Prediction of Heart Disease using Machine Learning

Nagaraj M. Lutimath, Chethan C, Basavaraj S Pol

Abstract: ***Machine learning is one of the fast growing aspect in current world. Machine learning (ML) and Artificial Neural Network (ANN) are helpful in detection and diagnosis of various heart diseases. Naïve Bayes Classification is a vital approach of classification in machine learning. The heart disease consists of set of range disorders affecting the heart. It includes blood vessel problems such as irregular heart beat issues, weak heart muscles, congenital heart defects, cardio vascular disease and coronary artery disease. Coronary heart disorder is a familiar type of heart disease. It reduces the blood flow to the heart leading to a heart attack. In this paper the UCI machine learning repository data set consisting of patients suffering from heart disease is analyzed using Naïve Bayes classification and support vector machines. The classification accuracy of the patients suffering from heart disease is predicted using Naïve Bayes classification and support vector machines. Implementation is done using R language.***

Keywords—Naïve Bayes Classification, Support Vector Machines, UCI machine learning repository data set, R Studio

I. INTRODUCTION

Machine Learning plays a vital role in diagnosing a heart disease. Some of the machine learning techniques are decision trees, neural networks, Naïve Bayes classification, genetic algorithms, regression and support vector machines. The decision tree algorithm is used for extracting rules in predicting heart disease. C5.0 decision tree procedure was accomplished using Cleveland data set. Its accuracy value of 85.33% was compared to the rest of the algorithms[1] [2]. It found to better than other machine learning algorithms. A graphical user based interface was used to input the patient data and predict whether the patient is suffering from heart disease or not, using Weighted Association rule based Classification. Results showed that Weighted Associative Classification was providing improved accuracy as compare to other already existing Associative Classifiers. Naïve Bayes is a probability based classification [3]. Medical attributes such as blood pressure, age, sex were used for prediction of heart disease. MatLab was used for implementation. A prediction model that uses combination of both pre pruning and post pruning of decision tree learning improved the classification accuracy by reducing the tree size [4].Other techniques in machine learning such as regression, neural networks, support vector machines and genetic algorithms can also be utilized for prediction.

This paper provides a comparison of support vector machines with Naïve Bayes classification and radial kernel support vector machine. The dataset used is UCI machine learning data set repository.

This paper is structured as follows, related works is explained in section II, methodology and data set analysis is described in section III section IV illustrates the feature engineering, section V presents prediction analysis and lastly section VI with conclusion.

II. RELATED WORKS

Heart disease is a vital disease explored by the researchers, in predicting the patients suffering from this disease. Machine learning is the art of construction of programs that learn from experience for a given problem. The approaches utilized in machine learning are decision trees, neural networks, Naïve Bayes classification, support vector machines and genetic procedures [5] for classification of the data set. Decision tree C4.5 and Fast Decision trees were studied [6] using a suitable medical data. Medical data sets was used from UCI repository. An accuracy of 69.5% was achieved for decision trees and an accuracy 78.54% was achieved for fast decision tree.

Analyses and predicting of coronary artery heart disease was done, utilizing a data set consisting of 335 records indicating the various 26 attributes [7]. The data set was pre-processed using correlation concept. Features selection and extraction was completed using particle swarm optimization (PSO) approach. The neural network, regression, fuzzy and decision tree models were modeled. The data set was applied to neural network model. An accuracy of 77% was found. It was further applied to regression model. This resulted in the accuracy of 83.5%. The other fuzzy and decision tree model did not show any major changes.

The data set was then optimized by utilizing the pre-processing approach. Correlation, feature selection and extraction with PSO, K-means clustering were used. The classification of the data set using one of the procedure or a combination of them was done. An accuracy of 88.4% was result for the regression model. The data set was further applied to hybrid model. The accuracy of

Revised Manuscript Received on September 25, 2019

Nagaraj M. Lutimath, Department of Computer Science and Engineering, Sri Venkateshwara College of Engineering, Bengaluru, India
Email: nagarajlutimath@gmail.com
Chethan C, Department of Information Science and Engineering, Sri Venkateshwara College of Engineering, Bengaluru India
Email: chethac123@gmail.com
Basavaraj S Pol, Department of Computer Science and Engineering, R L. Jalappa College of Engineering, Doddaballapur, India
Email: basavaraj.pol@gmail.com

classification procedures improved from 8.3 % to 11.4 %.
Another work on prediction of the heart disease was accomplished by pre-processing of the data by feature selection utilizing Ginni Index and support vector machine [8]. Classification of the data was further completed using suitable classification techniques. The algorithms used for classification were the Naive Bayes probability classification, Sequential Minimal Optimization (SMO) algorithm. SMO with bagging and artificial neural network models were also added for analysis. An accuracy of 93.4% was obtained for SMO with bagging. 75.51% accuracy for Naive Bayes probability classification. 94.08 accuracy for SMO and 88.11 accuracy for the neural network models. Verification of the results was completed by 10-cross fold validation method.
An Apriori procedure using the Transaction Reduction Method (TRM) was applied to in diagnosing the heart disease using a suitable medical data set [9] [10]. The obtained results were compared it with some of the classical methods. An accuracy of 93.75% was achieved using the algorithm. When SMO was utilized 92.09% an accuracy was obtained. When SVM was used 89.11% accuracy was achieved. C4.5 decision tree resulted in 83.85% accuracy and Naïve Bayes probability classification an accuracy of 80.15% accuracy was the result.
All the techniques mentioned above deal with the predictive analysis using classical methods. The classification approaches like decision trees, Naive Bayes, Support Vector Machines or neural networks are the models for consideration using suitable medical data sets.

III. Methodology and Dataset Analysis

Multi-dimensional data is collected from various sources and pre-processed and transformed into a suitable format. Then machine learning approach is applied on this data for further classification.

A. Experimental Procedures

SVM is a significant method for supervised classification. A hyperplane is utilized in classification of the target classes. Classification is performed by identifying the hyperplane that divides one class with the other classes. Training time for the SVM is very slow but it is very accurate in predicting the target classes.

IV. Feature Engineering

For studying classification process data set from UCI machine learning repository for heart disease at Cleveland is considered. The dataset is divided into two sets, the test data set and the training data set. The related feature engineering is done on the training data, and model thus obtained is utilized on the test data to predict the results.
The problem statement is defined as, *"To predict and analyze the value for the patients suffering from heart disease using support vector machine"*
To group the features with heart disease data set in order to analyze the number of patients with heart disease disorder.
Data Set used is the *"Heart disease diagnosis from the Cleveland dataset taken from UCI Machine Repository"*.
The variables are defined as data features as shown below.
f_age- age attribute given in years
f_sex- sex attribute categorized into male values 1 and female with value 0.
f_cp- chest pain attribute is categorized into values 1, 2, 3 and 4 in for angina, atypical angina, non-anginal pain, asymptomatic respectively.
f_trestbps-resting blood pressure (BP) attribute expressed in mm Hg, when the person is admitted to the hospital.
f_chol- serum cholesterol expressed in mg/dl
f_fbs- Fasting blood sugar > 120 mg/dl attribute with true and false indicated numerically by 1, 0.
f_restecg- attribute for resting electrocardiographic outcome expressed with values 0,1 for normal and S T-T wave abnormality(T wave inversions and/or ST elevation or depression of > 0.05 mV), 2= showing probable or definite left ventricular hypertrophy by Estes' criteria)
f_thalach- attribute for maximum heart rate of the patient.
f_exang- attribute for exercise induced angina indicated numerically by 1 and 0 for yes and no categorical values.
f_oldpeak- attribute for ST depression induced by exercise relative to rest
f_slope- attribute for the slope of the peak exercise ST segment expressed in terms of up sloping, flat and down sloping with values 1, 2 and 3 respectively.
f_ca- attribute for count of major vessels with a range from (0-3) with flourosopy coloring.
f_thal- attribute for type of heart defect with value 3 for normal, 6 for fixed defect and 7 for reversable defect
f_num- attribute for predicting the patients suffering from heart disease.
The input data set of 303 tuples is distributed into 258 tuples for training data set and 45 tuples into test data set. The dataset for training is executed in R and is taken using the equation 1 and 2.

```
split <-subset (dataset, SpiltRatio=0.85)          (1)
training_set=subset (dataset, split=TRUE)          (2)
```

The test data set is then calculated using equation 2.
The formula is computed using the equation 3 below,

```
formula = (f_num~f_age+f_sex+f_cp+ f_trestbps+ f_chol+
f_fbs+f_restecg+f_thalch+f_exang+f_oldpeak+f_slope+
f_ca+ f_thal)                                      (3)
```

In the equation (3) f_num is the predictor attribute, f_age, f_sex, f_cp, f_trestbps, f_chol, f_fbs, f_restecg, f_thalach, f_exang, f_oldpeak, f_slope, f_ca and f_thal are the response attributes.
The naïve bayes model is then designed using the equation 4.

```
fit1=naivebayes (as.factor (formula), data=training_set)
(4)
```

The parameters used in the naivebayes function of equation 4 are the formula which is used from equation 3, data is the training_set calculated from equation 2, naivebayes function utilized for classification analysis using Naïve Bayes classification. The function as.factor is used in the naivebayes is used because Naïve Bayes returns the

posterior conditional probabilities for the class variable v_num.

The classification accuracy results are further analyses with SVM using radial kernel using the following SVM model equation in equation 5.

fit2 = svm (formula, data = training_set, type = 'C-classification', kernel = 'radial') (5)

A. *Performance Measures*

Some of the important parameters used in the performance analyses of the data set are the Mean Absolute Error (MAE), Sum of Squared Error (SSE) and Root Mean Squared Error (RMSE). MAE is the square root mean of the absolute value of actual values subtract the predicted values of the instances in the data set. SSE is summation of the squares of the actual instance values minus the predicted instance values of the data set RMSE is the Root mean of the squares of the actual instance values minus the predicted values in the data set.

V. PREDICTION ANALYSIS

Before prediction analyses the data is preprocessed and missing data are evaluated using mean of the attribute. The MAE, SSE and MSE are calculated for the test dataset heart disease data set and are listed in Table I.

In the Table I the values of MAE, SSE and RMSE are calculated for Naïve Bayes classification and radial kernel. The values of MAE, SSE and RMSE are lower in case of radial kernel than Naïve Bayes classification. Now analyzing Table II, we find the lowest value of MAE is 1.42 for f_sex is female. RMSE 1.85 for f_sex is female. We also observe that SSE is lower when f_sex is female, which also supports the evidence that the model predicts with higher accuracy when f_sex is female.

TABLE I. MAE, SSE AND RMSE FOR OVERALL TEST DATA SET

Error Type	Using Naïve Bayes	Radial Kernel
MAE	1.98	0.78
SSE	309	65
RMSE	2.62	1.20

TABLE II. MAE , SSE AND RMSE FOR MALE AND FEMALE FOR f_sex using Naïve Bayes Classification

f_sex	MAE	SSE	RMSE
male	2.18	268	2.85
female	1.42	41	1.85

TABLE III. . MAE , SSE AND RMSE FOR MALE AND FEMALE FOR f_cp using Naïve Bayes Classification

Type of Error	Value of f_cp=1	Value of f_cp=2	Value of f_cp=3	Value of f_cp=4
MAE	1.33	1	2.62	2.16
SSE	28	6	79	196
RMSE	2.16	1	3.14	2.8

TABLE IV. . MAE , SSE AND RMSE FOR MALE AND FEMALE FOR f_slope for Naïve Bayes Classification

Type of Error	f_slope=1	f_slope=2	f_slope=3
MAE	1.81	2.3	1.25
SSE	126	174	9
RMSE	2.45	2.95	1.5

TABLE V. MAE , SSE AND RMSE FOR MALE AND FEMALE FOR f_sex for radial kernel

f_sex	MAE	SSE	RMSE
male	0.70	51	1.24
female	1	14	1.08

TABLE VI. . MAE , SSE AND RMSE FOR MALE AND FEMALE FOR f_cp for radial kernel

Type of Error	Value of f_cp=1	Value of f_cp=2	Value of f_cp=3	Value of f_cp=4
MAE	0.67	1	1.23	0.64
SSE	4	6	15	40
RMSE	0.82	1	1.37	1.26

TABLE VII. . MAE , SSE AND RMSE FOR MALE AND FEMALE FOR f_slope for radial kernel

Type of Error	f_slope=1	f_slope=2	f_slope=3
MAE	0.81	0.7	1
SSE	25	34	10
RMSE	1.26	1.30	1.22

Now observing Table III we see that minimum value of MAE and RMSE is 1. This occurs when f_cp has 2 as its value. Thus the model predicts better in this case. We also observe that the highest values of MAE and RMSE are 2.62 and 3.14 respectively. Thus the prediction model deviates from the actual values in this case.

Now observing Table IV we see that the lowest values of MAE and RMSE are 1.25 and 1.5 respectively. This occurs when the f_slope has 3 as its value. Hence the prediction accuracy of the model is better in this case. The highest value of the MAE and RMSE in Table IV are 2.3 and 2.95, this happens when the value of f_slope is 2, thus the prediction model deviates from the actual values in this case. The prediction model behaves moderately when the value of f_slope is 1.Using the tables Table II, Table III and Table IV, we find that the minimum MAE and RMSE considering the attributes f_sex, f_cp and f_slope we get 1. This occurs for attribute f_cp for value 2.Thus the model predicts better for this value of the attribute f_cp.

We now analyse the svm for radial kernel. Consider Table V, we find the lowest value of MAE is 0.7 for f_sex is male. RMSE 1.08 for f_sex is female. We also observe that SSE is lower when v_sex is female, which also supports the evidence that the model predicts with higher accuracy when f_sex is female.

Now observing Table VI we see that minimum value of MAE is 0.64 for f_cp is 4, and RMSE is 1.23 for f_cp is 2. The minimum value for SSE is 6

when f_cp is 2. Thus the model predicts when f_cp is 4. We also observe that the highest values of MAE and RMSE are 1.23 and 1.37 respectively.

Thus the prediction model deviates from the actual values when f_cp is 3.

Now observing Table VII we see that the lowest value of MAE is 0.7 for f_slope 2. Observing the lowest value of RMSE we find RMSE is 1.22. This occurs when the f_slope is 3. Hence the prediction accuracy of the model is better in this case. The highest value of SSE is 34, this happens when the value of f_slope is 2.The prediction model deviates from the actual values in this case. Using the tables Table V to Table VII, we find that the minimum MAE considering the attributes f_sex, f_cp and f_slope we get 0.7. This occurs for f_cp for value 1.Thus the model predicts better for this case. From tables Table II to Table VII, the minimum value of SSE is 6 for f_cp and f_sex with value 2 and f_cp for Naïve Bayes and radial Kernel respectively. The lowest value for MAE with radial Kernel is 0.67 for f_cp 1 for radial kernel classification. The lowest value for RMSE, considering Naïve Bayes and radial classification is 1 for f_cp 2 for Naïve classification. Thus model predicts better for Naïve Bayes classification in this case.

For the attribute f_sex the minimum value for MAE is 0.7. This occurs when f_sex is male for radial kernel. The minimum value for RMSE for the f_sex is 1.08. This occurs when f_sex is female for radial kernel SVM. The SVM with radial kernel performs better in this case. The minimum value for SSE is 14 for female f_sex. This occurs when SVM is radial. Thus SVM with radial kernel predicts better f_sex. Now for f_cp the minimum value for MAE is 1 for Naïve Bayes classification and radial kernel SVM. The lowest value of RMSE is 1. This is true for both Naïve Bayes and radial kernel SVM when f_cp is 2. Considering the minimum value of MAE for f_slope, we get 0.7 for radial kernel. The minimum value of RMSE for f_slope, we get 1.30 for radial kernel. Considering f_sex, f_cp and f_slope, radial kernel performs better for f_sex and f_slope. We find radial kernel SVM better than Naïve Bayes classification in terms these attributes. Observing the values RMSE for f_sex for Naïve Bayes and radial kernels, we find that RMSE is lower for female than male. Hence we predict females are affected by heart disease than the males.

VI. CONCLUSION

In this paper the Naïve Bayes classification and radial kernel SVM is used for prediction for the heart disease taking the UCI Cleveland data set. MAE, SSE and MSE are calculated utilizing suitable attributes of the data set using suitable features. In comparison with Naïve Bayes classification and radial kernel. SVM with radial kernel offers better accuracy than Naïve Bayes classification. The data set containing male and female attributes is also analyzed. We find female are more affected by the heart disease than male using the consistency measures RMSE. In future other machine learning techniques such as deep learning, association rule analysis and genetic algorithms will be studied in predicting the accuracy with suitable performance parameters.

REFERENCES

[1] Moloud Abdar,"Using Decision Trees in Data Mining for Predicting Factors Influencing of Heart Disease", Carpathian Journal of Electronic and Computer Engineering 8/2 , 2015,pp. 31-36.

[2] Jyoti, S., U. Ansari and D. Sharma, Sunita Soni, "Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers", .International Journal on Computer Science and Engineering (IJCSE),3: 23852392, 2011, pp. 2385-2392.

[3] Rupali, M and R.Patil, "Heart Disease Prediction System using Naive Bayes and Jelinek-mercer smoothing",.International Journal of Advanced Research in Computer and Communication Engineering, May 2014. Vol. 3, Issue 5,pp. 6787-6789.

[4] 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. 2429.

[5] 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.

[6] R. El-Bialy, M. A. Salamay, O. H. Karam, and M. E. Khalifa, "Feature Analysis of Coronary Artery Heart Disease Data Sets", Procedia Computer Science, ICCMIT 2015, vol. 65, pp. 459–468, doi: 10.1016/j.procs.2015.09.132.

[7] L. Verma, S. Srivastaa, and P.C. Negi, "A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data", Journal of Medical Systems, vol. 40, no. 178, 2016, doi: 10.1007/s10916-016-0536-z.

[8] R. Alizadehsani, J. Habibi, M. J. Hosseini, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, B. Bahadorian, and Z. A. Sani, "A data mining approach for diagnosis of coronary artery disease", Computer Methods and Programs in Biomedicine, vol. 111, no. 1, 2013, pp. 52-61, doi: 10.1016/j.cmpb.2013.03.0.

[9] Ch. Yadav, S. Lade, and M. Suman, "Predictive Analysis for the Diagnosis of Coronary Artery Disease using Association Rule Mining", International Journal of Computer Applications, vol. 87, no. 4, 2014, pp. 9-13.

[10] 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.