Acta Medica Europa

ISSN Online: 2996-1211 doi: 10.5281/zenodo.15365220

Prediction of Heart Disease by Naive Bayesian Classification Algorithm Based on Statistical Modeling

Ayşe Banu Birlik

[1.] Istinye University, Department of Medical Services and Techniques, Istanbul, Türkiye

Article Info

Received: 23 April 2025

Revised: 8 May 2025

Accepted: 8 May 2025

Published: 8 May 2025

Keywords:

Classification, forecasting, heart disease, machine learning, naive bayes.

Corresponding author:

Ayşe Banu Birlik

Istinye University, Department of Medical Services and Techniques, Istanbul, Türkiye

banu.birlik@istinye.edu.tr

ABSTRACT

It is predicted that cardiovascular diseases, which are critical health problems, will continue to be the most important cause of death in the world. Management strategies and diagnosis of the disease in the early stages play an important role in increasing survival. The amount of data collected in the healthcare field has increased significantly with the advancement of technology. Therefore, several research have been conducted to evaluate medical data using various data mining and machine learning approaches. A prediction based on machine learning techniques can be useful in detecting cardiovascular disease with maximum sensitivity and accuracy. In the literature, there are many studies in which the Naive Bayes algorithm, which is one of the machine learning methods, is widely used to model and predict the early diagnosis of heart disease. In this study, demonstrates the statistical modeling of the Naive Bayes algorithm, presenting a classificatory approach for the prediction of heart disease. In the study, a data set consisting of 14 variables belonging to 303 patients accessed from the Kaggle site was used. As a result of the study, it was determined that the classification accuracy of the Naïve Bayes algorithm was 89.47%.

Cite as: Birlik AB. Prediction of Heart Disease by Naive Bayesian Classification Algorithm Based on Statistical Modeling. Acta Med Eur. 2025;7(3):4-11.

INTRODUCTION

Cardiovascular disease (CVD) refers to a group of disorders that affect the heart or blood arteries. Coronary artery diseases (CAD) including angina and myocardial infarction are examples of CVD. Plaque accumulation on the arterial walls causes the disease known as atherosclerosis. This accumulation makes the arteries smaller and restricts blood flow. A blood clot can impede blood flow and result in a heart attack or stroke, according to research [1]. Coronary artery diseases are the leading causes of death globally and have been reported to significantly increase the overall health care burden [2]. More than four million people each year in the European Region of the World Health Organization and more than 1.9 million in the European Union (EU) die from cardiovascular disease, which continues to be the leading cause of death. Additionally, it is predicted to cost the EU economy 196 billion euros annually [3].

Artificial intelligence is a broad concept that refers to technologies or systems that can display human-like intelligence. A subset of artificial intelligence known as "machine learning -ML" refers to the capability of making predictions based on fresh data and learning from past experience. [4]. The availability of big data has made data mining and ML a research area in decision-making processes in many fields, including engineering, finance, management and medicine [5]. There is an expanding literature on machine learning-based algorithms and their potential clinical utility in healthcare systems. Clinicians need enormous amounts of information to analyze data in a cost-effective and time-efficient manner for the diagnosis and treatment of diseases. Medical science, big data, uses machine learning methods in areas such as outcome prediction, diagnosis, medical image interpretation, and treatment [6, 7].

Machine learning is a key model for predictive analytics and innovation in medical science and is leading the digital healthcare transformation. Clinicians will have the potential to decide on the most appropriate treatment plan for a particular patient, considering the risk score generated by a predictive model in addition to their clinical assessment. These methods assist clinicians in the planning and maintenance of the diagnosis, the best possible outcomes, cost reduction in diagnosis, and patient satisfaction. In the literature, several studies are using ML methods in the discipline of heart diseases (HD). They utilized a variety of data mining approaches to diagnose the problem and got varied results for each strategy. A solution model is proposed to detect cardiovascular disease using ML classification algorithms. In addition, studies presenting a wide range of case approaches including predictive machine learning models developed for the prediction of cardiac diseases have been reported [8].

The Naive Bayes (NB) method has been widely used to model and predict the early detection of heart disease [9]. NB and Decision Tree (DT) classification algorithms were analyzed on the dataset to estimate a patient's probability of heart disease. It has been reported that the heart disease patient was predicted with an accuracy of 91% in the DT model and 87% in the NB classifier [10]. NB classifier has been created as a decision support system to assess the risk of cardiovascular disease. After the variables were selected, data cleaning was performed to eliminate missing or incorrect data. Numerical data were categorized and data generalization was performed [11]. The potential contributions of various diagnostic techniques and the development of an automated decision support system to assist professionals in predicting the onset of heart disease have been explored. It has been reported that after analyzes that take into account the decision-making time, Bayesian algorithms turned out to be capable enough to provide good accuracy for cardiovascular diseases in an acceptable time [12]. To forecast CVDs, the algorithms Logistic Regression (LR), k-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest (RF) and NB were examined. The accuracy of the outcomes was used to evaluate these algorithms. According to the study, the NB model makes the best predictions based on data. This model was said to display 46 false cases prior to feature extraction, but only 36 false cases afterwards with an accuracy of 88.16% [13]. Predictive models for CVD are presented with the Gaussian NB, Bernoulli NB and RF by considering the risk factors associated with the disease. A dataset of general coronary patients from the Cleveland database of the UCI repository was used. The accuracy of the Gaussian NB, Bernoulli NB, and RF algorithms is 85%, 85%, and 75%, respectively. Additionally, the accuracy, F-measure, and recall of Gaussian and Bernoulli NB are better than those of the RF algorithm, highlighting the value of this method in predicting early disease diagnosis [9]. Comparing NB classification methods to others like SMO (Sequential Minimal Optimization), Bayes Net, and MLP (Multi-Layer Perception). Smart Heart Disease Prediction was created using the NB algorithm to predict heart disease risk factors, taking into account previous data and information. In comparison to other techniques, the proposed NB model performed better, with an accuracy value of 89.77% [14]. In addition, several academics suggested combining SVM with NB in a hybrid technique. They used the identical settings from the dataset from the UCI repository again and this time they were 100% accurate [15]. These findings suggest that NB is among the most effective algorithms for categorizing and forecasting cardiac disease. In this study, the NB algorithm was applied to data in order to contribute to the early detection of heart diseases, and the classification success was demonstrated using mathematical modeling.

METHODS

Dataset

This study used a database of 303 samples (Table 1) with and without heart disease risk supplied by the University of Cleveland (UCI machine learning repository) in the Kaggle database. The data set consists of 14 different columns. Pycharm was used to process the data, which was programmed in Python (JetBrains individual licenses). Algorithms for processing data use a variety of Python libraries, including numpy, scikit-learn, matplotlib, and pandas. There are eight categories (Sex, Cp, Fbs, Restecg, Exang, Slope, Thal, Target) and six numerical (Age, Trestbps, Chol, Thalach, Oldpeak, Ca) qualities.

Data Preprocessing

Data cleansing is a procedure that must be completed prior to data analysis. It includes processes like filling in missing data, reducing discrepancies, and recognizing outliers [16]. For the HD dataset utilized in this investigation, no missing data values of various characteristics were found. One of the most significant data transformations is feature scaling. The numeric characteristics utilized in the input should not have multiple scales for ML algorithms to work successfully [17]. As a result, the min/max normalization approach was used to rescale the data set so that the values in distinct scales varied in the range of 0-1. Using the following formula, the approach changes a number in the range 0 to 1.

$$x_{new} = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

Table 1. Features for data prediction.

	age	sex	cp	trestbps	chol	fbs	…	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	…	0	2.3	0	0	1	1
1	37	1	2	130	250	0	…	0	3.5	0	0	2	1
2	41	0	1	130	204	0	…	0	1.4	2	0	2	1
3	56	1	1	120	236	0	…	0	0.8	2	0	2	1
4	57	0	0	120	354	0	…	1	0.6	2	0	2	1
…	…	…	…	…	…	…	…	…	…	…	…	…	…
298	57	0	0	140	241	0	…	1	0.2	1	0	3	0
299	45	1	3	110	264	0	…	0	1.2	1	0	3	0
300	68	1	0	144	193	1	…	0	3.4	1	2	3	0
301	57	1	0	130	131	0	…	1	1.2	1	1	3	0
302	57	0	1	130	236	0	…	0	0.0	1	1	2	0
	[303 rows x 14 columns]												

Modeling

Preprocessed data split into 75% training and 25% testing for training and testing purposes. Then these data were tested with Naïve Bayesian ML classifiers (Figure 1).

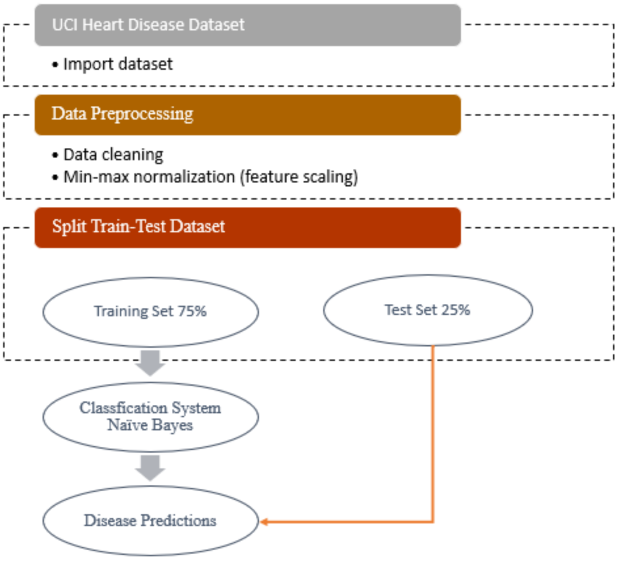


Figure 1. Flow chart of modeling.

Naïve Bayes Classification Algorithm

The probability theorem known as Bayes' theorem or Bayes' rule provides a widely utilized foundation for categorization. Let's first go through the two fundamental principles of probability theory.

$$P(X) = \sum_{Y} P(X,Y) \tag{2}$$

$$P(X,Y) = P(Y|X)P(X) \tag{3}$$

The first equation is the sum rule, while the second is the product rule. In this case, $p(X,Y)$ is a joint probability, $p(Y|X)$ is a conditional probability, and p(X) is a marginal probability.

Using the product rule and the symmetry property as a foundation $p(X,Y)$= $p(Y,X)$, it is easy to obtain the following Bayes' theorem,

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{4}$$

which plays a central role in ML, especially classification [18].

Let X = ($x1$, $x2$, …, xn) be the set of examples, C = ($C1$, $C2$, …, Cm) the set of classes. The object is to maximize $P(Ci|X)$. Maximizing $P(Ci|X)$ for the class Ci is termed the maximum posterior hypothesis. From Bayes' Rule;

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \tag{5}$$

btained. Since $P(X)$ is equal for all categorization, it is considered constant and only concerned with maximizing the numerator. Since it is difficult to calculate $P(C_i|X)$ in data sets with a large number of attributes, there is a presumption of class conditional independence. According to so presumption, there is no dependent relationship between the attributes. Thus $P(X|Ci)$;

$$\begin{aligned} &P(X|C_i) \\ &= \prod_{k=1}^{n} P(x_k|C_i) \\ &= P(x_1|C_i)P(x_2|C_i) \dots P(x_n|C_i) \end{aligned} \tag{6}$$

the form is obtained. The calculation of $P(X|Ci)$ depends on whether the attribute is categorical or numerical. If the features are numeric (assuming a normal distribution), it is assumed that the feature shows a Gaussian distribution accompanied by μ mean and σ standard deviation. Probability Density Function (PDF) of Gaussian distribution for an input value x_k is shown as follows [18];

$$(mean)\mu_{c_i} = \frac{1}{n}\sum_{k=1}^{n} x_k \tag{7}$$

$$\begin{aligned} &P(standard\ deviation)\sigma_{c_i} \\ &\quad = \frac{1}{n-1}\sum_{k=1}^{n}\left(x_k - \mu_{c_i}\right)^2 \end{aligned} \tag{8}$$

$$\begin{aligned} &g\left(x_k, \mu_{c_i}, \sigma_{c_i}\right) \\ &= \frac{1}{\sigma_{c_i}\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(\frac{x_k - \mu_{c_i}}{\sigma_{c_i}}\right)^2\right\} \end{aligned} \tag{9}$$

$$P(x_k|C_i) = g\left(x_k, \mu_{c_i}, \sigma_{c_i}\right) \tag{10}$$

Heart Disease Classification

From Bayes Theorem;

The posterior probability of an event C_i is able to express as accepted:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \tag{11}$$

where the vector $X = (X_1, X_2, \dots X_n)$ represents some n features that are assumed to be independent with each other, By changing the variables the formula is applied to the problem in the study, to make the theorem more concrete.

$$\begin{aligned} &Posterior\ Probability \\ &= \frac{(Likelihood)(Class\ Pirior\ Probabil}{Predictor\ Pirior\ Probability} \end{aligned} \tag{12}$$

$$\begin{aligned} &P(Heart\ disease|data) \\ &= \frac{P(data|Heart\ disease)P(Heart\ dis}{P(data)} \end{aligned} \tag{13}$$

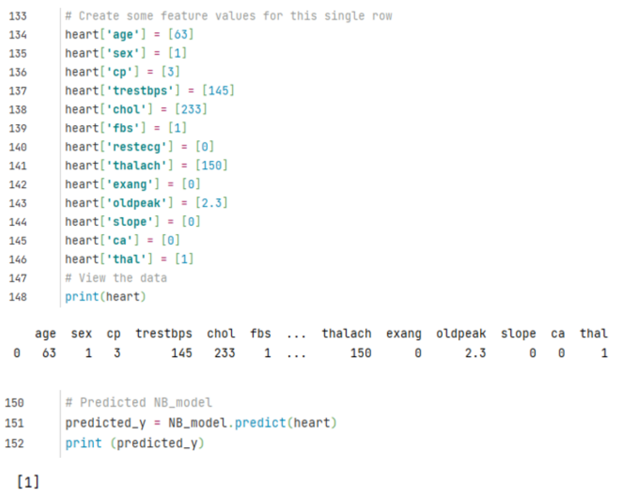
If choose $X = (age, sex, cp, \dots thal)$ as the associated features influencing the heart disease, the posterior probabilities concerning the presence (disease) and absence (not disease) states can be expressed as:

$$\begin{aligned} &P(disease|age, sex, cp, \dots thal) \\ &= \frac{P(disease)P(age|disease)P(sex|di}{Predictor\ Pi} \end{aligned} \tag{14}$$

$$\begin{aligned} &P(not\ disease|age, sex, cp, \dots thal \\ &= \frac{P(not\ disease)P(age|not\ disease)I}{Pr} \end{aligned} \tag{15}$$

$$\begin{aligned} &Predictor\ Pirior\ Probability \\ &= P(disease)P(age|disease)P(sex|di \\ &+ P(not\ disease)P(age|not\ disease)I \end{aligned} \tag{16}$$

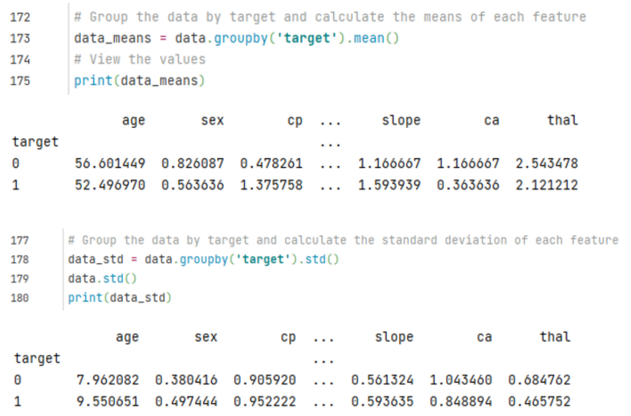
Below is an example (Figure 2) of the first row of our dataset to be classified as either disease or not disease (target).



```
# Create some feature values for this single row
heart['age'] = [63]
heart['sex'] = [1]
heart['cp'] = [3]
heart['trestbps'] = [145]
heart['chol'] = [233]
heart['fbs'] = [1]
heart['restecg'] = [0]
heart['thalach'] = [150]
heart['exang'] = [0]
heart['oldpeak'] = [2.3]
heart['slope'] = [0]
heart['ca'] = [0]
heart['thal'] = [1]
# View the data
print(heart)

   age  sex  cp  trestbps  chol  fbs  ...  thalach  exang  oldpeak  slope  ca  thal
0   63    1   3       145   233    1  ...      150      0      2.3      0   0     1

# Predicted NB_model
predicted_y = NB_model.predict(heart)
print (predicted_y)

[1]
```

Figure 2. First row of our dataset.

Parameters associated with the probability density function were obtained by grouping the data according to the target and for each feature, the standard deviation and mean are calculated (Figure 3).



```
# Group the data by target and calculate the means of each feature
data_means = data.groupby('target').mean()
# View the values
print(data_means)

              age       sex        cp  ...     slope        ca      thal
target                                 ...
0       56.601449  0.826087  0.478261  ...  1.166667  1.166667  2.543478
1       52.496970  0.563636  1.375758  ...  1.593939  0.363636  2.121212

# Group the data by target and calculate the standard deviation of each feature
data_std = data.groupby('target').std()
data.std()
print(data_std)

             age       sex        cp  ...     slope        ca      thal
target                                ...
0       7.962082  0.380416  0.905920  ...  0.561324  1.043460  0.684762
1       9.550651  0.497444  0.952222  ...  0.593635  0.848894  0.465752
```

Figure 3. Standard deviation and mean of each property.

A probability distribution for heart disease can be determined based on the data set: P(disease) and P(not disease) are the prior probabilities.

$$P(disease) = \frac{165}{303} = 0.54 \qquad (17)$$

$$P(age|disease)P(sex|disease)$$

$$P(cp\,|\,disease) \ldots P(thal\,|\,disease)$$

is the likelihood. The mean and standard deviation values for the sample are calculated using the probability density function. The mean and standard deviation values for the sample are calculated using the probability density function.

$$P(age|disease) = \frac{1}{\sigma_{c_i}\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(\frac{x_k - \mu_{c_i}}{\sigma_{c_i}}\right)^2\right\} \qquad (18)$$

$$P(age|disease) = \frac{1}{52.497\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(\frac{63 - 52.49}{9.551}\right)^2\right\} \qquad (19)$$

$$\vdots$$

$$P(thal|disease) = \frac{1}{2.121\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(\frac{1 - 2.121}{0.466}\right)^2\right\} \qquad (20)$$

As above, each probability is calculated separately to estimate heart disease. When the first row of the dataset was tested with the seted model, it gave the true positive value.

RESULTS

Correlation Matrix

The correlation between the variables is presented in Figure 4. According to the color scale, the correlation relationship between the columns close to red is high, while the correlation relationship decreases gradually in the colors towards the blue. In the data set, it is seen that cp, thalach, and slope variables have the highest correlation with the patient's heart disease status. In addition, a correlation above ±0.5 was not detected between the dependent variable (target) and other variables. This shows that no variable can be used independently in the estimation.

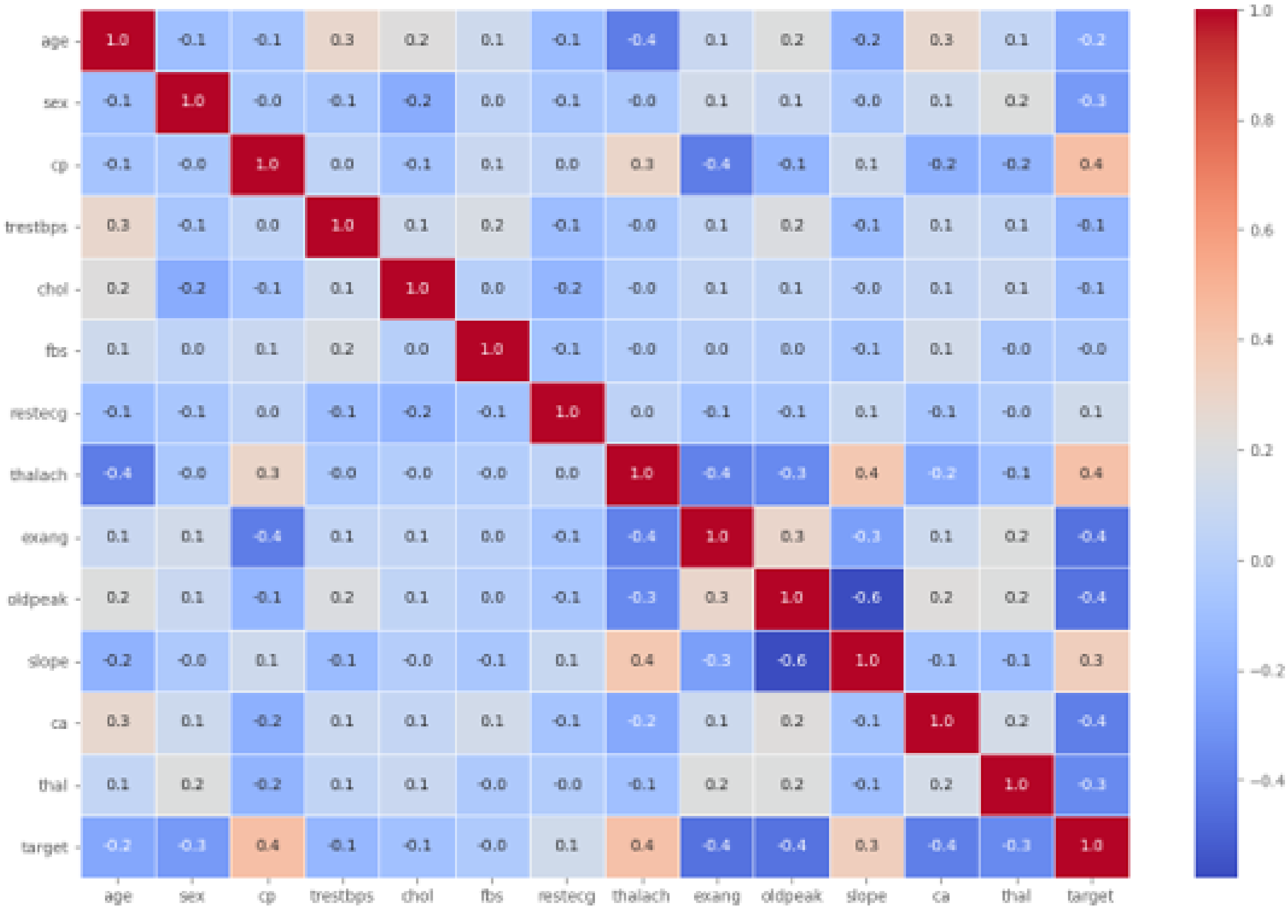


Figure 4. Correlation matrix.

Performance Evaluation

To interpret the achievement of the Naive Bayes algorithm used in the classification of heart disease diagnosis, the classification outcomes were transferred to the confusion matrix. A diagram of a confusion matrix is presented in the figure below in Figure 5.

Prediction → **↓Actual Output**	**no heart disease (0)**	**has heart disease (1)**
no heart disease (0)	True negatives (TN)	False positives (FP)
has heart disease (1)	False negatives (FN)	True positives (TP)

Figure 5. Diagram of a confusion matrix.

According to the confusion matrix, the Naïve Bayes algorithm correctly predicted 68 of 76 test data. However, 4 people are not predicted to have heart disease and 4 people are considered to not heart patients. The computed accuracy for Naïve Bayes classification is 89.47% as depicted in Figure 6.

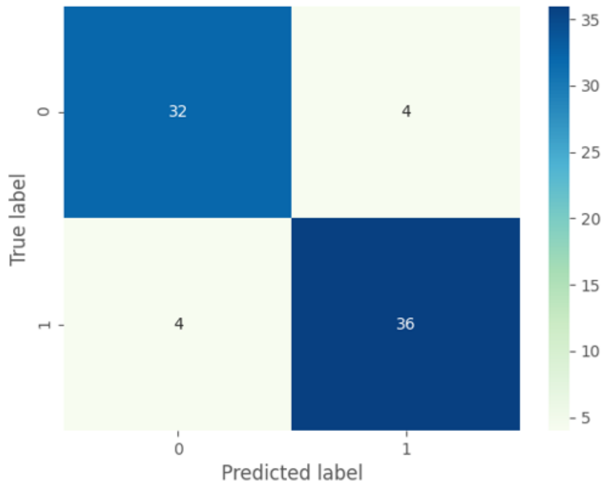


Figure 6. Naïve Bayes confusion matrix.

The AUC value obtained was 0.95, with the corresponding ROC graph displayed below in Figure 7. That is, the area under the model's ROC curve can be used as a criterion for measuring how well the test is in a particular clinical situation.

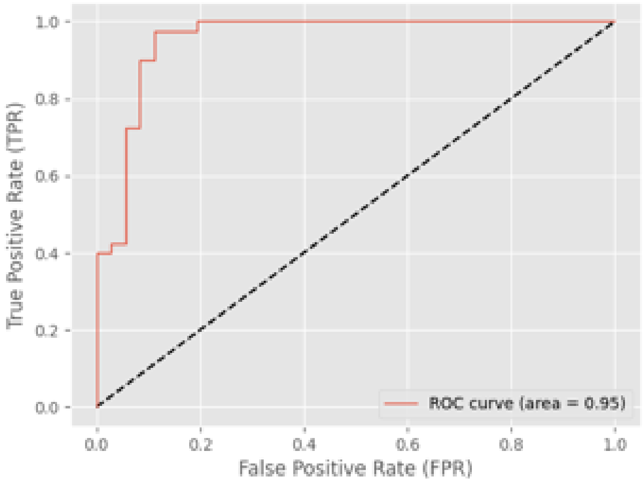


Figure 7. ROC (Receiver Operating Characteristic) curves from the Naïve Bayes mode.

DISCUSSION

Data mining algorithms are crucial in detecting heart disease and determining risk factors. In this study, Naïve Bayesian algorithms were used as ML to determine the risk of heart disease. Naïve Bayesian classification algorithm compared in terms of correlation and confusion matrix, accuracy, and ROC. When the data mining Naïve Bayes classification algorithm was examined in terms of accuracy, an accuracy rate of 89.47% was calculated. There are several limitations in this work. To begin, the NB algorithm assumes that each risk factor for heart disease is independent. In theory, this may not be the case because one risk factor may be linked to another. Furthermore, because a small HD dataset's variance is restricted, the accuracy of a small dataset may be computed extremely similarly and close to each other.

Multiclass classification of cardiovascular disease datasets could be considered in the future by integrating coronary artery angiography and coronary artery calcium score characteristics in the dataset, which are crucial in the early detection of cardiovascular disorders. To cope with real-life problems in hospitals and medical institutes, class imbalanced datasets can be investigated.

Ethics committee approval

Ethics committee approval is not required for this study.

Conflict of interest declaration

There is no conflict of interest with any person/institution in this study.

CC BY Licence

REFERENCES

1. Thomas H, Diamond J, Vieco A, et al. Global Atlas of Cardiovascular Disease 2000-2016: The Path to Prevention and Control. Glob Heart. 2018;13(3):143-163. doi:10.1016/j.gheart.2018.09.511
2. Naghavi M, Abajobir AA, Abbafati C, Abbas KM. Global, regional, and national age-sex specific mortality for 264 causes of death, 1980-2016: A systematic analysis for the Global Burden of Disease Study 2016. Lancet. 2017;390(10100):1151-1210. doi:10.1016/S0140-6736(17)32152-9
3. Nicholls M. Funding cardiovascular research in Europe. Eur Heart J. 2019;40(2):80-82. doi:10.1093/eurheartj/ehy817
4. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science. 2015;349(6245):255-260. doi:10.1126/science.aaa8415
5. Cioffi R, Travaglioni M, Piscitelli G, et al. Artificial intelligence and machine learning applications in smart production: Progress, trends, and directions. Sustainability. 2020;12(2):492. doi:10.3390/su12020492
6. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: Past, present and future. Stroke Vasc Neurol. 2017;2(4):230-243. doi:10.1136/svn-2017-000101
7. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347-1358. doi:10.1056/nejmra1814259
8. Shamsollahi M, Badiee A, Ghazanfari M. Using combined descriptive and predictive methods of data mining for coronary artery disease prediction: A case study approach. J AI Data Min. 2019;7(1):47-58. doi:10.22044/JADM.2017.4992.1599
9. Bemando C, Miranda E, Aryuni M. Machine-learning-based prediction models of coronary heart disease using Naïve Bayes and Random Forest algorithms. In: 2021 Int Conf Softw Eng Comput Syst 4th Int Conf Comput Sci Inf Manag ICSECS-ICOCSIM 2021. IEEE; 2021:232-237. doi:10.1109/ICSECS52883.2021.00049
10. Krishnan JS, Geetha S. Prediction of heart disease using machine learning algorithms. In: 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT). IEEE; 2019:1-5.
11. Miranda E, Irwansyah E, Amelga AY, et al. Detection of cardiovascular disease risk's level for adults using naive bayes classifier. Healthc Inform Res. 2016;22(3):196-205. doi:10.4258/hir.2016.22.3.196
12. Mahalingam PR, Dheeba J. Using machine learning and data analytics for predicting onset of cardiovascular diseases—An analysis of current state of art. In: Springer Nature Singapore Pte Ltd; 2020.
13. Gupta A, Kumar L, Jain R, Preeti N. Heart disease prediction using classification (Naive Bayes). In: Springer Nature Singapore Pte Ltd; 2020.
14. Repaka AN, Ravikanti SD, Franklin RG. Design and implementing heart disease prediction using Naive Bayesian. In: Proc Int Conf Trends Electron Informatics, ICOEI 2019. IEEE; 2019:292-297. doi:10.1109/icoei.2019.8862604
15. Jabbar MA, Samreen S. Heart disease prediction system based on hidden naïve bayes classifier. In: 2016 Int Conf Circuits, Control

Commun Comput I4C 2016. IEEE; 2017. doi:10.1109/CIMCA.2016.8053261

16. Krishnani D, Kumari A, Dewangan A, et al. Prediction of coronary heart disease using supervised machine learning algorithms. In: IEEE Reg 10 Annu Int Conf Proceedings (TENCON). IEEE; 2019:367-372. doi:10.1109/TENCON.2019.8929434

17. Roth GA, Mensah GA, Fuster V. The global burden of cardiovascular diseases and risks: A compass for global action. J Am Coll Cardiol. 2020;76(25):2980-2981. doi:10.1016/j.jacc.2020.11.021

18. Shin S, Austin PC, Ross HJ, et al. Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality. ESC Heart Fail. 2021;8(2):106-115. doi:10.1002/ehf2.13073