Diagnosis of cardiovascular diseases using classification algorithms

Accepted 21st May, 2022

ABSTRACT

Heart diseases are the most common diseases in the world, and it will continue to be the leading cause of death for a long time. Each year 17.9 million people die due to cardiovascular diseases (CVDs) which means an estimated 32% of all deaths worldwide. However, many of the heart disease factors are preventable or treatable. If these factors are prevented or treated, it is a good opportunity to reduce the loss of life as a result of heart diseases. At present, data science is actively used by people and the importance of data science is increasing daily. It is an important fact for humanity that heart diseases and similar medical problems can be predicted using data science. For this reason, early detection of the diseases aimed with the studies carried out by applying statistical methods in the field of medicine. This research determines the relationship between heart diseases and other characteristics of the human body to early diagnosis of heart diseases. In this research, data mining approaches were used and different data science algorithms were applied to predict the heart diseases of patients. Naïve Bayes, Logistic Regression, Multilayer Perceptron and Random Forest algorithms were used for the classification and diagnosis of cardiovascular diseases prediction done. Naïve Bayes algorithm had accuracy of 88.52%. Thus, the accuracy of Naïve Bayes is the best accuracy among all other algorithms.

Key words: Heart disease prediction, Naïve Bayes, logistic regression, random forest, machine learning, classification.

INTRODUCTION

The heart is one of the important components of the circulatory system and the human body. Also, it is the strongest muscle in the body despite its small size. The heart begins to beat while a fetus in the mother’s womb performs an average of hundred thousand beats per day in a healthy adult body. The heart beats approximately two and a half billion times in an average human life and pumps the necessary clean blood to every part of the body with each beat (Body basics, 2021). Approximately five liters of blood per minute passes through the heart and distributed throughout the body. However, the heart is not fed by the blood passing through it. There are veins that are feeding the heart which is intense and peg away at its job. These vessels are called coronary vessels. The disease that occurs as a result of narrowing or occlusion of the coronary vessels is called coronary artery disease (Quertermous et al., 2016). Coronary artery disease is the root of most of the cardiovascular diseases (Sweis et al., 2021). According to the data of the World Health Organization (WHO) (n.d.), deaths as a result of cardiovascular diseases take the first place among all deaths with 32%.

Myocardial infarction (MI) is one of the coronary artery diseases and popularly called a heart attack. Myocardial infarction (MI) occurs as a result of blood building up in a heart or impaired oxygenation in the heart. In other words, it is due to the inability of the heart muscle to sufficiently pump blood to the heart and thus sufficiently pump oxygen to the heart. Damage occurs in a heart when there is no enough oxygen and death occurs in a heart when there is no sufficient oxygen for a long time (Storrow and Gibler, 2000). The main causes of this disease, which poses a risk of death or adversely affects human health, can be listed as
follows (Hajar, 2017):

- Obesity
- Hypertension
- Diabetes
- Cholesterol
- Gender
- Age
- Tobacco use
- Alcohol use
- Family history

In addition to these substances, drug use, stress and intense lifestyle are among the factors that trigger the disease (Vassalle et al., 2004). The diagnosis of heart attack can be made according to the results of physical examination and tests such as creatinine kinase, troponin, myoglobin and electrocardiography (ECG) (Al-Hadi and Fox, 2009). At present, diagnosis of the disease is done only by doctors. But we can expect data scientists to help doctors in this regard with the power of data mining approaches. Medical world is very rich in data. But, usage of the data and studies about medical data are not sufficient as expected. In recent years, many algorithmic and statistical studies have been conducted on CVDs.

The presented studies can be shown as an example of how data processing and information extraction can yield useful results that can be used in the field of medicine. As the number of studies on cardiovascular diseases increases, it will provide more support for the prevention of heart diseases.

The aim of this research was to determine the relationship between age, gender, chest pain type, resting blood pressure, cholesterol value, fasting blood glucose, resting electrocardiographic result, maximum heart rate, angina, ST (Sinus Tachycardia) depression, ST segment slope and diagnosis of heart attack. In this study, the data were taken from UCI database for preprocessing the data with the data mining approaches and editing for further analysis. Naïve Bayes, Logistic Regression, Multilayer Perceptron and Random Forest algorithms were applied for diagnosis prediction.

The present study is organized as follows: related works about heart disease predictions; materials and methods; the result of the analysis; conclusion of the study; and future works about the study.

RELATED WORKS

In recent years, Data Mining and Machine Learning techniques have been applied to many areas in health informatics. Data mining can be applied to many industries. Valuable information can be identified through the application of data mining techniques in the healthcare system. The applications of artificial decision support systems in the field of health informatics are increasing daily. The use of artificial intelligence-based decision support systems is also becoming widespread to help diagnoses and treatments made by data analyst to prevent human-induced errors. In addition to algorithmic studies aimed at detecting cardiovascular diseases in advance, various statistical studies have also been carried out. There are many different studies conducted under the discipline of artificial intelligence and pattern recognition for CVDs predictions.

Ramya and Kaladevi (2020) applied prediction model to UCI- Cleveland heart disease dataset of 303 instances and 14 attributes. Proposed prediction system used PCA as a feature reduction and standardization technique. Heart disease prediction model is developed through feature standardization and feature reduction using PCA to train the machine learning classifiers. They found out that Logistic Regression classifier accuracy was 87% and the Support Vector Machine was 85% which is close to Logistic Regression accuracy. And also, with the K-Nearest Neighbor classifier they found out the model accuracy of 69% which is less as compared with other classifiers.

Latha and Jeeva (2019) applied comparative analysis of various classification techniques to increase the predictive accuracy of MI risk prediction using ensemble algorithms and techniques on the UCI-Cleveland heart disease dataset of 303 instances and 14 attributes. They applied Bayes Net, Naïve Bayes, Random Forest, C4.5, Multilayer Perceptron and PART as ensemble algorithms. At the same time, they applied Boosting, Bagging, Stacking and Majority Vote to increase the predictive accuracy of the classifier algorithms. They successfully increased the accuracy of a weak classifier by 7.26% based on ensemble algorithm. They got 85.48% accuracy using nine attributes with Majority Vote for Naïve Bayes, Random Forest, Bayes Net and Multilayer Perceptronalgorithms.

Ananey-Obriet and Sarku (2020) applied three different classification algorithms which are Logistic Regression, CART and Gaussian Naïve Bayes Model on UCI-Cleveland heart disease dataset of 303 instances and 14 attributes. As conclusion of the study, they found out that both Logistic Regression and Gaussian Naïve Bayes Model had accuracy of 82.75% and CART algorithm accuracy was 79.31%. However, the AUCROC value for the Gaussian Naïve Bayes Model was higher than the Logistic Regression algorithm.

CART algorithm accuracy was less as compared with the other two algorithms due to the sample size of the dataset.

Gupta et al. (2020) preprocessed the dataset with Data Imputation, Data Standardization and Data Stratification to replace the missing values for UCI-Cleveland heart disease dataset of 303 instances and 14 attributes. In the study, MIFH, a Machine Intelligence Framework, includes Data Imputation and Partitioning, Feature Extraction using FAMD, Features Normalization, Machine Learning Approach and Performance Metric Evaluation. It is used to
develop a model for classifying the MI risk prediction model and trained Logistic Regression, K-Nearest Neighbor, Support Vector Machine, CART and Random Forest classifiers using FAMD method. Random Forest achieved the best accuracy of 93.44% compared with other classifiers.

Mohan et al. (2019) applied HRFLM which is hybrid Random Forest algorithm with Linear Model to increase the accuracy of MI risk prediction for UCI-Cleveland heart disease dataset of 303 instances and 14 attributes. They found that Random Forest and Linear Model are the best among 8 classifier algorithms. The Linear Model method was the best compared to CART and Random Forest methods. And, they combined Random Forest and Linear Model methods to propose HRFLM method to improve the results.

Kodati et al. (2019) applied various data mining classifying algorithms to develop MI risk prediction system for UCI-Cleveland dataset of 297 instances and 13 attributes. Thus, in the study, Orange and Weka data mining tools were used. The main idea was to compare different data mining tools according to their classification precision and recall. Naïve Bayes precision and recall were the highest when compared with Support Vector Machine, Random Forest and K-Nearest Neighbor algorithms. Naïve Bayes precision was 0.824 with Orange and 0.837 with Weka. Also, recall was 0.806 with the Orange and 0.837 with the Weka. As a result, Weka data mining tool gave higher precision and recall as compared with Orange data mining tool.

Gazeloglu (2020) applied 18 different machine learning algorithms for UCI-Cleveland dataset of 303 instances and 14 attributes. He also used three different feature selection techniques which are Correlation-Based Feature Selection, Fuzzy Rough Set and Chi-Square algorithms for feature selection algorithm. Based on these results, SVM (PolyKernel) had ratio of approximately 90%.

Tama et al. (2020) developed a two-layer stacked architecture to do MI risk prediction for UCI-Cleveland (303 instances and 14 attributes), UCI-Hungarian (294 instances and 14 attributes), Statlog (261 instances and 14 attributes), Z-Alizadeh Sani (303 instances and 55 attributes) dataset. Proposed system has two-tier ensemble built upon three different classifier ensembles such as Random Forest, Gradient Boosting Machine and Extreme Gradient Boosting Machine in a stacked manner. Best prediction performance was on the Z-Alizadeh Sani dataset with an accuracy of 83.905% by running different numbers of particles in PSO.

**MATERIALS AND METHODS**

This part discusses the software, dataset, techniques and methods for this study. In this study, Weka software was used for all the applications for UCI-Cleveland heart disease dataset (UCI, n.d.). Data mining techniques that are used in this study explained below in details which are preprocessing the dataset and creating the models for this particular dataset.

**WEKA data mining tool**

Weka developed for the purpose of machine learning and data mining at the University of Waikato. This software includes machine learning algorithms and methods that are in use for the machine learning and data mining purposes today. The software was developed with Java language and it can be easily integrated into different application written in Java. Data mining operations can be done easily with Weka. Weka can operate Classification, Clustering, Association Rule Mining with its pre-written algorithms inside. In addition, Weka can be used for data-preprocessing and visualization purposes (The University of Waikato, n.d.).

**Dataset**

In the present study, dataset was gathered from UCI Machine Learning repository. The dataset name is “heart disease dataset” which consists of four different datasets. Each dataset contains information from different patients. These patient’s information was obtained from four different hospitals which are Cleveland Clinic Foundation, Hungarian Institute of Cardiology, Budapest, V.A. Medical Center, Long Beach, CA, University Hospital, Zurich, Switzerland. Each dataset has the same instance format. Each dataset contains 76 attributes such as age, sex, cp (chest pain type) etc. Most of the attributes eliminated in different studies in order to get better results. Also, in this study, some of the attributes were eliminated and the final attribute size is 14. Cleveland dataset considered in this study instead of 4 different datasets at the same time.

Cleveland dataset has 76 attributes in original dataset as mentioned above. But in this study, 14 attributes were used for 303 different instances. Descriptions of 14 attributes are given in the Table 1.

**Techniques and methods**

This study is completed in two stages which are data-preprocessing and application of different data mining models to train and test our models to obtain results as a conclusion of this study. In data-preprocessing part, some of the data mining techniques are applied to obtain better and workable dataset for our models instead of having dataset with too many outliers and missing values. In the application of model part Naïve Bayes, Logistic Regression, Multilayer Perceptron and Random Forest are applied to
Table 1: Heart disease features of Cleveland dataset.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description of attributes</th>
<th>Type</th>
<th>Values</th>
<th>Description of Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Age in years</td>
<td>Numeric Attribute</td>
<td>29-77</td>
<td>In range between</td>
</tr>
<tr>
<td>Sex</td>
<td>Gender</td>
<td>Nominal Attribute</td>
<td>0,1</td>
<td>0=female 1=male</td>
</tr>
<tr>
<td>Cp</td>
<td>Chest pain type</td>
<td>Nominal Attribute</td>
<td>1,2,3,4</td>
<td>1=typical angina 2=atypical angina 3=non-anginal pain 4=asymptomatic</td>
</tr>
<tr>
<td>Trestbps</td>
<td>Resting blood pressure (in mm Hg on admission to the hospital)</td>
<td>Numeric Attribute</td>
<td>94-200</td>
<td>In range between</td>
</tr>
<tr>
<td>Chol</td>
<td>Serum cholesterol in mg/dl</td>
<td>Numeric Attribute</td>
<td>126-564</td>
<td>In range between</td>
</tr>
<tr>
<td>Fbs</td>
<td>Fasting blood sugar &gt; 120 mg/dl</td>
<td>Nominal Attribute</td>
<td>0,1</td>
<td>0=false 1=true</td>
</tr>
<tr>
<td>Restecg</td>
<td>Resting electrocardiographic results</td>
<td>Nominal Attribute</td>
<td>0,1,2</td>
<td>0=normal 1=ST-T wave abnormality 2=definite left ventricular hypertrophy</td>
</tr>
<tr>
<td>Thalach</td>
<td>Maximum heart rate achieved</td>
<td>Numeric Attribute</td>
<td>71-202</td>
<td>In range between</td>
</tr>
<tr>
<td>Exang</td>
<td>Exercise induced angina</td>
<td>Nominal Attribute</td>
<td>0,1</td>
<td>0=false 1=true</td>
</tr>
<tr>
<td>Oldpeak</td>
<td>ST depression induced by exercise relative to rest</td>
<td>Numeric Attribute</td>
<td>0-6.2</td>
<td>In range between</td>
</tr>
<tr>
<td>Slope</td>
<td>The slope of the peak exercise ST segment</td>
<td>Nominal Attribute</td>
<td>1,2,3</td>
<td>1=upsloping 2=flat 3=down sloping</td>
</tr>
<tr>
<td>Ca</td>
<td>Number of major vessels (0-3) colored by fluoroscopy</td>
<td>Nominal Attribute</td>
<td>0-3</td>
<td>In range between</td>
</tr>
<tr>
<td>Thal</td>
<td>Heart status</td>
<td>Nominal Attribute</td>
<td>3,6,7</td>
<td>3=normal 6=fixed defect 7=reversible defect</td>
</tr>
<tr>
<td>Class</td>
<td>Diagnosis</td>
<td>Nominal Attribute</td>
<td>0,1</td>
<td>0=false 1=true</td>
</tr>
</tbody>
</table>

determine the best accuracy among them.

**Data preprocessing**

Data preprocessing is the process to handle missing values and outliers in order to get rid of them from dataset. This process should be done before any methods applied to our dataset. Process contains techniques such as completing missing data, eliminating inconsistencies and removing noise to detect outliers. Following techniques can be used to complete the missing data:
Instances with missing values can be deleted from the dataset.

- Missing values can be replaced with average of the attribute instead of missing values.
- Missing values can be replaced with median of the attribute instead of missing values.
- Missing values can be replaced with the average of their classes instead of missing values.
- Missing values can be replaced with the appropriate value produced using methods such as regression instead of missing values.

Following techniques can be used to detect outliers:

- Binning
- Clustering
- Regression Methods

There were missing values in UCI-Cleveland heart disease dataset. Nominal attributes of “ca” and “thal” had some missing values before data-preprocessing techniques were applied. These missing values were replaced based on majority mark. The value of the “ca” was ‘0’ as the majority mark and had four missing values. This value was gathered from 172 observations out of 303. The value of the “thal” was ‘3’ as the majority mark and had two missing values. This value gathered from 168 observation out of 303. At the end, missing values of “ca” and “thal” replaced with the values of ‘0’ and ‘3’ according to the majority mark. Figure 1 shows the attributes of the UCI-Cleveland heart disease dataset after the data-preprocessing techniques.

**Naive Bayes algorithm**

Naive Bayes is a frequently used classification technique for data mining approaches. Naive Bayes is predicated on Bayes Theorem. This statistical algorithm assumes that there is no any dependency between the dataset classes which are attributes. This approach makes this algorithm to build models easily. But, making the classes independent is not accurate for all approaches. At the same time, this algorithm works well with big datasets if we are talking about the speed to build models.

Equation (1) is Naive Bayes algorithm formula and shown as:

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

In this equation, \( P(A) \) is the independent probability (class prior probability) of the class \( A \), \( P(B) \) is the independent probability (predictor prior probability) of the class \( B \), \( P(A|B) \) is if \( P(B) \) satisfies the situation then it’s the probability (posterior probability) of the class \( A \) and the \( P(B|A) \) is the likelihood (Marathe et al., 2021).

**Logistic regression algorithm**

Logistic Regression is used for machine learning approaches to make discriminative models. This classification technique counts the posterior probabilities. Logistic Regression algorithm does not make any assumption about attributes conditional probabilities. So, this algorithm is comprehensive and can be used in various applications. Also, this algorithm can be magnified to multi-class classification models such as multinomial regression. However, the limit of this algorithm is that the algorithm learns only in linear decision boundaries. Since Naive Bayes can handle with missing values, logistic regression cannot handle with missing values because those missing values can be deleted from the training set to build a model with Logistic Regression. However, if there will be missing values in test instances, then logistic regression can fail to make a prediction of class labels (Mythili et al., 2013). For one observation of \( x \) with \( t \) features, \( m \) will be given in the formula. Equation (2) is Logistic Regression algorithm formula and shown as:

\[ m = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_t x_t)}} \]  

**Multilayer perceptron algorithm**

Multilayer Perceptron uses an artificial neural network to make more complex architecture of nodes. These nodes can learn in nonlinear decision boundaries. Multilayer Perceptron contains input layer, output layer and hidden layer. Multilayer Perceptron uses back propagation for training the data. Multilayer Perceptron algorithm process is input layer is used to represent instances from attributes (numerical or nominal attributes) as single nodes in input layer. These single nodes fed into different layers which are called hidden layers. These nodes are called hidden nodes in hidden layer. Hidden nodes produce activation values to transmit to next layer which is called output layer. Final layer is the output layer which processes values from its preceding layer and makes predictions (Yan et al., 2006). Figure 2 shows an example of Multilayer Artificial Neural Network (ANN).

**Random Forest**

Random Forest algorithm is developed by Breiman (2001) (Fawagreh et al., 2014). Random Forest is used for classification and regression tasks in data mining. The purpose of the random forest algorithm is to combine and show tree decisions trained in different training sets rather than one single decision tree. Whereas in determining the standard at every level, the standard is set by creating some calculations altogether trees. Then, the most used feature is
selected by combining the features in other trees. The selected attribute is included in the tree and the process is repeated at all levels (Jabbar et al., 2015). To start the algorithm, the number of variables and trees to be used in each node must be determined by the user.

RESULTS AND DISCUSSION

This research has been performed using machine learning classification algorithms such as Naïve Bayes, Logistic Regression, Multilayer Perceptron and Random Forest on WEKA tool. Accuracy, TP rate, FP rate, precision and recall scores have been used to evaluate the models' accuracies for this research. This scores and formulas used to calculate model's accuracies for this study is explained as:

Equation (3) corresponds to the number of positive examples correctly predicted by the classifier (Tan et al, 2019):

\[
\text{True Positive Rate} = \frac{TP}{TP + FN}
\]  

(3)
$TP$ (Number of True Positive Predictions)  
$FN$ (Number of False Negative Predictions)  

Equation (4) is corresponding to the number of negative examples wrongly predicted as positive by the classifier (Tan et al., 2019).

\[
False \ Positive \ Rate = \frac{FP}{FP + TN} \quad (4)
\]

$FP$ (Number of False Positive Predictions)  
$TN$ (Number of True Negative Predictions)  

Precision indicates the number of correct positive predictions. On the other hand, recall identifies the number of correct actual positive examples.

Data classification methods depend on the accuracies. Equation (5) is the closeness of measurements to the true value of the quantity being measured (Tan et al., 2019):

\[
Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (5)
\]

Data set is divided into two parts such as training set (80% of the dataset) and testing set (20% of the dataset). To determine the performance of the classifiers that have been used, confusion matrices of algorithms were made. Figure 3 shows the confusion matrices of the used algorithms. Naïve Bayes algorithm predicted 54 correct predictions and 7 wrong predictions. Logistic Regression predicted 53 correct predictions and 8 wrong predictions. Multilayer Perceptron predicted 52 correct predictions and 9 wrong predictions. And, Random Forest predicted 51 correct predictions and 11 wrong predictions. Confusion matrices showed that Naïve Bayes algorithm predictions were made correctly more than other algorithms performed in Figure 3.

Correlation analysis was carried out to determine the correlation between the characteristics of the dataset and
Figure 4: Correlation matrix of dataset.

Table 2: Detailed accuracy by class.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy (%)</th>
<th>TP Rate (%)</th>
<th>FP Rate (%)</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>88.52</td>
<td>88.5</td>
<td>11.2</td>
<td>89.6</td>
<td>88.5</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>86.88</td>
<td>86.9</td>
<td>12.8</td>
<td>88.4</td>
<td>86.9</td>
</tr>
<tr>
<td>Multilayer Perceptron</td>
<td>85.24</td>
<td>85.2</td>
<td>14.7</td>
<td>85.3</td>
<td>85.2</td>
</tr>
<tr>
<td>Random Forest</td>
<td>83.60</td>
<td>83.6</td>
<td>16.2</td>
<td>84.2</td>
<td>83.6</td>
</tr>
</tbody>
</table>

the diagnosis. The correlation matrix is shown in Figure 4.

As shown in Figure 4, there is a high positive correlation between chest pain type, exercise-induced angina and exercise-induced ST depression in the correlation of the characteristics with the diagnosis status. So, we can say that chest pain type, exercise-induced angina and exercise-induced ST depression affect the diagnosis.

After dataset performed training and testing processes for specific algorithms in Weka, detailed accuracies were gathered for different algorithms. Thus, the results of the algorithms used in this study are shown in Table 2.

By comparing the results in Table 2 and confusion matrices of the algorithms, Naïve Bayes was found to be the best data mining algorithm with accuracy of 88.52 % to predict heart diseases for UCI-Cleveland dataset. However, Logistic Regression (86.88 %) and Multilayer Perceptron’s (85.24 %) accuracies were pretty close to Naïve Bayes but obviously they were not much as accurate as Naïve Bayes.

In this study, Naïve Bayes algorithm gave the best accuracy due to the it’s learning mechanism. In our dataset, attributes were not much dependent on each other, and Naïve Bayes algorithm expects the features to be independent. As a result, Naïve Bayes algorithm performed the best accuracy due to independent attributes in our dataset.

CONCLUSION

The aim of this study is to predict heart diseases accurately
as possible by determining the relationship between different characteristics of patients and heart diseases. Data mining classification algorithms such as Naïve Bayes, Logistic Regression, Multilayer Perceptron and Random Forest have been used to find relationship between patient's characteristics and cardiovascular diseases by using UCI-Cleveland heart disease dataset. WEKA tool was used to run algorithms for the classification tasks. On the other hand, confusion matrices for the classification tasks were visualized using Python programming language.

To evaluate the models for this classification research, accuracy, TP rate, FP rate, precision and recall scores were used. Originally, UCI-Cleveland dataset has 76 attributes. But in this study, most of the attributes were eliminated and only 14 attributes were used to obtain better results. After that the values in the UCI-Cleveland dataset preprocessed and was made suitable for the analysis stage. Dataset was divided into two parts such as training set (80% of the dataset) and testing set (20% of the dataset) for all the algorithms. Thereafter, the dataset was analyzed and Naïve Bayes was found to be the best data mining algorithm with accuracy of 88.52% to predict heart diseases for UCI-Cleveland dataset. However, more algorithms and parameters can be used to classify heart disease more accurately.

FUTURE WORK

There are different research papers that focus on the same dataset with different classifiers, algorithms, or techniques to achieve the heart disease prediction. These studies show that different methods can improve the prediction of heart diseases. Therefore, by improving and proposing new methods, prediction accuracies would become more accurate than ever. Furthermore, due to technological developments and relevant studies on this subject, early diagnosis of heart diseases will be easier and more accurate.

REFERENCES


