Comparison of the performance of machine learning algorithms in predicting heart disease

Sajad Yousefi*

*Department of Electrical Engineering, Technical and Vocational University (TVU), Tehran, Iran

Article Info

Article type: Research

Article History:
Received: 2021-10-01
Accepted: 2021-12-23
Published: 2021-12-27

* Corresponding author:
Sajad Yousefi

Department of Electrical Engineering, Technical and Vocational University (TVU), Tehran, Iran

Email: syosefi_1980@yahoo.com

Keywords: Machine Learning, Heart Disease, Dataset, Decision Tree, Logistic Regression

ABSTRACT

Introduction: Heart disease is often associated with conditions such as clogged arteries due to the sediment accumulation which causes chest pain and heart attack. Many people die due to the heart disease annually. Most countries have a shortage of cardiovascular specialists and thus, a significant percentage of misdiagnosis occurs. Hence, predicting this disease is a serious issue. Using machine learning models performed on multidimensional dataset, this article aims to find the most efficient and accurate machine learning models for disease prediction.

Material and Methods: Several algorithms were utilized to predict heart disease among which Decision Tree, Random Forest and KNN supervised machine learning are highly mentioned. The algorithms are applied to the dataset taken from the UCI repository including 294 samples. The dataset includes heart disease features. To enhance the algorithm performance, these features are analyzed, the feature importance scores and cross validation are considered.

Results:
The algorithm performance is compared with each other, so that performance based on ROC curve and some criteria such as accuracy, precision, sensitivity and F1 score were evaluated for each model. As a result of evaluation, Accuracy, AUC ROC are 83% and 99% respectively for Decision Tree algorithm. Logistic Regression algorithm with accuracy and AUC ROC are 88% and 91% respectively has better performance than other algorithms. Therefore, these techniques can be useful for physicians to predict heart disease patients and prescribe them correctly.

Conclusion: Machine learning technique can be used in medicine for analyzing the related data collections to a disease and its prediction. The area under the ROC curve and evaluating criteria related to a number of classifying algorithms of machine learning to evaluate heart disease and indeed, the prediction of heart disease is compared to determine the most appropriate classification. As a result of evaluation, better performance was observed in both Decision Tree and Logistic Regression models.
patterns between data can be used to predict [1]. In machine learning, accuracy is processed for data sets. The researchers used various data mining methods such as classification and clustering to create a model for predicting heart disease.

Researchers have applied different data mining methods such as association rules, classification and clustering to build a model for the prediction of heart disease. Shiva Kazempour Dehkordi et al. [2] proposed a prediction model based on the prescription using the data mining method. They proposed an algorithm called Skating to enhance the accuracy of the system. Skating is an ensemble method similar to Boosting and Bagging. They compared four classification algorithms such as DT, Naive Bayes (NB), K-Nearest Neighbors (KNN) and Skating in a different label. They showed that the most accurate given classifier is Skating. This classification algorithm gave 73.17% accuracy. However, this is a comparatively low performing method compared to other classification algorithms and methods. Mezzatesta et al. utilized Machine Learning techniques to predict the mortality and hospitalization in heart failure subjects [3]. They have used five methods, LR with forwarding selection variable and LASSO regularization variable selection, Random Forest, Gradient Descent Boosting and SVM. Three years follow up was done and validated using 5-fold cross validation. Random Forest gave the best results with 0.72 mean C-statistic value for predicting mortality and 0.76 for predicting hospitalization. The inclusion of time-to-event analysis could ameliorate the outcome of the proposed model.

Jan et al. in 2018 implemented an ensemble data mining approach using two benchmark datasets collected from a UCI repository (namely Cleveland and Hungarian) where the ensemble of five different classification algorithms such as RF, neural network, NB, classification via regression analysis and support vector machines (SVM) were employed [4]. They observed that the lowest performing algorithm was regression methods, while in contrast, RF provided a very high accuracy of 98.136%.

In 2011, Jyoti Soni et al. applied DT with a genetic algorithm to improve the classification performance, and this was compared with other two algorithms such as NB and classification via cluster methods. They found 99.2% accuracy for the proposed system [5].

Le et al., in 2018 employed three classification methodologies for the listed 58 attributes in the dataset collected from UCI Machine Learning Repository [6]. They showed that a support vector machine (SVM) with a linear kernel gave a superior performance, with 89.93 accuracy.

Tarawneh and Embarak proposed a hybrid approach using 12 features and compared the performance with KNN, J48, GA, DT, artificial neural network (ANN), SVM and NB [7]. The proposed hybrid method generated 89.2% accuracy, which is the best performance compare to other applied algorithms.

Latha and Jeeva implemented an ensemble classification technique using the Cleveland dataset and ensemble majority vote with MP, RF, BN and NB using the feature selection method to improve the accuracy of the classifier [8]. The performance was evaluated from six sets of attributes. They built different ensemble models and compared the performance to find the best ensemble model. They found that the majority vote with MP, RF, BN and NB employing attribute selection method offered the best performance with 85.48% accuracy.

Chohan in 2020 compared several machine learning algorithms on a data set from the Kegel site, the regression algorithm performed with 89% accuracy [9]. Ghosh et al. Also achieved an accuracy of 99.05% in 2020 using the Random Forest Bagging Method model on the five data sets that were integrated [10].

### MATERIAL AND METHODS

#### Classification

Classification is a method of machine learning and is used to learn how to assign a class tag to an input instance. For example, classification can determine whether a person is ill. Class labels here are sick and healthy that must be converted to numeric values. Two classes are considered: class zero (healthy) and class one (sick). Classification is actually a predictive issue that predicts class labels. The data set records under analysis are divided into two categories: Training Set and Test Set. The individual records that make up the test dataset are randomly sampled from the r set under analysis. Test data set records are independent of training records.

#### Supervised Learning Algorithm

These algorithms include a target variable that is predicted from a set of independent variables or attributes. Using this set of variables, a function is created that maps the inputs to the desired outputs. The training process continues until the model achieves the desired level of accuracy on the training data. In this case, the data set is labeled, meaning that the algorithm clearly identifies the attributes and makes predictions or classifications based on them. By performing the training step, the algorithm is able to identify the relationships between the two variables so that a new result can be predicted [11]. The following is a list of supervised learning algorithms.

**K Nearest Neighbors (KNN)**

K-Nearest Neighbors (KNN) is probably one of the
simplest algorithms utilized in machine learning for both regression and classification techniques. KNN calculations use the data and characterize new data points dependent on resemblance measures (e.g., distance function). Basically, The KNN calculation accepts that comparative things are close to one another. In KNN, classification occurs by considering the majority vote to its neighbors. The data point goes to the class that has the most intimate neighbors. As we increment the number of nearest neighbors, the estimation of k and accuracy may increment. The predictor variables anticipate the target variable or the dependent variable [12].

**Logistic Regression (LR)**

Algorithm to predict the probability of the response variables by given a set of explanatory independent variables, is a supervised binary classification algorithm. The response variables are coded as binary values as 0 and 1, based on the values of the binary target variables the data will be classified. The logistic regression types are Binary or Binomial, Multinomial, and Ordinal. The values of the target variable in the Binary LR have two possible types 0 and 1. Multinomial Logistic regression is the response variable that has unordered 3 or more possible values to represent the data such as the class types (A, B, C, D). The Ordinal LR is the classification when the response variable has ordered 3 or more values such as students’ scores (High, Mid, Low). The LR algorithm is sensitive to outliers.

**DT or Decision Tree (DT)**

Decision Tree Algorithm is known as the supervised learning algorithm. Moreover, in supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems. The Decision Tree Algorithm is a decision support tool that uses a tree-like model. The goal of using a Decision Tree is to create a training model that can use to predict the target variable by learning simple decision rules inferred from training data [13, 14].

**RF Random Forest (RF)**

Random Forest Algorithm is used for supervised and classification, but mostly it's used for classification problems. It generates decision trees based on data samples and then gets the prediction from each of them. After prediction, it selects the most suitable solution by means of voting. It is an aggregate method that is better than a single decision tree because it decreases the over-fitting by averaging the result [14, 15].

**Support Vector Machine (SVM)**

The algorithm of SVM is primarily used for classification techniques. SVM generates a hyperplane separating the data into various classes. SVM can resolve either linear or non-linear issues. The main objective of the SVM is to discover a hyperplane in N-dimensional space (N matches the features number) that clearly classifies the datasets or points of data. The accuracy of the outcome is related directly to the hyperplanes that select. It should discover a plane with the maximum distance among data points of both classes. The hyperplane is illustrated graphically as a line that divides one class and another. Data points located come on different sides of the hyperplane are assigned to different classes. The hyperplane dimension dependent on the feature quantity. If the input has two features, then the hyperplane is a line and be a 2D plane if the features of input are three. [16, 17].

**Multi-Layer Perceptron (MLP)**

MLP is a neural network-based classification algorithm and is considered a class of artificial feed neural networks. It consists of three or more types of layers: an input layer, an output layer, and one or more hidden layers between the input and output layers. Each layer contains a number of "neurons" that are the computational units that connect all the layers together. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP uses a supervised learning technique called repayment for training. Its multiple layers and nonlinear activation distinguish MLP from a linear perceptron. It can actually distinguish data that is not linearly separable.

**Bagging (bootstrap aggregation)**

The bagging algorithm is a machine learning meta-algorithm designed for statistical classification and regression. The bagging algorithm is an ensemble model that tries to improve the stability and accuracy of prediction and reduce the variance of the model by combining predictions from separate trained models on randomly generated training samples. The final prediction of the hybrid model is obtained by calculating the average of all the predictions of the individual estimators.

**Extra Trees**

Extra tree classifier generates randomized multiple decision trees with different sub-samples without bootstrapping. It avoids the problem of over-fitting and results in better accuracy. Three important parameters for Extra tree classifies are: (i) $M$ represents the total number of trees to be generated, (ii) $K$ represents the number of attributes chosen for tree construction and (iii) $n_{\min}$ denotes minimum required samples. When compared with random forest mechanism, extra tree classifier differs by choosing random Kattribute and random split values for generating the tree. The attribute which shows lesser bias-variance is identified as the best split
attribute.

Comparison Metrics

For comparing the different classifiers performance, the metrics such as accuracy, precision, recall and F1 scores are used. Equation 1, 2, 3 and 4 shows the formula for computation of the above mentioned performance measures. True positive and True Negative are the instances rightly predicted as positive and negative samples respectively. False positive instances are predicted as positive whereas really it is a negative sample and the False Negatives are predicted as negative samples but really these are positive samples.

\[
Accuracy = \frac{TP + TN}{TP + FN + FP + TN}
\]

\[
Sensitivity = \frac{TP}{TP + FN}
\]

\[
Precision = \frac{TP}{TP + FP}
\]

\[
F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}
\]

Receiver operating characteristic (ROC) curve, is a graphical way to show how good the performance of a classifier is. Basically, it is a plot of a true positive rate against a false positive rate.

That is, the number of correct predictions is divided by the number of actual positive results, and the rate of positive predictions is calculated. FPR, on the other hand, indicates the number of positive identifications among negative observations. This ratio is also used as a false positive rate in the ROC chart. Area under the curve (AUC) is used as a criterion for evaluating the performance of a classifier. Therefore, the closer the area under the graph to the number one, the better the classifier performance.

Dataset

The heart disease dataset from the UCI repository is used in this paper. The data set worked on is related to the information of 294 people (sample) which includes a series of characteristics such as age, gender, type of chest pain, static blood pressure, blood cholesterol level, blood sugar level. Attribute values for each instance are stored in the dataset. The age range of people in the data set is between 28 and 66 years.

The data set contains 11 attributes according to Table 1 which lists the range of values for each feature. For example, for gender, female gender is 0 and male is 1. The target variable determines whether a person is sick or not. The value of the target variable is a binary value so that a value of 1 means sick and a value of zero means no disease.

In order to build up a more accurate Machine Learning model, data preprocessing is required. Data pre-processing is the process of cleaning the data. It will remove all the NAN values from our data. This process is also known as Data Wrangling. This includes the identification of missing data, noisy data and inconsistent data.

<table>
<thead>
<tr>
<th>SN</th>
<th>Attribute name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Age</td>
<td>Age in years</td>
</tr>
<tr>
<td>2</td>
<td>Sex</td>
<td>Male=1, Female=0</td>
</tr>
</tbody>
</table>
| 3  | Cp             | cp: chest pain type
Value 1: typical angina
Value 2: atypical angina
Value 3: non-anginal pain
Value 4: asymptomatic |
| 4  | Trestbps       | resting blood pressure (in mm Hg on admission to the hospital) |
| 5  | Chol           | serum cholesterol in mg/dl |
| 6  | Fbs            | fasting blood sugar > 120 mg/dl (1 = true; 0 = false) |
| 7  | Restecg        | resting electrocardiographic results
-- Value 0: normal
-- Value 1: having ST-T wave
-- Value 2: left ventricular hypertrophy |
| 8  | Thalach        | maximum heart rate achieved |
| 9  | Exang          | exercise induced angina (1 = yes; 0 = no) |
| 10 | Oldpeak        | ST depression induced by exercise relative to rest |
| 11 | Target         | 0=no disease and 1=disease |

RESULTS

In this paper, using the classification algorithms in the test phase, the value of the target variable is predicted according to other features. After obtaining and predicting the variable target, this predicted value is compared with its actual value in the test set. The degree of proximity of the prediction values to the real values in the test set is calculated.

In this study, k-fold cross validation was used to train and test the model. The dataset is divided into a number of groups. K refers to the number of groups, also known as ‘fold’. Cross validation is an approach to evaluate a machine learning model. K-fold cross validation is such a technique, where the dataset is split in to k number of groups and the model is trained by (k-1) groups and the other group participates to test or evaluate the trained model. In this approach, the model is trained k number of times and each time, different fold participates to evaluate the model. It indicates that each fold participates to train and test a model in k-fold cross validation [18].

K-Fold Cross Validation is used to prevent overfitting in forecasting models. In this paper, k is considered
equal to 10. Feature Importance Score is also used in classification algorithms. Fig 1 shows the dependency values between all attributes in the dataset. Values below zero indicate a negative dependence and values above zero indicate a positive dependence.

According to Fig 1, the properties of exang, oldpeak and cp have a high relationship with target variable. For example, the dependency coefficient of target and exang is equal to 0.58, the dependence coefficient of target and oldpeak is equal to 0.55 and the dependence coefficient of target and cp is equal to 0.51. In the data set used, the restecg and thalach properties have a negative dependence on the target attribute and also have the least amount of dependence on the target variable. Taking into account the significance score of the features in the algorithms (in algorithms such as KNN and MLP the feature importance score cannot be calculated), the values of the evaluation criteria are calculated.

Fig 2 to 5 shows the performance outcome parameters of the classification algorithms employed, namely accuracy, precision, recall and f1 score. The Logistic Regression algorithm outperformed others by producing 88% accuracy. The Decision Tree and MLP delivered 83% and Random Forest made it with 80%.

The value of the precision criterion for the decision tree algorithm is equal to 94%, which has the highest value compared to the rest. Also, the value of this parameter is equal to 90% for MLP and 86% for Decision Tree algorithm. The highest recall criteria for decision tree and logistic regression algorithms are 73% and 71%, respectively.

The Logistic Regression algorithm outperformed others by producing 81% F1 score. The Decision Tree made it with 79%.
Fig 6 to 11 are for a better understanding of feature ranking and importance according to classification algorithms. Each of these figures is a graphical representation of Table 2. These figures show the feature ranking based on feature importance and coefficient scores for all the applied classification algorithms except MLP and KNN. These figures also tend to represent the highly responsible attributes for heart disease. Table 2 shows the five most significant features according to feature importance and correlation value. According to Table 2, it is found that exang is the significant feature or factor for heart disease identification and prediction. Besides chest pain type (cp), the number of maximum heart rates achieved (thalach), ST depression induced by exercise relative to rest (oldpeak), and chol are also significant factors predicting heart disease.

**Table 2: Feature ranking for heart disease**

<table>
<thead>
<tr>
<th>Feature Ranking</th>
<th>JT</th>
<th>RF</th>
<th>XGBoost</th>
<th>ET</th>
<th>LR</th>
<th>Bagging</th>
</tr>
</thead>
<tbody>
<tr>
<td>first</td>
<td>exang</td>
<td>cp</td>
<td>exang</td>
<td>exang</td>
<td>restecg</td>
<td>Cp</td>
</tr>
<tr>
<td>second</td>
<td>thalach</td>
<td>exang</td>
<td>oldpeak</td>
<td>cp</td>
<td>oldpeak</td>
<td>oldpeak</td>
</tr>
<tr>
<td>third</td>
<td>trestbps</td>
<td>oldpeak</td>
<td>cp</td>
<td>oldpeak</td>
<td>thalach</td>
<td>Chol</td>
</tr>
<tr>
<td>forth</td>
<td>oldpeak</td>
<td>chol</td>
<td>sex</td>
<td>sex</td>
<td>exang</td>
<td>Exang</td>
</tr>
<tr>
<td>fifth</td>
<td>age</td>
<td>thalach</td>
<td>fbs</td>
<td>thalach</td>
<td>chol</td>
<td>Thalach</td>
</tr>
</tbody>
</table>

Table 3 shows the value of area under ROC for all the applied classification algorithms. The area under ROC represents a common area of true positive rate and false positive rate. Decision Tree shows higher performance than them. On the other hand, LR provided the good result. Fig. 12 represents the ROC curve, which is built by the value of the true positive.
rate and false positive rate. It is a graphical representation of the area under ROC.

DISCUSSION

Heart disease is a complication that can lead to death, so it is a very important issue. Machine learning techniques in the field of medicine can be used to analyze a set of data related to a disease and to predict the disease. In this paper, machine learning algorithms are used to predict heart disease. The data set is related to heart disease collected from the UCI repository.

Table 3: Value of area under ROC

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLP</td>
<td>0.886</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.897</td>
</tr>
<tr>
<td>SVM</td>
<td>0.500</td>
</tr>
<tr>
<td>Bagging</td>
<td>0.888</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.995</td>
</tr>
<tr>
<td>SGD</td>
<td>0.500</td>
</tr>
<tr>
<td>LR</td>
<td>0.906</td>
</tr>
<tr>
<td>KNN</td>
<td>0.658</td>
</tr>
<tr>
<td>Extra Tree</td>
<td>0.870</td>
</tr>
<tr>
<td>XGBOOST</td>
<td>0.829</td>
</tr>
</tbody>
</table>

CONCLUSION

In algorithms, the importance of features and cross-validation are the result of effective performance and increase accuracy. The value under the ROC curve and evaluation criteria such as accuracy, sensitivity, accuracy and F1 score are compared to a number of machine learning classification algorithms to assess heart disease risk and actually predict heart disease to identify the best appropriate classifier. As a result of evaluation, Decision Tree algorithm with accuracy value equal to 83%, AUC ROC equal to 99%, Logistic Regression algorithm with accuracy value equal to 88% and AUC ROC equal to 91% have a better performance than other algorithms.

FINANCIAL DISCLOSURE

No financial interests related to the material of this manuscript have been declared.

REFERENCES


