Heart Disease Prediction Using Naïve Bayes and SVM

Shubham Pakale
Information Technology

Pimpri Chinchwad College Of Engineering Pune
Pune,India

shubhampakale4992@gmail.com

Vaishnavi Jadhav
Information Technology

Pimpri Chinchwad College Of Engineering Pune
Pune,India

jadhavvaishnavi3107@gmail.com

Nivedita Birajdar
Information Technology

Pimpri Chinchwad College Of Engineering Pune
Pune,India

niveditavb05@gmail.com

Roshni Raut

Information Technology

Pimpri Chinchwad College Of Engineering Pune
Pune,India

roshani.raut@pccoepune.org

***Abstract*— The burden of heart disease can be significantly decreased via early detection and prevention, despite the fact that it is a major cause of morbidity and mortality worldwide. In this article, we present a heart disease prediction model employing Support Vector Machine and Naive Bayes are two popular machine learning techniques (SVM).A sizable dataset of patient records, which includes demographic, clinical, and lifestyle characteristics, is used to train the model. We first preprocess the data and select the most relevant features using feature selection techniques. We next train and test our model using Naive Bayes and SVM methods, and we evaluate the performance of each approach using a wide variety of measures as accuracy, precision, recall, and F1-score. According to our findings, both Naive Bayes and SVM models are highly accurate at predicting heart disease, with SVM slightly outperforming Naive Bayes..According to our study, healthcare practitioners may find the proposed heart disease prediction model to be a useful tool for identifying individuals who are at a high risk of acquiring the condition and for providing prompt therapies to enhance patient outcomes.**

***Keywords*—Naive Bayes, Support Vector Machine, Machine learning,Classification,Coronary artery disease (CAD)**

I. INTRODUCTION

More people die from cardiovascular problems than from any other cause. Cardiovascular diseases are more likely compared to any other condition to result in death. A World Health Organization research claims that it causes roughly one in every three fatalities globally each year. All of our body's other organs will stop functioning normally if our heart stops beating normally. Many nations, including Bangladesh, are seeing an increase in the mortality rate due to cardiovascular disease. The World Health Organization estimates that cardiovascular disease kills 17.9 million people annually, accounting for 31% of all fatalities globally, and that it was to blame for 14.31% of all mortality in Bangladesh in 2017 [1].

Information from hidden patterns in datasets can be extracted via data mining in the medical field. For now, everything medical information that we receive comes on paper. This data orientation necessitates the use of an embodied structure. To obtain a more accurate result, data must be pre-processed before applying machine learning algorithms. This extractive data will aid in the prediction of medical diagnoses by using data mining techniques. The future prediction-based technique will also help clinicians take the proper actions to treat the patient on time by drawing on prior patterns in the dataset. The reliable prediction of disease is made possible by methods of data mining and forecasting models.[2]

Finding the best technique to forecast cardiovascular disease is the aim of this study. Heart disease is another of the main killers in our country. The need of educating individuals about cardiovascular disease risk factors cannot be overstated.

The remainder of the text is organised as follows. In Part II, a summary of existing literature on Naive Bayes and SVM-based heart disease predictions is provided. Section III describes the methodology which has been used in our study. The performance of the Naive Bayes and SVM algorithms is compared in Section IV along with the findings of our investigation. The project is over, and Section V discusses potential future research topics.

II. RELATEDWORK/LITERATURE SURVEY

The largest cause of preventable mortality worldwide among non-communicable diseases is now heart disease, which has recently been expanding at an accelerated rate. The results suggest that countries in South Asia, in particular, may be more susceptible to heart disease. Heart disease is difficult to anticipate, and only doctors with extensive experience have the knowledge needed to do so. We have access to a lot of data that can reveal hidden facts, therefore we can use data analysis techniques to make some useful decisions in this field. This would relieve some of the load on the medical community while also giving us a dependable way to forecast cardiac disease.[3]

For a very long time, diagnosing and predicting heart disease has been a crucial and difficult duty for medical practitioners. Hospitals and other organisations provide costly medicines and procedures to address heart diseases. So, being able to recognise cardiac disease in its early phases would allow individuals all around the globe to take the essential actions before it worsens. The main contributors to heart disease, a serious issue in recent years, are alcoholism, smoking, and inactivity. Machine learning has been used to successfully make choices and forecasts from the massive volumes of data generated by the healthcare sector over time. [4]

Hearts are important to all living things. Heart-related disorders demand more precision, correctness, and accuracy in diagnosis and prognosis since even a small mistake might lead to tiredness or even death. Heart-related deaths are common, and the number of these deaths are rising rapidly. A method for illness awareness prediction is necessary to address the issue. Using information from real-world events, machine learning, a subset of artificial intelligence (AI), provides exceptional aid in forming predictions about any type of event. In this study, we use the UCI source dataset for both training and testing to determine the precision of the support vector machine (SVM) and naive bayes techniques for heart disease prediction. [5]

The risk of developing cardiovascular disease can be quickly determined by a patient's vascular resistance without the requirement of time-consuming blood tests (CVD). The patient's CVD risk can be assessed by a quick measurement of their digital volume pulse, which is the volume pulse at the point of their finger, utilising the infrared light absorption detector put on their index finger. A support vector machine (SVM) algorithm has been demonstrated to give an accurate (>85%) forecast of either high or low vascular resistance as revealed by the aortal pulse wave speed if the right attributes are retrieved from the waveform (PWV). The unique method offers potential as a tool to help doctors prevent cardiovascular problems because doing so would ordinarily need a time-consuming and difficult operation. [6]

III. METHODOLOGY

The data used for this problem is a UCI Machine Learning Repository Multivariate Heart Disease Data Collection.[11] Even though the set of data data set contains 76 features, all published research only mention employing a subset of 14 of these. The Cleveland database in particular is the only one that ML researchers have utilised up until this point. The "target" field makes reference to the patient's heart condition.

Table 1.0 provides a detailed overview of the dataset's characteristics.

Attribute	Description	Type
Age	Patient's age in years	numerical
Sex	sex (1 = male; 0 = female)	nominal/binomial
cp	Chest pain type(1=typical angina,2=atypical angina,3=non-anginal pain,4=asymptotic)	nominal
trestps	Resting blood pressure	numerical
chol	Serum cholestrol in mg/dl	numerical
fbs	(fasting blood sugar>120 mg/dl)(1=true,0=false)	nominal/binomial
restecg	resting electrocardiographic results (normal; abnormal; ventricular hypertrophy)	nominal
thalach	maximum heart rate achieved	numerical
exang	exercise induced angina (1 = yes; 0 = no)	nominal/binomial
oldpeak	ST depression induced by exercise relative to rest	numerical
slope	the slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = downsloping)	nominal
ca	number of major vessels colored by fluoroscopy (0 = mild; 1 = moderate; 3 = severe)	nominal
thal	Status of the heart (1 = normal; 2 = fixed defect; 3 = reversible defect)	nominal
target	(1 = heart disease; 0 = healthy)	nominal/binomial

1. ALGORITHMS

A. Naïve Bayes:

Naive Bayes is a popular and widely used classification algorithm in machine learning. It is a probabilistic approach based on the Bayes theorem that calculates the probability of a situation based on information about potential confounding variables. For applications such as text categorization, spam detection, sentiment analysis, and many others, the Naive Bayes algorithm is a straightforward and effective choice.

The assumption that almost all features are autonomous of one another is what makes the algorithm "naive". This means that the algorithm assumes that the presence or absence of one feature does not affect the probability of any other feature being present or absent. This assumption simplifies the calculation of probabilities, making Naive Bayes fast and efficient, but it can also lead to inaccuracies in cases where the features are not truly independent.

A given collection of features' probabilities for each class are calculated using Naive Bayes, and the category with the greatest chance is chosen as the class label. Using Bayes' theorem, the algorithm determines the likelihood of each class depending on the likelihood for every feature given

the class. The algorithm is trained on a labeled dataset, where each instance is labeled with its corresponding class.

Naive Bayes is an all-around effective and powerful algorithm that may be applied to a variety of classification applications. It is easy to implement, requires minimal training data, and can handle large datasets. However, the assumption of feature independence can lead to inaccuracies in some cases, and other algorithms may perform better on certain types of data.

B. Support Vector Machine:

SVM are a popular and successful classification method in machine learning. Finding a hyperplane which divides the data into distinct groups is the foundation of SVMs. The hyperplane is selected to optimise the distance between it and the closest data points in each class.

SVMs function by projecting the input data into a more complex space, where it is simpler to locate a separating hyperplane. A kernel function is used in this process to determine how comparable two sets of data points are in a higher-dimensional space. SVMs are capable of handling non-linear decision boundaries, making them suitable for a wide range of classification tasks.

SVMs are trained on a labeled dataset, where each instance is labeled with its corresponding class. The method learns the ideal values for the hyperplane's defining parameters, such as the slope and intercept, during training. Finding the hyperplane with the highest margin between classes and the lowest classification error is the objective.

SVMs are an effective and adaptable method that may be applied to a variety of classification jobs. They are particularly useful for high-dimensional data and non-linear decision boundaries. However, SVMs can be sensitive to the choice of kernel function and the parameters that define the hyperplane, and may require careful tuning to achieve optimal performance.

- *Using Naïve Bayes*

  Experimental Setup of Naïve Bayes for Heart Disease Prediction

  A probabilistic ML method called the Naive Bayes method is employed to address classification problems. It's especially handy for forecasting the likelihood of an occurrence based on a collection of characteristics.

  1. Data Collection: Gathering pertinent information on the numerous risk factors for heart disease. Data on blood pressure, cholesterol, smoking status, age and gender, as well as other pertinent medical details, may be provided.

  2. Data Preparation is the process of preprocessing and cleaning data so that it may be used in the Naive Bayes algorithm. This might include activities like eliminating missing values, standardising data, and converting category data to numerical data.

  3. The practise of splitting information into testing and training sets is known as data splitting. Naive Bayes model is trained using the training set, and its performance is evaluated using the testing set.

  4. Model Training: Using the training data, train the Naive Bayes model. Given the characteristics, the Naive Bayes method computes the probabilities of each feature and the probability of each class (heart disease or no heart disease).

  5. Model testing is the process of evaluating the performance of a trained model on testing data. The accuracy of the model is evaluated using metrics such as precision, recall, and F1-score.

  6. Model improvement is the process of improving the performance of the Naive Bayes model by adjusting hyperparameters or employing new feature selection approaches.

  7. Prediction: Utilising the trained Naive Bayes model to determine a patient's risk factors and determine whether or not they have heart disease.

- *Using SVM*

  Experimental Setup of SVM for Heart Disease Prediction

  A machine learning technique called SVM is used for categorization jobs like predicting diseases. The experimental setup for SVM-based disease prediction typically involves the following steps:

  1. Gather information from patients with and without the illness. Demographic information, medical history, clinical symptoms, and laboratory test results are all examples of data.

  2. Clean and preprocess the obtained data to eliminate unnecessary or missing information, normalise the features, and guarantee that the data is acceptable for the SVM algorithm.

  3. Using approaches such as correlation analysis or feature ranking, identify the most significant traits that can distinguish between patients with and without the condition.

  4. Use labelled data, i.e., data with known outcomes, to train the SVM algorithm on the specified features. The SVM method seeks to find the optimum dividing line, known as the decision boundary or hyperplane, that can divide a given dataset into two different groups.

  5. The effectiveness of a model's predictive capabilities can be evaluated using performance metrics such as accuracy, sensitivity, specificity, and area under the curve.

  6. Test the trained SVM model's generalizability and adaptability to fresh patient data using an independent dataset.

  7. Use the trained SVM model in clinical practise to forecast and monitor illness progression.

2. System Design:

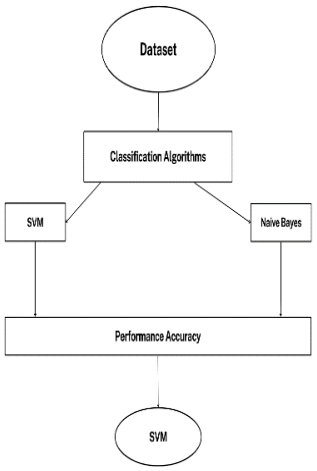

Fig.1. A Heart Disease Prediction Machine Learning Algorithm

3. Model Evaluation

The confusion matrix is a popular andappropriatetool for assessing a classifier. The confusion matrix is a table that compares predicted and actual values to describe the performance of a classifier. Table 2 depicts a more general form of the confusion matrix

Table 2.General Confusion Matrix

	Predicted class 1	Predicted class 2
Actual class 1	True Positive	False Negative
Actual class 2	False Positive	True Negative

This method enables us to obtain overall assessment measures, and they encompass the following parameters:

Accuracy: It is a statistic used to assess a classifier's overall accuracy. It indicates how well the classifier anticipated the outcome. Eq. 1 gives the accuracy of a classifier.

$$accuracy(\mathrm{a}) = \frac{tp + tn}{tp + tn + fp + fn}(1)$$

Precision: It is a statistic for determining the precision of a model's positive predictions. In layman's words. Precision refers to the model's accuracy in making positive predictions, which is the fraction of correct positive predictions produced out of all positive forecasts made. The precision formula is provided in eq. 2.

$$precision(p) = \frac{tp}{tp + fp}$$

Recall: It is a statistic that is used to evaluate a model's ability to recognise all positive cases. It shows the proportion of accurate positive predictions generated by the model out of all real positive data points. The recall formula is provided by eq. 3.

$$recall(r) = \frac{tp}{tp + fp}$$

F1-Score:It is a statistic used to assess the accuracy of a model that combines precision and recall into a single score. This score is determined by averaging accuracy with recall, with such a maximum score of 1 as well as a minimum value of 0. The model's performance is assessed using the FI score in order to determine the best combination of accuracy and recall. It is given by equation 4. F1 grade Equals=

$$\text{F1 Score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}(4)$$

IV. RESULT

In this section results are analyzed based on the features passed to them

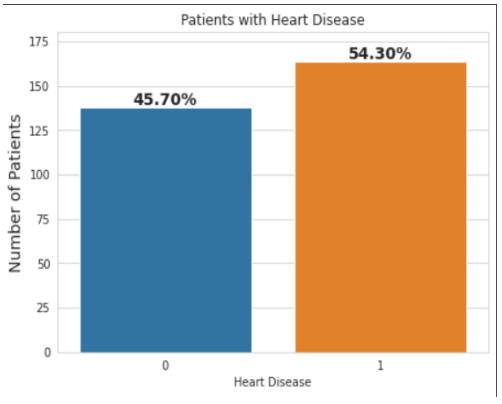

Figure 2..Exploratory Data Analysis

A. AccuracyMeasure:

The performance of classification algorithms can be compared across different metrics as shown in table 3 and 4

Table 3.accuracy measures of naive bayes

	Precision	Recall	f1-score	support
0	0.84	0.80	0.82	138
1	0.84	0.87	0.85	164
accuracy			0.84	302
Macro avg	0.84	0.83	0.84	302
Weighted avg	0.84	0.84	0.84	302

Table 4.accuracy measures of SVM

	Precision	Recall	f1-score	support
0	0.91	0.78	0.84	138
1	0.83	0.94	0.88	164
accuracy			0.86	302
Macro avg	0.87	0.86	0.86	302
Weighted avg	0.87	0.86	0.86	302

V. CONCLUSION

Early identification and prevention are crucial to improving patient outcomes since heart disease is a significant cause of mortality globally. Machine learning algorithms such Naive Bayes and SVM have already been recommended as useful methods for predicting the likelihood of coronary artery disease using medical, demographic, and lifestyle data.

A potential method for the early identification and avoidance of heart disease is to employ Naive Bayes and Svm classifiers for heart disease prediction. The combination of these algorithms with proper data preprocessing and feature selection can provide accurate predictions and aid healthcare professionals in identifying high-risk patients and taking necessary preventive measures.

For future work, more complex ML algorithms should be created to improve disease prediction efficiency. For improved performance, the calibration of the learning models must be done more regularly following the training period. To reduce overfitting and improve accuracy, the datasets should also be broadened to cover more diverse demographics. The effectiveness of learning models should also be improved by the use of pertinent feature selection techniques. Once the disease is predicted, the required medical resources could bemanaged efficiently, resulting in lower costs for treating the disease.

VI. REFERENCES

[1] *Cardiovascular Diseases (CVDs)*. 11 June 2021, www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] Wu, Ching-seh Mike, Mustafa Badshah, and Vishwa Bhagwat, "Heart Disease Prediction Using Data Mining Techniques." In Proceedings of the 2019 2nd International Conference on Data Science and Information Technology, pp. 7-11. 2019.

[3]S. R. Alty, S. C. Millasseau, P. J. Chowienczyc and A. Jakobsson, "Cardiovascular disease prediction using SVM," 2003 46th Midwest Symposium on Circuits and Systems, Cairo, Egypt, 2003, pp. 376-379 Vol. 1, doi: 10.1109/MWSCAS.2003.1562297.

[4]A. Singh and R. Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," 2020 International Conference on Electrical and Electronics Engineering (ICE3), Gorakhpur, India, 2020, pp. 452-457, doi: 10.1109/ICE348803.2020.9122958.

[5]R. Katarya and P. Srinivas, "Predicting Heart Disease at Early Stages using Machine Learning: A Survey," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 302-305, doi: 10.1109/ICESC48915.2020.9155586.

[6]D. K. Chohan and D. C. Dobhal, "A Comparison Based Study of Supervised Machine Learning Algorithms for Prediction of Heart Disease," 2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), Greater Noida, India, 2022, pp. 372-375, doi: 10.1109/CISES54857.2022.9844328.

[7] CORTES,C.,&VLADIMIRVAPNIK.((1995)).SupportVectorNetworks.

[8] Breiman, L., et al. "Classification and regression trees(Wadsworth,Belmont,ca,1984)."ProceedingsoftheThirteenth InternationalConference,Bari,Italy.1996

.

[9] Quinlan,J.Ross."Combininginstance-basedandmodel-basedlearning."Proceedingsofthetenthinternationalconferenceonmachine learning. 1993.

[10]Li, J., Chen, Y.,, & Wang, B. (2016). "An efficientmulticlassSVMalgorithmfortextcategorization".InternationalJournalofComputationalIntelligenceSystems

[11]https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset