IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

HEART DISEASE PREDICTION USING MACHINE LEARNING

B.Venkata Harish, S.Vishnu Koushik, K.Surendra Kumar

Under the Guidance of

Mr.R.Ramesh Kumar (asst.professor)

Department of Electronics and Communication Engineering

Madanapalle Institute of Technology & Science, Madanapalle, India

KEYWORDS: Decision Tree, K-Nearest Neighbour, Coronary Heart Disease

1.ABSTRACT

As indicated by a new WHO study, coronary illness is on the ascent. 17.9 million people die each year. As the population grows, it becomes increasingly difficult to diagnose early and start treatment. However, with recent technological advances, machine learning technology has accelerated the healthcare sector through several studies. Thusly, the motivation behind this paper is to make a ML model for foreseeing coronary illness in view of the pertinent boundaries. The methodology includes supplanting invalid qualities, resampling, normalization, standardization, arrangement, and expectation. This work plans to anticipate the danger of CHD utilizing AI calculations like Random Forest, Decision Trees, and K-Nearest Neighbors. Additionally, a similar report among these calculations based on expectation exactness is performed. Further, K-overlay Cross Validation is utilized to produce arbitrariness in the information. These calculations are tested over "Framingham Heart Study" dataset, which is having 4240 records. In our exploratory examination,Decision Tree, and K-Nearest Neighbor accomplished a precision of 80% above respectively. This result shows that Random Forest provides more accurate predictions in less time compared to other ML techniques. This model may be useful to clinic physicians as a decision support system.

2.INTRODUCTION

Machine learning techniques are around us and are compared and used in the analysis of many types of data science applications. The main motivation behind this research-based project was to investigate the feature selection methods, data preparation, and processing behind machine learning training models. With direct models and libraries, the challenges we face today are data with large variations in the accuracy of training, testing, and actual validation, as well as their abundance and cooked models. Therefore, this project is motivated by exploring the model, further implementing the logistic regression model to a training the retrieved data. Since all machine learning aims to develop appropriate computer-based systems and decision support that can contribute to the early detection of heart disease, this project uses a variety of characteristics to allow patients to be 10 years old. We have developed a model to classify whether or not you have heart disease. Whether to use logistic regression for years (i.e., potential risk factors that can cause heart disease). Therefore, an early prognosis of cardiovascular disease can help guide lifestyle changes in high-risk patients, thereby reducing complications. This could be a major milestone in the field of medicine.

3.SYSTEM MODEL AND ANALYSIS

3.1 Introduction

Tom Mitchell states machine learning as "A computer program is said to learn from experience and from some tasks and some performance on, as measured by, improves with experience". Machine Learning is combination of correlations and relationships, most machine learning algorithms in existence are concerned with finding and/or exploiting relationship between datasets. In Machine Learning there are various types of algorithms such as Regression, Linear Regression, Logistic Regression, Naive Bayes Classifier, Bayes theorem, KNN (K-Nearest Neighbor Classifier), Decision Tress, Entropy, ID3, SVM (Support Vector Machines), K-means Algorithm, Random Forest and etc.

Within the held of information examination, Machine learning is a technique used to gadget complex models and calculations that loan themselves to expectation; in business use, this is known as prescient investigation. These scientific models permit specialists, information researchers, architects, and examiners to "produce solid, repeatable choices and results" and uncover "stowed away experiences" through gaining from verifiable connections and patterns in the information.

3.2 System Architecture

An architecture diagram is a graphical representation of a set of concepts, that are part of an architecture, including their principles, elements and components. The diagram explains about the system software in perception of overview of the system.

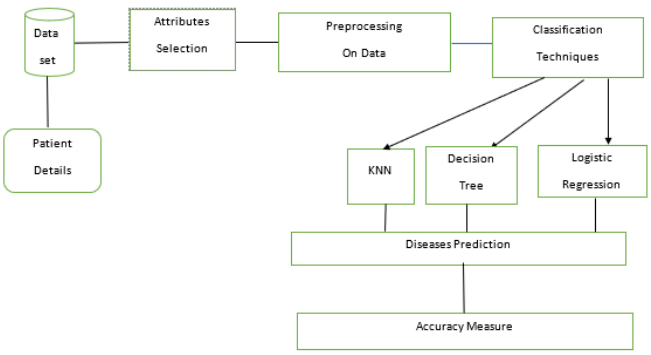

Fig 3.1 System architecture

The figure 3.1 shows the system architecture initially it selects attributes from data set then after it will preprocess the data by handling missing values. Different classification techniques in supervised algorithms.

3.3 Data Flow Diagram

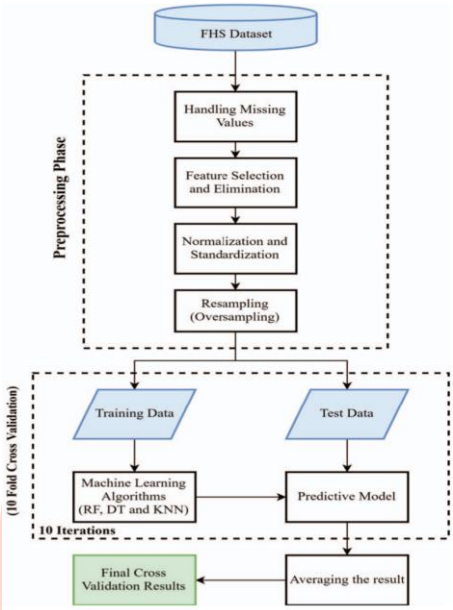

Fig: 3.2 Data flow diagram

Fig 3.2 shows the Data flow diagram. The FHS dataset is used for preprocessing phase which contains the stages of handling missing values, feature selection and elimination, normalization and standardization and resampling. After the preprocessing stage, the data is used for training and testing. Finally, the trained model is used for prediction of heart diseases.

4.4Methodology of System

This section illustrates various resources and approaches that are used in this work. Primarily, the description of dataset is provided to understand how to work on it, followed by the preprocessing steps involved. Finally, the internal working and understanding of the analytical models used are explained.

Dataset Description

We have practiced a dataset which is a subset of Framingham Heart Study (FHS) dataset, it is made publicly available through Framingham Heart Institute. The available section of FHS dataset used in this paper contains records of 4240 participants.

The dataset is generated by long term study on a population of Framingham.

The study is based on the cause and origin that lead to cardiovascular heart disease and it comes under one of the best public health disease management domain. The Framingham Heart study focused mainly to retrieve the risk factors that have an effect on the health of a person in perceiving a coronary heart disease. The dataset contains 16 different features that affect Coronary Heart Disease.

B. Preprocessing

Preprocessing is a method to obtain complete, consistent, interpretable data. The data quality affects the mining results that are obtained using machine learning algorithms. Quality data results in a quality decision. Therefore, the FHS dataset is integrated using the following preprocessing steps.

- Irrelevant features can decrease the performance of the model and reduces the learning rate. Therefore, feature selection plays a major role in preprocessing in which those features are selected that contributes the most in predicting the desired results. In the FHS dataset, using an automatic feature selection would have eliminated important features as well. Therefore, an analytical approach gives better performance.
- The mean is the most probable value that tends to occur in any attribute. Also, mean preserves the extremes of an attribute, therefore, missing values in the FHS dataset are replaced by the attribute mean, as shown in equation.

$$\text{Attribute Mean} = \frac{\sum_{i=0}^{l}(attribute\ value)}{l}$$

where, l is the total number of values in an attribute.

- Class imbalance of dataset is a major problem in data mining applications. Most of the machine learning algorithms fail to perform well on a dataset where classes are imbalanced.
- Sampling is an effective method to balance an imbalanced dataset. Sampling is of two types: oversampling and undersampling. Undersampling involves removing instances from the majority class to balance the class distribution. Oversampling involves replicating instances from minority class to balance the class distribution. Fig 1. illustrates the resampling mechanism.
- The target class in the dataset predicts the risk of coronary heart disease (CHD). The instances with the risk of an individuals those are more likely to suffer from CHD is 15.2% (644 out of 4240 entries) and that of individuals those are not suffering from CHD is 84.8% (3596 out of 4240 entries). In order to balance this class distribution, we used random oversampling to replicate the instances in the minority class, that is, individuals suffering from CHD.

C. Preprocessing of data

- Preprocessing needed for achieving prestigious result from the machine learning algorithms. For example, Random forest algorithm does not support null values dataset and for this we have to manage null values from original raw data. For our project we have to convert some categorized value by dummy value means in the form of “0”and “1” by using.

D. Data Balancing

- Data balancing is essential for accurate result because by data balancing graph we can see that both the target classes are equal. represents the target classes where “0” represents with heart diseases patient and “1” represents no heart diseases patients.

Experimental Analysis

Parameters Used

Performance evaluation of the proposed work is done based on the following measures:

Confusion Matrix is a matrix that is used to evaluate the performance of a model. The four terms associated with the confusion matrix which is used to determine the performance matrices are:

True Positive (TP): An outcome when the positive class is correctly predicted by the model

True Negative (TN): An outcome when the negative class is correctly predicted by the model

False Positive (FP): An outcome when the positive class is incorrectly predicted by the model

False Negative (FN): An outcome when the negative class is incorrectly predicted by the model

Accuracy: is the ratio of a number of correct predictions given by the model to the total number of instances.

$$\text{Accuracy} = \frac{(T\,P + T\,N)}{(T\,P + F\,P + F\,N + T\,N)} \quad (1)$$

Precision: Precision in this work measures the proportion of individuals predicted to be at risk of developing CHD and had a risk of developing CHD

$$\text{Precision} = \frac{(T\,P)}{(T\,P + F\,P)} \quad (2)$$

Recall/Sensitivity: Recall, in this work, measures the proportion of individuals that were at a risk of developing CHD and were predicted by the algorithm to be at risk of developing CHD.

$$\text{Recall} = \frac{(T\,P)}{(T\,P + F\,N)} \quad (3)$$

Specificity: Specificity here measures the proportion of individuals who were not at risk of developing CHD and were predicted by the algorithm to be not at risk of developing CHD.

$$\text{Specificity} = \frac{(T\,N)}{(T\,N + F\,P)} \quad (4)$$

F1 Score: F1 Score is the harmonic mean of precision and recall.

$$\text{F1 Score} = \frac{2(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (5)$$

ROC (Receiver Operator Characteristic): It is a probability curve indicating the capability of a model to distinguish between classes.). According to, AUC (Area Under the Curve) closer to 1 would be able to perfectly differentiate the two classes in the case of binary classification. Therefore, AUC closer is 1 is better predictive measure.

$$\text{TRP} = \frac{(T\,P)}{(T\,P + F\,N)}$$

$$\text{FRP} = \frac{(F\,P)}{(F\,P + T\,N)}$$

CONCLUSION

We propose a preprocessing extensive work where Random Forest is the most compatible contender for prediction model and gives the highest performance measure among K-Nearest Neighbour and Decision Tree. The accuracy, recall, precision, specificity and F1 score of RF on the proposed work are above 90 % respectively. The Decision Tree, however, gives lesser performance set against that of Random Forest though in quite lesser time (0.8138). The accuracy for K- Nearest Neighbour is the highest among all, however, the performance measures are quite similar to that of the Decision Tree. Thus, in an environment similar to that of the used dataset, if all the features are preprocessed such that they acquire normal distribution, Random Forest is a good selection to obtain a robust prediction model. And, such models provide a valuable assistant to the society for health care management domain.

Further, as an extension to this work, a more real-time and bigger dataset is required to obtain a better training model.

REFERENCES

I. **H. M. S. U. Marjia Sultana,** "Analysis of Data Mining Techniques for Heart Disease Prediction," 2018

II. **M. I. K., .A. I., .S. Musfiq Ali,** "Heart Disease Prediction Using Machine Learning Algorithms"

III. **K. Bhanot,** "towarddatascience.com," 13 Feb 2019. [Online].

IV. **M. A. K. H. K. M. a. V. P. M Marimuthu,** " A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach

V. **Rajliwall, Nitten S., Rachel Davey, and Girija Chetty,** "Machine learning based models for Cardiovascular risk prediction", IEEE International Conference on Machine Learning and Data Engineering (ICMLDE), 2018

VI. **M. A. Jabbar, Shirina Samreen,** "Heart disease prediction system based on Hidden Nave Bayes classifier", IEEE International Conference on Circuits, Controls, Communications and Computing (I4C), pp. 1-5, 2016.

VII. **Santhana Krishnan J and Geetha S,** "Prediction of Heart Disease using Machine Learning Algorithms" ICIICT, 2019

VIII. **Aditi Gavhane, Gouthami Kokkula, Isha Panday, Prof. Kailash Devadkar,** "Prediction of Heart Disease using Machine Learning", Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA), 2018

IX. **Amandeep Kaur and Jyoti Arora,** "Heart Diseases Prediction using Data Mining Techniques: A survey" International Journal of Advanced Research in Computer Science, IJARCS 2015-2019

X. **Patel, J., Upadhyay, P. and Patel,** "Heart Disease Prediction Using Machine learning and Data Mining Technique" Journals of Computer Science & Electronics, 2016.