EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

Prediction of Heart Disease using Biomedical Data through Machine Learning Techniques

Nagaraj M. Lutimath[1], Neha Sharma[2,*] and Byregowda B K[3]

[1]Associate Professor, Department of Information Science and Engineering, East Point College of Engineering and Technology, Bangalore, India
[2]Assistant Professor, Department of AIT Computer Science and Engineering, Chandigarh University, Punjab, India
[3]Assistant Professor, Department of Information Science and Engineering, Sir M Visveswaraya Institute of Technology, Bengaluru, India

Abstract

INTRODUCTION: Random Forests are an important model in machine learning. They are simple and very effective classification approach. The random forest identifies the most important features of a given problem.
OBJECTIVES: The heart disease is cardiovascular disease, with a set of conditions affecting the heart. During heart disease, there will be heartbeat problems with congenital heart disorders and coronary artery defects. A coronary heart defect is a heart disease, which decreases the flow of blood to the heart. When the flow of blood decreases heart attack occurs. It is necessary to analyse the prediction of heart attack based on the symptoms.
METHODS: The available data set of patients with heart defects symptoms is taken and analysed in this paper using the random forest and decision tree regression models. The missing data is updated using mean value of the attribute. Python language is used to predict the accuracy.
RESULTS: Three performance measures are taken for analysing the available UCI Cleveland data set for heart disease. The performance measures are the Mean Absolute Error, Mean Squared Error and Root Mean Squared Error. Vital attributes of the data set are taken for analyses using the random forest regression model and decision tree regression model. The analyses shows that the slope attribute provides the better prediction for the heart disease. The results are shows that the females are more prone to heart attack.
CONCLUSION: Prediction of heart disease using the UCI machine learning data set at Cleveland repository is analysed using random forest regression model and decision tree regression models. We find random forest regression model provides better accuracy than decision tree regression model.

Keywords: Machine Learning, Random Forest, Python, Bio-Medical data, Information Retrieval, Predictive Analysis

Received on 11 December 2020, accepted on 23 August 2021, published on 30 August 2021

doi: 10.4108/eai.30-8-2021.170881

1. Introduction

Machine Learning is the technology to improve the performance of the machines by developing efficient algorithms so that they learn by experience for a given task. Classification is one such method that makes the machines to learn. The well-known procedure for classification are decision trees, regression, random forests, artificial neural networks and support vector machines. Machine learning techniques are suitably utilized with data set for classification accuracy is predicted. The data set used for implementation are heart diseases data set, Diabetes Mellitus dataset and Liver disease dataset [1]. Big data analytics is used with the healthcare data set in medical domain after pre-processing the data set for prediction of

*Corresponding author. Email: nehasharma0110@gmail.com

diseases with no extra overhead [2] The heart disorder data set is scaled for validation. Spark language frame work is utilized for prediction of heart disease. Data set consisted of 600 dataset records. 98% prediction accuracy was attained.

Kumar, et.al have analysed and compared with various machine learning techniques [3]. The best predicting approach was learned using suitable parameters. Different machine techniques are used with diabetes dataset. The techniques used are Random forests, Support Vector Machines with decision tree approaches. The dataset is pre-processed by using R language in R studio. Hadoop Distributed File System (HDFS) is also used. HIVE is used for further data cleaning of the data set. Prediction and analysis are done using the R studio integrated development environment. Rapidminer application was utilized to design and execute the experiment outcomes in Kumar et. al [4]. The data set consisted of 400 records. The data set was taken from UCI machine learning repository for analysis. Classification of Kidney disorder using Artificial neural networks, Naïve Bayes learning procedures were applied. The results showed that Naïve Bayes approach was better in classification, as it classified with 100% prediction accuracy compared to the Artificial Neural Network algorithm with 72.3% accuracy. Many related works are done taking the medical data for diagnosing heart attack. Coronary heart disease (CHD) is vital heart disability in grownups that causes to death in most of the developed countries in the world. Prediction model taking C4.5 decision tree was used on the heart data set for classification [5]. Avoiding over fitting of data set is a vital ingredient of decision tree. A prediction accuracy of a model can be increased by mixing pre pruning and post pruning approach with decrease in the size of the tree [6]. It increased the prediction accuracy when compared with other classification learning methods such as the CART, ID3 and C4.5 decision trees. Other decision learning approaches such as the neural networks, Bayesian networks, association rule learning have their own importance. These learning techniques were compared using available data set for heart disease [7]. In this paper a data set on heart disease is analysed using random forest learning. The random forest regression model is compared with decision tree regression model. The paper is organized as follows section 2 discusses classification. Section 3 describes classification methods. Section 4, 5, 6 and 7 briefs feature engineering, performance measures, prediction analysis and conclusion.

2. Classification

Classification is an important procedure for machine learning. Supervised, unsupervised and semi-supervised learning are the three forms of classification. In supervised learning approach examples consists of labelled classes. The classification learning approach considers categorical values but the regression procedure takes numerical values. In the unsupervised learning approach, the classes are not labeled but are grouped into clusters as per their attribute characteristics. Semi-supervised learning utilizes both labeled and unlabelled class data. The classification learning is normally a supervised procedure that takes an example in the data set and identifies to it to a class attribute. An example has two parts the predictor attribute values and target attribute values respectively. The predictor attribute values are used to predict the values of target attribute value and also to predict the class of an example. The training and test data set are disjoint sets obtained by segregating the data set during the classification procedure [8]. The classification process consists of two stages. During the training stage the model is trained on the training data. After training stage, we use the trained model on the test data for the prediction of the target attribute value. Using the classification model, we can obtain the relationship between predictor attribute value and the corresponding class to which it belongs. In the testing phase, the actual class of the just classified example is predicted. If the data in the test data set were not seen during training the prediction accuracy can be improved. The knowledge learned by a classification model can be described in the form decision tress, association rule learning and artificial neural networks to name a few.

3. Classification Methods

The paper discusses two important methods for classification, the decision tree and random forest.

3.1. Decision Tree

The decision tree model is more powerful for classification problems. It is a representation of a decision procedure for determining the class of a given instance. The process of this model is divided into two steps, the training phase, and the prediction phase. In the training phase, the decision tree is built using the training data set with calculation of statistical measure such as entropy and information gain of the attributes in the data set. In the prediction phase, the decision tree is used to predict the target class for the given test data.

3.2. Random Forest

Random forest is one of the supervised learning approaches which uses classification and regression methods. A group of decision trees form a random forest. Random forest uses Bootstrap aggregating. Variance is reduced in random forest and it helps to avoid overfitting. Bootstrap aggregating or bagging uses random selection of features along with replacement. Voting concept is used to select the root. Root is the node with highest votes. Random forests are not pruned or reduced into branches. Prediction accuracy is normally high in random forests.

4. Feature Engineering

Data set repository for heart disorder from UCI is used for the process of classification. Training and test data sets are obtained by segregating the data set. During feature engineering we consider suitable attributes for training the model. The trained classification model is then used to predict the class of the examples in the test data. The problem statement is described as follows:" *To predict the value for the patients suffering from heart disease using Random Forests*" The prediction analysis is done on the Data Set from
"Heart disease diagnosis from the Cleveland dataset taken from UCI Machine Repository". The attributes are defined as data fields as shown below.
The data set has the following attributes as data fields,
c_age- This characteristic feature of the attribute indicates the age in terms of number of years.
c_sex- The characteristics of this feature indicates the sex of the patient., specified in male and female with a value of 1 and 0 respectively.
c_cp- The characteristics of this feature indicates the chest pain category for typical angina, atypical angina and asymptomatic category with values 1, 2 and 3 respectively.
c_trestbps- the value of BP at rest when the patient is admitted. It is measured in mm/Hg.
c_chol- The feature is used for serum cholesterol measured in mg/dl.
c_fbs- Represents the level of fasting blood sugar. It is true or 1 when the measured fasting blood sugar is more than 120 mg/dl otherwise considered to be false or 0.
c_restecg- Specified as 0 for normal and 1 for wave abnormality in ST-T with inversion in T wave and/or evaluation or depression in ST with > 0.05mV. Definite observation of left ventricular hypertrophy by Estes' criteria and is represented by a value of 2.
c_thalach- This attribute is specified for person suffering from maximum heart rate. c_exang-The characteristic feature of this attribute with values of 1 and 0 where 1 for exercise induced angina with categorical values of yes and no. **c_oldpeak-**The characteristic feature of this attribute is for ST depression made by exercise relative to rest.
c_slope- Which represents the peak exercise slope in ST segment were up sloping, flat and down sloping is represented by values 1,2 and 3 respectively.
c_ca- The count of major vessels from 0 to 3 for fluoroscopy coloring.
c_thal- This attribute represents the type of heart disease with values of 3 for normal person, 6 and 7 for fixed disorder and reversible defect respectively.
c_num- This attribute represents the prediction of the persons with heart disorder."
The heart disease UCI data set repository of the Cleveland contains 303 examples and is segregated into 211 examples for training set and remaining 92 examples for test data. The training and testing data set are created using Python code as shown in the equation (1).

train=df. iloc [0:211:] (1)

The random forest regression model is then constructed using equation (2) below,

c=RandomForestRegressor (bootstrap=True, criterion='mse', max_depth=3, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=100, n_jobs=None, oob_score=False, random_state=0, verbose=0, warm_start=False) (2)

The regression tree is constructed using equation (3) below with 8 estimators, depth size of 2 and random state 0 in Python.

c=RandomForestRegressor (n_estimators = 8, max_depth=2, random_state = 0) (3)

The random forest is then fitted with training size in equation (4).
r=c.fit (train, target) (4)
where train is training data set and target is target class label values.
The decision tree regression model is then constructed using equation (5) below,
dt=DecisionTreeRegressor (criterion='mse', max_depth=None, max_features=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, presort=False, random_state=None, splitter='best') (5)

The decision tree is constructed using equation (6) below, min_samples_leaf=2, max_depth=2, random_state = 0 in Python.

dt=DecisionTreeRegressor (min_samples_leaf=2, max_depth=2, random_state = 0) (6)

The decision forest is then fitted with training data set and target class values in equation (7).

d=dt.fit (train, target) (7)

Analysis is done for both the random forest regression and decision tree regression models.

5. Performance Measures

The three performance measures used in this work are Mean Absolute Error (MAE) which is obtained by calculating the difference of the mean between absolute actual and predicted values. Mean of Squared Error (MSE) and Root Mean Squared Error (RMSE) for predictive analysis.

MAE is given by,

$$MAE=(\boldsymbol{y_i} - \boldsymbol{o_i}) \quad (8)$$

In the equation 8, MAE is the Mean Absolute Error, y_i is the ith actual data set and o_i is the ith predicted data set value.
MSE is calculated using the mean of the squares of the actual and predicted values of the data set.

It is given by,

$$MSE-=\frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{f_i} - \boldsymbol{p_i}) \quad (9)$$

In the equation 9, MSE is the Mean Squared Error, f_i is the ith value of an instance in the data set, p_i is the ith predicted instance value of data set and n indicates the number of test samples.

RMSE is defined as the square root of the average of squared errors

It is given by,

$$RMSE=\sqrt{\frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{x_i} - \boldsymbol{y_i})} \quad (10)$$

In the equation 10, RMSE is the Mean Squared Error, x_i is the ith value of an instance in the data set, y_i is the ith predicted instance value of data set and n indicates the number of test samples

6. Predictive Analysis

In this study of prediction analysis, preprocessing of data set is done first. After preprocessing we evaluate the mean of the attribute values for representing the missing data. The performance measures used during the prediction process are namely MAE, RMSE and MSE. These measures are calculated using the training and test models on the heart disease data using random forest regression model and decision tree regression models. Observing Table 1 the value of MAE is lesser than MSE and RMSE for random tree regression model than decision tree regression model. In Table 2 and Table 3, we find the MAE, MSE and RMSE are minimum when c_sex is female with values of 0.020, 0.004 and 0.070 respectively for random forest regression model. Thus, when c_sex attribute with female value, performs better prediction. Now looking at Table 4 and Table 5, we observe that when c_cp equals 2 MAE, MSE and RMSE have a lower value of 0.020,0.003 and 0.054 respectively for random forest model. Thus, the prediction accuracy of the model is better in this case of c_cp attribute. Now analysing the contents of Table 6 and Table 7 and Table 8 we observe that MAE is minimum value of 0.011 when c_slope attribute is 1 for random forest regression model. Now analysing the contents of Table 9 and Table 10 we observe that MAE is minimum with values are 0.011 when c_age attribute analysed for minimum training data set age. Thus c_age performs better4 in this case. We obtain a low value for MAE, MSE and RMSE considering the attributes c_sex, c_cp, c_slope and c_age. The values are 0.011, 0.003 and 0.051 respectively This occurs when c_slope has a value of 1. Thus, the attribute c_slope provides better prediction.
Thus, we see that random forest regression model provides better accuracy than decision tree regression model. The analyses also shows that there is less deviation for the females, so there is more chance that females are affected by heart disease than males for the given parameter measures.

Table 1. Values of MAE, MSE, RMSE for overall data set for the Random Forest and Decision Tree Models

Error_Type	Value for Random Forest Model	Value for Decision Tree Model
MAE	0.040	0.043
MSE	0.018	0.019
RMSE	0.135	0.138

Table 2. Values of MAE, MSE, RMSE for overall data set for the Random Forest and Decision Tree Models for c_sex=1

Error_Type	Value for Random Forest Model for c_sex=1	Value for Decision Tree Model for c_sex=1
MAE	0.0520	0.0550
MSE	0.0260	0.0260
RMSE	0.1637	0.1639

Table 3. Values of MAE, MSE, RMSE for overall data set for the Random Forest and Decision Tree Models for c_sex=0

Error_Type	Value for Random Forest Model for c_sex=0	Value for Decision Tree Model for c_sex=0
MAE	0.020	0.023
MSE	0.004	0.006
RMSE	0.070	0.081

Table 4. Values of MAE, MSE, RMSE for overall data set for the Random Forest and Decision Tree Models for c_cp=2

Error_Type	Value for Random Forest Model for c_cp=0	Value for Decision Tree Model for c_cp=2
MAE	0.020	0.023
MSE	0.003	0.004
RMSE	0.054	0.063

Table 5. Values of MAE, MSE, RMSE for overall data set for the Random Forest and Decision Tree Models for c_cp=4

Error_Type	Value for Random Forest Model for c_cp=4	Value for Decision Tree Model for c_cp=2
MAE	0.076	0.081
MSE	0.035	0.035
RMSE	0.180	0.180

Table 6. Values of MAE, MSE, RMSE for overall data set for the Random Forest and Decision Tree Models for c_slope=1

Error_Type	Value for Random Forest Model for c_slope=1	Value for Decision Tree Model for c_cp=1
MAE	0.011	0.012
MSE	0.003	0.003
RMSE	0.051	0.051

Table 7. Values of MAE, MSE, RMSE for overall data set for the Random Forest and Decision Tree Models for c_slope=2

Error_Type	Value for Random Forest Model for c_slope=2	Value for Decision Tree Model for c_slope=2
MAE	0.055	0.060
MSE	0.022	0.022
RMSE	0.140	0.140

Table 8. Values of MAE, MSE, RMSE for overall data set for the Random Forest and Decision Tree Models for c_slope=3

Error_Type	Value for Random Forest Model for c_slope=3	Value for Decision Tree Model for c_slope=3
MAE	0.190	0.180
MSE	0.140	0.140
RMSE	0.380	0.380

Table 9. Values of MAE, MSE, RMSE for overall data set for the Random Forest and Decision Tree Models for age greater than mean c_age

Error_Type	Value for Random Forest Model for c_age>mean c_age	Value for Decision Tree Model for c_age>mean c_age
MAE	0.053	0.055
MSE	0.028	0.028
RMSE	0.160	0.160

Table 10. Values of MAE, MSE, RMSE for overall data set for the Random Forest and Decision Tree Models for age greater than minimum c_age

Error_Type	Value for Random Forest Model for c_age>min c_age	Value for Decision Tree Model for c_age>min c_age
MAE	0.040	0.043
MSE	0.018	0.018
RMSE	0.135	0.135

7. Conclusion

In this paper prediction of heart disease using the UCI machine learning data set at Cleveland repository is analysed using random forest regression model and decision tree regression models. We find random forest regression model provides better accuracy than decision tree model. We have calculated the performance measures namely MAE, RMSE and MSE and it was observed that MAE, MSE and RMSE for the male is more than the female for the given attribute in the data set. We find female are more affected to heart disease than male for the

given parameters as there is less deviation from actual values. From the analysis of random forest regression, we find the prediction model performs better for c_slope attribute value in the data set in terms of MAE MSE and RMSE. In future the accuracy of the prediction can be improved by utilizing other machine learning methods such as multiple regression model, deep neural networks, association rule mining, deep learning and genetic algorithms.

References

[1] Shweta Ganiger, K.M.M. Rajashekharaiah, "Chronic Diseases Diagnosis using Machine Learning", IEEE, 2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET), Kottayam, India, Dec. 21-22, 2018, pp.1-6

[2] Rashmi G Saboji, Prem Kumar Ramesh, "A Scalable Solution for Heart Disease Prediction using Classification Mining Technique", IEEE, International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS-2017). Chennai, India, Aug 1-2,2017pp.1780-1785.

[3] Kumar, P. Suresh, and S. Pranavi. "Performance analysis of machine learning algorithms on diabetes dataset using big data analytics." Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS), 2017 International Conference on. IEEE, 2017.

[4] Kunwar, Veenita, et al. "chronic kidney disease analysis using data mining classification techniques." Cloud System and Big Data Engineering (Confluence), 2016 6th International Conference. IEEE, 2016

[5] Minas A. Karaolis, Member, IEEE, Joseph A. Moutiris, Demetra Hadjipanayi, Constantinos S. Pattichis, "Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees", IEEE transactions on information technology in biomedicine, Vol. 14, No. 3, May 2010, pp.559-566.

[6] Ali Mirza Mahmood1, 2* Mrithyumjaya Rao Kuppa, "Early detection of clinical parameters in heart disease by improved decision tree algorithm", Second Vaagdevi International Conference on Information Technology for Real World Problems, 2010, pp. 24-29.

[7] František Babič, Jaroslav Olejár, Zuzana Vantová, Ján Paralič, "Predictive and Descriptive Analysis for Heart Disease Diagnosis", Proceedings of the Federated Conference on Computer Science and Information Systems, Prague, 2017, ISSN 2300-5963 ACSIS, Vol. 11, DOI: 10.15439/2017F219, pp. 155–163

[8] Y. Yan, "Gingivitis detection by Fractional Fourier Entropy and Biogeography-based Optimization," 2020 IEEE/ACM 13th International Conference on Utility and Cloud Computing (UCC), 2020