HEART DISEASE PREDICTION USING DATA MINING CLASSIFICATION ALGORITHMS

Deepanshu Sharma and Siddhartha Chauhan

Department of Computer Science and Engineering,
NIT Hamirpur, H.P., India

ABSTRACT

Heart diseases, also referred to as "cardiovascular diseases," are a group of disorders that affect the heart. This illness can cause a heart attack, stroke, and other symptoms. After examining a few research papers on the subject, it became clear that the majority of them used a single machine learning algorithm to predict heart disease. A few of them state that they are unable to enhance their model's performance through optimization techniques. As a result of these findings, they encountered some difficulties in effectively predicting heart disease using their suggested method. In an earlier study PCA was also used, but it failed to provide considerable accuracy for such a sensitive research area, i.e., medical diagnosis. Data for this method was gathered from the "Heart Disease UCI" UCI repository, which was accessible on Kaggle. Working upon the given dataset we used various dimensionality reduction techniques, using various classifiers and found out their effectiveness. Thus, we were able to get considerably higher accuracy (98%) by using certain techniques to de-noise data (checking correlations, outliers, removing them etc.), using the MLP classifier.

KEYWORDS

PCA, MLP, cardiovascular diseases, ML, Data mining

1. INTRODUCTION

A country's prosperity is largely dependent on its citizens' health. A class of illnesses known as heart diseases affect the heart. It is also referred to as "cardiovascular diseases." This disease may narrow blood vessels, increasing the risk of heart attacks, strokes, and other issues. Heart attacks and strokes are usually acute events caused by a blockage that prevents blood flow to the heart or brain. The most common cause of this is fatty deposits building up on the inner walls of the blood vessels supplying the heart and brain. Strokes can be caused by blood clots or bleeding from brain vessels. Heart diseases also include certain other conditions such as heart states that affect the heart's muscle valves or pattern. Sometimes heart disease is "silent," not recognized until a patient presents with symptoms suggestive of an arrhythmia, heart failure, or heart attack. The main behavioral risk factors for heart disease and stroke are excessive alcohol consumption, poor eating habits, smoking, and inactivity. Individuals with behavioral risk factors may experience elevated blood pressure, blood sugar, blood cholesterol, and obesity. In primary care settings, these "intermediate risks factors" can be measured to identify patients who are more susceptible to heart attacks, strokes, heart failure, and other complications.

We call the process of finding patterns in massive amounts of data "data mining." Data mining uses a variety of algorithms, including classification for supervised learning and clustering for unsupervised learning.

David C. Wyld et al. (Eds): AISO, SIP, SOFTFM, NLDM, CRBL, BIOM, COMIT– 2024
pp. 125-132, 2024. - CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.141511

1.1. Problem Statement

ML techniques were employed to predict heart disease in earlier research studies [11]. However, instead of employing any optimization of techniques for improvement, these studies focused on the distinctive outcomes of machine learning techniques [1] [2]. Some studies [3] were unable to provide reliable accuracy results for heart disease prediction using data mining techniques because they do not offer optimization strategies for enhancing the classification model or assess the model's performance in the obtained result [11]. Therefore, in order to achieve significantly greater accuracy, we must combine specific classifiers with appropriate optimization strategies. To improve our results, we might filter our dataset using various optimization tools.

1.2. Objective

The main objective of this work is to determine whether ML classification algorithms can reliably predict the onset of heart disease. The goal of this research is to use data mining classification techniques to predict heart diseases by looking at earlier studies [11]. Lastly, to determine how to enhance the classifier's performance and accuracy.

2. LITERATURE SURVEY

Prior research [1] [2] concentrated on the distinct results of specific machine learning methods without applying any optimization strategies for improvement. Because some studies [3] do not provide optimization strategies for improving the classification model or evaluate the model's performance in the obtained result, they are unable to provide trustworthy accuracy results for heart disease prediction using data mining techniques [11]. According to the World Health Organization, this specific illness affects both the heart and the body's vascular system, which can result in a number of heart-related infections, such as hypertension, cerebrovascular disease, and coronary heart disease. [4]. Prior studies found that the risk of coronary heart disease was 34.9 percent for men, 24.2 percent for women, and 42.4 percent for women over the age of 70 [5][11]. Orthopnoea, fatigue, weariness, and inflammation in the ankles are some of the early signs of heart disease. and diagnosis includes costly procedures such as chest radiographs, echocardiograms, electrocardiograms (ECGs), etc. [6]. These days, data permeates practically every part of our existence. It is present, for instance, in social media, retail establishments, and healthcare facilities where enormous volumes of data, including sales and medical records, are collected. This data can be turned into knowledge with the help of data mining [7]. Among the many problems in the real world that data mining classification algorithms can address is heart disease. Additionally, they can be applied to any data set to predict its labelled class [7] [8].

3. METHODOLOGY

The study's objective is to assess the effectiveness of different classification algorithms by examining their performance, and it provides a framework for predicting heart diseases.

The proposed system loads the datasets and encrypts them. The data is then divided in half, with half of the sample being used for training and the other half for model evaluation [11].

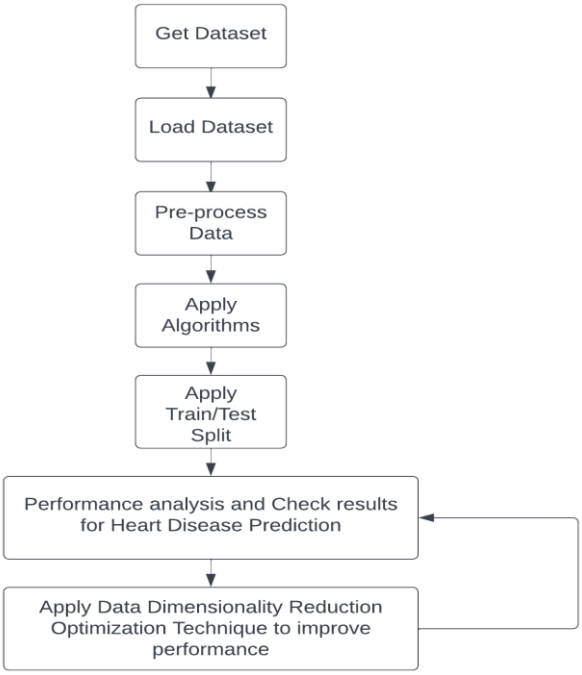

Figure 1. Proposed Work

3.1. Dataset

3.1.1. Kaggle Heart disease prediction dataset [9].
3.1.2. is made up of four databases: Long Beach V, Cleveland, Hungary, and Switzerland. Together with the anticipated attribute, it has 76 attributes [9].
3.1.3. Every published experiment refers to utilizing 14 of them as a subset [9].
3.1.4. The patient's heart condition is indicated in the "target" field. 0 indicates no disease, and 1 indicates disease [9].
3.1.5. This dataset includes 303 original data instances, 1 class label attribute (heart diease diagnosis), 13 attributes (age, gender, cp, trest bps, cholesterol, fbs, Rest ECG, Thalach, exang, old peak, slope, CA, and Thal) [9].

3.2. Loading, Preprocessing Data

After the data has been gathered, the second step entails loading the data by reading the.csv file. The.csv file was read using the pd.read command, and pd was imported as import pandas as pd. The data was then entered into the notebook for initial processing [11].

After the data has been loaded, pre-processing occurs. Datasets are extremely vulnerable because of the possibility of missing or inconsistent data due to natural error or large datasets. Preprocessing data is the way to address these issues [11]. During this stage, a variety of preprocessing methods can be employed inside the dataset. Examples include data reduction, data integration, and data cleaning. Before visualizing the target data, we need to ascertain whether it is balanced.

3.2.1. How did we de-noise data?

3.2.1.1. checked the correlation between all the features and the target.

3.2.1.2. checked the F(variance) and P(probability) values for each one of the categorical features by using ANOVA (*Basically we wanted the F value to be as high as possible, and the P value to be as low as possible). Thus fbs (fasting blood sugar) column was removed.*

3.2.1.3. checked for **outliers**, by using box plots and removed any data outside the minimum and maximum lines.

3.2.1.4. At last, finally removed those outliers (Out of the original 1025 rows, 947 still remain thus we did not lose much data).

3.2.2. Performing Train-Test Split

The dataset will be pre-processed before using train-test split. Test data must be used to make a prediction, and training data must be used to train the data. Stated differently, the performance of the model is essentially assessed using the test data.

10% of the dataset was used for testing and 90% was used for training in this study. Consequently, out of the 947 instances, roughly 95 samples will be used for testing and the rest for training [11]. A random state of 42 is also specified for data splitting so that the model can work with repeated data each time it runs. The first step in finishing all of this is to install the Scikit-Learn library and use the train_test_split() function of the Python library [11].

3.2.3. Checking Perfomance and Results

The process of evaluating each algorithm's performance will come in the second to last phase, following its application.

The accuracy score needs to be ascertained initially in this phase. The accuracy ranking will determine how many correctly estimated instances there are compared to all instances. Next, a confusion matrix will be used to examine each algorithm's performance in greater detail. The classification report function will produce the following metrics: accuracy, recall, f1 score, and precision.

Classification Report:

	precision	recall	f1-score	support
0	0.95	1.00	0.98	42
1	1.00	0.96	0.98	53
accuracy			0.98	95
macro avg	0.98	0.98	0.98	95
weighted avg	0.98	0.98	0.98	95

Figure 2. Classification Report

4. IMPLEMENTATION AND EVALUATION

4.1. Algorithm

A feedforward multilayer perceptron (MLP) is a type of artificial neural network model that maps input data sets to a set of appropriate outputs [10]. Each of the layers that make up an MLP is fully connected to the layer above it. The nodes of the layers are neurons with nonlinear activation functions, apart from the nodes of the input layer. Between the input and output layers, there could be one or more nonlinear hidden layers.

In our study, we used an MLP classifier with 64 hidden layers, the "adam" solver, and "relu" activation. Additionally, a random state of 42 with values for learning rate and alpha of 0.001 each is specified.

Before we made the adjustments for dataset optimization, the classifier had a 96% accuracy rate. However, our efforts paid off, as we went on to attain an incredibly high accuracy of 98%.

4.2. Perfomance Measurement Using Classification Report

The classification report (Fig. 2) provides an overview of the overall performance of the classification algorithms used in this study and can be consulted for further details regarding the accuracy performance of the algorithms.

4.2.1. Accuracy: Confusion matrix uses accuracy measurement as one of its common performance detection techniques. It determines how frequently the algorithm can yield accurate results. It is measurable as the ratio of the total number of predictions received by the classifier to the number of corrected predictions it receives.

$$\text{Accuracy: } (T\,P + T\,N)/(T\,P + T\,N + F\,P + F\,N) \qquad (1)$$

4.2.2. Precision: Precision measures the percentage of patients with heart disease who are actually expected to have the illness among those who receive a diagnosis. It can be calculated as follows:

$$\text{Precision: } T\,P/(T\,P + F\,P) \qquad (2)$$

4.2.3. Recall: Recall calculates the proportion of patients with heart disease who have the condition in reality as opposed to those who have only made the diagnosis. It can be calculated as follows:

$$\text{Recall: } T\,P/(T\,P + F\,N) \qquad (3)$$

4.2.4. F1-Score: The f1-score is determined by taking the harmonic mean of the recall and accuracy. The maximum value is shown when the accuracy and reminder value match.

$$\text{F1}-\text{score: } (\text{Precision} * \text{Recall})/(\text{Precision} + \text{Recall}) * 2 \qquad (4)$$

4.3. Results

The algorithmic results are shown in Table 1 before optimizations, and the results after optimizations are shown in Table 2.

Table 1. Before optimizations

	Accuracy	**Precision**	**Recall**	**F1-Score**
Decision Tree	0.79	0.87	0.71	0.78
Logistic Regression	0.92	0.87	0.94	0.90
K-Nearest Neighbor	0.90	0.82	0.94	0.88
Naïve Bayes	0.87	0.87	0.85	0.86
MLP	0.96	0.96	0.96	0.96

Table 2. After optimizations

	Accuracy	**Precision**	**Recall**	**F1-Score**
Decision Tree	0.79	0.78	0.77	0.79
Logistic Regression	0.78	0.79	0.76	0.78
K-Nearest Neighbor	0.83	0.83	0.83	0.83
Naïve Bayes	0.81	0.81	0.81	0.81
MLP	**0.98**	**0.98**	**0.98**	**0.98**

5. FINDINGS

Initially, we looked for any patterns in the data. Heat maps were our first tool, but they weren't very helpful. Then, we employed the ANOVA tool to eliminate any unnecessary data; in essence, we wanted the F, or variance value, to be as high as possible and the P, or probability value, to be as low as possible. As a result, the fbs column was eliminated. Next, we used boxplots to further filter out any additional noise in the data. Ultimately, we were left with 947 rows out of the original 1025 rows (so not a lot of data was lost).

Next, we looked for different classifiers, such as decision trees, logistic regression, naïve bayes, MLP, etc. The outcomes for every other technique were fairly similar both before and after optimization, but there was a significant improvement in accuracy, going from 96% to 98% ("Fig. 2" displays the classification report with 0.98 accuracy).

5.1. Limitations and Future Directions

We were only allowed to use an old dataset from Kaggle for this study, and it was also a small dataset. It only has 303 data features in total. The pandemic has kept us apart, making it more difficult for us to get together and gather any recent data that we can use. Larger data sets are preferable for prediction because it is harder to gauge how quickly an algorithm will operate on smaller data sets.

In future, we could ourselves collect data and expand the dataset by collecting information from various hospitals in India. In order to provide more accurate and useful heart disease prediction, we are also excited to develop a hybrid model that incorporates some of the algorithms we used in this investigation. Additional feature selection optimization techniques are also available to us. To obtain additional results that can be compared in subsequent projects, we can also employ additional ML and DL algorithms.

6. CONCLUSION

In order to predict heart diseases, we employed a variety of classifiers in this study, along with a range of optimizations. We then compared the classifiers' performances with and without the optimizations. In order to maximize and denoise data and obtain the best results, heat maps, ANOVA, and box plots were also employed. It can be deduced that the MLP classifier outperformed the others and produced significantly better results than the previous one (i.e. without any optimization).

ACKNOWLEDGMENT

My supervisor, Dr. Siddhartha Chauhan, is deserving of recognition and my sincere gratitude for making this work possible. His direction and counsel saw me through this study from beginning to end. I also want to express my gratitude to my friends for their wise counsel and support at all times.

REFERENCES

[1] Khourdifi, Y., Bahaj, M. (2019). Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization. International Journal of Intelligent Engineering and Systems, 12(1), 242–252. https://doi.org/10.22266/ijies2019.0228.24 Ding, W. and Marchionini, G. 1997 A Study on Video Browsing Strategies. Technical Report. University of Maryland at College Park.

[2] Browsing Strategies. Technical Report. University of Maryland at College Park. 2. Patel, J., Tejalupadhyay, S., Patel, S. B. (2016). Heart Disease prediction using Machine learning and Data Mining Technique. Journal - IJCSCVolume: 7.

[3] Rajesh, N., T, M., Hafeez, S., Krishna, H. (2018). Prediction of Heart Disease Using Machine Learning Algorithms. International Journal of Engineering Technology, 7(2.32), 363.

[4] South Eastern Health and Social Care Trust. (2013). Cardiovascular Disease. Communications Department

[5] Lloyd-Jones, D. M., Larson, M. G., Beiser, A., Levy, D. (1999). Lifetime risk of developing coronary heart disease. The Lancet, 353(9147), 89–92.

[6] Ministry of Health Kenya. (2015). Kenya National Guideline for Cardiovascular Diseases Management. DIVISION OF NONCOMMUNICABLE DISEASES.

[7] Bramer, M. (2020). Principles of Data Mining (Undergraduate Topics in Computer Science) (4th ed. 2020 ed.). Springer.

[8] Han, J., Kamber, M., Pei, J. (2011). Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems) (3rd ed.). Morgan Kaufmann.

[9] Heart Disease UCI. (2018, June 25). Kaggle. https://www.kaggle.com/ronitf/heart-disease-uci

[10] Multilayer Perceptron and Neural Networks (2009) by MARIUS-CONSTANTIN POPESCU,VALENTINA E. BALAS,LILIANA PERESCU-POPESCU,NIKOS MASTORAKIS.

[11] Journal of Cardiovascular Disease Research(2021) by S M RAHID HAQUE, MD. ATIK FOYSAL, ARUPKUMAR DAS.

[12] ANOVA Analysis of Student Daily Test Scores in Multi-Day Test Periods(2016) by Matthew L. Mouritsen, Jefferson T. Davis, & Steven C. Jones.

[13] The Box Plot: A Simple Visual Method to Interpret Data(1989) by David F. Williamson, PhD; Robert A. Parker, DSc; and Juliette S. Kendrick, MD

AUTHORS

Deepanshu Sharma, is currently pursuing his Dual Degree(B.Tech. + M.Tech.) in Computer Science and Engineering from National Institute of Technology, Hamirpur.
Dr. Siddhartha Chauhan is currently an associate professor in the Department of Computer Science and Engineering, National Institute of Technology, Hamirpur.