Predictive Analytics Using Rule Based Classification And Hybrid Logistic Regression (HLR) Algorithm For Decision Making

S. Clement Virgeniya, E. Ramaraj

Abstract: Human Beings are facing a lot of health problems due to changes in food habits. One such major problem is heart disease. The more prevalent death now-a-days is due to heart attack. Apart from food habits, lifestyle and genetic family history of the individual also plays a major role in heart disease. Variations in heart beat may definitely lead to heart related problems. Today's healthcare industry generates huge volumes of data every second. This data needs to be mined for taking proper decision. The accuracy of existing data mining and machine learning logistic regression algorithm in predicting heart disease is moderate. It is essential to propose a rule based classification technique with machine learning to predict the heart disease. The proposed work is implemented in python using secondary sources of data collected from Kaggle. A set of rules is applied to the dataset and model is developed. The model is fit into the testing dataset and given as input into the logistic regression algorithm. The proposed hybrid model builds a user defined classifier which shows higher accuracy of 86% compared to an existing logistic regression algorithm. The strength of the proposed model is compared with an existing system in terms of precision, accuracy, recall and F1 score. The obtained results help the physicians to accurately predict the heart disease and take appropriate decision in advance.

Index Terms: Accuracy, Healthcare, Heart disease, Logistic Regression, Machine learning, Predictive analytics, Rule based classifier.

1. INTRODUCTION

THE medical industry has changed a lot in the 21st century. This is mainly due to new inventions and innovations in science and technology. Apart from all these deaths due to heart attack is increasing year by year constantly. This is mainly due to certain risk factors like diabetes, high amount of bad cholesterol i.e. LPL (Low density lipoprotein), abnormal pulse rate, smoking habits, physical inactivity, nutrition and lifestyle of individuals. Studies reveal that deaths due to cardiovascular disease have increased from 2.57 crore in 1990 to 5.75 crore in 2018 [1]. Tamil Nadu stands third most prevalent state in India and nearly half of the total cardiovascular deaths is among people less than 70 years. Quitting Smoking and doing physical activity helps one to keep away from heart disease. In India BMI (Body Mass Index) has nothing to do with heart disease. BMI has no impact on heart disease. In recent years, many healthcare organizations are very keen on storing patients' details like age, sex, previous medical history, living lifestyle, food habits, physical activities, smoking patterns of patients. Though these details are true to some extent, they are not used for making any decision. This data cannot infer the probability of patients getting heart disease unless it is properly used. Many machine learning algorithms is used in predicting the probability of heart disease [2]. Classification is an important task in machine learning and data mining for predicting heart disease. Earlier a classifier is built based on the predefined outcome of the data. A classification algorithm is then applied to the training dataset to build a classifier that is used to predict the outcome on test dataset. For the purpose, mainly Support Vector Machines (SVM), Decision tree algorithm, Neural network were used. In the proposed system, a rule based classification technique is implemented along with logistic regression. Rule based classifier makes use of IF-THEN rules classification [3]. Rule based classification is very clear and provides a precise idea to the users. It predicts the outcome so that users have a clear logic behind the outcome. Rule based classification shows better accuracy when compared to other classification techniques. There are two popular rule based classification − rule Induction and classification based on association rule mining [3], [4]. Classification based on association rule mining is widely used by many researchers. Many algorithms are developed to detect association rule from large amounts of data. One such rule is Class Association Rules (CAR). The organization of this paper is as follows. Section 2, presents the related works in heart disease prediction. Section 3, presents Proposed Methodology. Section 4, presents Experimental Results of the proposed methodology. And Section 5, ends with Conclusion and Future Enhancement.

- *S. Clement Virgeniya is currently pursuing Ph.D. in Computer Science in Alagappa University, Karaikudi, Tamil Nadu, India.*
- *Dr. E. Ramaraj is currently Professor & Head in Dept. of Computer Science in Alagappa University, Karaikudi, Tamil Nadu, Ind*

2 RELATED STUDY

Research shows various studies regarding diagnosing heart disease using different data mining and machine learning algorithms. Different techniques like SVM, Logistic Regression, Decision Tree were used in predicting heart disease. Heart related issues is more prevalent in people having diabetes [5]. Data collected from National Heart Association reveal that 65 % of people with diabetes are facing heart disease mainly type 2 diabetes. The Framingham study was the first proof to reveal that the major cause of heart disease is with people having diabetes. Apart from diabetes, high blood pressure, smoking, high cholesterol levels, and family history of early heart disease show positive sign to Heart disease. Anam Mustaqeem et al. [6] developed a predictive model for classification of arrhythmias. He used a wrapper algorithm in random forest and other machine learning classifiers. The results conclude that MLP (Multi-Layer perceptron) beats others classifiers with an average accuracy of 78.26%. Prathibhamol Cp et al. [7] suggested clustering approach and regression methodology for predicting the type of cardiac arrhythmia disease. The author used a

clustering approach, namely DBSCAN and regression methodology like multiclass logistic regression and acquires an overall accuracy of 80%. Liaqat Ali et al. [8] proposed a hybrid model combining two SVM models based on their linearity. It optimizes two SVM algorithm simultaneously and its performance metrics are increased by 3.3% compared to existing SVM models. Raid Lafta et al. [9] Developed a Bagging Based Ensemble Model (BBEM) combining three classifiers namely artificial neural network, least square SVM, and Naive Bayes to predict heart disease in advance. The BBEM provides a promising tool for analyzing time series data and provides accurate results. Liaqat Ali et al. [10] proposed a statistical model based on deep neural network to overcome the conventional ANN and DNN models. The model is used by the physicians to accurately predict heart disease. Carlos Ordonez et al. [11] proposed a system where association rules are applied to predict heart disease. Unfortunately, association rules produce a large number of rules which are irrelevant and time consuming. The author proposed search constraints and test set validation with lesser number of association rules and with high predictive accuracy.

3 PROPOSED METHODOLOGY

The Architecture of the proposed system is given in Fig. 1

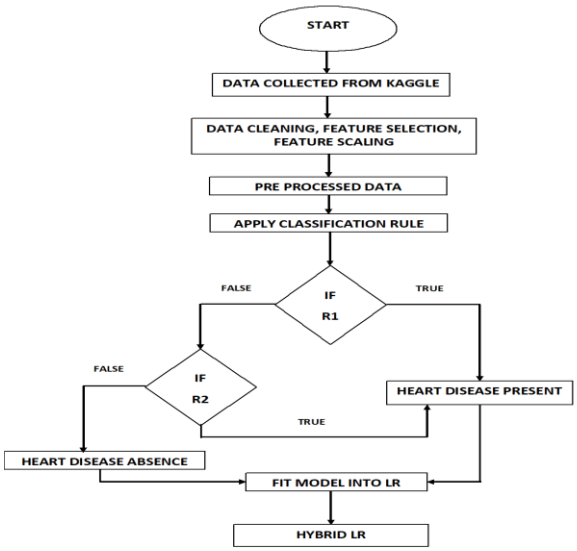

Fig. 1*. Predictive Analytics Using Hybrid Logistic Regression Model to Support Medical Practitioner for Heart Disease*

In this research work, secondary source of data collected from Kaggle is taken [12]. The dataset contains many missing values which affect the quality of data for decision making. Steps involved in proposed methodology include Data Pre-processing, Rule Base Classification and Hybrid Logistic Regression(HLR). The following Rule Based Algorithm illustrates the rules given to the training dataset as input and predicts the outcome to the test dataset.

Required Input: Training Dataset X_train –Attributes with target Class

Output: Predict heart disease in test dataset
Step 1: Define user defined function for X_train
Step 2: for ∀ Attributes find highest correlation
Step 3: for ∀ Attributes with highest correlation Apply the rules do
Step 4: If attribute REST_ECG in dataset X_train greater than zero
Return 1
Step 5: Else if Check for other attributes
If True
Return 1
Else
Return 0
Step 6: Add Outcome Attribute to Existing X_train
Step 7: Repeat for Testing Dataset
Step 8: Save Predicted Outcome to Test Dataset

3.1 Data Preprocessing

The raw dataset needs to be analyzed first. Analyzing data is the first step in Data Pre-processing. There are 920 subjects, out of this 85% of data is split into training data and 15% as test data. 779 subjects as training dataset and 141 subjects as test dataset. Out of this 779 subjects, 522 instances have missing values which are 67% of the total value and out of 141 subjects 30 instances have missing values which are 4% of the total value. Missing data are inaccurate and will not give any valuable information regarding decision making. There are many missing values, especially in Peak Exercise, Vessels coloured by Fluoroscopy, thallium attributes. These attributes will not give any meaning full value. They are removed and training dataset contains 633 subjects and testing dataset contains 111 subjects with 9 attributes. After analyzing and Splitting, correlations of each feature with predictive variable in the dataset are found. Attribute with the highest correlation is selected and rules are applied. The attributes like Resting ECG, Age, Chest Pain Type, Resting BP, Fasting Glucose show highest correlation to the diagnosis of heart disease compared to other features which show lesser value i.e., there exists a negative correlation between the attributes. The following Table 1 shows the correlation of attributes with predictive attribute. A good attribute is one that contains features highly correlated with predictive attribute and uncorrelated with non-predictive attribute. Data is normalized before feeding into the model. After undergoing dimensionality reduction, feature selection, and feature scaling the dataset undergoes rule based classification.

TABLE 1 *CORRELATION WITH PREDICTIVE VARIABLE*

Attribute Name	Correlation with Predictive Variable
Resting ECG	0.109829
Age	0.288526
Sex	0.307464
Chest Pain Type	0.473108
Serum Cholesterol	-0.136262
Heart Rate	-0.400902
Resting BP	0.145840
Fasting Glucose	0.095242
Heart Disease(Predictive Variable)	1.000000

3.2 Rule Based Classification

Rule Based Systems [13] uses a small number of attributes for decision making. Rule based algorithms provide logical conclusions. Rule-based classification adds new rules to the

existing problem. In rule based classification, the attributes are classified based on multiple if-then rules. A rule R covers an instance X if the attributes of the instance satisfy the condition of the rule. It provides a very good data model that can be easily understood by human beings. The following rules are generated to predict the presence of heart disease with the given dataset.

R1: (rest_ecg>0) → Heart Disease
R2: (age>=45) ∧ (chest_pain_type>1) ∧ (resting_bp>120) ∧ (fast_glucose>=0) → Heart disease.

Rule R1 checks whether Resting ECG is greater than 0. Resting ECG is the ECG that is taken during the admission of the patient. If it is 0 then the patient is normal and if it is greater than zero, it indicates risk of heart disease. Rule R2 checks the condition given above and then predicts the occurrence of heart disease. When Rule 1 is greater than zero, it confirms the risk of heart disease. The Class Association Rule as Features (CAR) [3] are applied to the existing dataset and fed into traditional data mining algorithm. Conditional statements are developed to make use of classification rules as attributes. The training and testing dataset is checked with classification rules and the attributes for classification rule is chosen by considering the correlation between the features. Resting ECG, Age, Chest Pain Type, Resting BP and Fasting Glucose are chosen as main attributes and classification rules are generated to these attributes by defining a user defined function termed as Condition. Rule 1 and Rule 2 explains the criteria for user defined function.

3.3 Hybrid Logistic Regression(HLR)

Logistic Regression is a Regression analytical technique which is used to predict the outcome of dichotomous dependent variables. It is used to find the probability of a certain class or event likely to happen i.e., occurrence or outcome of an event. In proposed work, a hybrid logistic regression is used. The Rule based classification technique is applied to training and testing dataset and it is fed into a logistic regression algorithm to form hybrid technique and results are evaluated. The following Hybrid Logistic Regression Algorithm is applied to training and testing dataset, and fit into logistic regression classifier to predict the target variables accuracy. Required Input: Training and Testing Dataset Output: Predict heart disease in the test dataset using HLR

Step 1: Apply Rule 1 on training dataset Else Rule 2
Step 2: Apply Rule 1 on testing dataset Else Rule2
Step 3: fit training model into logistic regression classifier
Step 4: Predict target variable with test dataset
Step 5: Calculate Confusion Matrix, Accuracy, Precision, Recall and F1 Score
Step 6: Save Predicted Outcome to test Dataset

4 EXPERIMENTAL RESULTS

Proposed Methodologies are carried over in Python using jupyter notebook. The dataset contains 920 subjects with 14 attributes and has many missing values which affect the quality of data for decision making. The Table 2 list the attribute names with their description.

TABLE 2 *ATTRIBUTE NAME AND DESCRIPTION*

S. No	Attribute Name	Description
1	Age	In Years
2	Sex	0 ➔Female, 1➔ Male
3	Chest Pain Type	1➔Typical Angina, 2➔Atypical Angina, 3➔Non Anginal Pain, 4➔Asymptotic
4	Resting blood pressure	mmHg(At the time of admission)
5	Serum cholesterol	mg/dl
6	Fasting blood sugar > 120 mg/dl	0➔False, 1➔True
7	Resting ECG	0➔Normal, 1➔ST-T Wave Abnormality, 2➔Left Ventricular Hypertrophy
8	Max-heart rate achieved	0➔Absence
9	Exercise induced angina	1➔Yes, 0➔No
10	ST depression induced by exercise relative to rest	-
11	Peak exercise ST segment	1➔Unsloping, 2➔Flat, 3➔Down sloping
12	Number of major vessels coloured by fluoroscopy	0, 1, 2, 3
13	Thallium scan	3➔Normal, 6➔Fixed Defect, 7➔Reversible Defect
14	Diagnosis of heart disease/ angiographic disease status	0➔Absence, 1➔Presence.

The first and foremost is Data Pre-processing. The dataset contains numerous missing values which will not give any meaningful data when built. After applying data cleaning techniques missing instances are removed and the dataset shows a clear and meaningful dataset without null values. The results are shown in the following Fig. 2. and Fig. 3. The lines represent the presence of missing values and the other without missing values.

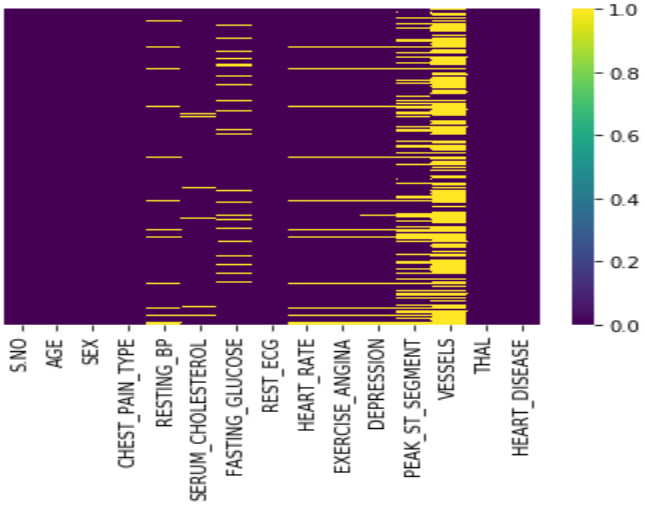

Fig. 2*. Data with missing values*

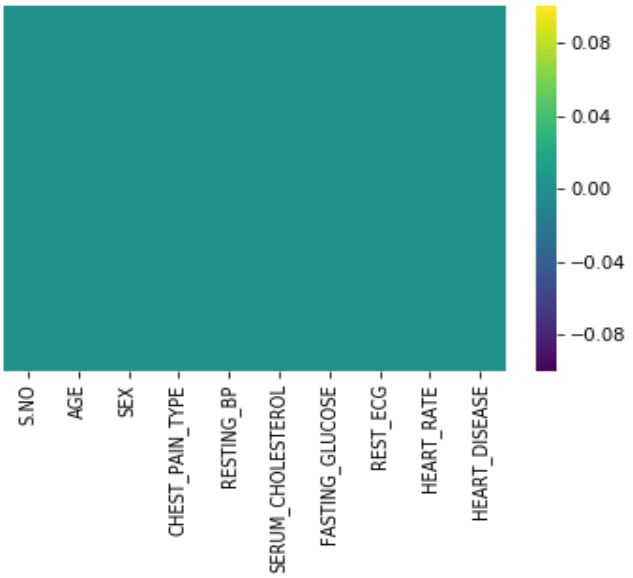

***Fig. 3**. Data without missing values*

Once pre-processing is done, the next step is Feature Selection. The attributes with highest correlation is chosen and undergoes Rule based Classification. Out of 15 attributes, only 9 are selected and evaluated. The Correlation of attributes with highest correlation is shown in the following Table 3

***TABLE 3** CORRELATION OF ATTRIBUTES WITH HIGHEST CORRELATION*

Attribute Name	Correlation with Predictive Variable
Resting ECG	0.109829
Age	0.288526
Chest Pain Type	0.473108
Resting BP	0.145840
Fasting Glucose	0.095242

Algorithms like Rule based Classification Algorithm, Existing Logistic Regression Algorithm and Hybrid Logistic Regression Algorithm are evaluated. The evaluation of the model is performed with confusion matrix, precision, recall and F1 Score. Totally four outcomes are generated, namely True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN).

(i) Accuracy for Rule based Classification $= \frac{TN + TP}{TN+TP+FN+FP}$

$= \frac{22+55}{22+55+10+24}$

$= 0.6936 * 100$

$= 69\ \%$

The following Table.4 shows precision, recall, F1 score and support values of Rule Based Classification

***TABLE.4** CLASSIFICATION REPORT FOR RULE BASED CLASSIFICATION ALGORITHM*

Classification Report	Precision	Recall	F1 Score	Support
0	0.69	0.48	0.56	46
1	0.70	0.85	0.76	65

(ii) Accuracy for Existing Logistic Regression Algorithm $= \frac{TN + TP}{TN+TP+FN+FP}$

$= \frac{31+56}{31+56+12+12}$

$= 0.78 * 100$

$= 78\ \%$

The following Table.5 shows precision, recall, F1 score and support values Logistic Regression algorithm

***TABLE.5** CLASSIFICATION REPORT FOR EXISTING LOGISTIC REGRESSION ALGORITHM*

Classification Report	Precision	Recall	F1 Score	Support
0	0.72	0.72	0.72	48
1	0.62	0.82	0.70	63

(iii) Accuracy for Hybrid Logistic Regression Algorithm $= \frac{TN + TP}{TN+TP+FN+FP}$

$= \frac{23+73}{23+73+6+9}$

$= 0.8648 * 100$

$= 86\ \%$

The following Table.6 shows precision, recall, F1 score and support values of Hybrid Logistic Regression Algorithm.

TABLE.6 *CLASSIFICATION REPORT FOR HYBRID LOGISTIC REGRESSION(HLR) ALGORITHM*

Algorithm	Accuracy (in %)
Rule Based Classification	69
Logistic Regression Algorithm	78
Hybrid Logistic Regression Algorithm	86

Table.7 provides an overall accuracy of all the three algorithms, where the Hybrid logistic regression algorithm provides higher accuracy compared to other two.

TABLE.7 *CLASSIFICATION REPORT FOR HYBRID LOGISTIC REGRESSION(HLR)*

Classification Report	Precision	Recall	F1 Score	Support
0	0.79	0.72	0.75	32
1	0.89	0.92	0.91	79

V. CONCLUSION

Heart disease is more common now-a-days. Prediction of heart disease at an early stage is a challenging task. Due to the advancement of machine learning techniques, prediction has become easier. It provides a novel technique to predict heart disease in advance. Rule based classification applied into Logistic Algorithm provides a Hybrid variety of Logistic Algorithm called HLR which shows 86% of accuracy compared with Rule Based Classification which is 69% and Existing logistic regression algorithm which is 78%. The Proposed HLR model shows better accuracy. Further extension of this work will be focused on real datasets collected through sensors in real time. Furthermore, neural network techniques would be implemented to show better accuracy.

ACKNOWLEDGMENT

This Paper has been written with the financial support of RUSA - Phase 2.0 grant sanctioned vide Letter No. F. 24-51/2014-U, Policy(TNMulti-Gen), Dept. of Edn. Govt. of India, Dt.09.10.2018

REFERENCES

[1] The Hindu BusinessLine Statistics report. September 2018. Available Online: https://www.thehindubusinessline.com/news/science/heart-disease-stroke-among-top-killers-in-india/article24935985.ece

[2] Senthilkumar Mohan, Chandrasegar Thirumalai, And Gautam Srivastava, “Effective Heart Disease Prediction using Hybrid Machine Learning Techniques”, IEEE ACCESS ,2017.

[3] Xiao-Li Li Institute for Infocomm Research Singapore, Bing Liu University of Illinois at Chicago, “Rule-based Classification”.

[4] R. Agrawal, T. Imieliski, and A. Swami, “Mining association rules between sets of items in large databases”, Proceedings of ACM SIGMOD International Conference on Management of Data (SIGMOD-1993), pages 207–216, 1993.

[5] WebMD 2018 Available Online: https://www.webmd.com/diabetes/type-2-diabetes-guide/heart-blood-disease#1.

[6] Anam Mustaqeem, Syed Muhammad Anwar, Muhammad Majid, Abdul Rashid Khan, “Wrapper Method for Feature Selection to Classify Cardiac Arrhythmia”, IEEE Conference, 2017.

[7] Prathibhamol Cp, Anjana Suresh, Gopika Suresh, “Prediction of Cardiac Arrhythmia type using Clustering and Regression Approach (P-CA-CRA)”, IEEE Conference,2017.

[8] Liaqat Ali, Awais Niamat, Javed Ali Khan, Noorbakhsh Amiri Golilarz, And Xiong Xingzhong, “An Expert System Based on Optimized Stacked Support Vector Machines for Effective Diagnosis of Heart Disease”, IEEE ACCESS, 2017.

[9] Raid Lafta, Ji Zhang, Xiaohui Tao, Yan Li, Xiaodong Zhu, Yonglong Luo and Fulong Chen, “Coupling a Fast Fourier Transformation with a Machine Learning Ensemble Model to Support Recommendations for Heart Disease Patients in a Telehealth Environment”, IEEE ACCESS, 2017.

[10] Liaqat Ali, Atiqur Rahman, Aurangzeb Khan, Mingyi Zhou, Ashir Javeed, And Javed Ali Khan, “An Automated Diagnostic System for Heart Disease Prediction Based on χ2 Statistical Model and Optimally Configured Deep Neural Network”, IEEE ACCESS, 2019.

[11] Carlos Ordonez, “Association Rule Discovery with the Train and Test Approach for Heart Disease Prediction”, in IEEE Transactions on Information Technology in Biomedicine, vol. 10, no. 2, April 2006.

[12] Dataset source. Available Online: https://www.kaggle.com/datasets

[13] T. Mythili, Dev Mukherji, Nikita Padalia, and Abhiram Naidu, “A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL)”, in International Journal of Computer Applications (0975 – 8887), Volume 68– No.16, April 2013.