International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 2 Feb 2022, pp: 324-328 www.ijaem.net ISSN: 2395-5252

Heart Disease Prediction Using Feature Selection and XGBoost

PriyankaKriplani, Prof. Prateek Gupta, Prof. Nagendra Singh
[1]*Student, Shri Ram Institute of Sc. & Tech,, Jabalpur, MP*
[2] *Prof,Shri Ram Institute of Sc. & Tech,, Jabalpur, Jabalpur, MP*
[3] *Prof,Shri Ram Institute of Sc. & Tech,, Jabalpur, Jabalpur, MP*

Submitted: 25-01-2022 Revised: 05-02-2022 Accepted: 08-02-2022

ABSTRACT: Micro blogging websites like Twitter and Facebook, in this new era, is loaded with opinions and data. One of the most widely used micro-blogging site, Twitter, is where people share their ideas in the form of tweets and therefore it becomes one of the best sources for sentimental analysis. Opinions can be widely grouped into three categories good for positive, bad for negative and neutral and the process of analyzing differences of opinions and grouping them in all these categories is known as Sentiment Analysis. Data mining is basically used to uncover relevant information from web pages especially from the social networking sites. Merging data mining with other fields like text mining, NLP and computational intelligence we are able to classify tweets as good, bad or neutral. In order to improve classification results in the domain of sentiment analysis, we are using ensemble machine learning techniques for increasing the efficiency and reliability of proposed approach. For the same, we are using Linear Support Vector Machine and experimental results prove that our proposed approach is providing better classification results in terms of f-measure and accuracy in contrast to individual classifiers.

KEYWORDS: Sentiment Analysis, Twitter, Adjective Analysis, Naïve Bayes, SVM.

I. INTRODUCTION

The World Health Organization (WHO) [1] lists cardiovascular diseases as the leading cause of death globally with 17.9 million people dying every year. The risk of heart disease increases due to harmful behaviour that leads to overweight and obesity, hypertension, hyperglycaemias, and high cholesterol [1]. Furthermore, the American Heart Association [2] complements symptoms with weight gain (1–2 kg per day), sleep problems, leg swelling, chronic cough and high heart rate. Diagnosis is a problem for practitioners due the symptoms' nature of being common to other conditions or confused with signs of aging. The growth in medical data collection presents a new opportunity for physicians to improve patient diagnosis. In recent years, practitioners have increased their usage of computer technologies to improve decision-making support. In the health care industry, machine learning is becoming an important solution to aid the diagnosis of patients. Machine learning is an analytical tool used when a task is large and difficult to program, such as transforming medical record into knowledge, pandemic predictions, and genomic data analysis. In this research work, the supervised machine learning concept is utilized for making the predictions. A comparative analysis of the three data mining classification algorithms namely Random Forest, Decision Tree and Naïve Bayes are used to make predictions. The analysis is done at several levels of cross validation and several percentage of percentage split evaluation methods respectively. The Stat Log dataset from UCI machine learning repository is utilized for making heart disease predictions in this research work. The predictions are made using the classification model that is built from the classification algorithms when the heart disease dataset is used for training. This final model can be used for prediction of any types of heart diseases [2][3].

A key challenge confronting healthcare organization (hospitals, medical centers) is the facility of quality services at reasonable prices. Quality amenities suggest diagnosing patients accurately and regulating medications that are effective. Poor clinical choices can prompt deplorable results, which are in this manner unsatisfactory. Hospitals should limit the cost of clinical tests. They can accomplish these outcomes

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 2 Feb 2022, pp: 324-328 www.ijaem.net ISSN: 2395-5252

by utilizing fitting PC based data and additionally choice emotionally supportive networks [4][5]. The heart is the essential piece of our body. Life is itself reliant on effective working of the heart. If task of the heart isn't legitimate, it will influence the other body parts of human, for example, cerebrum, and kidney and so on. Coronary illness is a sickness that effects on the activity of the heart. There are several elements which builds danger of Heart ailment [6]. Because of a wide accessibility of superlative measure of information and a need to change over this accessible huge measure of information to helpful data requires the utilization of information mining strategies. Information Mining and KDD (learning disclosure in the database) have turned out to be prominent as of late [7][8]. The popularity of information mining and KDD (information revelation in database) shouldn't be an amazement since the measure of the information increases that are accessible are extremely extensive to be analyzed physically and even the techniques for programmed information investigation in view of established insights and machine adapting frequently threaten issues when preparing large, dynamic information increases comprising of complex items [9]. Information Mining is the center piece of Knowledge Discovery Database (KDD). Numerous individuals regard Data Mining as an equivalent word for KDD since it's a key piece of KDD process [10][11]. There are sure stages of information mining that you will need to get comfortable with, and these are exploration, pattern identification, and deployment. Information mining is an iterative procedure that commonly includes the accompanying stage [12].

II. LITERATURE REVIEW

Liaqat Ali et.al. [13]introduces an expert system that stacks two support vector machine (SVM) models. The first SVM model is linear and L1 regularized. It has the capability to eliminate irrelevant features by shrinking their coefficients to zero. While the second SVM model is L2 regularized. It is used as a predictive model. To optimize the two models, we propose to use a hybrid grid search algorithm (HGSA) which is capable of optimizing the two models simultaneously. The effectiveness of the proposed method is evaluated using six different evaluation metrics including accuracy, sensitivity, specificity, MCC, ROC charts and area under the curve (AUC). The first model has the capability to eliminate irrelevant features by shrinking their coefficients to zero. Performance comparison is done using accuracy, ROC chart and AUC evaluation metrics. The proposed method is efficient in terms of time complexity.

AshirJaveedet. al. [14] develops a novel diagnostic system. The proposed diagnostic system uses random search algorithm (RSA) for features selection and random forest model for heart failure prediction. The proposed diagnostic system is optimized using grid search algorithm. Two types of experiments are performed to evaluate the precision of the proposed method. In the first experiment, only random forest model is developed while in the second experiment the proposed RSA based random forest model is developed. Experiments are performed using an online heart failure database namely Cleveland dataset. In addition, the proposed method achieved classification accuracy of 93.33% while improving the training accuracy as well. We develop only random forest model which is implemented in Python programming package. The model hyper parameters are tuned using grid search algorithm. It was shown that the proposed RSA-RF learning system improves the performance of random forest model by 3.3%. It was also observed that the proposed system reduces the time complexity of the machine learning models by reducing the number of features.

Authors have proposed different classification algorithms, each with its own advantage on three separate databases of disease (Heart, Breast cancer, Diabetes) available in UCI repository for disease prediction. The feature selection for each dataset was accomplished by backward modeling using the p-value test. [15]

After data munging and attributes selection, machine learning algorithms including Logistic Regression, Decision Trees, Random Forest, Support Vector Machine(SVM) and Adaptive Boosting, are used for prediction of the above-mentioned diseases, and a comparison of their accuracy is done for selecting best model for that disease dataset. The feature selection from 13 input parameter by backward elimination resulted in a total of 11 significant input parameters which include gender, type of chest pain, blood pressure, blood sugar level, electrocardiograph result, maximum heart rate, exercise induced angina, old peak, Slope, number of vessels colored, thal.

Logistic Regression was found to have the highest accuracy among all.The prediction accuracy of our proposed method reaches 87.1% in Heart Disease detection using Logistic Regression.

Anjan Nikhil Repakaet. al. [16] Smart Heart Disease Prediction) is built via Navies Bayesian in order to predict risk factors concerning heart disease. For predicting the chances of heart

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 2 Feb 2022, pp: 324-328 www.ijaem.net ISSN: 2395-5252

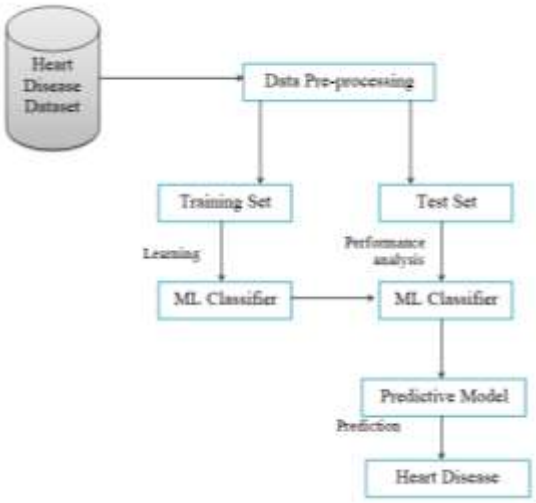
disease in a patient, the following attributes are being fetched from the medical profiles, these include: age, BP, cholesterol, sex, blood sugar etc... The collected attributes acts as input for the Navies Bayesian classification for predicting heart disease. This classification algorithm basically employs conditionalindependence; this implies that value of an attribute for anavailable class is not dependent on other attribute values sincethe algorithm relies upon the Bayesian theorem.Proposed classification techniquesperformance which is compared with prevailing techniques ofSMO (Sequential Minimal Optimization), Bayes Net and MLP(Multi-Layer Perception). Effective outcome is exhibited by theproposed Navies Bayesian with greater performance in contrastto rest of the techniques.

AmoghPowaret. al. [17] Data mining gears are expecting destiny traits therefore permitting making understanding driven selections. There are many statistics mining techniques like Decision Tree, Naïve Bayes, & Neural Network. This paper proposes to use three of those techniques for predicting the coronary heart ailment. Each of these techniques will be tested in the various parameters and different accuracies will be obtained with the different parameters. In this exploration we have endeavoured to overviewtwenty research papers, which fundamentally talksabout the different normal information mining andcomputerized reasoning systems that can be utilized forforecast of heart maladies and shows which techniqueis the most precise.

III. PROPOSED SYSTEM

Figure below represents the architecture of the proposed prediction system. First step is to collect heart disease dataset. After this data pre-processing will be performed. After pre-processing the dataset, it is split into training set and testing set. Training set is used to train the algorithm and testing set is used for testing purpose. Proposed algorithm takes training dataset as the input to train on various samples and produces a trained model based on proposed algorithm. Testing data is than applied on the model for performance evaluation. For predicting the result, best performed model will be applied. The proposed system uses XGBoost method.

PROPOSED SYSTEM



Proposed classifier is based on ensemble machine learning method. XGBoost (Extreme Gradient Boosting) is also referred by gradient boosting, stochastic gradient boosting, multiple additive regression trees or simple gradient boosting machines. It is a kind of supervised machine learning algorithm which can be regarded as an improved version of gradient boosting machine. Boosting as also is an ensemble technique which leverage the errors made by existing models by correcting them until no errors can be corrected by adding models sequentially. XGBoost models are based on the technique wherein we predict the errors of models are predicted by newer models which are then added together to make a final assessment of the prediction. XGBoost algorithm is called gradient boosting because it particularly minimizes the loss when adding new models using the gradient decent algorithm.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 2 Feb 2022, pp: 324-328 www.ijaem.net ISSN: 2395-5252

IV. RESULT

For this thesis, data were collected from Kaggle Data Set. This dataset consists of 11 features and a target variable. It has 6 nominal variables and 5 numeric variables. The detailed description of all the features is as follows:

1. Age: Patients Age in years (Numeric).
2. Sex: Gender of patient (Male - 1, Female - 0) (Nominal).
3. Chest Pain Type: Type of chest pain experienced by patient categorized into 1 typical, 2 typical anginas, 3 non-anginal pains, 4 asymptomatic (Nominal).
4. Resting bps: Level of blood pressure at resting mode in mm/HG (Numerical).
5. Cholesterol: Serum cholesterol in mg/dl (Numeric).
6. Fasting blood sugar: Blood sugar levels on fasting > 120 mg/dl represents as 1 in case of true and 0 as false (Nominal).
7. Resting ecg: Result of electrocardiogram while at rest are represented in 3 distinct values 0 : Normal 1: Abnormality in ST-T wave 2: Left ventricular hypertrophy (Nominal).
8. Max heart rate: Maximum heart rate achieved (Numeric).
9. Exercise angina: Angina induced by exercise 0 depicting NO 1 depicting Yes (Nominal).
10. Oldpeak: Exercise induced ST-depression in comparison with the state of rest (Numeric).
11. ST slope: ST segment measured in terms of slope during peak exercise 0: Normal 1: Up sloping 2: Flat 3: Down sloping (Nominal).

The classification performance can be evaluated in terms of accuracy. Accuracy explains correctly classified instances of the symptoms with respect to heart disease.

Classifier	Accuracy (in %)
GBM	85.10
CART	83.40
KNN	80.85
SVC	82.55
Random Forest	90.63
Adaboost	83.40
SGD	79.14
Proposed	**91.9**

AS we can see from above results, Proposed XGBoost Classifier is best performer as it has highest test accuracy of 91.9.

V. CONCLUSION

The main motivation of this thesis is to provide an insight about detecting and curing heart disease using data mining technique. For this thesis, data were collected from Kaggle Data Sets. All attributes are numeric-valued. The data was collected from the four locations. It is integer valued from 0 (no presence) to 4 to predict the likelihood of patient getting heart diseases. These attributes are fed in to Naive Bayes, SVM, KNN, Decision Tree and Random forest, in which Random Forest gave the best result with the highest accuracy. Valid performance is achieved using Random Forest algorithm in diagnosing heart diseases and can be further improved by increasing the number of attributes.

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.

REFERENCES

[1] Cardiovascular diseases (CVDs) retrieved from http://www.who.int/cardiovascular_diseases/en/ (2019, July 16) Google Scholar.
[2] American Heart Association Classes of heart failure Retrieved from https://www.heart.org/en/health-topics/heart-failure/what-is-heart-failure/classes-of-heart-failure (2018, August 11), Google Scholar.
[3] P. Melillo, N.D. Luca, M. Bracale, L. Pecchia "Classification tree for risk assessment in patients suffering from congestive heart failure via long-term heart rate variability", IEEE J. Biomed Health Inf, 17 (3) (2013), pp. 727-733, 10.1109/jbhi.2013.2244902View Record in ScopusGoogle Scholar.

[4] G. Parthiban, S.K. SrivastavaApplying machine learning methods in diagnosing heart disease for diabetic patientsInt J ApplInfSyst, 3 (7) (2012), pp. 25-30, 10.5120/ijais12-450593.

[5] M. Yang, Y. Nataliani, "A feature-reduction fuzzy clustering algorithm based on feature-weighted entropy",IEEE Trans Fuzzy Syst, 26 (2) (2018), pp. 817-835, 10.1109/tfuzz.2017.2692203.

[6] R. Chen, N. Sun, X. Chen, M. Yang, Q. Wu "Supervisedfeature selection with astratified feature weighting method", IEEE Access, 6 (2018), pp. 15087-15098, 10.1109/ACCESS.2018.2815606.

[7] B. Dun, E. Wang, S. MajumderHeart disease diagnosis on medical data using ensemble learning (2016).

[8] R.S. Singh, B.S. Saini, R.K. Sunkaria, "Detection of coronary artery disease by reduced features and extreme learning machine",Clujul Med, 91 (2) (2018), p. 166, 10.15386/cjmed-882.

[9] B.M. Asl, S.K. Setarehdan, M. Mohebbi, "Support vector machine-based arrhythmia classification using reduced features of heart rate variability signal", ArtifIntell Med, 44 (1) (2008), pp. 51-64, 10.1016/j.artmed.2008.04.007.

[10] R. Rajagopal, V. Ranganathan, **"**Evaluation of effect of unsupervised dimensionality reduction techniques on automated arrhythmia classification", Biomed Signal Process Contr, 34 (2017), pp. 1-8, 10.1016/j.bspc.2016.12.017

[11] D. Zhang, L. Zou, X. Zhou, F. He, "Integrating feature selection and feature extraction methods with deep learning to predict clinical outcome of breast cancer",IEEE Access, 6 (2018), pp. 28936-28944, 10.1109/access.2018.2837654.

[12] S. Negi, Y. Kumar, V.M. Mishra, "Feature extraction and classification for EMG signals using linear discriminant analysis"**,** 2016 2nd international conference on advances in computing, communication, & automation (ICACCA) (2016), 10.1109/icaccaf.2016.7748960.

[13] Liaqat Ali, AwaisNiamat, Javed Ali Khan, NoorbakhshAmiriGolilarz, And XiongXingzhong," An Expert System Based on Optimized Stacked Support Vector Machines for Effective Diagnosis of Heart Disease', 2169-3536 (c) 2018 IEEE.

[14] AshirJaveed, Shijie Zhou, Liao Yongjian, IqbalQasim, Adeeb Noor, RedhwanNour, SamadWali and Abdul Basit," An Intelligent Learning System based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection", DOI10.1109/ACCESS.2019.2952107, IEEE Access.

[15] Pahulpreet Singh Kohli, ShriyaArora," Application of Machine Learning in Disease Prediction", 2018 4th International Conference on Computing Communication and Automation (ICCCA), IEEE.

[16] Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin," Design And Implementing Heart Disease Prediction Using Naives Bayesian" , Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019), IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978-1-5386-9439-8.

[17] AmoghPowar, SeemaShilvant,VarshaPawar," Data Mining &Artificial Intelligence Techniques for Prediction of Heart Disorders: A Survey", 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN), IEEE-2019.