International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 04, April-2023

Prediction of Heart Disease by Machine Learning

A. Mounika Rajeswari
(Asst. Professor)
Department of Computer Science & Engineering

D. V. Sathwik Reddy
(B.Tech Computer Science & Engineering)
CMR College of Engineering & Technology
Hyderabad, Telangana

***Abstract*— One of the leading causes of death worldwide is heart disease. Medical professionals find it tough to forecast because it is a complex task that calls both experience and advanced knowledge. An automated system for diagnosing illnesses might improve medical effectiveness while simultaneously cutting costs. Based on the parameters provided regarding the patients' health, we will create a system that can effectively identify the rules to estimate the risk level of the patients. The aim is to anticipate the presence of heart illness in patients where the presence is evaluated on a scale and to identify hidden patterns by applying data mining techniques, which are notable to heart disorders. Large amounts of data that are too complex and numerous to collect and analyze using traditional methods are needed for the prediction of cardiac disease. Our goal is to identify a machine learning method that can accurately forecast cardiac disease while also being computationally efficient. Data mining is a technique for extracting hidden patterns and relationships from sizable databases by combining statistical analysis, machine learning, and database technology. A variety of data mining strategies are tested using the Cleveland heart disorders data set from the University of California, Irvine (UCI) machine learning repository.**

Keywords*— *Machine Learning (ML), Support Vector Machines (SVM), Supervised Learning

1. INTRODUCTION

A. Basics and Foundations

In the entire world, heart disease is regarded as one of the leading causes of death. Medical professionals find it tough to forecast because it is a complex task that calls both experience and advanced knowledge. An automated system for diagnosing illnesses might improve medical effectiveness while simultaneously cutting costs. Based on the provided parameters regarding the patients' health, we will create a system that can effectively identify the rules to estimate the risk level of the patients. In order to forecast the presence of heart disease in patients where the presence is valued on a scale, data mining techniques are used to uncover hidden patterns that represent notable heart disorders.

B. Literature Review

Mohan et al. described a way to find hidden facts for effective decision-making. Finding underlying linkages and patterns is frequently untapped. Advanced ML methods can help to solve this problem. Several models that can be used for prediction are presented as the research's conclusion.

Repaka et al. described the prediction performance for two categorization models, which was examined and contrasted with earlier work. Experimental findings demonstrate that our suggested strategy has a higher accuracy % for risk prediction than other works.

Using a variety of data mining approaches, Gavhane et al. solves the problem of predicting heart disease based on input attributes and presents the results with their accuracy in tabular format. It proposes to create a program that, given basic symptoms like age, sex, pulse rate, etc., may forecast the vulnerability of a cardiac condition.

The suggested system uses the machine learning algorithm neural networks because it has been shown to be the most accurate and dependable algorithm.

Krishnan et al. foresees the potential for heart disease. The results of this system provide a percentage likelihood of developing heart disease. Medical parameters are utilized to categorize the datasets. This system uses a data mining classification algorithm to analyze such parameters. The datasets are analyzed using the four main machine learning algorithms, Decision Tree, Logistic Regression, Support Vector Machine, and Naive Bayes Algorithm, in the Python programming language. These algorithms are shown to have the highest accuracy level for heart disease among these two.

2. PROPOSED SYSTEM

It is a web-based machine learning tool that was developed using data from UCI. The user enters their unique medical information to receive a forecast of heart disease. The algorithm will determine the likelihood that cardiac disease is present. The outcome will be shown directly on the website. reducing the expense and time involved in disease prediction.

The format of the data is important in this application. The user data upload application will examine the file format to ensure it is appropriate, and if it is not, an ERROR dialogue box will be prompted.

The following four algorithms will be assigned to use:

- Support Vector Machine (SVM)
- Decision Tree
- Naïve Bayes Algorithm
- Logistic Regression

These algorithms' operation has been described in the parts that follow.

The University of California, Irvine provided a data set that was used to train the algorithms. 25% of the data set's entries were utilized to test the algorithm's accuracy, while the remaining 75% were used for training. Additionally, steps have been taken to optimize the algorithms, increasing accuracy. The dataset must be cleaned up as well as the data must be processed. The accuracy of the algorithms was evaluated, and it was found that the SVM, with a 64.4% efficiency, was the most accurate of the three.

As a result, it was chosen for the primary application. The main application is a web application that computes the outcome after

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

receiving input from the user's various parameters. The outcome and prediction accuracy are both displayed.
Data set and user data inputs; effective disease prediction outputs.

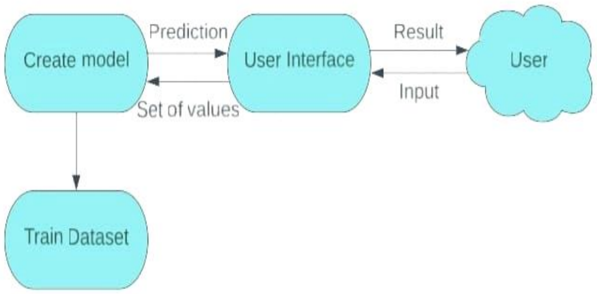

Fig. Architectural Design

PROCESS FLOW

1. Begin
2. Enter the information
a. Verify the details' format
b. Analyze the details
c. Avoid using spaces as delimiters.
3. Train Dataset
4. Using the Support Vector Machine algorithm, Decision Trees, Naive Bayes Algorithm, Logistic Regression, and others, predict the outcome.
5. Show results
6. END

3. METHODOLOGY

A. Data preprocessing

We did not receive a flawlessly accurate and error-free dataset. Therefore, we initially performed the following procedures on it:

Data Cleaning

NA values from the dataset was a huge setback for us because it significantly decreased the prediction's accuracy. As a result, we deleted the variables with NA values. We replaced it with the column's mean value. In this manner, we eliminated all NA values from the data set.

Feature scaling

Without feature scaling, several machine learning algorithms' objective functions will not perform as intended because of the broad range of values in the raw data. For instance, the vast majority of classifiers uses the Euclidean distance to compute the separation between two spots. The distance will be determined by a specific feature if it has a wide range of values among the features. To ensure that each feature contributes about equally to the final distance, the range of all features should be scaled. Therefore, we scaled the individual fields to bring their values closer to one another.

Factorization

In this section, we gave the values definitions so that the algorithm won't mix them up. Giving 0 and 1 in the age part, for instance, means would prevent the algorithm from seeing 1 as larger than 0 in that section.

B. Support vector machine

Support vector machines (SVMs) are supervised learning techniques that examine data used in regression and classification studies. An SVM training procedure produces a model that assigns new samples to one category or the other after being given a set of training data that has been designated as belonging to either one of two categories, making it a non-probabilistic binary linear classifier. An SVM model is a mapping of the examples as points in space with as much space between the examples of the various categories as possible.
Then, based on which side of the gap they fall, new samples are projected into that same area and predicted to belong to a category. Based on the hyper plane that separates the points, they are divided.

C. Decision Tree

By learning simple decision rules inferred from prior data (training data), a decision tree is used to generate a training model that may be used to predict the class or value of the target variable. In decision trees, we begin at the tree's root when anticipating a record's class label. We contrast the root attribute's values with the record's attribute. We follow the branch that corresponds to that value and go on to the next node based on the comparison.

D. Naive Bayes

The Naive Bayes family of probabilistic algorithms uses Bayes' Theorem and probability theory to predict the tag of a text (such as a news article or customer review). Because they are probabilistic, they determine the probabilities of each tag for a given text and output the tag with the highest likelihood. The Bayes Theorem, which estimates the likelihood of a feature based on prior knowledge of circumstances that may be related to that feature, is the method they use to arrive at these probabilities. When dealing with conditional probabilities, Bayes' Theorem is helpful; the following formula is used:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

E. Logistic Regression

The algorithm for logistic regression also predicts a value using a linear equation and independent predictors. Anywhere from negative infinity to positive infinity can be the expected value. The algorithm's output must be class variable, i.e., 0-no, 1-yes. As a result, we condense the linear equation's output into the range [0, 1]. We employ the sigmoid function to compress the projected value between 0 and 1.

4. RESULT

The patient's risk of developing the disease is assessed using the heart disease prediction system. On the basis of precision in percentage factor, HDPS provides results. This percentage displays a model's accuracy for a specific set of user-provided data values. The accuracy of each algorithm is shown in the following chart. In the diagram below, the green, red, yellow, and blue pillars indicate the results of SVM, Decision Tree, Naive Bayes, and Logistic Regression, respectively. As expected, the HDPS provides a result that is bi-valued, or (Yes/No).

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

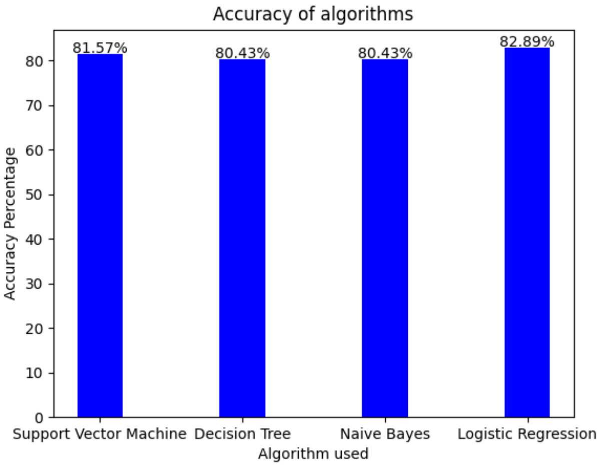

Fig. Accuracy of algorithms

5. CONCLUSION AND FUTURE WORK

The four algorithms were initially put into practice. Individual datasets were trained for each algorithm. All of them were then put to the test. On the basis of a number of factors, the most effective algorithm had to be chosen. With an accuracy of 82.89%, we discovered that the logistic regression approach is the most effective of the four. SVM had an accuracy of 81.57%, whereas Decision Tree and Naive Bayes had accuracy of 80.43% and 80.43%, respectively. In order to further implement these four methods, a better user interface was used. This was created using the Python Django framework, HTML, and CSS to create an interactive online application. And this website's application uses machine learning to create a reliable model for heart disease prediction.

Using the machine learning idea, a newly trained dataset can be utilized to create a prediction system that is even more precise. Each user can register an account, after which their heart condition can be tracked by looking at past choices to see whether it has improved or gotten worse.

REFERENCES

[1] Repaka, A. N., Ravikanti, S. D., & Franklin, R. G. (2021, January 1). Design and Implementing Heart Disease Prediction Using Naive Bayesian | Semantic Scholar. Design and Implementing Heart Disease Prediction Using Naive Bayesian | Semantic Scholar. https://www.semanticscholar.org/paper/Design-And-Implementing-Heart-Disease-Prediction-Repaka-Ravikanti/d1038f406d8662d07b4d95c22ff008f9307043c0

[2] Mohan, S., Thirumalai, C., & Srivastava, G. (2023, January 1). [PDF] Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques | Semantic Scholar. [PDF] Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques | Semantic Scholar. https://www.semanticscholar.org/paper/Effective-Heart-Disease-Prediction-Using-Hybrid-Mohan-Thirumalai/2bc3644ce4de7fce5812c1455e056649a47c1bbf

[3] J, S. K., & S, G. (2020, January 1). Prediction of Heart Disease Using Machine Learning Algorithms. | Semantic Scholar. Prediction of Heart Disease Using Machine Learning Algorithms. | Semantic Scholar. https://www.semanticscholar.org/paper/Prediction-of-Heart-Disease-Using-Machine-Learning-SanthanaKrishnan.-G./78291d0478ca3fc9d3b4abccb7b42e34b731ab58

[4] Gavhane, A., Kokkula, G., Pandya, I., & Devadkar, P. K. (2022, January 1). [PDF] Prediction of Heart Disease Using Machine Learning | Semantic Scholar. [PDF] Prediction of Heart Disease Using Machine Learning | Semantic Scholar. https://www.semanticscholar.org/paper/Prediction-of-Heart-Disease-Using-Machine-Learning-Gavhane-Kokkula/828d61fd204bebf70a506ecc0720e4287bec4fd0

[5] Kohli, P. S., & Arora, S. (2021, January 1). Application of Machine Learning in Disease Prediction | Semantic Scholar. Application of Machine Learning in Disease Prediction | Semantic Scholar. https://www.semanticscholar.org/paper/Application-of-Machine-Learning-in-Disease-Kohli-Arora/09ddf6771c7d946eef42b4fcf9fee7abec968699

[6] T, C., & Choudhary, A. (2018, January 1). Heart Disease Diagnosis using a Machine Learning Algorithm | Semantic Scholar. Heart Disease Diagnosis Using a Machine Learning Algorithm | Semantic Scholar. https://www.semanticscholar.org/paper/Heart-Disease-Diagnosis-using-a-Machine-Learning-C.-Choudhary/ac00cf5b5ec590f9953ae11681b6de0d9f576bdf

(This work is licensed under a Creative Commons Attribution 4.0 International License.)