SSRG International Journal of Computer Science and Engineering
ISSN: 2348–8387 / https://doi.org/10.14445/23488387/IJCSE-V10I12P102

Volume 10 Issue 12, 5-9, December 2023
© 2023 Seventh Sense Research Group®

Original Article

Prediction of Cardiovascular Disease using NaïveBayes with Confusion Matrix

Ramesh[1], R. Rathidevi[2], Priyanandhini[3]

[1, 2]*Department of Computer Appli and Technology, SRM Arts and Science College, Tamilnadu, India.*
[3]*Department of Computer Science, SRM Arts and Science College, Tamilnadu, India.*

[1]*Corresponding Author : rameshcat@srmasc.ac.in*

Received: 07 October 2023 Revised: 18 November 2023 Accepted: 03 December 2023 Published: 18 December 2023

Abstract *- Cardiovascular disorders significantly contribute to reduced life expectancy. Factors such as obesity, elevated cholesterol levels, smoking habits, hypertension, diabetes, among others, can precipitate these conditions. Data from the World Health Organization reveal that cardiac-related afflictions, including myocardial infarctions and angina, are responsible for the annual demise of millions globally. The current system is designed to evaluate and benchmark historical patient outcomes against new diagnoses to forecast an individual's risk of developing heart disease in the future. By putting the aforementioned concept into practice, the proposed system is more accurate at predicting the likelihood that a new patient will have a heart attack. The Heart Attack Prediction System utilizes deep learning algorithms and techniques. However, there is very little precision while using all these methods in the existing systems. The proposed system's objective is to identify the people who are all suffering heart attacks by using important metrics. Deep learning algorithms and methods are applied to this system to improve performance and accuracy. The performance of a machine learning algorithm on test data can be quantified using a confusion matrix, which is commonly utilized to evaluate the accuracy of classification models. These models aim to assign a categorical class to each input sample. The matrix displays the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) produced by the model when applied to the test dataset.*

The following algorithm can be used in machine learning

1. *Logistic regression*
2. *Naïve Bayes*

Keywords *- Deep Learning, Confusion matrix, True negatives, False Positives.*

1. Introduction

The heart is considered the most vital organ in the human body, primarily responsible for pumping blood throughout the entire body. This four-chambered organ effectively segregates oxygenated and deoxygenated blood to ensure proper circulation. The human heart features a network of five distinct types of blood vessels: arteries, veins, capillaries, arterioles, and venules. Structurally, the heart roughly matches the size of an individual's clenched fist and has an average weight of about 300 grams. Typically, a woman's heart is nearly 25% lighter than a man's heart. The heart contains both arteries and veins, serving the crucial function of collecting blood from various body parts, purifying it, and then distributing it to all parts of the body. This blood carries essential nutrients and oxygen to different body tissues while also aiding in the removal of metabolic waste products. In the present day, the human lifespan has been diminishing due to the prevalence of heart diseases. Various factors to heart disease exist, encompassing excessive body weight, heightened cholesterol, tobacco use, increased blood pressure, and diabetes, among additional risk factors. The World Health Organization (WHO) reports that heart-related ailments, including myocardial infarctions and angina, claim the lives of millions each year globally. The proposed system operates by utilizing extensive data gathered from large hospitals. Data comparison is carried out using a confusion matrix, which proves highly valuable in disease prediction and outcome assessment. The suggested system operates with substantial data sourced from extensive hospital datasets. Data comparisons are executed using a confusion matrix, which proves highly effective in disease prediction. Based on attribute classification, accurate and inaccurate predictions are tallied, enabling the analysis of the percentage of correct predictions.

In machine learning, the algorithms that can be applied are Logistic Regression and Naïve Bayes.

1.1. Logistic Regression

Logistic regression involves the modeling of the probability of a discrete result based on input variables. In the most typical application, logistic regression models binary outcomes, which are situations where there are only two possible values, like true/false, yes/no, and similar

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

cases. In contrast, multinomial logistic regression is employed when there are more than two potential discrete outcomes to be modeled. Logistic regression, often referred to as the sigmoid function, offers a convenient graphical representation. It is recognized for its capacity to deliver high accuracy. In this algorithm, the initial step involves importing and training the data. Through the use of an equation, the logistic regression algorithm is depicted graphically, illustrating attribute variations. Subsequently, the best and approximate coefficients are estimated based on the training data and presented.

In the context of logistic regression with Sklearn, a matrix is employed for comparison and to assess confusion. Logistic regression, commonly known as the sigmoid function, simplifies graph representation and delivers high precision.

In this approach, the initial steps involve data importation and subsequent training. Equations are utilized to depict the logistic regression process within graphs, highlighting attribute differences. It is essential to derive the most accurate and approximate coefficients from the training data and represent them accordingly.

Logistic regression, often visualized through a sigmoid curve, is conducive to graphical representation and yields reliable accuracy. This method involves initially importing the data set and then training it. The algorithm utilizes an equation to delineate the distinction between various features graphically. It involves deriving the most suitable and close-fitting coefficients from the training data to depict the model accurately. Matrix for comparison and confusion Logistic regression with Sklearn is used.

The logistic regression is also referred to as the sigmoid function, which makes graph representation simple. It also offers excellent precision. For this approach, the data should be imported first, followed by training. The graphs displaying the distinction between the qualities use equations to describe the logistic regression process. To determine the most accurate approximation coefficient from the training data and represent it.

Naïve Bayes

The widely employed Naive Bayes algorithm plays a crucial role in classification tasks and can be especially beneficial for large datasets. This classification method operates under the assumption of attribute independence, making it suitable for scenarios where the removal of correlated data can offer advantages.

2. Existing System

The current system integrates both Deep Learning and Data Mining techniques and relies on a comprehensive report stating that all methods employ a robust prediction algorithm. The primary objective of this system is to compare and assess the medical records of a new patient with those of previous patients who have had the same disease. This comparison aids in determining the likelihood of a future heart disease occurrence for a particular patient. By implementing the model described above, the accuracy of predicting the likelihood of a new patient experiencing a heart attack is significantly improved. The proposed Heart Attack Prediction System utilizes Deep Learning algorithms and approaches. In contrast, the accuracy of the existing system is notably lower.

2.1. Limitations

- The heart is safeguarded by the rib cage and enclosed within a dual-layered membranous sac known as the pericardium. This four-chambered organ segregates oxygen-rich blood from oxygen-poor blood.
- Risk factors that can precipitate heart disease include excessive body weight, elevated cholesterol levels, tobacco usage, rising blood pressure, and diabetes, among others.

3. The Proposed System

The NCM (Naive Bayes with Logistic Regression) system proposed here is designed to classify patients into those with or without heart disease based on various features. The primary goal of this system is to leverage available data to create a predictive model that determines whether a patient is affected by this condition through comprehensive data exploration and analysis. This approach involves the use of logistic regression, a classification algorithm, with the sklearn library used for calculating the model's accuracy score.

Additionally, the Naive Bayes algorithm is integrated to assess accuracy results. The final phase entails analyzing the outcomes by comparing different models and applying a Confusion Matrix.

It's essential to structure and categorize the dataset into distinct data segments.

- This proposed system has data which classifies if patients have heart disease or not according to its features.
- This proposed system can try to use this data to create a model which tries to predict (reading data and data Exploration) if a patient has this disease or not.
- The data should be classified into different structured data based on the features of the patient's heart. From the availability of the data.

3.1. Phases of NCM

3.1.1. Phase 1

Data Retrival and Correlation Analysis

3.1.2. Phase 2

Data Prediction with NCM method.

3.1.3. Phase 3

Data Validation and Analyzation

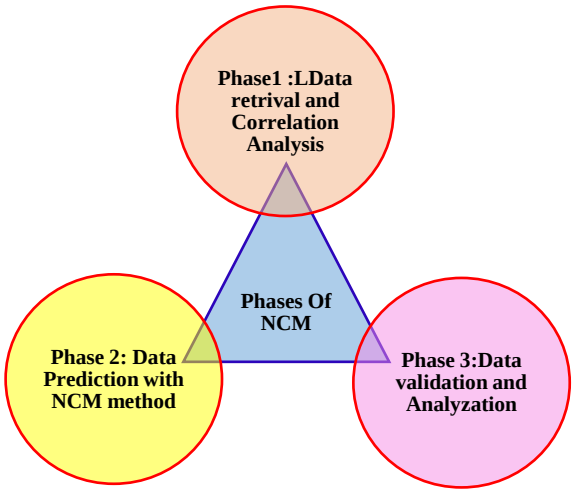

Fig. 1 Phases of NCM

Phase 1: Data Retrival and Correlation Analysis

Data retrieval comes first in the process. Heart Disease UCI Dataset will be applied during this procedure. This information will be imported into Py Charm. Both category and numerical data were acquired. Additionally, all variables in the imported dataset will be visualized as a histogram to make it easier to comprehend the data overall and to aid in data analysis. Analyzing the Correlation between Variables is a step in the process that looks at the correlation between variables to show that the logistic regression model is the best model to utilize. A matrix will be used to visualize the relationships between the variables in the provided dataset. Additionally, this is done to examine the database's variables for multicollinearity.

Phase 2: Data Prediction with NCM Method

Training data and testing data will be separated from the dataset imported into PyCharm. Models are constructed based on training data. Testing data is utilized in the interim as a foundation for validating or testing the model. The information will then be split into train and test data. The data that was partitioned in the previous procedure will be used in this one. When making predictions, the logistic regression method will generate a variety of data that can be used as a foundation.

Phase 3: Data Validation and Analyzation

The confusion matrix approach and K-fold cross-validation with 10-fold are the techniques used to validate the results. The effectiveness of the application of the logistic regression model can be determined by employing a confusion matrix. Additionally, the K-fold cross-validation method generates estimates of potential errors when employing a logistic regression. PyCharm imports the dataset that the researcher used as the basis for their investigation. To show the value of each variable used in the overall study analysis, the data retrieval procedure is also carried out in the data visualization.

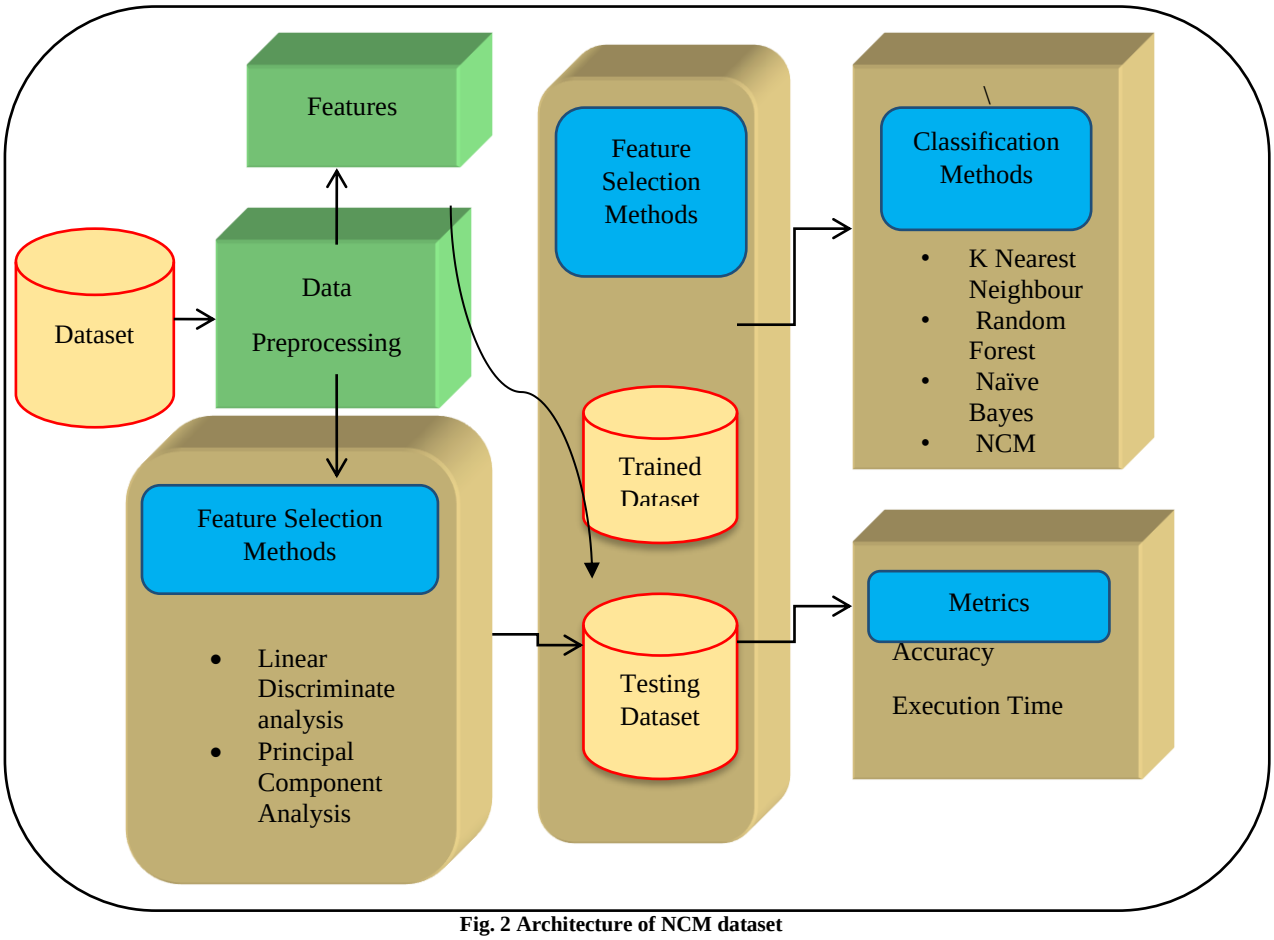

Fig. 2 Architecture of NCM dataset

Machine Learning Repository has access to the dataset. 14 input features, 1 output feature, and 303 samples make up the dataset. Financial, personal, and social characteristics of loan applicants are described by the features. There were 300 cases of negative credit in the sample. Features expressed in the dataset are nominal, ordinal, or interval scales. The table provides a list of all those features.

Table 1. Features for disease

S.No.	Features	Feature id	Description
1.	Age	AGE	Age in years
2.	Sex	SEX	Male -1 Female - 0
3.	Chest pain types	CPT	Atypical angina – 1 Typical angina – 2 Asymptomatic – 3 Nonanginal pain - 4
4.	Blood Pressure	RBP	mm Hg admitted at the hospital in mg/dl
5.	Serum Cholesterol	SCH	
6.	Blood sugar in fasting	FBS	Fasting blood sugar>120mg/dl(1 – true, 0 – false)
7.	Resting electrocardiographic results	RES	0 – normal 1 – having ST-T 2 – hypertrophy
8.	Heart rate in maximum	MHR	
9.	Exercise-induced Angina	EIA	1 – yes 0 – No
10.	Old peak – ST depression induced by exercise relative to rest	OPK	-
11.	Slope of the peak exercise ST segment	PES	1 – up sloping 2 – flat 3 – down sloping
12.	Fluoroscopy coloured major Vessels (0-3)	VCA	-
13.	Thallium Scan	THA	3 – normal 6 – fixed defect

In this method, a dataset of 210 records with 8 attributes was employed. The data mining tool was utilized to conduct the experiments and implementations. Comparative results have been drawn from the experiments.

Table 2. Performance analysis

Classification Techniques	**Accuracy**	**Timing Taken**
NCM	55.67%	608ms
Naïve Bayes	53.35%	610ms
Decision List	52%	719ms
KNN	45.6%	1000ms

The graphical representation is given for the above performance analysis table. In both, accuracy is high, and the time taken for execution is low in the proposed NCM method when compared with other algorithms.

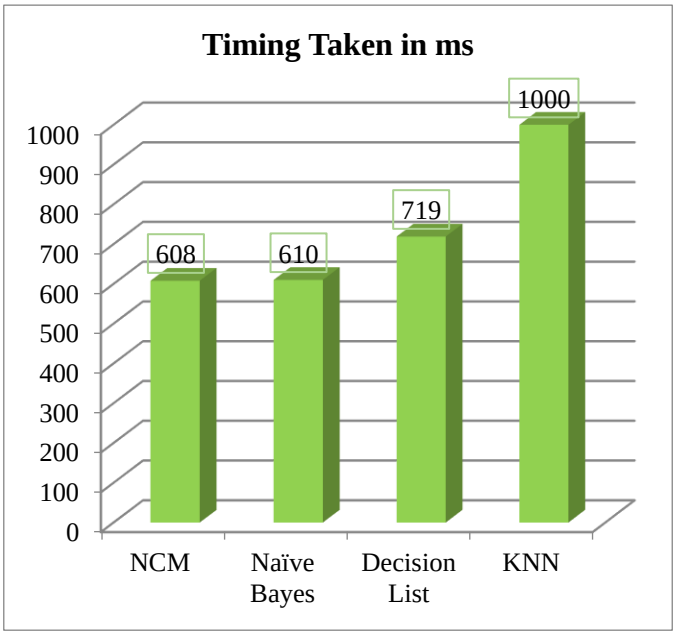

Fig. 3 Time taken in milliseconds

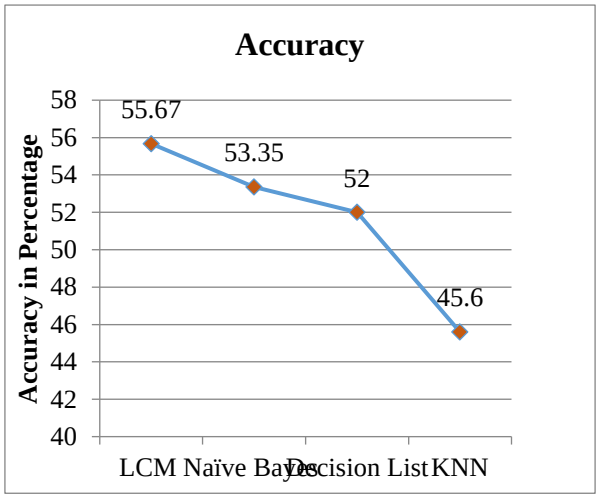

Fig. 4 Accuracy in percentage

4. Conclusion

This study offers a thorough understanding of machine learning methods for categorizing cardiac disorders. In order to predict the treatment that can be given to patients, classifiers play a key role in the healthcare business. In order to identify the effective and precise methods, the existing methodologies are examined and contrasted. Machine learning approaches dramatically increase the accuracy of cardiovascular risk prediction, allowing for the early diagnosis of patients who can then receive preventative care. Conclusion: Machine learning algorithms have enormous potential for predicting heart-related or cardiovascular disorders. The proposed NCM gives good results in prediction when compared with others. Each of the aforementioned algorithms has done incredibly well in some situations while failing miserably in others.

References

[1] Karl-Patrik Kresoja et al., "A Cardiologist's Guide to Machine Learning in Cardiovascular Disease Prognosis Prediction," *Basic Research in Cardiology*, vol. 118, no. 10, pp. 1-12, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Xin Qian et al., "A Cardiovascular Disease Prediction Model Based on Routine Physical Examination Indicators Using Machine Learning Methods: A Cohort Study," *Frontiers in Cardiovascular Medicine*, vol. 9, pp. 1-11, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[3] PE Rubini et al., "MCIDS-Multi Classifier Intrusion Detection System for IoT Cyber Attack using Deep Learning Algorithm," *Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, Tirunelveli, India, pp. 354-360, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[4] Belal Abuhaija et al., "A Comprehensive Study of Machine Learning for Predicting Cardiovascular Disease Using Weka and SPSS Tools," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 2, pp. 1891–1902, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[5] Abdullah Alqahtani et al., "Cardiovascular Disease Detection using Ensemble Learning," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1-9, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[6] Ahmed M. Alaa et al., "Cardiovascular Disease Risk Prediction Using Automated Machine Learning: A Prospective Study of 423,604 Uk Biobank Participants," *Plos One*, vol. 14, no. 5, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[7] Ruby Hasan, "Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction," *ITM Web of Conferences*, vol. 40, pp. 1-7, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[8] Archana Singh, and Rakesh Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," *2020 International Conference on Electrical and Electronics Engineering (ICE3)*, Gorakhpur, India, pp. 452-457, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[9] Apurv Garg, Bhartendu Sharma, and Rijwan Khan, "Heart Disease Prediction using Machine Learning Techniques," *IOP Conference Series: Materials Science and Engineering*, vol. 1022, pp. 1-10, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[10] Chayakrit Krittanawong et al., "Machine Learning Prediction in Cardiovascular Diseases: A Meta-Analysis," *Scientific Reports*, vol. 10, no. 16057, pp. 1-11, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[11] Agmir Javaid et al., "Medicine 2032: The Future of Cardiovascular Disease Prevention with Machine Learning and Digital Health Technology," *American Journal of Preventive Cardiology*, vol. 12, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[12] M. Ramesh, and C. Vijayakumaran, "Review of Big Data Analytics and Machine Learning Models for Contract Tracking and Detecting of COVID-19 Pandemic Cases," *Proceedings - 6th International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, pp. 825-831, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[13] M.D. Amzad Hossen et al., "Supervised Machine Learning-Based Cardiovascular Disease Analysis and Prediction," *Mathematical Problems in Engineering*, vol. 2021, pp. 1-10, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[14] Karan Bhanot, "Predicting Presence of Heart Diseases using Machine Learning," *Medium*, 2019. [Google Scholar] [Publisher Link]

[15] Shadman Nashif et al., "Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System," *World Journal of Engineering and Technology*, vol. 6, no. 4, pp. 854-873, 2018. [CrossRef] [Google Scholar] [Publisher Link]