Machine Learning Methods for Heart Disease Prediction

Rashmi G.O, Ashwin kumar .U.M

Abstract: ***Machine learning is utilized to empower a program to analyze information, understand correlations and make utilization of bits of knowledge to take care of issues or potentially enhance information and for prediction. The American Heart Association Statistics 2016 Report shows that coronary illness is the main source of death for people, responsible for 1 in every 4 deaths. Machine learning algorithms play a very important role in medical area. We use machine learning to understand, predict, and prevent cardiovascular disease using numeric data. The end goal is to produce an approved machine learning application in healthcare. In an effort to refine the search for a useful and accurate method with the dataset, the results of several algorithms will be compared. The front-runners will be analyzed and used to develop a unique, higher-accuracy method. Machine learning methods inclusive of Logistic Regression, Naïve Bayes, Decision tree(CART). We use ensemble learning for better accuracy which includes algorithms like Random Forest, XGBoost, Extra trees classifier. Also, our work adds to the present literature by giving a far reaching review of machine learning algorithms on sickness prediction tasks. Our goal is to perform predictive analysis with these machine learning algorithms on heart diseases using ensembles like bagging, boosting, stacking. Machine Learning algorithms used and conclude which techniques are effective and efficient. A huge medical datasets are accessible in different data repositories which used in the real world application.***

***Index Terms*: Machine learning, Cardiovascular disease, Decision tree(CART), Ensemble learning.**

I. INTRODUCTION

The paper focuses to predict conceivable number of heart attacks patients from the dataset utilizing Machine learning strategies and determines which model gives the most percentage of correct predictions for the diagnoses. Machine learning is a region of computer science in which the computer predicts the following undertaking to perform by investigating the information gave to it. The calculations of machine learning are built so as to take in and make expectations from the information not at all like the static programming calculations that require unequivocal human guidance. Machine learning enables framework with the capacity to adapt consequently and show signs of improvement with experience without being expressly modified [2].

Manuscript published on 30 May 2019.
***** Correspondence Author (s)

Rashmi .G.O, Department of Computer Science and Information Technology, REVA University, Bangalore, India.
Ashwinkumar .U.M , Department of Computer Science and Information Technology, REVA University, Bangalore, India.

A. Motivation

The American Heart Association Statistics 2016 Report indicates that coronary illness is the leading cause of death for humans, responsible for 1 in every 4 deaths. Even modest improvements in prognostic models of heart events and complications could save hundreds of lives and help to significantly reduce the cost of health care services, medications, and lost productivity.

B. Methods

Progression and development of fresher advances, for example, machine learning, artificial intelligence, analytics have affected numerous parts, such as health care, automotive and so forth. The World over, heart attack Disease is influencing a huge number of individuals. Different machine learning procedures incorporate ensemble classifiers specifically Random Forest, Extra Trees and XGBoost can be utilized in medicinal services for enhancing prediction accuracy [3].

II. RELATED WORK

In paper [1], adaptable start to finish tree boosting framework XGBoost, scales past billions of precedents utilizing far less assets than existing frameworks. The most critical factor behind the achievement of XGBoost is its scalability in all situations. The framework runs in excess of multiple times quicker than existing well known arrangements on a single machine and scales to billions of precedents in conveyed or memory-restricted settings. Parallel and distributed computing makes learning quicker which enables quicker model investigation. They planned and assembled a profoundly versatile start to finish tree boosting framework. hypothetically justified weighted quartile draw for proficient proposition count and presented a novel sparsity-mindful methods for parallel tree learning. They proposed a compelling cache-aware block structure for out-of-core tree learning.

Retrieval Number: E10450585S19/2019©BEIESP
DOI:10.35940/ijeat.E1045.0585S19
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

In paper [2], different machine learning techniques including Na¨ıve Bayes (NB), Decision Tree (DT), K-Nearest Neighbour (K-NN), Multilayer Perceptron (MLP), Radial Basis Function (RBF), Single Conjunctive Rule Learner (SCRL).Utilizing ensemble machine learning approaches, the proficiency of every individual classifier, and furthermore such classifiers in mix, by utilizing the ensembles bagging, boosting, and stacking methods, has been assessed. They watched a few upgrades now and again in the wake of applying the ensembles like bagging, boosting, and stacking approaches. SVM outflanked all the others when the boosting strategy was applied.

In paper [5], it is assumed that though most experts are using various classifier techniques, for instance, Logistic regression, random forest, Naive Bayes, gradient boosting and SVM in the conclusion of heart disease, applying logistic regression and Naïve Bayes provides better results in the discovery of coronary sickness and good precision when stood out from various classifiers. Accuracy emerges by the attributes gathered from the data. The accuracies can in like manner be extended with the assistance of higher quality, datasets and greater computational frameworks. In view of the exactness created by the calculations the optimum calculation which grant accuracy is picked to find the prediction.

In paper [6], Prediction possibility of disease using data mining or machine learning techniques in order to enhance the accuracy of the disease detection system. study of different approaches such as neural network, Naïve Bayes, SVM, KNN, FCN, etc and it is concluded that SVM gives the best performance as compared to the other existing techniques. The research is designed using SVM and RF algorithm. The accuracy is compared on diabetes, kidney and liver diseases.

In the paper [7], The survey about various classification strategies utilized for anticipating the hazard dimension of every individual deploy on pulse rate, age, cholesterol, gender, ,blood pressure etc. The patient danger level is grouped utilizing data mining classification techniques for example, Decision Tree Algorithm, KNN, Naïve Bayes, Neural Network, etc. Accuracy of the danger level is high while utilizing more attributes. According to the analysis made, it is seen that numerous authors utilize different technologies and different number of attributes for their investigation. Therefore, different technologies provide different precision be based on how many attributes considered. Utilizing ID3 and KNN algorithm the endanger rate of coronary illness was discovered and accuracy level also afford for diverse number of attributes.

III. METHODOLOGY

Machine learning utilizes algorithms to discover patterns in information and then utilizes a model that perceives those examples to make prediction on new data. Machine learning incorporates some of the accompanying methodologies like Supervised learning algorithms, Ensemble Machine Learning Algorithms.

IV. SUPERVISED LEARNING ALGORITHMS

Supervised learning algorithms use labelled data. In this paper we are explaining some popular supervised algorithms used in health care.

A. Logistic Regression :

Logistic regression is widely known for binary classification where it standout amongst the greatest quantity effective machine learning methods. Because of its straightforwardness, which encounter its utilization on a wide scope of issues and gives appropriate arrangements . It takes a shot at the needy variable which is absolute. The factors are binary dependent variables, for example, 1s or 0s, yes or no and so on[5].

B. Naive Bayes :

Naive Bayes classifier uses the Bayes theorem. Naive Bayes classification algorithm is emphatically versatile of essential variables linear as indicator factors in an issue proclamation. It is comparable regression and classification which does intense challenge by SVM. It distinguishes the forte of the ill people identified with the disease. It demonstrates the likelihood of each info property for the anticipated state and gives the likelihood of occasion happen. A contingent likelihood is the probability of some decision A, given some proof/perception, B, where a reliance correlation exists among A and B. This likelihood is signified as P (A|B) where, P(A) is the likelihood of occasion A, P(B) is the likelihood of occasion B, P(B|A) is the likelihood of occasion B with the condition that occasion A has occurred P(A|B) = (P(B|A)*P(B)/P(A) [5].

C. Decision tree(CART) :

Decision Trees (DTs) are not involving any assumptions as to the parameters of a frequency distribution supervised learning method utilized in favour of classification and regression. The intent is to build a model that estimates a predict variable by learning basic decision rules induced from the data features. Among the various decision tree algorithms CART (Classification and Regression Trees) is very alike to C4.5, however it varies in that it reinforce numerical predicting variable (regression) and does not compute rule sets. CART assemble binary trees utilizing the feature and threshold feature that yield the high information gain at individual node. Gini Metric is used in CART. Gini impurity is a measurement of how regularly an graciously picked element from the set would be inaccurately named if it was arbitrarily labeled in a manner corresponding to the distribution of labels in the subset.

V. ENSEMBLE MACHINE LEARNING ALGORITHMS

Lately, ensemble learning models led the pack and wound up well known among machine learning professionals. Ensemble learning model utilizes numerous machine learning calculations to conquer the potential shortcomings of a single model. The three most famous techniques for joining the predictions from various models are :

A. Bagging :

Building up of various models normally of a similar type from various subsamples with replacement of training dataset. The last output prediction is found the middle value of over the prediction of all the sub-models. Examples of bagging models are:

- *Random Forest :* In machine learning Random Forest is a procedure which consolidates decision trees and ensemble classifier. Forest produces N number of decision trees by utilizing randomly chosen data samples as their input. Samples of the training dataset are taken with substitution, the trees are built in a way that diminishes the connection between's individual classifiers. In Random Forest, the oblique remains the identical, however the quantity of trees increments. Additionally, the inconsistent of the model reductions as various trees in the forest grow. Random forest creates an accumulation of trees with controlled fluctuation. The prediction from all the trees can be decided by majority voting. Each tree in random forest, is increased by taking into consideration that every tree is developed on a bagging representative, the persist information are said to be “out-of-bag”, the "out-of-bag" information fill as a test set of the tree[3].
- *Extra Trees Classifier* **:** An extra trees classifier, also called as an Extremely randomized trees classifier, is a variation of a random forest. Unlike a random forest, at each progression the whole sample is utilised and decision boundaries are taken at random, as opposed to the best one.

B. Boosting :

Assembling numerous models generally of the same type, each of which learns to fix the prediction errors of a prior model in the chain. An example of boosting which is famous boosting ensemble machine learning algorithm is :

- *eXtreme Gradient Boosting :* Gradient boosting is a machine learning method for classification and regression issues, which regularly utilizes decision trees to create a model in the form of an ensemble of weak prediction models. In boosting as the name proposes, one is learning from other which in turn boosts the learning. XGBoost is short for eXtreme gradient boosting, which delivers an expectation model as that of a joint exertion of feeble prediction models. Like other boosting strategies, it makes the model in a consistent structure, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. In many supervised learning issues there is a yield variable y and a vector of input variables x which are related together through a joint likelihood appropriation P(x, y). Utilizing a training set (x1,y1),...,(xn,yn) with known values of x and the related values of y, the goal is to find an approximation Fˆ(x) to a function F´(x) which limits the expected value of a predefined loss function L(y,F(x)):[4]

C. Voting :

Building numerous models normally of various sorts and by basic statistics like computing the mean are utilized to consolidate predictions.

D. Stacking *:*

It is an ensemble learning type that joins various grouping or relapse models by means of a meta- regressor or a meta-classifier. This base dimension models are prepared dependent on a total preparing set, at that point the meta-show is prepared on the yields of the base dimension display as highlights. The base dimension regularly comprises of various learning calculations and consequently stacking ensembles are frequently heterogeneous [9].

VI. TOOLS AND TECHNIQUES

IBM Watson Machine learning service can be used for deploying the machine learning algorithms. The service helps in deployment and integration of Artificial Intelligence(AI) into our applications. Watson machine learning empowers to deploy, monitor and optimize models quickly and easily, as it is Ubiquitous in which the service used on cloud environment. It is simple to manage models in production, and its seamless workflow enable continuous retraining to maintain and improve model accuracy. Watson machine learning benefits by scale, speed and simplicity. The Watson on IBM cloud allows us to integrate the AI into our application to manage, store, train and test our data. This project use a network of remote services hosted on the internet rather than a local server or a personal computer. Sklearn library plays a very important role in carrying out the experiment.

VII. RESULTS AND DISCUSSION

The experiment is carried out using Cleveland dataset for heart disease in UCI Machine Learning Repository. Some eminent classifiers are considered. The accuracy measure is considered for validation. Table 1 shows the accuracies with different features for different classifiers. Our experiment demonstrate ensembles performs better than single classifiers where as the stacked ensembles gives higher accuracy. Figure 1 shows the Comparison of accuracy with different techniques in bar chart.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

Table 1. Accuracy of Proposed stacked model with other techniques.

Machine learning Algorithms	**Accuracy %**		
	Accuracy with all 14 feathers	***Accuracy with selected 8 feathers***	***Accuracy by Stacking ensembles selecting threshold feathers***
Logistic Regression	87.77	84.44	98.0
Naive bayes	87.77	86.17	97.80
Decision Tree(CART)	75.55	74.44	95.60
Extra Trees	83.33	82.22	98.90
Random Forest	83.33	82.22	97.80
XGBoost	82.22	81.11	97.80

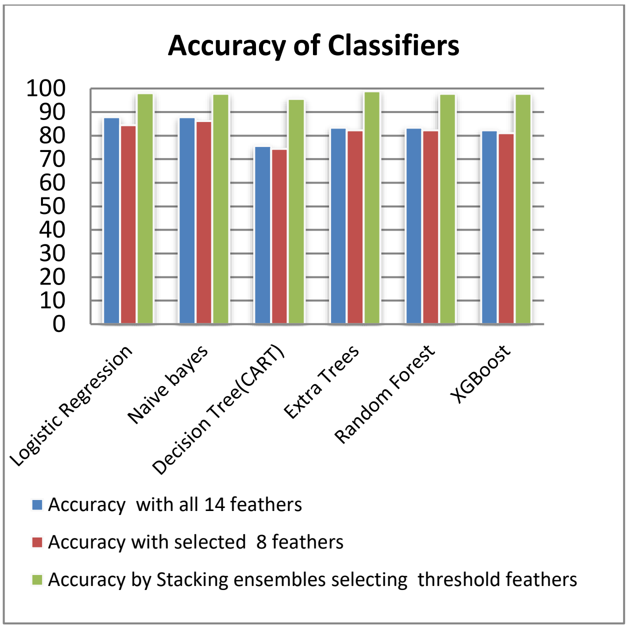

Figure 1. Comparison of accuracy with different techniques.

VIII. CONCLUSION AND FUTURE SCOPE

Our work related to machine learning techniques such as classification in health domain. The literature review is done on various machine learning techniques. The classification using ensembles gives better results. Hence we used different ensembles like bagging, boosting and stacking to get better accuracy results in predicting heart disease numeric dataset. Ensemble strategies have been extremely effective in setting record execution on testing datasets. Here the results are showing great result in stacking ensemble. The project highlights applications, challenges and future work of Machine learning in healthcare to get the higher accuracy by using for image, ECG and signalling data.

REFERENCES

1. *Tianqi Chen, Carlos Guestrin " XGBoost: A Scalable Tree Boosting System " KDD '16, August 13-17, 2016, San Francisco, CA, USA, 2016 ACM.*
2. H K.Gianey, R.Choudhary " *Comprehensive Review On Supervised Machine Learning Algorithms* " 2017 International Conference on Machine learning and Data Science.
3. I.Yekkala, , et al. "*Prediction of heart disease using Ensemble Learning and Particle Swarm Optimization*" 2017 International Conference On Smart Technology for Smart Nation.
4. A.Batra, et al. "Classification of Arrhythmia Using Conjunction of Machine Learning Algorithms and ECG Diagnostic Criteria" International Journal of Biology and Biomedicine.
5. Dinesh K.G, et al. "*Prediction of Cardiovascular Disease Using Machine Learning Algorithms*" Proceeding of 2018 IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore, India.
6. S.Jatav, V.Sharma "*An Algorithm For Predictive Data Mining Approach In Medical Diagnosis* " International Journal of Computer Science & Information Technology (IJCSIT) Vol 10, No 1, February 2018 DOI:10.5121/
7. T.Princy. R , J. Thomas " *Human Heart Disease Prediction System using Data Mining Techniques* "2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT]
8. UCI Machine Learning Repository :http://archive.ics.uci.edu/ml/datasets.html
9. https://blog.statsbot.co/ensemble-learning-d1dcd548e93
10. N.Bhargava "An Approach for Classification using Simple CART Algorithm in Weka" 2017 11 th International Conference on Intelligent Systems and Control (ISCO)- *978-1-5090-2717-0/171$31 .00* ©2017 IEEE.

AUTHORS PROFILE

Mrs. Rashmi G O pursed Bachelor of Engineering Degree from Sri Jagadguru Murugharajendra Institute of Technology, Chitradurga, affiliated to Visveswaraya Technological University in 2009. She is currently pursuing Master of Technology in Data Engineering and Cloud Computing, Department of C&IT, REVA University Banglore since 2017.

***Dr.* Ashwinkumar.U.M.** received B.E. Degree from Karnataka University, M.Tech. Degree from the University of Mysore Sri J C College of Engineering, Mysore, Awarded Ph.D degree in Computer Science and Engineering, Visvesvaraya Technological University Belgaum, India. He is currently working as Associate Professor in School of Computing and Information Technology. He is involved in research of Data Mining. Artificial Intelligence and Image Processing. He has published over 45 papers in national and international conferences and 5 papers in national and international journals. He presented papers at various International conferences in India and Abroad like Thailand, Phuket. He has received the Best paper Award at Third International Conference in Advances in Computing and Communication Technologies held at Rohtak. He is reviewer for nearly more than 5 international/national reputed Journals. Some of the Journals which he has reviewed are Scientific and Academic publishing Journals He has been reviewer, session chair and programme committee member for more than 5 national/international conferences. Some of conferences where he was programme committee member are as follows. Third International conference on Advances in Recent Technologies in Communication and Computing ARTCom 2011 ,Bangalore, ITU Kaleidoscope Academic Conference Tokyo Japan. He has presented many invited talks on Computer Graphics and Visualisation and Machine learning. He has participated in the Live Transmission of VTU EDUSAT Programme-17 and delivered Lectures on Data base management Systems. He has worked on DRDO Project as Junior research fellow Sanctioned of Rs 14.40 Lacs Title A Unique medical data analysis platform Using Data Mining Techniques for Armed Forces Personnel. At SJBIT Computer Science Dept, Research Center, Bangalore. He has conducted several national workshops and seminars for faculty and students.