IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Effective Study of Machine Learning Algorithms for Cardiovascular Disease Prediction

Naidu Subhasri M.Tech,

Prof.M.Sampath Kumar

DEPARTMENT OF INFORMATION TECHNOLOGY AND COMPUTER APPLICATIONS
Andhra University, Visakhapatnam

Abstract:

Cardiovascular Disease (CVD) is most common disorder of heart and blood vessels, that rapidly increasing death rate every year. Heart is an important organ in human body used for pumping blood throughout the body. To predict cardiovascular disease and to minimize the cost of clinical tests various Machine Learning algorithms and techniques are applied to different datasets used from Heart disease diagnosis and Health care industries. This project analyzes how effectively and accurately the machine learning algorithms works to predict classifier models accuracy such as Ensemble Learning, Random Forest, Adaptive boosting, Decision Tree, Support Vector Machine, K-Nearest Neighbor, Naive Bayesian Classifier.

Keywords: Cardiovascular Disease, Ensemble Random Forest, Adaptive boosting, Decision Tree, Support vector machines, k-Nearest Neighbor, Naive Bayesian Classifier, Accuracy.

I. INTRODUCTION:

Machine learning is an application of Artificial Intelligence that provides systems with the ability to learn automatically and improve from experience without being explicitly programmed. Machine Learning applies on different algorithms to solve data problems. The process of Learning begins with observations or data or instructions in order to look patterns in data and to make better decisions in the future. The primary aim of machine learning is to allow computers learn automatically without human intervention or assistance.

Machine learning is categorized into three categories:

• Supervised learning: In this type of learning, the machine is provided with a given set of inputs with their desired outputs. The machine needs to study those given sets of inputs and outputs and find a general function that maps inputs to desired outputs.

• Unsupervised learning: Here, the goal is to find a good internal representation of the input. Labeled examples are not available in unsupervised learning.

• Reinforcement learning: Here, the algorithm learns a policy of how to act given an observation of the world without knowing whether it has reached the goal or not.

II. SUPERVISED LEARNING:

Supervised learning is the task of machine learning that infers a function from a given set of data that maps an input to a desired output based on example input-output pairs. The training data supervises and produces a general rule (function), with set of examples which the computer is trained [2].

Supervised learning comes in two different flavors:

We consider each training case consists of an input vector x and a target output t.

• Regression: The target output is a real number or a whole vector of real numbers such a price of stock in 6 months' time or the temperature at noon tomorrow.

• Classification: The target output is a class label like in the simplest case choosing between positive and negative. We can also have multiple alternative levels [3].

Naive Bayes:

It is a classification technique-based on Bayes Theorem and Assumes that features are statistically independent. This theorem makes an assumption that the variables are independent of each other and still proves itself to be a good classifier. Naïve Bayes focuses on the classification of the text and is primarily used for performing clustering and classification [4]. It is a probabilistic classifier and makes use of the probability to classify a given object.

The formula for the Bayes Theorem is $P(L|M) = P(M|L)\ P(L) / P(MP)$

P (L|M) signifies the Posterior Probability, the probability of hypothesis L over the occurred event M.

P (M|L) is the probability of likelihood, Probability of the evidence given that the probability of a hypothesis is true.

P (L) signifies the probability of hypothesis before the evidence is observed.

P (M) signifies the probability of evidence.

Decision Tree:

Decision Tree is a supervised learning algorithm and this technique is mostly used in classification problems. It performs effortlessly with continuous and categorical attributes. This algorithm performs effortlessly with continuous and categorical attributes. Decision Tree algorithm, first calculates the entropy of each and every attribute. Then the dataset is split with the help of the variables or predictors with maximum information gain or minimum entropy. These two steps are performed recursively with the remaining attributes.

Support Vector Machine:

Support Vector Machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. The model is trained with a set of training data points that belong to either of the two classes of a binary classifier. Based on this, the SVM then implements the learned model on the newer data points belonging to the test sample and place them into either of the two classes. The SVM is a non-probabilistic binary classifier. The idea behind this classification model is to implement an (n-1)-dimensional hyperplane to linearly classify n-dimensional feature vectors into two separate classes that achieves largest distance of two classes. It overcomes the high dimensionality problems [4].

K-Nearest Neighbors (k-NN):

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique. This algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. It is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Ensemble Learning:

Ensemble learning is the process by which multiple models, such as classifiers or experts, are strategically generated and combined to solve a particular computational intelligence problem. Ensemble learning is primarily used to improve the performance of a model, or reduce the likelihood of an unfortunate selection of a poor one. Two most popular ensemble methods are Bagging and Boosting.

Bagging: Bagging or bootstrap aggregating is applied where the accuracy and stability of a machine learning algorithm needs to be increased. It is applicable in classification and regression. It also decreases variance and helps in handling overfitting.

Boosting: Boosting‟ refers to a family of algorithms which converts weak learner to strong learners. It is a technique in ensemble learning which is used to decrease bias and variance.

Random forest Algorithm:

Random Forest algorithm is a supervised learning algorithm. There is a direct relationship between the number of trees in the forest and the results it can get. It uses a number of decisions trees and predicts the more accurate

Result by averaging in case of regression and voting in case of classification. Random forest is an ensemble model using bagging as the ensemble method and decision tree as the individual model. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

AdaBoost (Adaptive Boosting)

AdaBoost is a boosting ensemble model and works especially well with the decision tree used to boost the performance of decision trees on binary classification problems. AdaBoost learns from the mistakes by increasing the weight of misclassified data points. This makes a new prediction by adding up the weight (of each tree) multiply the prediction (of each tree). Obviously, the tree with higher weight will have more power of influence the final decision.

III. IMPLEMENTATION:

Heart diseases have emerged as one of the most prominent cause of death all around the world. According to World Health Organization, heart related diseases are responsible for the taking 17.7 million lives every year, 31% of all global deaths. In India too, heart related diseases have become the leading cause of mortality [5]. Heart diseases have killed 1.7 million Indians in 2016, according to the 2016 Global Burden of Disease Report, released on September 15, 2017. Heart related diseases increase the spending on health care and also reduce the productivity of an individual. Estimates made by the World Health Organization (WHO), suggest that India have lost up to $237 billion, from 2005-2015, due to heart related or cardiovascular diseases [6].

The risk parameters associated with heart risk are age, family history, hypertension, high cholesterol, diabetes, smoking, tobacco, alcohol consumption, obesity, poor diet and chest pain. Thus, feasible and accurate prediction of heart related diseases is very important.

Data set and attributes:

The data is collected from the UCI machine learning repository. The data set is named Heart Disease Dataset found in the UCI machine learning repository. The UCI machine learning repository contains a vast and varied number of datasets which include datasets from various domains. These datasets are widely used by machine learning community from novices to experts to understand data empirically. Various academic papers and researches have been conducted using this repository. This repository was created in 1987 by David Aha and fellow students at UCI Irvine. Heart disease dataset contains data from four institutions [4].

1. Cleveland Clinic Foundation.

2. Hungarian Institute of Cardiology, Budapest.

3. V.A. Medical Centre, Long Beach, CA.

4. University hospital, Zurich, Switzerland.

For the purpose of this study, the data set provided by the Cleveland Clinic Foundation is used. This dataset was provided by Robert Detrano, M.D, Ph.D. Reason to choose this dataset is, it has fewer missing values and is also widely used by the research community [10].

Table 1. Attributes of the Heart disease dataset

Attribute	Representation	Information	Description
Age	Age	Interger	Age in years(29 to 77)
Sex	Sez	Interger	Gender instance(0= Female, 1= Male)
ChestPain	Cp	Interger	Chest pain type
Rest Blood Pressure	Trestbps	Interger	Resting blood pressure in mm Hg [94, 200]
SerumCholestoral	Chol	Interger	Serum cholesterol in mg/dl[126, 564]
FastingBloodSugar	Fbs	Interger	Fasting blood sugar > 120 mg/dl (0 = False, 1= True)
RestElectrocardiographic	Restecg	Interger	Resting ECG results
MaxHeartRate	Thalach	Interger	Maximum heart rate achieved[71, 202]
ExerciseInduced	Exang	Interger	Exercise induced angina (0: No, 1: Yes)
Oldpeak	Oldpeak	Real ST	depression induced by exercise relative to rest[0.0, 62.0]
Slope	Slope	Interger	Slope of the peak exercise ST segment
Major Vessels	Ca	Interger	Number of major vessels colored by fluoroscopy (values 0 - 3)
Thal	Thal	Interger	Defect types: value 3: normal, 6: fixed defect
Class	Class	Interger	Diagnosis of heart disease (1: Unhealthy, 2: Healthy)

Data Preprocessing:

The preprocessing of data is necessary for efficient representation of data and machine learning classifier which should be trained and tested in an effective manner. Preprocessing techniques such as removing of missing values, standard scalar, and Min Max Scalar have been applied to the dataset for effective use in the classifiers. The standard scalar ensures that every feature has the mean 0 and variance 1, bringing all features to the same coefficient. Similarly, in Min Max Scalar shifts the data such that all features are between 0 and 1.

Dimensionality Reduction:

Dimensionality Reduction involves selecting a mathematical representation such that one can relate the majority of the variance within the given data by including only most significant information. The data considered for a task or a problem, May consists of a lot of attributes or dimensions, but not all of these attributes may equally influence the output. A large number of attributes, or features, may affect the computational complexity and may even lead to overfitting which leads to poor results. Thus, Dimensionality Reduction is a very important step considered while building any model.

A. Feature Extraction: A new set of features is derived from the original feature set. Feature extraction involves a transformation of the features. This transformation is often not reversible as few, or maybe many, useful information is lost in the process. In [7] and [8]Principal Component Analysis (PCA)is used for feature extraction. Principal Component Analysis is a popularly used linear transformation algorithm. In the feature space, it finds the directions that maximize variance and finds directions that are mutually orthogonal. It is a global algorithm that gives the best reconstruction.
B. Feature Selection: A subset of original feature set is selected. In [9], key features are selected by CFS (Correlation based Feature Selection) Subset Evaluation combined with Best First Search method to reduce dimensionality.

PERFORMANCE EVALUATION METHODS:

Performance of the algorithm can be evaluated on the basis of various parameters like accuracy, Confusion matrix, precision and many more.

Accuracy: Accuracy measure used to calculate model's efficiency to perfectly label the unknown data. If data is categorical, accuracy can be measured as the rate at which data will be labeled with true category of data. If data is continuous, accuracy can be measured by the distance between predicted value and the correct value.

Confusion Matrix: It is measured by the count of right and wrong prediction made by model in comparison with real classifiers in case of test dataset. This matrix would of n*n, n is count of classes.

Conclusion:

Machine Learning Algorithms plays a vital role in predicting cardiovascular diseases or heart related diseases. Each of the above-mentioned algorithms has performed extremely well. Random Forest Ensemble model have performed well with good accuracy because they solve the problem of overfitting by employing multiple algorithms (multiple Decision Trees in case of Random Forest).

References:

[1] M. Welling, "A First Encounter with Machine Learning".
[2]. R. Sathya, Annamma Abraham, "Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification", (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 2, No. 2, 2013.
[3]. Thomas G. Dietterich, "Machine-Learning Research", AI Magazine Volume 18 Number 4 (1997).
[4] Gnaneswar B., EbenezarZebarani M.R., "A review on prediction and diagnosis of heart failure", 2017 (ICIIECS).
[5] Ramadoss and Shah B et al. "A. Responding to the threat of chronic diseases in India". Lancet. 2005; 366:1744–1749. doi: 10.1016/S0140-6736(05)67343-6.
[6] Global Atlas on Cardiovascular Disease Prevention and Control. Geneva, Switzerland: World Health Organization, 2011.
[7] Dhomse Kanchan B and Mahale Kishor M. et al. "Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis", 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication. [8] R. Kavitha and E. Kannan et al. "An Efficient Framework for Heart Disease Classification using Feature Extraction and Feature Selection Technique in Data Mining ", 2016.
[9] Shan Xu, Tiangang Zhu, Zhen Zang, Daoxian Wang, Jun Feng Hu and Xiaohui Duan et al. "Cardiovascular Risk Prediction Method Based on CFS Subset Evaluation and Random Forest Classification Framework", 2017 IEEE 2nd International Conference on Big Data Analysis.
[10] 5] U. H. Dataset, "UCI Machine Learning Repository", [online]. https//archive.ics.uci.edu/ml/machine-learningdatabases/heartdisease/ Hear