IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 25, Issue 5, Ser. 2 (Sept. – October. 2023), PP 52-57
www.iosrjournals.org

Determining The Accuracy Of NaïVe-Bayes Algorithm And Multi-Layer Perceptor In Predicting Heart Diseases

Sarthak Jain
Keshav Jain
Prince Jhalani
Divyannsh Pincha

Abstract
Advancements in healthcare techniques have increased lethal diseases, particularly heart-related issues, which have become leading causes of disability and premature death. Coronary heart disease, affecting the heart, blood vessels, and circulation, has become a global concern. Despite efforts to develop effective diagnostic and preventive methods, its complexity remains a challenge. Single classification-based data mining systems are insufficient, leading to ongoing research for better methods.
Machine learning algorithms have shown promise in predicting heart diseases using medical data. Multi-layer Perceptron and Naïve-Bayes algorithms are popular due to their efficiency and ease of use. However, they have limitations.

Date of Submission: 13-10-2023 Date of Acceptance: 23-10-2023

I. Analysis of Primary Research

I would conduct my primary research by gathering essential patient information, including age, gender, blood pressure, and other pertinent factors contributing to the likelihood of heart disease. These valuable data are acquired through interviews with medical professionals and careful analysis of medical records.

II. Analysis for Secondary Research

Through secondary research, this paper delves into previous literature reviews and research papers that focus on the application of machine learning algorithms in predicting heart diseases. Additionally, it draws upon the UCI library, a valuable machine-learning repository housing datasets curated for analytical purposes. This resource proves instrumental in evaluating the performance of Naive Bayes and MLP algorithms, shedding light on their respective strengths and weaknesses. This research holds major significance, as it addresses critical issues related to heart diseases. The utilization of data mining techniques in medical classification methods empowers researchers to accurately predict such conditions, enabling timely interventions for improved health outcomes. The primary aim of this research paper is to ascertain the extent to which the Naïve-Bayes Algorithm and Multilayer Perceptron Algorithm, can accurately predict heart diseases. To address this inquiry, each individual model's performance will be meticulously assessed and analyzed.

III. Methodology

1) Data collection-The primary phase of this study necessitates gathering the necessary data in order to create accurate models. The dataset is required to cover a range of risk factors including age group distribution, gender representation along with corresponding cholesterol levels and blood pressure values. Furthermore. It must also encompass an assessment of smoking habits as well as any familial precedent related to cardiovascular conditions.

2)Pre-Processing-In order to address missing values, normalize the data and convert categorical data to numerical data. Pre-processing will be conducted on the collected data. The following techniques will be utilized: a) The dataset will be divided into two parts, namely the train and test. Using the 10-fold cross-validation method. b)To determine the most suitable feature sets we will utilize two methods: Correlation-based Feature Selection (CFS) and Best First Search (BFS)10. CFS focuses on maximizing the correlation between the class variable and the selected features in order to assess their relevance. In contrast, BFS operates by analyzing the subset of features and selectively adding or removing them based on their individual predictive power.

3)Naive Bayes-We will implement the naive Bayes algorithm and train it using the training data. However, we will evaluate its performance using the testing data.
4)Machine learning- We will implement the MLP algorithm and determine its architecture based on the number of inputs and hidden layers. Similar to the naive Bayes algorithm. We will train it using the training data and evaluate it with the testing data.
5)Evaluation Matrix-To assess the performance of the model we will measure accuracy using both the confusion matrix and Root Mean Square Error (RMSE).
6)Results-The results of the models will be compared and analyzed in order to determine which model offers the most accurate prediction of heart disease.

IV. Background information

Ensuring timely detection, precise diagnosis, and effective treatment methods are paramount in combating the surge of individuals afflicted by heart diseases. The advancement of medical technology has bestowed computer science with a pivotal role in accurately predicting and treating these conditions. Although certain risk factors associated with heart disease can be addressed through curative measures. Some are beyond our reach for a complete cure. According to the American Heart Association (AHA). Age, genealogy details, tobacco consumption, high cholesterol levels, hypertension issues, obesity, and diabetes serve as key elements that elevate the likelihood of developing a cardiovascular condition.

Overview of techniques used

Numerous classical algorithms can be used for disease prediction, which involves accurately identifying connections or patterns across diverse fields in vast databases. Regularly updating databases and employing machine learning is essential to achieve this goal. Machine learning encompasses two primary techniques: supervised and unsupervised methods.

Supervised Learning Technique

Supervised learning is a widely utilized approach in data mining. It involves training a system to predict class labels for new instances based on provided data. This method, also known as a supervisory signal, utilizes training examples that include input objects and desired output values. By analyzing the underlying data patterns a deduced equation is generated, which is referred to as a classifier if the output is categorical or a regression equation if the output is numerical.

Unsupervised learning, an indispensable machine learning approach, entails the discernment of intricate patterns and structures within unlabeled data. In stark contrast to supervised learning, where predefined outputs or targets exist, unsupervised learning leverages algorithms to unearth relationships and underlying structures inherent in the data. This technique closely mimics the cognitive processes of the human brain, striving to derive meaningful representations from input patterns. Rather than relying on training data, unsupervised learning orchestrates the grouping of data into clusters, a technique often denoted as "cluster analysis." Particularly valuable in scenarios devoid of labelled data, unsupervised learning has exhibited remarkable prowess when it comes to forecasting fault-prone systems, surpassing the performance of supervised learning.

The machine learning methodologies of Naïve-Bayes and Multilayer Perceptron have experienced a surge in prominence within the medical domain, specifically in the context of predicting cardiovascular ailments. Naïve-Bayes, an algorithm rooted in probability theory, employs Bayes' theorem to ascertain the probabilities associated with assigning a given data point to a particular class. Conversely, the Multilayer Perceptron (MLP) represents a neural network variant capable of comprehending intricate correlations between input and output data.

Naive Bayes applies Bayesian inference, using prior and conditional probabilities, to estimate class likelihood based on input features for accurate predictions in classification tasks.

The formula below is used to determine the probability of the hypothesis ‘W’ provided the evidence ‘Z’:-

$$P(W|Z) = \frac{P(Z|W)P(W)}{P(Z)}$$

Where:

- P(W|Z) → probability of the hypothesis W given the evidence Z
- P(W|Z) → probability of the evidence Z given the hypothesis W
- P(W) → prior probability of the hypothesis W
- P(Z) → prior probability of the evidence Z

Here are the procedural steps involved in the functioning of the Naive Bayesian classifier:

1. **Data preprocessing:** The classifier necessitates the collection and preprocessing of the data set, which is subsequently partitioned into training and testing sets. Additionally, appropriate class labels are assigned to the data.
2. **Feature extraction**: The extraction of pertinent features from the data takes place, enabling subsequent predictions. This process may involve the selection of highly relevant features or transforming the data into a more suitable format.
3. **Probability estimation:** The classifier proceeds to estimate the probabilities associated with each class based on the data features. It also calculates the probabilities of each feature given each class. This estimation is commonly achieved by counting the occurrences of each feature within each class and dividing it by the total number of instances in that class.
4. **Prediction:** When presented with a novel, unseen instance, the classifier computes the probabilities of each class given the instance's features. Subsequently, it assigns the class with the highest probability to that instance.
5. **Evaluation**: Finally, the classifier's performance is assessed through the utilization of evaluation metrics such as accuracy and precision.

The Multilayer Perceptron (MLP) is a prominent artificial neural network in machine learning. It features feedforward data flow through interconnected layers for computations, trained using backpropagation.

The formula for computing the output of a single neuron in the MLP is:-

$$z = \sum \quad w_i x_i + b$$

Where:-
- Z is the output of the neuron
- b is the bias term (a constant value added to the weighted sum of inputs)
- w_i is the weight associated with each input
- x_i is the input value

This formula can be expressed in vector form as:-

$$z = b + w^T \times x$$

Where w^T is the transpose of the weight vector w.

The Multi-Layer Perceptron (MLP) operates by training a neural network to acquire a functional relationship between input data and corresponding outputs. The underlying process entails a series of fundamental procedures within the MLP algorithm:

1. **Input Data:** The data is typically structured into feature vectors, which represent numerical representations of the inputs.

2. **Network Architecture**: Determining the appropriate number of layers, neurons within each layer, and activation functions to employ.

3. **Initialization**: Initializing the network's weights and biases with small random values.

4. **Forward Propagation**: The network receives input data, and the output of each neuron is computed using a defined formula. The outputs of each layer are then transmitted as inputs to the subsequent layer until the output layer is reached.

5. **Activation Function**: Each neuron's output undergoes an activation function, such as sigmoid or ReLU, to introduce non-linearity into the model.

6. **Error Calculation**: The predicted network outputs are compared to the actual outputs, and the ensuing error between the two is computed. Various metrics, such as mean squared error, can be utilized to quantify the error.

7. **Backpropagation**: Employing the chain rule of differentiation, the error is retroactively propagated throughout the network. This necessitates calculating the gradient of the error with respect to the weights and biases of each neuron in the network.

8. **Weight Update**: Leveraging an optimization algorithm to update the network's weights and biases, thereby minimizing the error.

9. **Iteration**: Steps 4-8 are iteratively repeated for a defined number of iterations or until the error surpasses a predetermined threshold.

10. **Prediction**: Once the network is trained, it can be deployed to generate predictions for novel input data. This is accomplished by feeding the data through the network and retrieving the output from the output layer.

Measuring Accuracy

Confusion Matrix

In order to evaluate the performance of the algorithms. This study employs a widely used methodology called the confusion matrix. The confusion matrix presented in the table below demonstrates a two-class classification:-

		Actual Class	
		Positive (P)	Negative (N)
Predicted Class	Positive (P)	True Positive (TP)	Fake Positive (FP)
	Negative (N)	False Negative (FN)	True Negative (TN)

Figure 1- Confusion Matrix Table- www.ml-science.com/confusion-matrix

The values predicted by the models are displayed in the columns on the left most side while the correct predictions are shown on the diagonal. The matrix usually consists of four entries and is crucial for analyzing the ability of the classifier.

To determine the models' performance accuracy is calculated based on the equation provided below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

When evaluating a model, accuracy pertains to how many correct predictions it produces in relation to all predictions made. To derive these predictions, we shall employ the aforementioned formula and consider four essential metrics explained as follows:

1. True Positive (TP) refers to cases where the model correctly identifies the positive class by accurately predicting it as positive.
2. True Negative (TN) signifies instances where the model correctly identifies the negative class by accurately predicting it as negative.
3. False Positive (FP) pertains to situations where the model incorrectly predicts a positive class. Despite belonging to the negative class.
4. False Negative (FN) represents cases where the model incorrectly predicts a negative class. Even though it actually belongs to the positive class.

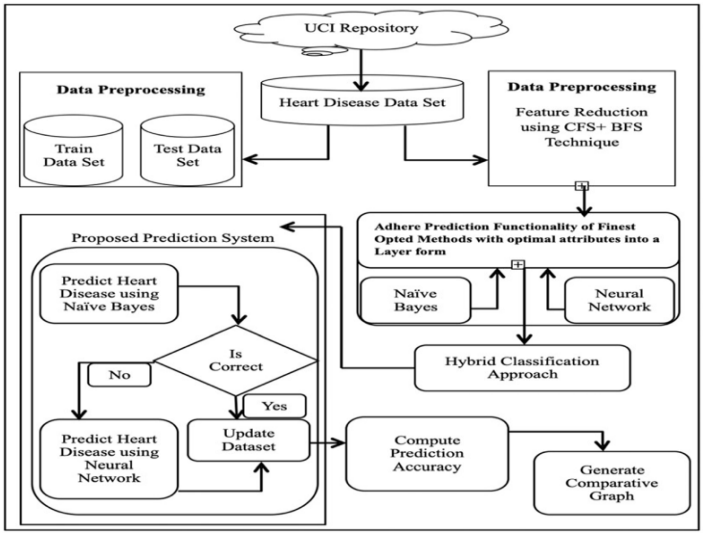
Root Mean Square Error (RMSE)

In the realm of evaluating regression models, the root mean square error (RMSE) holds a prominent position as a metric. Its utmost usefulness arises when contrasting various types of models. Calculating RMSE necessitates implementing the formula presented hereafter:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (Predicted_i - Actual_i)^2}{n}}$$

Predicted and observed data are crucial for unbiased assessments. Predictions stem from models, while observations come from real-life measurements. Calculating squared deviations between them and deriving a mean provides valuable insights. Evaluating RMSE, with its scale-independent properties, indicates a model's predictive power; lower values signify better data matching.

Figure 2- Demonstration of the overall working of the model



V. Data Collection

Standardization

Standardization involves transforming the data into a common scale and distribution to ensure its suitability for modelling. This process presented challenges, including data cleaning to remove missing, invalid values, and duplicates. Normalization was a crucial step, harmonizing data scales and converting categorical variables like gender into numerical values. Additionally, binary variables were created for each category. Finally, the dataset was split into test, training, and validation sets.

Data Set Used

To assess the efficiency of the models, the same dataset employed for Naive-Bayes and MLP analysis was used. The dataset's attributes were utilized by both techniques for accurate data prediction. For consistency, the same parameters as in other analyses were employed. The dataset selected from the UCI library was the Staglog (Heart) dataset, featuring 270 observations, 14 attributes, and 2 classes: presence and absence of heart disease.

VI. Evaluation and Analysis

A Java-based application was used for the comparative evaluation of algorithms due to its robust programming language. Additionally, WEKA, an open-source machine learning software tool, was employed to apply the models to the datasets (the code used is included in the appendix).

In this experiment, the evaluation focused on the Staglog dataset, which comprises 270 instances and 14 attributes. Before commencing the execution process, data pre-processing was applied to ensure accurate results. The goal of dataset pre-processing is to prepare it for a reliable experiment, as online-accessed data can contain errors such as incompleteness, noise, inconsistency, and missing attribute values.

Essentially, each record in the dataset indicates whether an individual has heart disease or not. This dataset includes records of 120 individuals with heart disease and 150 entries for individuals without heart disease. The table below (Figure 3) presents the complete set of attributes associated with the Staglog dataset on the left side and, on the right side, illustrates the selected attributes used in the proposed approach for data prediction purposes.

S. No.	Attribute Name	S. No.	Selected Attribute Name
1.	Age	1.	Chest
2.	Sex	2.	Thal
3.	Chest	3.	Oldpeak
4.	Resting_Blood_Pressure		

5.	Serum_cholestrole		
6.	Fasting_blood_sugar		
7.	Number_of_major_vessels		
8.	Thal		
9.	Excercise_induced_agina		
10.	Oldpeak		

The table clearly demonstrates that, for data prediction purposes, the designed approach utilized only 3 attributes instead of the original dataset's 10 attributes. Following the attribute selection process, the dataset underwent evaluation using both hybrid and standalone prediction techniques. The Naive-Bayes algorithm was employed to assess the dataset and calculate the probability of heart disease based on symptoms and risk factors. The results indicated that the Naive-Bayes algorithm achieved accurate predictions for heart disease in the majority of cases. Specifically, the algorithm correctly classified 226 instances and misclassified 44 instances, resulting in a confusion matrix accuracy of 83.7037% (refer to Figure 4) and an RMSE of 0.35.

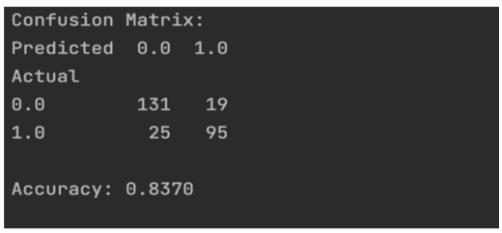


Figure 4- Confusion Matrix table for Naive-Bayes Algorithm

In contrast, the MLP algorithm employed a more sophisticated modelling approach to capture the relationships between inputs and outputs. However, the results of the MLP algorithm did not achieve higher accuracy in prediction compared to the Naïve-Bayes Algorithm. Specifically, MLP correctly classified 211 instances and misclassified 59 instances, resulting in a confusion matrix accuracy of 78.1481% (see Figure-5) and an RMSE of 0.4295.

The efficiency and accuracy of each algorithm are presented below in a table and graphical form

S. No.	Prediction Techniques	Accuracy
1.	Naive-Bayes	83.70%
2.	Multi-Level Perceptron	78.14%

VII. Conclusion-

In summary, accurately predicting heart diseases demands a profound understanding of medical concepts and statistical principles. This essay delved into two distinct algorithms, Naïve Bayes and Multi-Level Perceptron (MLP), and culminated in the creation of a powerful combined model. Naïve Bayes demonstrated simplicity and efficacy, relying on symptom occurrence probabilities and risk factor analysis. In contrast, MLP showcased complexity, leveraging neural networks to unravel intricate relationships. In conclusion, machine learning algorithms hold significant potential in accurate heart disease prediction, aiding early diagnosis and treatment. Further research is warranted to validate and explore alternative algorithms in this context.