IJHAS
International Journal of Health and Allied Sciences
Official Publication of JSS Academy of Higher Education and Research (JSS AHER), Mysuru

Manuscript 1045

Statistical Machine Learning Algorithm for Predicting the Risk Factors in Heart Disease

Chaithra N

Shalini H. Doreswamy

Pallavi N

ORIGINAL STUDY

Statistical Machine Learning Algorithm for Predicting the Risk Factors in Heart Disease

Chaithra Nagaraju [a,*], Pallavi Nagaraju [b], Shalini Doreswamy [c]

[a] Division of Medical Statistics, School of Life Sciences, JSS Academy of Higher Education & Research (JSS AHER), Mysuru 570015, Karnataka, India
[b] School of Life Sciences and Natural Sciences, JSS Academy of Higher Education & Research (JSS AHER), Mysuru 570015, Karnataka, India
[c] Center of Excellence in Molecular Biology and Regenerative Medicine (CEMR), Department of Biochemistry, JSS Medical College, JSS Academy of Higher Education & Research (JSS AHER), Mysuru 570015, Karnataka, India

Abstract

Heart disease is one of the major non-communicable disease and leading cause of mortality in the world. According to WHO heart disease is taking about nearly 17.9 million lives of people each year. Its mortality forecasts indicate a rise in global annual deaths to 20.5 million in 2020 and as high as 24.2 million by 2030. Risk factors are one of the most powerful predictors of heart disease. The study includes modified and non-modified risk factors that contribute to the disease such as Age, Gender, Family history, Hypertension, Diabetics, Obesity, Blood Pressure, Smoking, Alcohol intake, Exercise and Heart rate. Machine learning is one of the most useful techniques that can help researchers, entrepreneurs, and individuals for extracting valuable information from sets of data. The objective of this study is to highlight the utility and application of machine learning techniques for the prediction of heart disease to facilitate experts in the healthcare domain. A total of 336 patients were examined and their personal and medical data were collected in JSS hospital. This prospective study was consisting of 55% patients are free from the heart disease and 45% have heart disease. From the result, it has been determined that males are more likely to develop the heart diseases than females and very common in elderly persons. The accuracy of the Naïve Bayes model is found to be 94%, Obesity plays a vital role in getting the disease followed by hypertension, alcohol intake, smoking, exercise and age has more impact on developing the heart disease.

Keywords: Heart disease, Risk factors, Odd ratio, Naïve Bayes algorithm

1. Introduction

Heart disease is the main cause of death in the world [1]. According to WHO heart disease is taking about nearly 17.9 million lives of people each year. It has also reported that 4 out of 5 heart disease deaths are due to strokes and heart attack while 1/3rd of these deaths occurs in the people who are below 70 years of age [2]. Its mortality forecasts indicate a rise in global annual deaths to 20.5 million in 2020 and as high as 24.2 million by 2030. They constitute 31.5% and 32.5% of all global deaths respectively, for males it has been predicted that the percentage of male dying due to CHD rises from 13.1% in 2010 to 14.9% in 2030 while it drops for the female from 13.6% in 2010 to 13.1% in 2030. Stroke deaths rise from 9.2% to 10.4% for males and 11.5%−11.8% for females [3].

The 'heart disease' refers to disease of the heart and the fluctuating functions of the blood vessels in it [4]. Heart disease is often used interchangeably with the term cardiovascular disease (CVD). A major change has been made in the emphasis from communicable diseases to a new epidemic of Non-Communicable Diseases (NCDs) and their side effects. Cardiovascular Disease (CVD), Cancer, and

Received 16 February 2023; accepted 15 July 2023.
Available online 28 September 2023

* Corresponding author at: Division of Medical Statistics, School of Life Sciences, JSS Academy of Higher Education & Research (JSS AHER), Mysuru 570015, Karnataka, India.
E-mail address: chaithra.mstats@jssuni.edu.in (C. Nagaraju).

https://doi.org/10.55691/2278-344X.1045

Diabetes Mellitus (DM) have been the leading causes of morbidity and mortality worldwide and more than 75% of CVD deaths occur in low- and middle-income countries [5]. It has found that CVD affected Indians at least a decade prior and in their most active middle age years compared with the people of European ancestry [6].

Risk factors are one of the most powerful predictors of heart disease [1]. Although metropolitan areas in India have seen the fastest growth in cardiovascular illness, rural areas have also seen a significant increase in vascular fatalities. It seems reasonable to argue that people with changing lifestyles due to growing urbanization are associated with adverse CVD risk factors irrespective of their habitat [7]. Non modifiable and modifiable risk factors can increase the probability of developing CVD. The primary prevention strategy, which is focused on the early identification of cardiovascular risk factors in the general population, is the most efficient and effective technique to reduce the incidence and prevalence of CVD and the subsequent correction of risk factors namely physical inactivity, unhealthy diet, harmful effects of tobacco and alcohol and other habit forming are interfered with all of the disease related modifiable risk factors like obesity, hypertension, diabetes and high cholesterol [5].

Machine learning is the process of building a computer system that can autonomously gather data and combine that data to develop knowledge. It enables to discover new and interesting structures and formats about a set of data that are previously unknown. Machine learning is one of the most useful techniques that can help researchers, entrepreneurs, and individuals for extracting valuable information from sets of data [8]. It refers to the act of searching for huge information stores automatically, to identify patterns and trends which go beyond simple analytical procedures (thesis). The objective of this study is to highlight the utility and application of machine learning techniques for the prediction of heart disease to facilitate experts in the healthcare domain [9].

2. Methodology

This was a prospective descriptive study conducted at Department of cardiology, JSS hospital. A total of 336 patients were examined and their personal and medical data were collected. Simple random sampling technique was applied to gather information about gender, age, locality, education level, occupation, marital status. Health Insurance, SES. There are several factors affecting the disease. This study includes modified and non-modified risk factors that contribute to the disease. The risk factors of the study are Hypertension, Diabetics, Obesity, Blood Pressure, Smoking, Alcohol intake, Family history, Exercise and Heart rate which were analysed by using statistical and machine learning techniques. The dependent variable "Diagnosis" was identified as a predicted attribute with a value is equal to "1" for patients suffering from heart disease and value equal to "0" for patients not suffering from IHD. The work flow of prediction model is shown in Fig. 1. From the analysis using descriptive statistics it concludes that the data is normally distributed.

2.1. Supervised machine learning techniques

Machine learning is a vast area, there are several ways to implement machine learning techniques, however, the most commonly used ones are

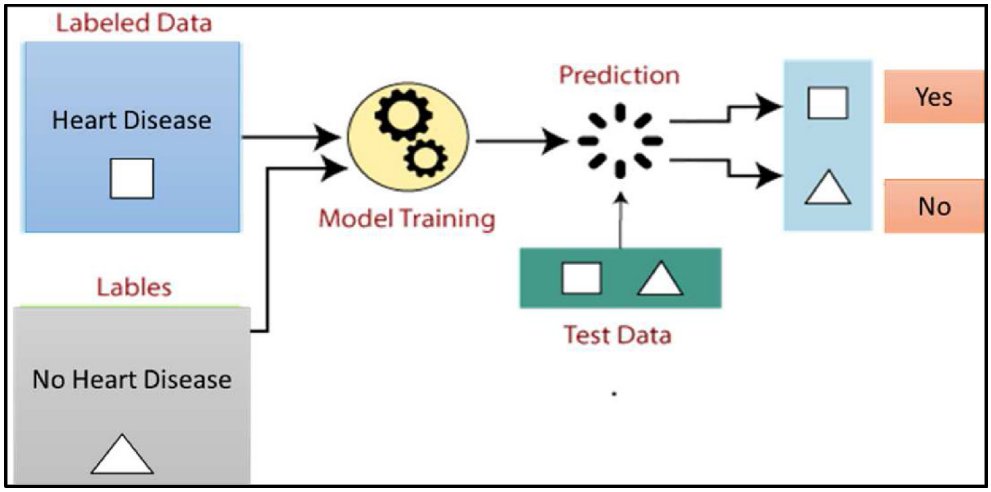

Fig. 1. Working flow of supervised Machine learning Models.

supervised and unsupervised learning. Supervised learning deals with learning a function from available training data. It examines the training data and generates an inferred function that can be applied to new examples. Logistic Regression, Decision Tree, Random Forest, Neural Networks and Naive Bayes are the popular supervised learning algorithms [10,11].

2.2. Odds ratio (OR)

The odds ratio can be used to assess if a specific exposure poses a risk for a specific outcome and to compare the relative importance of several risk variables for that outcome. It can also be applied as such to express the strength of the association between test results and sickness which is calculated by using Odds Ratio (OR) represented in Table 1. The range of OR is from 0 to infinity. When the value of OR increases or decreases from 1, the relation grows stronger and stronger. Further, the 95% confidence interval (CI) is used to measure the accuracy of the OR. A large CI indicates a low level of OR accuracy whereas a small CI indicates a higher OR accuracy [12].

$$\mathrm{OR} = \frac{\textit{Odds of the event in the exposed group}}{\textit{Odds of the event in the unexposed group}}$$

$$\mathrm{OR} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

95% CI = exp {ln (OR) ± 1.96 × SE (ln (OR))} Where, SE the standard error is given by, $\mathrm{SE} = \sqrt{(1/a) + (1/b) + (1/c) + (1/d)}$.

OR = 1 Exposure does not affect odds of outcome; OR>1 Exposure associated with higher odds of outcome; OR<1 Exposure associated with lower odds of outcome [13].

2.3. Naïve Bayes

The Bayesian Classification represents both a supervised method of learning and a statistical method of classification [17]. A Naive Bayes classifier is a simple probabilistic classifier based on applying theorem of Bayes with strong assumptions of independence [18]. The classifiers of Naive Bayes presume that the influence of a variable value on a particular class doesn't depend on the values of another variable. This presumption is referred to as conditional independence. Naive Bayes or Bayes Rule is the baseline for many methods of machine learning and data mining. In simple terms, a Naive Bayes classifier assumes that the presence (or absence) of a specific characteristic of a group is unrelated to the presence (or absence) of any other characteristic [19]. It has a power to solve the complex problems of the real world. Naïve Bayes is easy to construct, without any iterative complex estimation of parameter, making it explicitly helpful in the field of clinical science to diagnose cardiac patients [20,21].

Bayes theorem provides a way of calculating the posterior probability $P(C/X)$, from $P(C)$, $P(X)$, and $P(X/C)$.

$$\mathrm{P}(C/X) = \frac{\mathrm{P}(X/C) * \mathrm{P}(C)}{\mathrm{P}(X)}$$

where, P(C) and P(X) are probability of events C and X.

P(C) is the prior probability of class.

P(X) is the prior probability of predictor.

P(C/X) is the posterior probability of C (class) conditional on X (target).

P (X/C) is the likelihood which is the probability of predictor given class.

The technique was derived from the work of the 18th century mathematician Thomas Bayes, who evolved the basic principles to describe the probability of events. These principles laid the foundation for what is now identified as Bayesian methods [14–16]. by applying the Bayes' theorem to the class-specific conditional probabilities $P(Y = C_k \setminus X = x)$, thus we have.

$$P(Y = C_k \setminus X = x) = \frac{P(Y = C_k)P(X = x \setminus Y = C_k)}{P(X = x)}$$

$$P(Y = C_k \setminus X = x) = \frac{P(Y = C_k)\prod_{i=1}^{d} P(X_i = x_i \setminus Y = C_k)}{P(X_1 = x_1, X_2 = x_{2,\ldots,} X_d = x_d)}$$

$$\widehat{y} = g(x) = arg_{y\in[0,1]}\, max\widehat{P}(Y \setminus X) = arg_{y\in[0,1]}\, max\widehat{P}$$

$$(Y)\widehat{P}(X \setminus Y) = arg_{y\in[0,1]}\, max\widehat{P}(Y = y)\prod_{j=1}^{d}$$

$$\widehat{P}\left(X_j = x_j \setminus Y = y\right)$$

The difference between the expected and actual success rates is represented by the confusion matrix (Table 2) (see Table 1). Similar to that, it contrasts the expected and actual numbers of failures.

- Sensitivity (True Positive Rate)

The percentage of cases that were accurately categorised as positive, or those that were projected to be successful and actually turned out to be successful

Table 1. Calculating odds ratio (OR)

		Event	
		Yes	No
Exposure	Yes	A	B
	No	C	D

Table 2. Confusion matrix

Predicted Values	Actual Values	
	Positive	Negative
Positive	TP	FP
Negative	FN	TN

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

- Specificity (True Negative Rate)

The specificity also mention as true negative rate, is used to assess the rate of actual negatives that are accurately identified.

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

- Positive Predicted Value (PPV)

The positive predictive value (PPV) is the probability that patients who have a positive test result will actually have the disease. It is commonly used in medical tests where the "positive" result means that actually have a disease.

$$\text{PPV} = \frac{TP}{TP + FP}$$

- Negative Predicted Value (NPV)

The negative predictive value is the likelihood that people who get a negative test result do not actually have the disease. It is also known as negative predictive agreement.

$$\text{NPV} = \frac{TN}{TN + FN}$$

- Kappa (k):

Kappa Statistics is a standard benchmark, used specifically to compare the observed accuracy values with the excepted accuracy values; this statistical test can be conducted on both single classifiers and multiple classifiers.

$$\text{k} = \frac{TA - RA}{1 - RA}$$

where, TA = Total Accuracy, RA = Random Accuracy.

2.4. Statistical analysis

Demographic data and frequency of risk factors are reported by giving their percentages. Cross tabulation and Pearson chi-square test were used to identify the risk factors that affect heart disease at 5% level of significance. Independent Sample t-test was measuring the difference between the means of two groups including males and females with respect to clinical parameters. Odd ratios (OR) between heart disease and its risk factors. An OR < 1 indicates decreased event occurrence, OR >1 indicates increased event occurrence. Naïve Bayes was used to predict by keeping the disease as dependent variable and other factors such as Age, Gender, Family history, Hypertension, Diabetics, Obesity, Smoking, Alcohol intake and Exercise were taken as independent variables. Data was entered and analyzed in the Statistical Package for the Social Sciences (IBM SPSS statistics 22.0) and R software.

3. Results and discussion

The prevalence of heart disease related risk factors has increased in India, as indicated by studies over the last decade and as expected by future estimate projections. The major risks are modifiable and can be prevented, treated and controlled. There are significant health benefits for both men and women of all ages in quitting smoking, minimizing cholesterol and blood pressure, eating a balanced diet and improving physical activity. Deepa Shokeen et al. (2015). From the result of dataset, it has been confirmed that heart disease is very much common in elderly persons and those who are obese.

The Table 3 exhibits the frequencies and percentage of variables consists three risk factors specifically, non-modified risk factors, Disease related Modified risk factors and Life style related Modified risk factors. The dataset contains 336 individuals out of them 147 were having the disease and the remaining 189 were free from diseases. There were 57% male and 43% female are considered in this study, 55% and 45% patients are coming from urban and rural place, 88% of patients are above 39 years of age. On an average person whose age is 61 ± 13 years have high risk of getting the disease. The results show that the heart diseases and risk factors are increasing with a rapid pace in Indian population. Obesity (51%) and Hypertension (32%) are the major risk factors for CVD where as Diabetes, Blood pressure, alcohol, smoking, exercise and Heart rate are the causes contributing to heart disease.

Table 3. Frequency distribution of non-modifiable and modifiable risk factors

Risk Factors		Frequency	Percentage
Non modified risk factor			
Age	19–38	41	12.2
	39–58	153	45.5
	59–78	127	37.8
	79+	15	4.5
Gender	Male	191	56.8
	Female	145	43.2
Locality	Urban	184	54.8
	Rural	152	45.2
Family history	Yes	39	11.6
	No	297	88.4
Disease related - Modifiable Risk factor			
Hypertension	Yes	106	31.5
	No	230	68.5
Diabetes	Yes	89	26.5
	No	247	73.5
Obesity	Yes	172	51.2
	No	164	48.8
Blood pressure	High	78	23.2
	Normal	254	75.6
	Low	4	1.2
Life style related - Modifiable risk factor			
Smoking	Current	34	10.1
	Never	254	75.6
	Past	48	14.3
Alcohol	Current	32	9.5
	Never	263	78.3
	Past	41	12.2
Exercise	Regular	79	23.5
	Moderate	12	3.6
	Never	242	72
	Past	3	9
Heart rate	Normal	307	91.4
	T disease	20	6
	B disease	9	2.7

The Table 4 shows the analysis of odds ratio between the various risk factors and the disease. From the result, the odds ratio of getting the disease among female is 0.51, since the ratio is less than 1, hence we can say that males are more likely to get the disease compared to females. Out of 106 Hypertensive people it is seen that 53 of them have the disease and 94 are free from disease, and thus the odds of being diseased is 1.447 times higher in hypertensive people than non-hypertensive. Also, it can be observed that the number of persons who have cardiac disease is higher in obese group that is 88. However, the odds in favour of obese person developing disease is 1.864 times higher than to those who are not obese. Following, the odds ratio of obesity is more and it can be stated that obesity plays a significant role in getting the disease. By examining smoking variable, out of 82 members 43 have the disease and 39 are disease free, hence estimated odds of having disease is 1.590 times for smokers than non-smokers. Alcohol Intake and Exercise are the protective factors against the disease, which indicates an odds ratio of less than 1.0. Out of 39 people it is found that 21 are getting the illness due to family history and the odds ratio is found to be 1.583 which is greater than 1 and it implies that genetic factor has importance to heart disease.

The A priori probability and conditional probabilities are calculated using the training dataset shows in Tables 5 and 6. By A priori probability shows that 55% of the patients are free from the heart disease and 45% have heart disease. From the conditional probabilities result, the chance of male to have or develop an illness is more when compared to female. Further, with the above findings it is easy to make out that, the Cardio Vascular Disease have less influence of family history. The chance of getting the heart disease among Hypertensive, Diabetic, Current Smoker and Current Alcoholic people is 0.365, 0.307, 0.144, 0.105 respectively and the model shows that for non-hypertensive, non − diabetic, non-current smoker the probability of

Table 4. Odds Ratio analysis of Risk Factors in patients for the study subject

Factors		Heart Disease			Odds Ratio (OR)	95% Confidence Interval	
		Yes	No	Total		Lower	Upper
Gender	Female	50	95	145	0.51	0.327	0.796
	Male	95	94	191			
Hypertension	Yes	53	94	106	1.447	0.911	2.298
	No	53	136	230			
Diabetics	Yes	44	45	89	1.367	0.84	2.223
	No	103	144	247			
Obesity	Yes	88	84	172	1.864	1.204	2.887
	No	59	105	164			
Smoking	Yes	43	39	82	1.59	0.958	4.045
	No	104	150	254			
Alcohol Intake	Yes	30	43	73	0.871	0.414	1.821
	No	117	146	263			
Family History	Yes	21	18	39	1.583	0.81	3.095
	No	126	171	297			
Exercise	Yes	39	52	91	0.951	0.585	1.546
	No	108	137	245			

Table 5. A priori Probability

Disease	
No	Yes
0.55	0.45

Table 6. Conditional probability

Risk Factors		Disease	
		No	Yes
Gender	Female	0.51	0.36
	Male	0.49	0.64
Family History	No	0.88	0.85
	Yes	0.11	0.15
Hypertension	No	0.74	0.64
	Yes	0.26	0.36
Diabetics	No	0.75	0.69
	Yes	0.25	0.31
Obesity	No	0.54	0.40
	Yes	0.46	0.59
Smoking	Current	0.06	0.14
	Never	0.808	0.692
	Past	0.131	0.163
Alcohol Intake	Current	0.085	0.106
	Never	0.792	0.828
	Past	0.123	0.106
Heart Rate	B-Disease	0.015	0.067
	Normal	0.908	0.865
	T-Disease	0.077	0.067
Exercise	Moderate	0.046	0.019
	Never	0.731	0.712
	Past	0.000	0.019
	Regular	0.223	0.250

getting no disease is more that is 0.738, 0.746 and 0.061 Further it is found that the obesity has high impact on heart with 0.596 probability and is considered to be main cause of illness. Also, it has seen that the person who never exercise have the higher risk of getting the disease that is 71% than the one who does regular exercise that is 25%.

The Navie Bayes model correctly classifies 47 cases as 'No' and 27 cases as 'Yes' and 12 were wrongly

Table 7. Confusion matrix

Predicted Value	Actual Value	
	No	Yes
No	47	16
Yes	12	27

Table 8. Prediction performance

Accuracy	0.940
Sensitivity	0.816
Specificity	0.963
Positive Predictive Value	0.838
Negative Predictive Value	0.971
Kappa	0.810

predicted as they had the disease while they were disease free as well as 16 were predicted as they don't have the disease while they were suffering from the disease shown in Table 7. The accuracy of the Naïve Bayes model is found to be 94%. Further, the kappa value of 0.810 tells that there is strong agreement between the predicted and actual values of disease and the model reaches. The Sensitivity (true positive rate) value is 0.816, approximately 82% of diseased person are predicted correctly as getting the disease and the specificity (true negative rate) value of 0.963, this means that about 96% predictions were made of the people in the class 'no' that is, healthy people were correctly predicted healthy. Positive Predictive Value is 0.838 and Negative Predictive Value 0.971 is display in Table 8.

Naïve Bayes classifier shows that the model has a high specificity then the sensitivity. Thus, Age, Gender, Obesity, Hypertension, Smoking, Exercise, Heart rate, Diabetes, family history and Alcohol intake are key factors in determining patients with heart disease and strongest interaction with the response variable shown in Fig. 2.

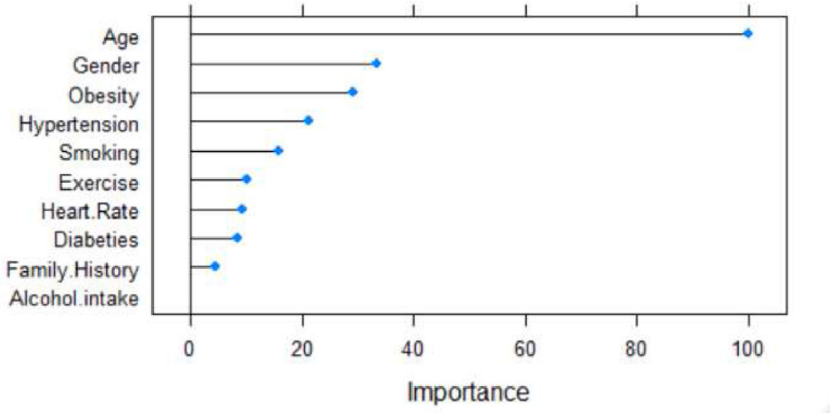

Fig. 2. Graphical representation of disease predictors according to their importance.

4. Conclusion

The results show that the heart diseases and risk factors are increasing with a rapid pace in Indian population. The risk factors are found higher in male population than Female. Obesity, smoking and alcohol intake showed the significant difference with gender. Obesity plays a vital role in getting the disease followed by hypertension, smoking, exercise. Among all the variables, age has more impact on developing the heart disease. The model is clearly more specific than sensitive, as shown by the data above, which means that negative values are predicted more precisely than positive ones. With a greater accuracy rate of 94%, Naïve Bayes model can help cardiologists make an accurate diagnosis of heart disease.

Conflict of Interest

There are no conflicts of interest.

Financial support and sponsorship

No Financial support and sponsorship from university or any other resources.

References

[1] Vembandasamy K, Sasipriya R, Deepa E. Heart diseases detection using Naive Bayes algorithm. Int J Innov Sci Eng Technol 2015;2(9):441−4.
[2] Who int. 2020. Health topics. [online], WHO article, Available at: https://www.who.int/health-topics/. [Accessed 14 August 2022].
[3] Kasliwal RR, Mahansaria K, Bansal M. Cardiovascular risk algorithms and their applicability to Indians. Chapter10. Cardiological Society of India, Elsevier (RELX India Pvt Ltd.). Section II — Preventive Cardiology; p. 75−84.
[4] Shyamala K, Marikani T. *A N*ew Cygnus Optimization Algorithm for Prediction of Cardio Vascular Disease. Int J Innovative Technol Explor Eng 2019;8(12):4351−5.
[5] Patil CR, Thakre SS, Thakre SB. A cross-sectional study on the risk factors for cardiovascular disease and risk profiling of adults in central India. J Clin Preven Cardiol 2017;6(3):104.
[6] Prabhakaran D, Jeemon P, Roy A. Cardiovascular diseases in India: current epidemiology and future directions. Circulation 2016;133(16):1605−20.
[7] Elsayad AM, Fakhr M. Diagnosis of Cardiovascular Diseases with Bayesian Classifiers. J Comput Sci 2015;11(2):274−82.
[8] Roman WP, Martin HD, Sauli E. Cardiovascular diseases in Tanzania: the burden of modifiable and intermediate risk factors. J Xiangya Med 2019;4(33):1−13.
[9] Silwattananusarn T, Tuamsuk K. Data mining and its applications for knowledge management: a literature review from 2007 to 2012. Int J Data Mining Knowledge Manag Process 2012;2(5):3−24.
[10] Lantz B. Machine learning with R: expert techniques for predictive modeling, 3rd Edition. Packt Publishing Limited; 2019 Apr 15.
[11] Jayakiran K, Pranavi M, Novika M, Tejaswini V, Rajesh N. Prediction of Heart Disease in Comparison with Different Machine Learning Algorithm. Int J Innovative Technol Explor Eng 2019;8(6):328−32.
[12] Szumilas M. Explaining odds ratios. J Canadian Acad Child Adolescent Psychiatry 2010;19(3):227.
[13] Kundu Jhumki, Kundu Sampurna. Cardiovascular disease (CVD) and its associated risk factors among older adults in India: Evidence from LASI Wave 1. Clin Epidemiol Global Health 2022;13:1−5.
[14] Medhekar DS, Bote MP, Deshmukh SD. Heart disease prediction system using naive Bayes. Int J Enhan Res Sci Technol Eng 2013;2(3):1−5.
[15] Manjusha KK, Sankaranarayanan K, Seena P. Prediction of different dermatological conditions using naive bayesian classification. Int J Adv Res Comput Sci Software Eng 2014;4(1).
[16] Arun R, Deepa N. Heart Disease Prediction System Using Naïve Bayes. Int J Pure Appl Math 2018;119(16):3053−64.
[17] Jayakiran K, Pranavi M, Novika M, Tejaswini V, Rajesh N. Prediction of Heart Disease In Comparison With Different Machine Learning Algorithm. Int J Innovative Technol Explor Eng 2019;8(6):328−32.
[18] Miranda E, Irwansyah E, Amelga AY, Maribondang MM, Salim M. Detection of cardiovascular disease risk's level for adults using naive Bayes classifier. Healthcare Inform Res 2016;22(3):196−205.
[19] Zhang Z. Naïve Bayes classification in R. Ann Transl Med 2016;4(12):241−5.
[20] Singh G, Bagwe K, Shanbhaga S, Singh S, Devi S. Heart disease prediction using Naïve Bayes. Int Res J Eng Technol 2017;4(3):1−3.
[21] Spino S, Sathik MM, Nisha SS. The prediction of heart disease using naive bayes classifier. Int Res J Eng Technol (IRJET) 2019;6(3):373−7.