Prediction of Heart Disease Using Hybrid Model

Rahul Bhole[1] Deepak Kakade[2] Apurva Khurud[3] Aishwarya Patodekar[4] Sukrut Mane[5]
[1,2,3,4,5]Department of Information Technology Engineering
[1,2,3,4,5]Zeal College of Engineering and Research, Pune, Maharashtra, India

Abstract— Heart disease is one among the foremost significant causes of mortality within the world today. Prediction of disorder may be a critical challenge within the area of clinical data analysis. Machine learning (ML) has been shown to be effective in assisting in making decisions and predictions from the huge quantity of data produced by the healthcare industry. We've also seen ML techniques getting used in recent developments in several areas of the web of Things (IoT). Various studies give only a glimpse into predicting heart disease with ML techniques. Within the proposed work, we propose a completely unique method that aims at optimized model leading to improving the accuracy within the prediction of disorder. We produce an optimized model with hybrid model combining two machine learning algorithms.

Keywords: Heart Condition, Decision Tree, Random, Cleveland Database, Forest Model

I. INTRODUCTION

Data mining (DM) is that the extraction of useful information from large data sets that results in predicting or describing the data using techniques like classification, clustering, association, etc. data processing has found extensive applicability within the healthcare industry like in classifying optimum, treatment methods, predicting disease risk factors and finding efficient cost structures of patient care. Research using processing models are applied to diseases like diabetes, asthma, cardiovascular diseases, AIDS, etc. Various techniques of data mining like naïve Bayesian classification, artificial neural networks, support vector machines, decision trees, logistic regression, etc. are used to develop models in healthcare research. Common cardiovascular diseases include coronary, heart. Condition, cardiomyopathy, hypertensive heard disease, heart failure, etc. Common causes of heart diseases include smoking, diabetes, lack of physical activity, hypertension, high cholesterol diet, etc. Research within the sector of cardiovascular diseases using processing has been an ongoing effort involving prediction, treatment, and risk score analysis with high levels of accuracy. Multiple CVD surveys are conducted with the foremost prominent one being the data set from the Cleveland Heart Clinic. The Cleveland heart disease Database (CHDD) intrinsically has been considered the de facto database for heart disease research. Recommending the parameters from this database, this paper proposes a framework to use logistic regression, support vector machines, and decision trees to achieve individual predictions which are successively utilized in rule based algorithms.

The methodology aims to accomplish of two goals: the first is to primarily present a predictive framework for heart disease, and thus the second is to match the efficiency of merging the outcomes of multiple models as against employing one model.

II. LITERATURE SURVEY

Heart disease may be a leading explanation for premature death within the world. Predicting the results of disease is that the challenging task. Data processing is involved to automatically infer diagnostic rules and help specialists to make diagnosis process more reliable. Several processing techniques are employed by researchers to help health care professionals to predict the center disease. Random forest is an ensemble and most accurate learning algorithm, suitable for medical applications. Chi square feature selection measure is used to guage between variables and determines whether or not they're correlated or not. During this paper, we propose a classification model which uses random forest as classifier, chi square and genetic algorithm as feature selection measures to predict heart disease. The experimental results have shown that our approach improve classification accuracy compared to other classification approaches, and thus the presented model are often successfully employed by health care professional for predicting heart condition .

Heart disease remains a growing global health issue. Within the health care system, limiting human experience and expertise in manual diagnosis rein accurate diagnosis, and thus the knowledge about various illnesses is either inadequate or lacking in accuracy as they're collected from various kinds of medical equipment. Since the proper prediction of a person's condition is of great importance, equipping bioscience with intelligent tools for diagnosing and treating illness can reduce doctors' mistakes and financial losses. During this paper, the Particle Swarm Optimization (PSO) algorithm, which is one of the foremost powerful evolutionary algorithms, is used to urge rules for heart disease. First the random rules are encoded then they're optimized supported their accuracy using PSO algorithm. Finally we compare our results with the C4.5 algorithm.

III. METHODOLOGY

A. Decision Tree

Decision tree could also be a kind of supervised learning algorithm that's mostly utilized in classification problems. It works for both categorical and continuous input and output variables. During this technique, we split sample into two or more homogeneous sets (or sub-populations) supported most significant splitter / differentiator in input variables. In decision tree internal node represents a test on the attribute, branch depicts the result and leaf represents decision made after computing attribute.

B. Random Forest Model

Given there are no cases within the training dataset. From these n cases, sub-samples are chosen randomly with replacement. These random sub-samples chosen from the training dataset are wont to build individual trees. Assuming there are k variables for input, variety m is chosen such m < k. m variables are selected randomly out of k variables at each

node. The split which is that the better of these m variables is chosen to separate the node. The worth of m is kept unchanged while the forest is grown. Each tree is grown as large as possible without pruning. The category of the new object is predicted based upon the bulk of votes received from the mixture of all the choice trees.

C. Hybrid Model

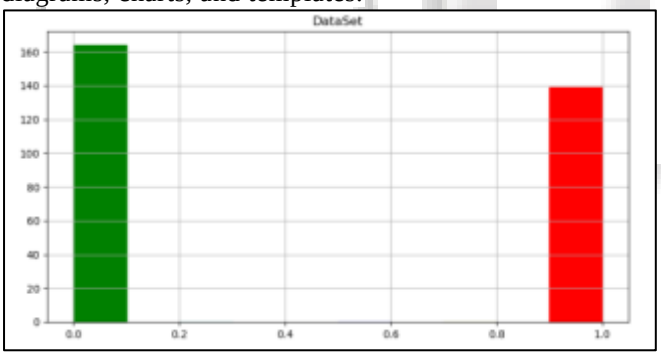
We develop hybrid model using decision tree and random forest algorithm. The combined model works based probabilities of random forest. The possibilities from random forest is added to coach data and fed to decision tree algorithm. Similarly decision tree probabilities are identified and fed to check data. Finally the values are predicted.

D. Data Collection

The data collection process involves the choice of quality data for analysis. Here we used Heart disease dataset taken from uci.edu for machine learning implementation. The job of a data analyst is to find ways and sources of collecting relevant and comprehensive data, interpreting it, and analyzing results with the help of statistical techniques.

E. Data Visualization

A large amount of data represented in graphic form is simpler to know and analyze. Some companies specify that a knowledge analyst must skills to make slides, diagrams, charts, and templates.



F. Data Preprocessing

The purpose of preprocessing is to convert raw data into a form that fits machine learning. Structured and clean data allows a knowledge scientist to urge more precise results from an applied machine learning model. The technique includes data formatting, cleaning, and sampling.

G. Model Training

After a knowledge scientist has preprocessed the collected data and split it into train and test can proceed with a model training. This process entails "feeding" the algorithm with training data. An algorithm will process data and output a model that's ready to find a target value (attribute) in new data a solution you would like to urge with predictive analysis. The purpose of model training is to develop a model.

H. Model Evaluation and Testing

The goal of this step is to develop the only model ready to formulate a target value fast and tolerably. A data scientist can achieve this goal through model tuning. That's the optimization of model parameters to realize an algorithm's best performance.

IV. Conclusion and Future Work

In conclusion, as identified through the literature review, there's a requirement for combinational and more complex models to extend the accuracy of predicting the first onset of cardiovascular diseases.

The proposed framework using combinations of Decision Tree and Random forest for heart disease prediction. Using the Cleveland heart condition database, train and test the system and thus attain the foremost efficient model. In future, we are interested to study some of the deep learning models such as CNN or DNN algorithm for heart disease prediction. Also we have interest to classify it as multi class problem to identify the level of the disease.

References

[1] Mackay,J., Mensah,G. 2004 "Atlas of heart condition and Stroke" Nonserial Publication, ISBN-13 9789241562768 ISBN-10 9241562765.
[2] Robert Detrano 1989 "Cleveland heart condition Database" V.A. Medical Center, Long Beach and Cleveland Clinic Foundation.
[3] Yanwei Xing, Jie Wang and Zhihong Zhao Yonghong Gao 2007 "Combination data processing 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 5 data processing Algorithms in Coronary Heart Disease" Life System Modeling and Simulation Lecture Notes in computing, pp 129-135.
[5] Jyoti Soni, Ujma Ansari, Dipesh Sharma 2011 "Predictive data processing for Medical Diagnosis: an summary of heart condition Prediction" International Journal of Computer Applications, doi 10.5120/2237-2860.
[6] Mai Shouman, Tim Turner, Rob Stocker 2012 "Using data processing Techniques In heart condition 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 replacement probability algorithm for the diagnosis of arteria coronaria 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.

