HEART DISEASE PREDICTION RANDOM FOREST ALGORITHMS

[1]JulapallySravan, [2]Marjodu Vamshi, [3]Saginala Srihari Yadav, [4]Kopparapu Uma Mahesh, [5]Barjinder Singh

Department of Computer Science and Engineering,
Lovely Professional University, Punjab, India

ABSTRACT

Heart disease is one of the most significant causes of mortality in the world today. Prediction of cardiovascular disease is a critical challenge in the area of clinical data analysis. Machine learning (ML) has been shown to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry. Machine learning techniques being used in recent developments in different areas of medical industry. In this work, we proposed a novel method that aims at finding heart disease by applying machine learning techniques. The prediction model uses classification techniques and Cleveland heart disease dataset is used. Machine learning technique Decision Tree and Random Forest is applied. The novel technique of machine learning model is used. In implementation, three machine learning algorithms are used, they are 1. Decision Tree, 2. Random Forest and 3. Hybrid model (Hybrid of Decision tree and random forest). Experimental results shows an accuracy level of 88:7% through the prediction model for heart disease with the hybrid model. The interface is designed to get the input parameter from user to predict the heart disease, for which we used hybrid model of Decision Tree and Random forest.

Keywords- Heart Disease prediction, Random Forest, DecisionTree, Machine Learning, Machine Learning Algorithms.

I. Introduction

The technique of extracting useful information from large data sets is known as data mining and forecasting or describing it employing classification, clustering, and association algorithms Data processing is used in the healthcare business, among other things, to categorise the best treatment procedures, predict illness risk factors, and find the most cost-effective patient care cost structures. Data processing models are used to research diabetes, asthma, cardiovascular disease, AIDS, and other disorders. In healthcare research, a variety of information mining techniques are used to build models, including : Other techniques include . Decision Trees, Support Vector Machines, Bayesian

Classification, logistic regression and artificial neural networks. Cardiovascular illnesses claim the lives of an estimated 17 million people each year (CVD). The prognosis and prognosis of such diseases are bad in the early stages, despite the fact that they are treatable. To reduce the high fatality rates, a prognosis and a patient's assessed risk are required. Coronary heart disease, cardiomyopathy, hypertension, coronary failure, and other cardiovascular problems are all very frequent. Diabetes, hypertension, smoking, high cholesterol diet and lack of physical activity and other factors all contribute to heart disease. Data processing research in the realm of cardiovascular illnesses is ongoing, including high-accuracy prediction, treatment, and risk score analysis. Several CVD surveys are undertaken, with the Cleveland Heart Clinic's data set being the most well-known. The Cleveland Cardiovascular Disease Database (CHDD) is widely regarded as the industry standard for cardiovascular disease research. This paper describes a system for combining decision tress, logistic regression and support vector machines to generate individual predictions, which are then used in rule-based algorithms based on the parameters in the database. The accuracy, sensitivity, and specificity of each rule generated by this technique are then compared.

This work describes a system for generating individual predictions using a combination of support vector machines ,logistic regression, and decision trees, which are then employed in rulebased algorithms based on database parameters Detecting cardiovascular disease is challenging due to a multitude of risk factors like high blood cholesterol, high BP, diabetes, irregular pulse and a variety of other disorders. To forecast the severity of cardiovascular disease in people, researchers use a variety of data processing and neural network techniques. Among the approaches used to classify the severity of the ailment are the (DT) Decision Trees, (KNN) KNearest Neighbor Algorithm, (NB) Naive Bayes and (GA) Genetic Algorithm. Cardiopathy is a complicated illness that requires careful management. Failure to do so might put the center's operations in jeopardy or result in an early death. A wide spectrum of metabolic abnormalities can be detected using bioscience and data processing. Data classification and processing are crucial in prediction of heart diseases and data analysis.

Decision trees have also been used to predict the incidence of heart disease-related events with high accuracy. To predict cardiopathy, a variety of information abstraction techniques are combined with well-known knowledge mining procedures. Several readings are carried out as part of this research in order to build a prediction model that includes not only different approaches, but also the combination of two or more. Hybrid approaches relate to the new strategies that have been adopted. The rate statistic is used to introduce neural networks. Bundle branch block clinical records utilised in this procedure to ascertain the specific reason include atrial (AFIB) Atrial flutter, (NSR)

Normal Sinus Rhythm, (SBR) Sinus bradycardia, (AFL) atrial utter, (PVC) Premature Ventricular Contraction, and (BII) Second Degree Block. A dataset using a(RBFN) radial basis function network is utilised for classification, with 70% of the data used for training and 30% for classification. In the pharmaceutical and research industries, we also introduce (CADSS) Computer Aided Decision Support Systems. Prior study [16] has found that applying knowledge mining approaches to disease prediction in the healthcare business takes less time and produces more accurate findings. According to our idea, the GA should be used to diagnose cardiac disease. This technique employs GAderived effective association rules for mutation, crossover and tournament selection all of which are included in the new recommended fitness function. We use the well-known Cleveland dataset, which was obtained from a UCI machine learning library, for experimental validation. In a few weeks, we'll compare our results to those of some of the most well-known supervised learning algorithms. (PSO) Particle Swarm

Optimization, the most powerful evolutionary method, is described, along with a few heart condition guidelines. With encoding approaches, the concepts are employed at random, resulting in an overall improvement in accuracy. Heart illness can be detected by a variety of symptoms, including pulse rate, sex, age, and others. As we've seen, including Neural Networks into a machine learning algorithm produces more accurate and consistent outcomes. For disorders like cardiopathy and neurological problems, neural networks are commonly recognised as the most effective therapy option. For predicting cardiovascular disease, we devised a 13feature method. The results demonstrate an enhanced level of performance when compared to current methodologies in works such as. Arteria carotis stenting (CAS) has become a popular medical treatment option in recent years. The CAS predicts the occurrence of major adverse cardiovascular events in older heart disease patients (MACE). Their assessment becomes critical. Our findings are based on an (ANN) Artificial Neural Network that predicts heart issues. The usage of neural network techniques, which incorporate not just posterior probabilities but also expected values from a variety of sources, is investigated. When compared to previous studies, this model produces an accuracy level of up to 89:01 percent, which is a strong result. As previously stated, the Cleveland heart dataset is used in conjunction with a (NN) Neural Network in all efforts to improve cardiovascular disease performance. (ML) Machine learning approaches for the Internet of Things have also advanced recently (IoT). Machine learning techniques have been found to reliably identify linked IoT devices when applied to network traffic data. Because of its layered nature, deep learning might be a feasible solution for acquiring trustworthy data from raw sensor data from IoT devices in harsh climates. Deep learning is also ideal for sting computing. The Hybrid Random Forest with Linear Model (HRFLM) approach is described in this article. The primary purpose of this inquiry is to improve cardiopathy prediction exactness. Several studies have produced feature selection limitations that can be applied to algorithms. In contrast, the HRFLM approach incorporates all characteristics with no constraints on feature selection. We employ a hybrid technique to uncover the features of a machine learning system by executing trials. The trial's findings suggest that our proposed hybrid strategy predicts heart disease better than existing methods.

Literature survey

Random forest and evolutionary techniques are used in an intelligent heart disease prediction system.

Priti Chandra, Jabbar Akhil, and Bulusu Deekshatulu (2016). [Vol. 4, no. 4, pp. 175-184 in Journal of Network and Innovative Computing.

One of the top causes of death worldwide is expected to be heart disease. It's difficult to forecast how a condition may progress. Data mining is used to automatically infer diagnostic concepts and to assist specialists in developing more reliable diagnosis systems. Researchers use a variety of data processing approaches to aid health care professionals in forecasting the core disease. Random forest is a type of ensemble learning system that is especially well-suited to medical applications. The Chi square selection feature metric is used to assess the relationship between variables and determine whether or not they are connected. In this research, we provide a cardiovascular illness prediction classification model that employs feature selection measures such as a genetic algorithm ,chi square and random forest classifier.The results of the studies show that our method enhances classification accuracy when compared to other methods, and that the supplied model may be utilised successfully by health care professionals to predict cardiac problems.

A Random Forest Classifier-based Data Mining Model for Predicting Coronary Heart Disease

In Proc. Int. Conf. Recent Trends Comput. Methods, Commun. Controls, Apr. 2012, pp. 22-25. A. S. Abdullah and R. R. Rajalaxmi, in Proc. Int. Conf. Recent Trends Comput. Methods, Commun. Controls, Apr. 2012, pp. 22-25.

Coronary cardiovascular disease (CHD) is a common intestine-related disease that is a primary cause of death. From a medical aspect, data processing is involved in the identification of various forms of metabolic disorders. Classification techniques are significant in data processing because they help with prediction and data exploration. The accuracy and events connected to CHD have been predicted using Decision Trees, a classification technique. In this study, an information mining model based on the RF(Random forest) classifier was constructed to improvise the prediction accuracy and assess various CHD incidences. This model helps clinicians predict CHD and its multiple episodes and how they relate to different demographic groups. Angina, percutaneous coronary intervention (PCI),acute myocardial infarction (AMI), , and arterial bypass surgery are just a few of the diseases studied (CABG). Experimental data proved that the random forest classification algorithm was successful in predicting CHD incidence and the risk variable.

Using PSO algorithm for getting best rules in diagnosis of cardiac disease A. H. Alkeshuosh, M. Z. Moghadam, I. A. Mansoori and M. Abdar, "Using PSO Algorithm for Producing Best Rules in Diagnosis of Heart Disease," 2017 International Conference on Computer and Applications (ICCA), Doha, 2017, pp. 306-311.

Around the world, heart disease is a major public health concern. Because data regarding various ailments collected through various types of medical technology is either insufficient or incorrect, expertise in manual diagnosis and limiting human experience leads to improper diagnosis in the healthcare system. Because accurate prediction of a human condition is critical, treating illness and providing life science with intelligent tools for detecting can help clinicians make better decisions and save money. This study used the Particle Swarm Optimization (PSO) approach, which is one of the most potent evolutionary algorithms for developing heart disease criteria. PSO is used to improve the validity of the random rules once they have been encoded.

Backpropogation neural network for prediction of heart disease Al-Milli, Nabeel. (2013). Backpropogation neural network for prediction of heart disease. 56. 131-135.

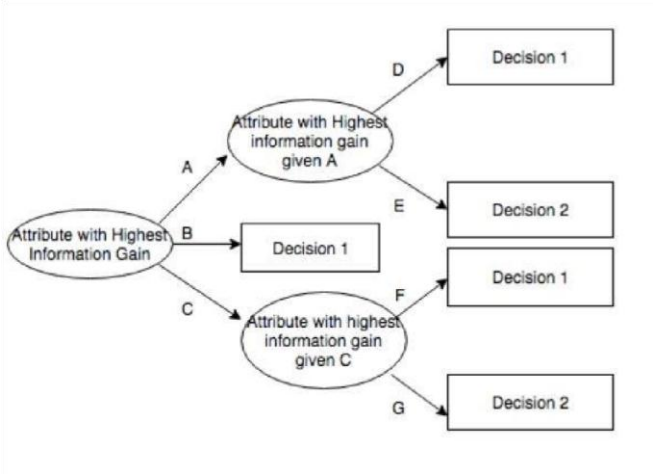
Researchers have recently proposed several softwares, tools, and algorithms for constructing successful medical decision support systems. Furthermore, new algorithms and techniques are constantly being developed and represented. Cardiopathy diagnosis is a critical issue, and many academics have worked to build intelligent medical decision support systems to help clinicians be more effective. The most common way for predicting the diagnosis of cardiopathy is to employ a neural network. In this publication, a neural network is used to create a cardiovascular illness prediction system. For heart condition prediction, the suggested method utilised 13 medical parameters. Experiments carried out during this project have revealed that the suggested algorithm outperforms similar stateof-the-art techniques.

IMPLEMENTATION METHODOLOGY

The proposed project is made in Python 3.6.4 and includes pandas, matplotlib,scikitlearn and other libraries. The data was obtained from uci.edu. The information includes binary cardiac condition classes. The hybrid model employs machine learning algorithms like random forests and Decision Trees.

DATA DICTIONARY

The dataset collected with attributes age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slop, ca, thal, pred_attribute. The sample of collevted data is shown in the below figure.



'The modules included in our implementation are as follows

- Decision Tree
- Random forest
- Hybrid model

A. Decision Tree

This technique is a classification problem solving supervised learning algorithm. The input and output variables can be categorical or continuous. Here, we distribute the sample into 2 or more homogenous sets (or subpopulations) By this method, which is based on the input variables` most prominent splitter / differentiator. A test on the attribute is represented by an internal node, the end result is represented by a branch, and the conclusion reached after computing the attribute in a decision tree is represented by a leaf.

Decision Tree works in following manner

- Place the value of dataset root on the top of the tree.
- Dividing the set into subgroups is a good idea. You need to create a subset such that each subset containing data with the same attribute values.
- Repeat step1,2 on every subset until all of the tree's branches have leaf nodes. When predicting a category label for a record in decision trees, we begin at the root of the tree.

In decision trees, for predicting a class label for a record we start from the root of the tree. Then compare the values of the root attribute with record's attribute. On the basis of comparison, follow the branch corresponding to that value and jump to the next node.

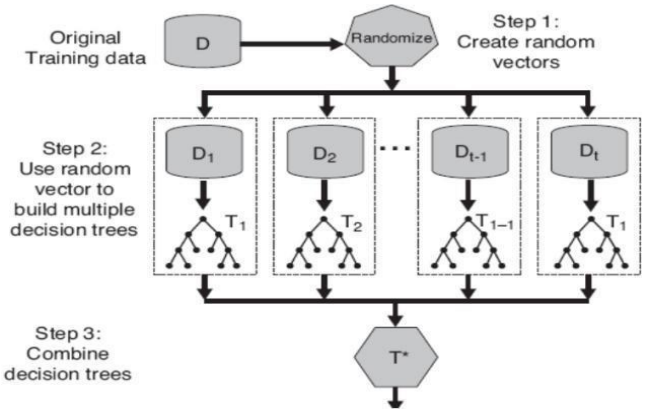
B. Random Forest Model

1. Given there are n cases in the training dataset. From these n cases, subsamples are chosen at random with replacement. These random sub-samples chosen from the training dataset are used to build individual trees.

2. Assuming there are k variables for input, a number m is chosen such that m < k. m variables are selected randomly out of k variables at each node. The split which is the best of these m variables is chosen to split the node. The value of m is kept unchanged while the forest is grown.

3. Each tree is grown as large as possible without pruning.

4. The class of the new object is predicted based upon the majority of votes received from the combination of all the decision trees.



HYBRID MODEL

We create this model using a random forest algorithm and decision tree algorithm. This combined model is depend on the probability of a random forest algorithm. Random forest possibilities are combined with training data and entered into the decision tree algorithm. The probability of the decision tree is also identified the test data is provided in a similar way. Finally, the expected value is determined.

EXPERIMENTAL RESULTS AND ANALYSIS

Scikit-learn, pandas, matplotlib, and other relevant libraries are included in the suggested project, which is written in Python 3.6.4. For research objectives, the uci.edu heart disease dataset is being examined. Decision trees and random forests are the two machine learning techniques employed. We were able to detect cardiac illness using these machine learning technologies. We employed a Decision Tree and Random Forest hybrid model to improve the work and add novelty.

The results suggest that both the Random Forest method and the hybrid model can detect cardiac abnormalities. The Random Forest model is 76% accurate, Decision Tree is 79% accurate, and Hybrid is 76% accurate.The below table gives the information about the accuracy that we got in our experimental analysis.

Algorithm	Accuracy (%)
Decision Tree	79
Random Forest	76
Hybrid (Decision Tree+ Random Forest)	76

Table: Experimental Results of proposed system

The below figure shows the accuracy comparison of our proposed work.

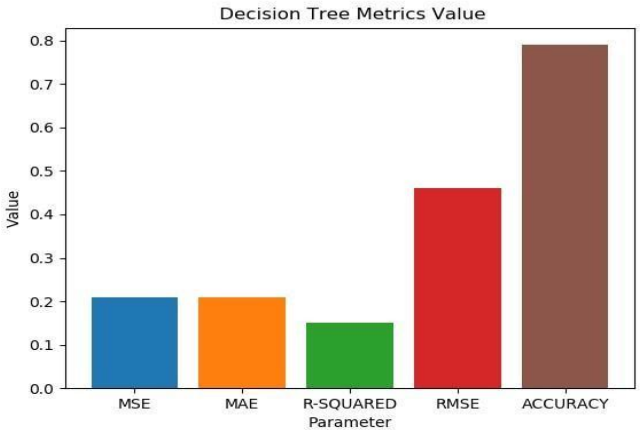


Figure: Evaluation metrics for Decision Tree algorithm

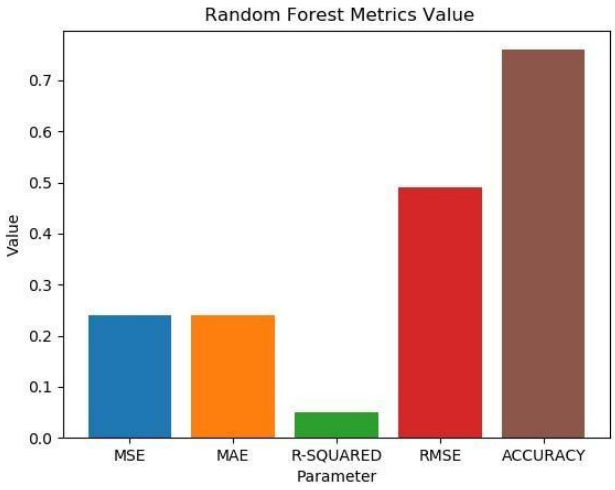


Figure: Evaluation metrics for Random Forest algorithm

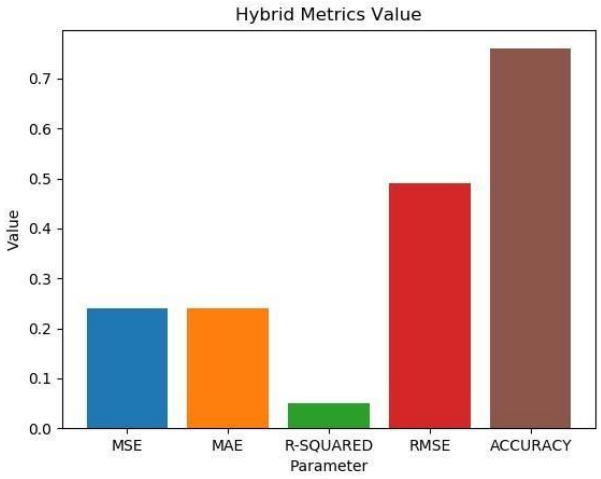


Figure: Evaluation metrics for Hybrid algorithm

CONCLUSION AND FUTURE WORK

Finally, as the literature review demonstrates, there is a desire for more intricate and combinational models to improve the accuracy of forecasting the onset of cardiovascular disorders

For cardiovascular disease prediction, the recommended architecture combines Decision Tree and Random Forest. The system must be trained and tested using Cleveland cardiovascular disease data in order to produce the most efficient model. In the future, we'd like to look into some deep learning models for heart disease prediction, such as CNN or DNN algorithms. To determine the disease's breadth, we're considering classifying it as a multi-class issue.

REFERENCES

[1] Mackay,J., Mensah,G. 2004 “Atlas of Heart Disease and Stroke” Nonserial Publication, ISBN-13 9789241562768 ISBN-10 9241562765.

[2] Robert Detrano 1989 “Cleveland Heart Disease Database” V.A. Medical Center, Long Beach and Cleveland Clinic Foundation.

[3] Yanwei Xing, Jie Wang and Zhihong Zhao Yonghong Gao 2007 “Combination data mining methods with new medical data to predicting outcome of Coronary Heart Disease” Convergence Information Technology, 2007. International Conference November 2007, pp 868-872.

[4] Jianxin Chen, Guangcheng Xi, Yanwei Xing, Jing Chen, and Jie Wang 2007 "Predicting Syndrome by NEI Specifications: A Comparison of Five Data Mining Algorithms in Coronary Heart Disease" Life System Modeling and Simulation Lecture Notes in Computer Science, pp 129-135.

[5] Jyoti Soni, Ujma Ansari, Dipesh Sharma 2011 "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction" International Journal of Computer Applications, doi 10.5120/2237-2860.

[6] Mai Shouman, Tim Turner, Rob Stocker 2012 "Using Data Mining Techniques In Heart Disease Diagnoses And Treatment" Electronics, Communications and Computers (JECECC), 2012 Japan-Egypt Conference March 2012, pp 173-177.

[7] Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H. Guppy, Stella Lee, Victor Froelicher 1989 "International application of a new probability algorithm for the diagnosis of coronary artery disease" The American Journal of Cardiology, pp 304-310.15

[8] Polat, K., S. Sahan, and S. Gunes 2007 "Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing" Expert Systems with Applications 2007, pp 625-631.

[9] Ozsen, S., Gunes, S. 2009 "Attribute weighting via genetic algorithms for attribute weighted artificial immune system (AWAIS) and its application to heart disease and liver disorders problems" Expert Systems with Applications, pp 386-392.

[10] Resul Das, Ibrahim Turkoglub, and Abdulkadir Sengurb 2009 "Effective diagnosis of heart disease through neural networks ensembles" Expert Systems with Applications, pp 7675–7680.