Research on Heart Disease Auxiliary Diagnosis based on Naive Bayes Classification

Zhaoxi Ye*

School of Mathematics, East China University of Science and Technology, Shanghai, 200237, China

* Corresponding Author Email: 20004001@mail.ecust.edu.cn

Abstract. Cardiovascular diseases (CVDs) have become a global health crisis, drawing widespread attention due to their severity and the significance of associated risk factors. The gravity of the situation necessitates a concerted effort to develop effective predictive tools and intervention methods to combat the escalating burden of heart diseases. As a result, this study addresses the urgent need for early prediction and intervention strategies for heart diseases. It collects various medical data related to heart disease patients, screens and extracts key datasets, and constructs a robust predictive model using the Naive Bayes method. The model is capable of analyzing and extracting predictive-relevant key features from a large set of diverse categorical features, thereby enhancing the efficiency and accuracy of the Naive Bayes model. With input from a relatively small amount of medical data, the model can assess an individual's risk of developing heart disease. This innovative approach not only conserves precious healthcare resources but also empowers healthcare institutions with the capability to monitor patient conditions and take timely, well-informed actions. Despite its inherent limitations, this research represents a significant step towards model-assisted heart disease prediction and highlights the potential application of machine learning models in the medical field, paving the way for future studies in personalized medicine.

Keywords: Cardiovascular diseases, heart disease prediction, Naive Bayes, risk assessment, feature selection.

1. Introduction

Cardiovascular diseases (CVDs) have become a global health crisis, drawing widespread attention due to their severity and the significance of associated risk factors [1]. These diseases take a toll on health and burden healthcare systems significantly [1]. According to the World Health Organization (WHO), CVDs are the leading cause of global mortality [2]. In 2019, an estimated 17.9 million lives were claimed by CVDs, accounting for a staggering 32% of all global deaths. Alarmingly, a substantial 85% of these CVD-related deaths were attributed to heart attacks and strokes [2]. What is particularly concerning is the rising prevalence of heart diseases not only among the elderly but also among younger populations, underscoring the urgent need for early prediction and intervention strategies.

Therefore, the core objective of this comprehensive research endeavor is to explore the multifaceted realm of potential factors contributing to heart diseases. While extensive research has shed light on the role of factors such as smoking [3], obesity [4], and alcohol consumption [5] in elevating the risk of heart diseases, numerous unidentified potential determinants warrant meticulous investigation. Simultaneously, early prediction of heart diseases has become a focal point in medical research.

This study will focus on collecting and examining a comprehensive array of patient physiological indicators. These encompass, but are not limited to, demographic variables such as gender and age, biochemical markers like cholesterol levels and fasting blood glucose, and lifestyle elements including dietary habits and descriptions of chest pain. The ultimate goal is to discern whether a discernible association exists between these diverse factors and the propensity for heart diseases, as well as to quantify their respective contributions to the overall risk of heart diseases. Within the international academic community, research dedicated to predicting heart diseases has accumulated a substantial body of literature.

Among the major contributions are four classical machine learning algorithms renowned for their effectiveness in disease risk prediction [6]. Wang et al. harnessed the power of the Artificial Bee Colony (ABC) algorithm to conduct classification and prediction tasks related to hypertension [7]. Jones et al. embarked on extensive epidemiological investigations, meticulously scrutinizing the intricate relationship between dietary factors and the incidence of heart diseases [8]. Furthermore, Naidu pioneered a cutting-edge heart disease prediction analysis method, amalgamating Artificial Intelligence (AI) and Machine Learning (ML) technologies, to deliver more accurate prognostic insights [9]. These studies have contributed a diverse array of methodologies and perspectives, enriching understanding of the multifaceted mechanisms and risk factors associated with heart diseases.

Nevertheless, despite the commendable progress achieved through these research endeavors, there remains a compelling need for deeper exploration of heart disease prediction. The Naive Bayes model, known for its probabilistic and statistical classification capabilities, stands as a promising frontier in the domain of heart disease risk prediction. Consequently, this research endeavor aims to amalgamate and synthesize the findings gleaned from previous studies. The ultimate goal is to harness the predictive capabilities of the Naive Bayes model, thereby constructing a robust and reliable heart disease risk prediction model. Such a model promises to provide individuals with tailored recommendations for early intervention and treatment based on their unique risk profiles.

By analyzing and extracting contributions from various feature categories to heart disease, this study will make a significant contribution to the field of heart disease prevention and management. The subsequent sections of this comprehensive research paper aim to provide a detailed overview and elucidate the research methodology, clarify the data sources, provide a comprehensive explanation of the complex process of building a predictive model using the Naive Bayes method, and most importantly, reveal research findings. The goal is to enhance the efficiency and reliability of heart disease prediction models by extracting crucial features, thereby offering possibilities for advancing the field of heart disease prediction. Through the exploration of the etiology and prediction of heart diseases in this research, the hope is to provide assistance to heart disease patients in the medical domain.

2. Methodology

2.1. Data source and description

The dataset used in this research study comprises a comprehensive collection of physiological and clinical attributes relevant to heart disease prediction. The data were obtained from two reputable sources: the "Heart Failure Prediction" dataset available on Kaggle and the "Mortality Rate of Heart Patients in Pakistan Hospital" dataset. To ensure data quality and consistency, extensive data cleaning procedures were applied to remove outliers and inconsistencies. The resulting dataset represents a valuable resource for investigating the multifaceted determinants of heart disease. Author affiliations

2.2. Variable description

The processed and filtered dataset consists of 1,111 test subjects and a total of 11 collected attributes. These metrics encompass demographic variables, biochemical markers, and clinical indicators, each of which plays a crucial role in heart disease risk assessment.

Demographic Variables: Age and Gender are fundamental demographic factors considered in this analysis. Age is a well-established risk factor for heart disease, with older individuals generally being at higher risk. Gender has also been shown to influence heart disease risk, with variations in prevalence and symptomatology between males and females. The statistical results of Age and Gender are presented in Table 1.

Table 1. Element Statistics: Age and Gender

Elements	Age						Gender	
Category	24-30	31-40	41-50	51-60	61-70	71-77	Female	Male
Total	16	95	262	487	230	21	265	846
Patients	11	41	134	337	184	15	123	599

Biochemical Markers: Cholesterol levels and Fasting Blood Sugar are important biochemical markers associated with heart disease risk. Elevated cholesterol levels and high fasting blood sugar levels are known risk factors for heart diseases. The statistical results of RestingBP, Cholesterol, and FastingBS are presented in Table 2.

Table 2. Element statistics: RestingBP, Cholesterol and FastingBS

Elements	RestingBP			Cholesterol			FastingBS	
Category	92-120	121-150	151-200	<140	140-200	201-300	>300	>120	Normal
Total	338	638	135	15	204	739	153	175	936
Patients	195	423	104	9	120	488	105	132	590

Clinical Indicators: Clinical indicators such as Chest Pain Type, Resting Blood Pressure, Resting Electrocardiogram, Max Heart Rate, Exercise-Induced Angina, Oldpeak, and ST Slope have been included to provide a comprehensive assessment of heart disease risk. The statistical results of RestingEC, MaxHR, ExerciseAngina, Oldpeak and ChestPainType are presented in Table 3 and Table 4.

Table 3. Element statistics: RestingEC, MaxHR, and ExerciseAngina

Elements	RestingECG			MaxHR			ExerciseAngina	
Category	LVH	Normal	ST	<101	101-150	151-200	0	1
Total	372	613	126	68	648	395	619	492
Patients	291	356	75	51	470	201	280	442

Table 4. Element statistics: Oldpeak and ChestPainType

Elements	Oldpeak						ChestPainType			
Category	0	0.1-0.9	1-1.9	2-2.9	3-3.9	>=4	NAP	ASY	ATA	TA
Total	399	138	311	156	70	37	204	667	181	59
Patients	159	71	249	140	67	36	82	571	36	33

These selected metrics will be used as input features for the predictive model, allowing us to assess their individual and collective contributions to heart disease risk.

2.3. Method introduction

The methodology employed in this research involves the development of a predictive model for heart disease risk assessment. The Naive Bayes algorithm, known for its probabilistic classification capabilities, serves as the foundation of the model. Below, the key components of the methodology are introduced, including a brief overview of the Naive Bayes principle.

The Naive Bayes algorithm is a probabilistic machine learning method based on Bayes' theorem. It is particularly suitable for classification tasks and has been widely utilized in various domains, including medical diagnosis. The core theory behind Naive Bayes is to train by inputting specific categorical or continuous features and output the probability of a data instance belonging to a particular class:

$$P(C|X) = \frac{P(X|C)\cdot P(C)}{P(X)} \quad (1)$$

In this formula, P(C|X) represents the posterior probability of class C given the observed features X. P(X|C) is the likelihood, representing the probability of observing features X given class C. P(C) is the prior probability of class C. P(X) is the evidence, representing the overall probability of observing features X.

Naive Bayes provides a method for dealing with uncertainty and probability in heart disease prediction. It accomplishes this by estimating the probabilities of individuals belonging to different heart disease risk categories based on observed features, offering a principled approach to modeling uncertainty. Its core assumption is the conditional independence of features. This method is relatively simpler compared to other models, yet it maintains excellent interpretability, aiding in the identification of the most critical features in heart disease prediction. Furthermore, Naive Bayes helps to comprehend the conditional probabilities of different features given a specific category, providing interpretability and facilitating the extraction of the most influential predictive features within the code [10].

In the following sections, through in-depth research and model development, we will unveil the crucial features for predicting heart disease and the practical effectiveness of this predictive model in heart disease assessment.

3. Results and Discussion

In order to streamline the feature extraction process, continuous features in the data are transformed into categorical features based on medical data categorization and sample quantity planning. The statistical analysis of the transformed categorical features for patients with and without the disease is presented in Figure 1, providing a visual representation of the disease prevalence across different features and facilitating the identification of significant features for further screening.

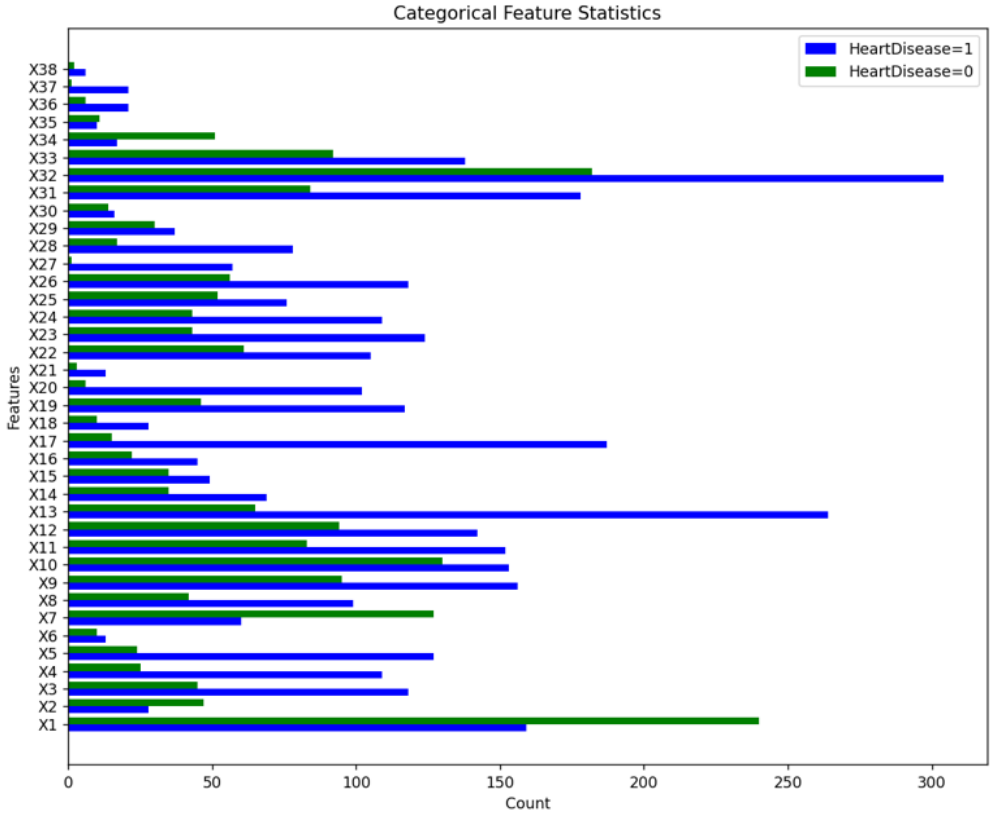


Figure 1. Transformed categorical features statistics

Then, in order to enhance the performance of the Naïve Bayes classifier, it was imperative to first filter out the significant factors contributing to the model. In our experiments, the average accuracy of categorical feature columns was assessed through cross-validation, Figure 2 illustrates the correlation performance of each factor in prediction.

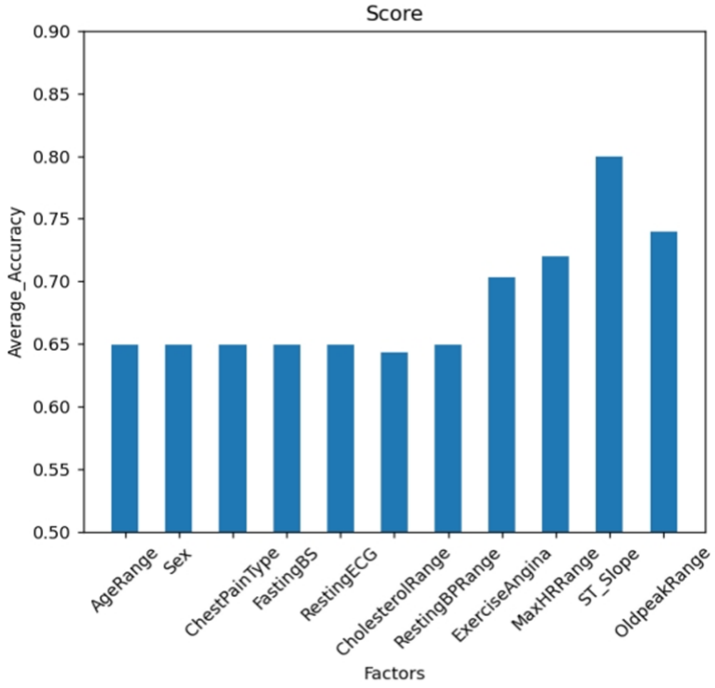


Figure 2. Correlation Performance of Factors Chart

It is evident that some features exhibited similar performance, indicating their limited relevance in the diagnosis of heart disease. Notably, 'ST_Slope,' 'OldpeakRange,' 'MaxHRRange,' and 'ExerciseAngina' consistently outperformed other factors, maintaining an average accuracy rate of approximately 70%. Among them, 'ST_Slope' emerged as the top-performing feature, achieving an accuracy rate of 80%. Therefore, these four features were selected as the most influential in classifying heart disease, serving as conditional inputs for training the classifier (Figure 2).

Following the transformation of categorical features into numerical form through one-hot encoding, further assessment was conducted to determine each feature's contribution to accuracy through chi-squared tests, subsequent feature scaling, and PCA analysis to reduce feature dimensions. Upon completing this series of data processing steps, the model's performance was evaluated and its generalization ability understood. The curves are illustrated in Figure 3.

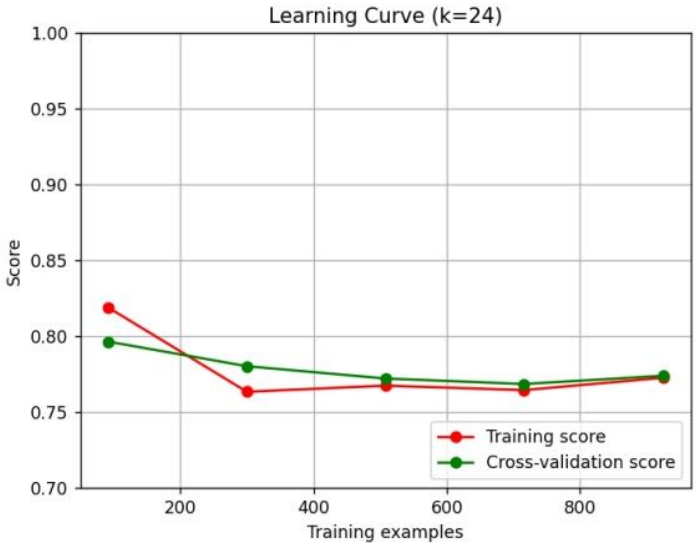


Figure 3. Performance on the Training set and Validation set

Observations reveal that training performance (red curve) gradually declines as the volume of training data increases, indicating the model's growing generalization to more diverse data. Given the substantial size of the dataset, this phenomenon is expected. In contrast, the cross-validation

performance (green curve) remains relatively stable. This suggests that due to the simplicity of the Naïve Bayes classification and the extraction of optimal features, the model demonstrates stability in performance even with smaller datasets. Additionally, the convergence of training performance (red curve) and cross-validation performance (green curve) signifies the model's proficiency in training on existing data and generalizing to unseen data. The confusion matrix of the model is depicted in Figure 4:

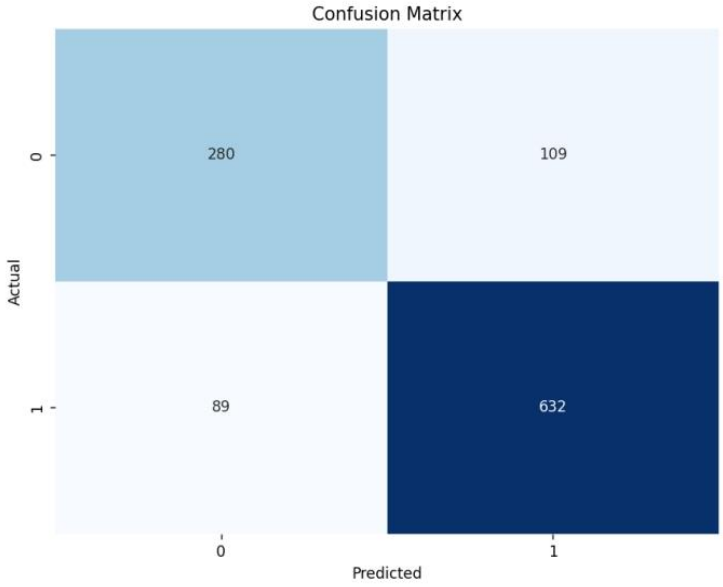


Figure 4. Confusion Matrix Chart

Here, '0' denotes patients predicted as 'No Heart Disease,' and '1' denotes patients predicted as 'Heart Disease.' As shown in the figure, the model predicted 'Heart Disease' in 632 cases, accounting for 56.9% of the total. It falsely predicted 'Heart Disease' in 109 cases, constituting 9.8% of the total. For 'No Heart Disease,' the model made 280 correct predictions, amounting to 25.2%, and 89 false predictions, representing 8% of the total. This implies an overall accuracy rate of approximately 82.2%. This accuracy holds significant promise for practical applications.

To analyze the model's real-world applicability further, the Chart below summarizes multiple performance metrics, including Precision, Recall, and F1-Score. These metrics are essential for a comprehensive evaluation of the model's performance.

The chart reveals a Precision of 85.29%, a Recall of 87.66%, and an F1-Score of 86.46%. In practical applications, as an auxiliary tool for heart disease prediction, the primary objective should be to capture as many actual heart disease cases as possible, resulting in a high Recall rate. This is crucial for the early detection of heart disease patients. Therefore, the model's Recall rate of 87.66% indicates its high utility. In conclusion, when considering the overall metrics, this model demonstrates commendable performance and practicality.

4. Conclusion

This research project has successfully delved into various aspects of predicting heart diseases. The study has unveiled the intricate network of factors contributing to heart diseases and extracted prominent features for heart disease prediction, leading to the development of a robust predictive model based on the Naive Bayes methodology. This model can be applied to individuals to assess their risk of heart disease and take informed preventive measures. It can also serve as a medical aid for early heart disease prediction to a certain extent.

Admittedly, due to limited data availability, the feature selection process from the dataset might not encompass all features relevant to heart disease detection, potentially introducing errors in practical application scenarios. Additionally, the model's accuracy might be affected by the concentration of data sources primarily from two locations, leading to variations in different

ethnicities and regions. Nevertheless, this study holds significant value. Firstly, through data preprocessing and feature selection, it amalgamates two datasets and classifies data, extracting critical features without requiring specialized medical knowledge in cardiology, thereby enhancing model efficiency and precision. Also, the adoption of the Naive Bayes classification in this experiment ensures independence among various features, mitigating the risk of overfitting and making such models readily applicable without the need for extensive customization based on specific contexts.

On the other hand, the limitations of this model, including its relatively low complexity, which makes it challenging to fine-tune for specific scenarios, and the limited variety of feature types, which makes it difficult to extract higher-precision features. Future research efforts should focus on optimizing the model's performance, bridging it with clinical practices, exploring its potential in personalized medicine, and identifying suitable application scenarios. By acquiring more relevant features from clinical data, further investigation into the interactions among risk factors can be conducted to enhance the overall accuracy of the model.

In summary, this research represents a significant step towards the use of model-assisted heart disease prediction. It also provides a possibility for future research in this direction. With the continuous advancement of medical experiments, this model can be updated to adapt to changes in features and data, further serving potential heart disease patients and healthcare systems.

References

[1] Yuan Wang, Xinxī Cao, Yabing Hou, et al. (2020). Comparative Study of Cardiovascular Disease Burden in China and Globally in 1990 and 2017. Chinese Journal of Chronic Diseases Prevention and Control, 28(1), 10-13, 19. DOI: 10.16386/j.cjpccd.issn.1004-6194.2020.01.003.

[2] World Health Organization. Cardiovascular diseases (CVDs), 2021, Information on https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[3] World Health Organization. Tobacco: health benefits of smoking cessation, 2020, Information on https://www.who.int/news-room/questions-and-answers/item/tobacco-health-benefits-of-smoking-cessation.

[4] Tian, Y., et al. A Prospective Study on the Association Between Central Obesity and the Risk of Ischemic Heart Disease in Chinese Adults," Chinese Journal of Epidemiology 39(9), 1172-1178 (2018).

[5] Fang, Y. H., et al. "Attributable Burden of Hemorrhagic Stroke and Hypertensive Heart Disease to Alcohol Consumption in China," Chinese Journal of Epidemiology 44(3), 393-400 (2023).

[6] Huang, G. C., et al. "Application and Comparison of Machine Learning Algorithms in Disease Risk Prediction," Chinese Health Resources 23(4), 432-436 (2020).

[7] Wang, S. Z. and Zhang, Y. H., "Chronic Disease Diagnosis Classification Study Based on ABC-NB," Computer Measurement and Control 25(11), 197-200 (2017).

[8] Ouyang, S., "Research of Heart Disease Prediction Based on Machine Learning," 2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), Wuhan, China, 315-319 (2022).

[9] Naidu, T. P., et al. "A Hybridized Model for the Prediction of Heart Disease using ML Algorithms," 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 256-261 (2021).

[10] Hong Fan. (2013). Research on the Application of Bayesian Methods in Medical Diagnostic Systems [Doctoral dissertation]. Sichuan: University of Electronic Science and Technology of China. DOI: 10.7666/d.D768282.