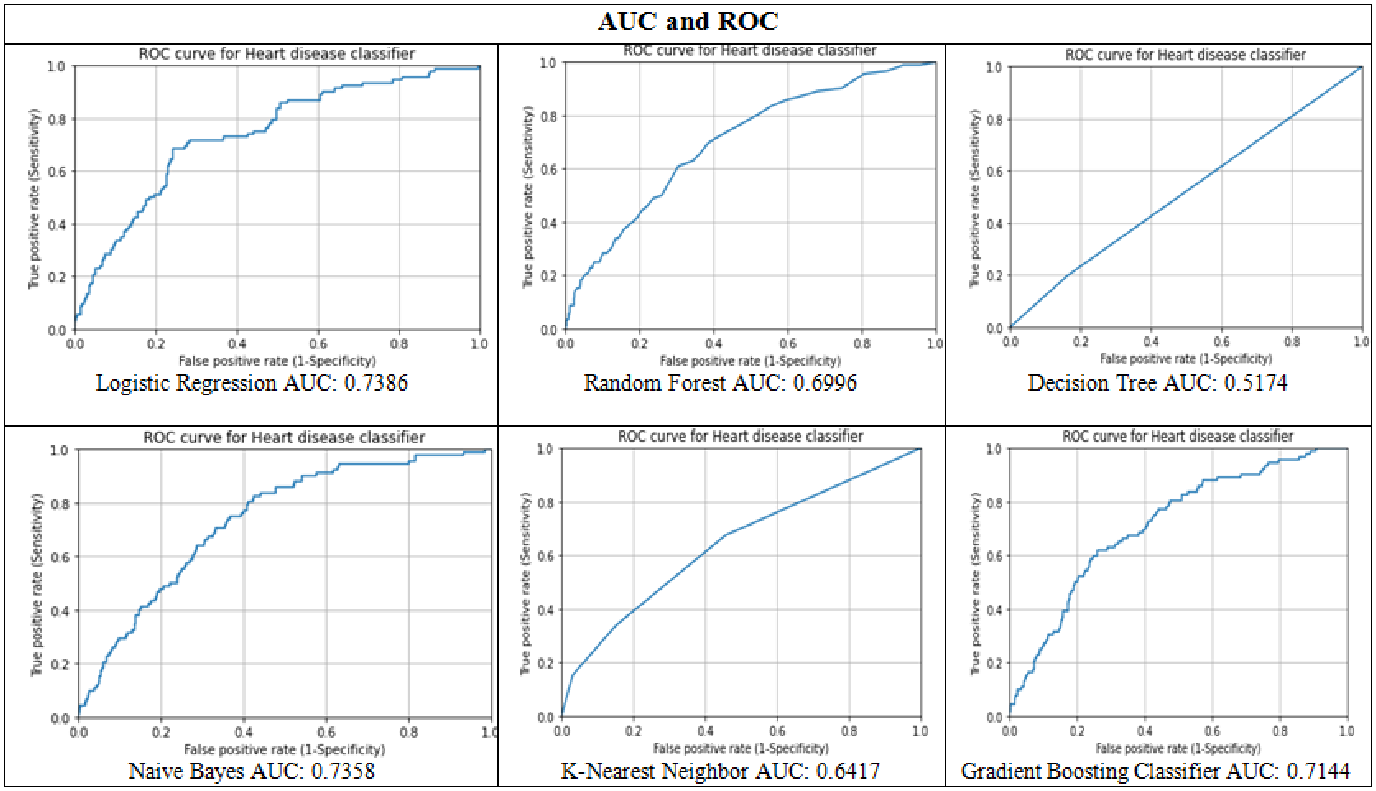


Figure 4. ROC and AUC of the Classifier

5. Discussion

LR is a relapse survey under measurement, which is used to expect a clear outcome of variables from a bunch of indicators or autonomous factors. In the calculated LR, the relevant variables are always parallel. Logistic regression is mostly used to envision and further compute the chance of finish (Barton and Miller 2015, Zinbarg et al. 2010, Stevenson et al. 2019). The outcomes in table 2 show section of properties with P-value superior than the favored alpha (= 5%) and in this way representing little measurably critical association with the likelihood of coronary illness.

This fitted model (Table 4) shows that, holding all various features consistent, the odds of getting resolved to have coronary ailment for people groups (sex = 1)over that of females (sex = 0) is 1.788687. With respect to transform, we can say that the odds for people groups are 78.8% higher than the odds for females.

The coefficient for age says that, holding all others predictable, we will see 7% additions in the odds of getting resolved to have CDH for a one year increase in age since 1.067644.

Furthermore, with each extra cigarette one smokes there is a 2% extension in the odds of CDH.

For Total cholesterol level and glucose level there is no basic change.

There is a 1.7% development in possibilities for every unit increase in systolic Blood Pressure.

Out of 15 features (Table 5), we have selected only six features by backward elimination P-value based features for analysis (Maldonado et al. 2014). Measurable investigation of the information was performed and unmistakable measurements were resolved for segment and illness explicit factors.

Since the model predicts heart disease, too many Type II errors are not suitable. In the current situation (Table 6), false negatives are more dangerous than false positives. Therefore, in order to expand the influence, the threshold can be lowered.

A run of the mill strategy to picture the trade offs of different thresholds is by using a ROC (Figure 3), a plot of the certified positive rate versus the false positive rate for all likely determinations of thresholds. Model with extraordinary portrayal accuracy should have in a general sense more apparent positives than false positives at all limits.

The ideal circumstance for roc curve is towards the upper left corner where the specificity and sensitivity are at ideal levels.

The territory under the ROC measures model characterization precision; the higher the region, the more noteworthy the dissimilarity among valid and false positives, and the more grounded the model in grouping individuals from the preparation dataset. A territory of 0.5 compares to a model that plays out no in a way that is better than arbitrary grouping and a decent classifier remains as distant from that as could reasonably be expected (Figure 4). The closer AUC is to 1, the better.

We compare the results with the earlier studies in Table 7. Our method achieves better results by using feature selection (p-value) techniques and six machine learning methods.

Table 7. Comparison of accuracy and number of features we found earlier

Author	Method	Accuracy	Features
Our Finding	Feature Selection (p-value) and LR, RF, DT, NB, KNN and GBC	87.61%	6
(Latha and Jeeva 2019)	Majority vote with NB, BN, RF and MP	85.48%	9
(Sarangam 2018)	Naive Bayes	83.70%	13
(Brisimi et al. 2018)	Random Forest	81.62%	212
(Pouriyeh et al. 2017)	SVM and MLP	84.15%	14
(Miao et al., 2016)	Adaptative Boosting	80.14%	29

In view of these results, our model beat those results in this article. The important thing is that experts can only handle three or more times instead of a given number of features, and can complete results compared to full features. Our strategy can help reduce meaningless features and increase the amount of information.

6. Conclusion

In this report, we propose a straightforward technique to survey the 10-year danger of hard CVD, which relies upon the danger factors assessed routinely during clinic visits. The result depends on more than 10 years of comprehensive development and determining the occurrence and passage of CVD. Our calculation takes into account the assessment of risk factors, which include uninterrupted and unmitigated risk factors. It also represents a competitive risk of non-cardiovascular death.

Our method is based on p-value based statistical feature selection and six ML classifiers. Table 5 is a performance table, in which GBC's performance is better than other classifiers. The performance of the classifier is measured by confusion matrix, ROC and AUC. The 10-year heart disease data set estimates the patient's future heart disease, so the threshold prediction is calculated as p = 0.5 in table 6.

The accompanying end has been assessed by this exploration are:

- All features chose after the end cycle demonstrate P-values inferior than 5% and accordingly proposing critical function in the Heart illness expectation.
- Men seem, to be more vulnerable to coronary ailment than women. Expansion in Age, number of cigarettes smoked each day and systolic Blood Pressure in like manner show growing odds of having coronary disease.
- Cholesterol shows no gigantic change in the odds of CHD. This could be a result of the presence of 'good cholesterol (HDL) in the absolute cholesterol reading. Glucose also causes a completely immaterial change in possibilities (0.2%).
- The model anticipated with 87.61% exactness by GBC. The specificity of the model is more sensitive.
- The Area under the ROC curve is 73.86 which are genuinely agreeable.
- Generally model could be improved with more data.

References

7. Adebayo AO, Chaubey MS. (2019). Data mining classification techniques on the analysis of student’s performance. GSJ.7(4):45-52.
8. Aggrawal R, Pal S. (2020). Sequential Feature Selection and Machine Learning Algorithm-Based Patient’s Death Events Prediction and Diagnosis in Heart Disease. SN Computer Science. 1(6):1-6.
9. Aggrawal R, Pal S. (2021). Multi-Machine Learning Binary Classification, Feature Selection and Comparison Technique for Predicting Death Events Related to Heart Disease. International Journal of Pharmaceutical Research, 13(1).
10. Balu R, McCracken L, Lancaster E, Graus F, Dalmau J, Titulaer MJ. (2019). A score that predicts 1-year functional status in patients with anti-NMDA receptor encephalitis. Neurology. 15;92(3):e244-52.
11. Barton YA, Miller L. (2015). Spirituality and positive psychology go hand in hand: An investigation of multiple empirically derived profiles and related protective benefits. Journal of religion and health. 54(3):829-43.
12. Besse JP, Coquery M, Lopes C, Chaumot A, Budzinski H, Labadie P, Geffard O. (2013). Caged Gammarus fossarum (Crustacea) as a robust tool for the characterization of bioavailable contamination levels in continental waters: towards the determination of threshold values. Water research. 1;47(2):650-60.
13. Brisimi TS, Xu T, Wang T, Dai W, Adams WG, Paschalidis IC. (2018). Predicting chronic disease hospitalizations from electronic health records: an interpretable classification approach. Proceedings of the IEEE. 6;106(4):690-707.
14. Chaurasia V, Pal S, Tiwari BB.(2018). Chronic kidney disease: a predictive model using decision tree. International Journal of Engineering Research and Technology.
15. Chaurasia, V., & Pal, S. (2021). Stacking-Based Ensemble Framework and Feature Selection Technique for the Detection of Breast Cancer. SN Computer Science, 2(2), 1-13.
16. Chaurasia V, Pal S. (2020). Machine learning algorithms using binary classification and multi model ensemble techniques for skin diseases prediction. International Journal of Biomedical Engineering and Technology. 34(1):57-74.
17. Chaurasia, V., Pandey, M. K., & Pal, S. (2021, March). Prediction of Presence of Breast Cancer Disease in the Patient using Machine Learning Algorithms and SFS. In IOP Conference Series: Materials Science and Engineering (Vol. 1099, No. 1, p. 012003). IOP Publishing.
18. Cole SA. (2004). More than zero: Accounting for error in latent fingerprint identification. J. Crim. l. & Criminology. 95:985.
19. Den Ruijter HM, Peters SA, Anderson TJ, Britton AR, Dekker JM, Eijkemans MJ, Engström G, Evans GW, De Graaf J, Grobbee DE, Hedblad B. (2012). Common carotid intima-media thickness measurements in cardiovascular risk prediction: a meta-analysis. Jama. 308(8):796-803.

20. Doody RS, Pavlik V, Massman P, Rountree S, Darby E, Chan W. (2010). Predicting progression of Alzheimer's disease. Alzheimer's research & therapy. 2(1):1-9.
21. Dudek G. (2015). Short-term load forecasting using random forests. In Intelligent System’s 2014 (pp. 821-828). Springer, Cham.
22. Einarson TR, Acs A, Ludwig C, Panton UH. (2018). Prevalence of cardiovascular disease in type 2 diabetes: a systematic literature review of scientific evidence from across the world in 2007–2017. Cardiovascular diabetology. 17(1):1-9.
23. Gao S, Tibiche C, Zou J, Zaman N, Trifiro M, O’Connor-McCourt M, Wang E. (2016). Identification and construction of combinatory cancer hallmark–based gene signature sets to predict recurrence and chemotherapy benefit in stage II colorectal cancer. JAMA oncology. 2(1):37-45.
24. Gao W, Wang W. (2018). Analysis of k-partite ranking algorithm in area under the receiver operating characteristic curve criterion. International Journal of Computer Mathematics. 95(8):1527-47.
25. Harrell Jr FE. (2015). Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer.
26. Hojati AT, Ferreira L, Washington S, Charles P, Shobeirinejad A. (2016). Reprint of: Modelling the impact of traffic incidents on travel time reliability. Transportation research part C: emerging technologies. 1;70:86-97.
27. Janket SJ, Baird AE, Chuang SK, Jones JA. (2003). Meta-analysis of periodontal disease and risk of coronary heart disease and stroke. Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology. 95(5):559-69.
28. Kakulapati V, Kirti A, Kulkarni V, Raj CP. (2017). Predictive Analysis of Heart Disease using Stochas-tic Gradient Boosting along with Recursive Feature Elimination. 6(5):909-912.
29. Kannel WB, Feinleib M, McNamara PM, Garrison RJ, Castelli WP. (1979). An investigation of coronary heart disease in families: the Framingham Offspring Study. American journal of epidemiology. 110(3):281-90.
30. Khourdifi Y, Bahaj M. (2019). Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization. International Journal of Intelligent Engineering & Systems. 12(1):242-52.
31. Krawczyk B. (2017). Active and adaptive ensemble learning for online activity recognition from data streams. Knowledge-Based Systems. 138:69-78.
32. Latha CB, Jeeva SC. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked. 16:100203.
33. Long NC, Meesad P, Unger H. (2015). A highly accurate firefly based algorithm for heart disease prediction. Expert Systems with Applications. 42(21):8221-31.
34. Maldonado S, Weber R, Famili F. (2014). Feature selection for high-dimensional class-imbalanced data sets using support vector machines. Information sciences. 286:228-46.
35. Miao KH, Miao JH, Miao GJ. (2016). Diagnosing coronary heart disease using ensemble machine learning. Int J Adv Comput Sci Appl (IJACSA).
36. Miyamoto K, Setsuie R, Osada T, Miyashita Y. (2018). Reversible silencing of the frontopolar cortex selectively impairs metacognitive judgment on non-experience in primates. Neuron. 97(4):980-9.
37. Narasimhan B, Malathi A. (2019). Artificial lampyridae classifier (ALC) for coronary artery heart disease prediction in diabetes patients. International Journal of Advance Research, Ideas and Innovations in Technology. 5(2):683-9.
38. Noyce AJ, R'Bibo L, Peress L, Bestwick JP, Adams-Carr KL, Mencacci NE, Hawkes CH, Masters JM, Wood N, Hardy J, Giovannoni G. (2017). PREDICT-PD: An online approach to prospectively identify risk indicators of Parkinson's disease. Movement Disorders. 32(2):219-26.
39. Park H. (2013). An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain. Journal of Korean Academy of Nursing. 43(2):154-64.
40. Pouriyeh S, Vahid S, Sannino G, De Pietro G, Arabnia H, Gutierrez J. (2017). A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In2017 IEEE symposium on computers and communications (ISCC) ; (pp. 204-207). IEEE.
41. Proust-Lima C, Dartigues JF, Jacqmin-Gadda H. (2016). Joint modeling of repeated multivariate cognitive measures and competing risks of dementia and death: a latent process and latent class approach. Statistics in medicine. 35(3):382-98.

42. Reddy NS, Nee SS, Min LZ, Ying CX. (2019). Classification and feature selection approaches by machine learning techniques: Heart disease prediction. International Journal of Innovative Computing. 9(1).
43. Sarangam Kodati DR. (2018). Analysis of heart disease using in data mining tools Orange and Weka. Global Journal of Computer Science and Technology.
44. Saxena K, Sharma R. (2016). Efficient heart disease prediction system. Procedia Computer Science. 1;85:962-9.
45. Singh M, Spertus JA, Gharacholou SM, Arora RC, Widmer RJ, Kanwar A, Sanjanwala RM, Welle GA, Al-Hijji MA. (2020). Comprehensive geriatric assessment in the management of older patients with cardiovascular disease. InMayo Clinic Proceedings (Vol. 95, No. 6, pp. 1231-1252). Elsevier.
46. Smith NJ, Levy R. (2013). The effect of word predictability on reading time is logarithmic. Cognition. 128(3):302-19.
47. Sofi F, Macchi C, Abbate R, Gensini GF, Casini A. (2014). Mediterranean diet and health status: an updated meta-analysis and a proposal for a literature-based adherence score. Public health nutrition. 17(12):2769-82.
48. Stamate D, Alghamdi W, Ogg J, Hoile R, Murtagh F. (2018). A Machine Learning Framework for Predicting Dementia and Mild Cognitive Impairment. In2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 671-678). IEEE.
49. Stevenson JC, Millings A, Emerson LM. (2019). Psychological well-being and coping: The predictive value of adult attachment, dispositional mindfulness, and emotion regulation. Mindfulness. 10(2):256-71.
50. Suguna R, Shyamala Devi M, Bagate RA, Joshi AS. (2019). Assessment of feature selection for student academic performance through machine learning classification. Journal of Statistics and Management Systems. 22(4):729-39.
51. World Health Organization. (2012). Effect of increased potassium intake on cardiovascular disease, coronary heart disease and stroke.
52. Ye L, Gao L, Marcos-Martinez R, Mallants D, Bryan BA. (2019). Projecting Australia's forest cover dynamics and exploring influential factors using deep learning. Environmental Modelling & Software. 119:407-17.
53. Zinbarg RE, Mineka S, Craske MG, Griffith JW, Sutton J, Rose RD, Nazarian M, Mor N, Waters AM. (2010). The Northwestern-UCLA youth emotion project: Associations of cognitive vulnerabilities, neuroticism and gender with past diagnoses of emotional disorders in adolescents. Behaviour research and therapy. 48(5):347-58.