XGB CLASSIFICATION TECHNIQUE TO RESOLVE IMBALANCED HEART DISEASE DATA

Kolluru VenkataNagendra[1], Dr. Maligela Ussenaiah[2],
[1]*ResearchScholar, VS University, Nellore*
[2]*Assistant Professor, VS University, Nellore*
(E-mail:kvnscholar@gmail.com)

***Abstract* --** In the most recent decade there has been expanding use of data mining procedures on medicinal data for finding helpful patterns that are utilized in Diagnosis and Decision Making. Data Mining methods, Clustering, Classification, Regression, Association Rule Mining, CART (Classification and Regression Tree) are broadly utilized in human healthcare sector. Data Mining algorithms, when fittingly utilized are fit for enhancing the nature of expectation, analysis and illness characterization. The main focus of this paper is to analyze Coronary Heart Disease (CHD) studied by collecting various risk factors. The experimental results demonstrate that Extreme Gradient Boosting (XGBoost) algorithm perform better than the remaining algorithms in the context of class imbalanced dataset. We evaluate the Data Mining techniques using statistical metrics Accuracy, Precision, Recall and F1 Score.

***Keywords*—** Coronary Heart Disease, Supervised learning, Neural Networks and Extreme Gradient Boosting

1. INTRODUCTION

Data mining is the process of digging data for discovering latent patterns which can be translated into valuable information. Data mining usage witnessed unprecedented growth in the last few years. Of late the usefulness of data mining techniques has been realized in Healthcare domain. Medical data mining can exploit the hidden patterns present in voluminous medical data which otherwise is left undiscovered. Data mining techniques which are applied to medical data include association rule mining for finding frequent patterns, prediction, classification and clustering. Traditionally data mining techniques were used in various domains. However, it is introduced relatively late into the Healthcare domain. Nevertheless, as on today lot of research is found in the literature. This has led to the development of intelligent systems and decision support systems in Healthcare domain for accurate diagnosis of diseases, predicting the severity of various diseases, and remote health monitoring. Especially the data mining techniques are more useful in predicting heart diseases, lung cancer, and breast cancer and so on.

Classification is a data mining technique that assigns categories to a collection of data in order to aid in more accurate predictions and analysis. Classification is one of the several methods intended to make analysis of very large data sets effectively. It is used to find out in which group each data instance is related within a given dataset. It is used for classifying data into different classes according to some constraints. Artificial Neural Network (ANN), Bayesian Networks (BN), Decision Tree (DT), Nearest Neighbor (NN), Support Vector Machine (SVM), Rough Sets, Fuzzy Logic, Genetic Algorithms are different classification techniques for discovering knowledge. The goal of classification is to accurately predict the target class for each case in the data. The major issue is preparing the data for Classification involves the Data cleaning, Relevance Analysis, Data Transformation and reduction, Normalization and Generalization activities.

In recent days, the study of heart disease is a challenging problem with ML approach. This work was studied from Framingham Heart Study (FHS). In 1948, the Framingham Heart Study was started during this period 939 subjects created CHD and 36337 kicked the bucket free of CHD. There was a step wise increment in mean hazard score with propelling age, in light of the fact that propelling age gives expanded hazard for CHD and on account of a more prominent weight of CHD chance elements with propelling age. The dataset has 85% of non heart attack and 15% Heart attack patients i.e., it is very imbalanced dataset has considered in our experimental study.

2. LITERATURE SURVEY ON HEART DISEASE ANALYSIS

Researcher & Year	Description
Merlet et al. 1992	Iodine-Metaiodobenzylguanidine (MIBG) has been utilized to think about cardiovascular adrener gic nerve action. This study was undertaken to examine the prognostic value of MIBG cardiac imaging of patients with heart failure in comparison with noninvasive markers.
Ordonez et al.	This paper chiefly clarified around two

2001	angles: mapping restorative information to an exchange organize reason able for mining affiliation leads and recognizing valuable imperatives. It demonstrates that the solid information on finding affiliation manages in therapeutic information to foresee coronary illness.
Turkoglu et al. 2002	Based on the pattern recognition an expert diagnosis system is presented for interpretation of the Doppler signals of the heart valve diseases. It manages the element extraction from estimated Doppler flag wave for matthe heart valve utilizing the Doppler Ultrasound.
Huang et al. 2007	In this paper the two preparing stages: learning to making stage and an information deducing stage are incorporated with CBR in a model of interminable maladies forecast and finding (CDPD) framework. In this they find the internal importance rules utilizing information mining strategies, the choice tree acceptance calculation and the case affiliation are embraced from well being examination information. The extricated decides that are put away in an administer base will be utilized for the particular incessant ailments forecast.
Pencina et al. 2009	The approach in this examination depends on cutting edge measurable strategies that permit staying away from predisposition in the appraisal of genuine outright hazard. Overlooking the contending danger of death blows up the appraisals by a normal of 1% to 2% on the total scale (or 10% on the relative scale), which prompts second rate alignment as exhibited in the statistics.
Xing et al. 2010	Arranging ECG time arrangement (the time arrangement of heart rates) clarified that the information recovered from a patient or from a solid individual. The estimations of a succession are gotten in time stamp rising request for Temporal emblematic arrangements and Time arrangement.
Tomar and Agarwal 2013	Classification rules are focused on class attributes and the Association rules are used to identify relationship between attributes. In Decision making the association rules plays a vital role. The Domain Experts consider the useful rules and omit the in consequences
Banaee et al. 2013	This article has uncovered patterns in the choice of the information handling strategies keeping in mind the end goal to screen well being parameters, for example, ECG, RR, HR, BP and BG. the audit laid out the more typical information mining assignments that have been connected, for example, irregularity recognition, expectation and basic leadership while considering specifically constant time arrangement estimations Banaee et al.
Song et al. 2014	ARM to give understanding on populace examples of conceivably preventable endless illness related antagonistic occasions. Framingham Heart Study associate information for displaying danger of CVD inside 10 long stretches of benchmark evaluation. The outcomes recognize a subset with extraordinarily raised CVD hazard, setting 13% of the cases into at least one of three hazard groups with more than over two times the likelihood of creating CVD when contrasted with an adjust of cases not related with any bunch Song et al.
Masethe and Masethe 2014	The analysts executed a mixture framework that utilization worldwide advancement advantage of hereditary calculation for in statement of neural system weights. The expectation of the coronary illness depends on chance factors, for example, age, family history, diabetes, hypertension, elevated cholesterol, smoking, liquor admission and heftiness. The prescient exactness controlled by J48, REPTREE and SIMPLE CART calculations recommends that parameters utilized are dependable pointers to anticipate the nearness of heart infections Masethe and Masethe.
Iskandar and Ujir 2015	In this paper, introduce a hypothetical structure to speak to the Cardiac MRI picture data bank in a cosmology which will expand existing medicinal ontologies. The fundamental objective is to investigate and examine the strategies on the most proficient method to connect the spatio-transient semantic hole issue in biomedical chart books utilizing semantic web innovation.
Hayashi and Yukita 2016	The convoluted illnesses like Heart malady, stroke, vision misfortune, kidney disappointment, and lower-appendage removals are instantly influenced with diabetes. Diabetics are at an expanded danger of cardiovascular

	disease. Good glucose control can help evade a few intricacies, especially smaller scale vascular eye, kidney, and nerve infection and early identification and treatment can help avert malady movement; consequently, checking that incorporates enlarged eye exams, pee tests, and foot exams is fundamental Hayashi and Yukita
Ali & Ghazal 2017	In this paper, a Real-time Heart Attack Mobile Detection Service (RHAMDS); an e-well being IoT benefit utilizing SDN controlled MECVANET engineering. RHAMDS intends to decrease and counter act vehicle impact through the recognition of heart assaults that drivers may experience the ill effects of. It exhibits the model of the administration empowered through SDN for IoT networks and its two varieties. They propose a voice controlled RHAMDS display and a signal controlled RHAMDS model. Both join sensors from the savvy; given its notoriety with clients and expanding accessibility. The principal variety of RHAMDS just considers that the client would utilize the administration in the vehicle, while the second variety helps the client even outside a vehicular system setting Ali and Ghazal.
Xiao and Fang 2017	In this examination, RF Miner, a hazard factor revelation and digging system for distinguishing critical hazard factors utilizing incorporated measures were proposed. In the showing of trial comes about distinguish cardiovascular sicknesses, for example, heart assaults. Particularly this structure predicts probability of heart assaults superbly. This system incorporate, a fell classifier to enhance the accuracy and review for the lopsided informational collection which beats the condition of-heart comes about; and furthermore locate a novel hazard factors by coordinating different intriguing quality measures Xiao and Fang
Nag et al. 2017	A capable approach is proposed in this paper can foresee the odds of heart assault when a man is bearing chest agony or proportional side effects. We have built up a model by coordinating clinical information gathered from patients conceded in various doctor's facilities assaulted by Acute Myocardial Infarction (AMI).

3. CLASSIFICATION ALGORITHMS

We utilized the following supervised algorithms which were implemented in caret R package.

- ✓ Feed-Forward NN
- ✓ SVM
- ✓ XG Boost
- ✓ Random Forest
- ✓ LDA

3.1. Feed-Forward NN

In this work, we have used Feed-Forward Neural Network, as it is very simple, effective and easy to understand when compared with other models. In FFNN, a set of random weights are initialized to pass the data from one layer to the other. The features are supplied to input nodes which in turn are connected to hidden layer nodes. The hidden layer represents the relation between the input and output layer. Each hidden node learns by least squares to fit the model. The output layer has an activation function where it can classify the inputs to Outputs. In general the activation function can be sigmoid function which produces a binary outcome.

3.2. Support Vector Machines (SVM)

Support Vector Machines (SVM) with linear or nonlinear kernels has become one of the most promising learning algorithms for classification. SVM (support vector machines) is a group of supervised learning techniques or methods, which is used to do for classification.SVM can be applied to both classification and regression

3.3. Extreme Gradient Boost (XGBoost)

Extreme Gradient Boosting (XGBoost) is a supervised classification algorithms and it is very popular in various data science competitions. The term "Gradient Boosting" come from greedy function approximation. It is similar to "gradient boosting" but more efficient. It supports various objective functions linear models, tree learning algorithms and ranking. The big prosperity and popularity of XGBoost is its scalability on a single machine by executing parallel computations which allow quicker model exploration.

3.4. Random Forest (RF)

Random forest is an ensemble method of Classification and Regression. It is a supervised learning algorithm. It constructs several decision trees on training examples and outputs the mean prediction of all class labels. It reduces variance error. The RF splits the training set randomly with replacement and fit the trees by averaging multiple decision trees or majority vote. The forest converges when the limit of trees in the forest becomes large.

3.5. Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is straight forward model to classify given dataset. LDA is closely related to principal component analysis (PCA) where both techniques examine for linear combination of variables of dataset. LDA attempts to separate the classes of data. PCA doesn't take into account of classes. It computes the statistical properties namely mean and the covariance matrix of every feature/variable. These statistical properties are supplied to LDA to make predictions. It works based on two intuitions. i). The independent variables follows Gaussian distribution. ii).Every variable has the same variance.

4. PERFORMANCE MEASURES

To evaluate the performance of a model, we use various metrics are computed from confusion matrix. Where TP - True Positive, FP - False Positive, TN - True Negative and FN - False Negative.

$$Accuracy=(TP+TN)/(TP+FP+TN+FN) \quad (1)$$

$$Recall=TP/(TP+FN) \quad (2)$$

$$Precision=TP/(TP+FP) \quad (3)$$

$$F1-Measure = 2 * (Precision * Recall)/(Precision + Recall) \quad (4)$$

5. FEATURE SELECTION ALGORITHMS

Feature selection is the process of reducing the number of features. The aim of removing those features from the learning algorithm which have low impact on the classification problem. The Primary motivation behind feature selection is that the training data contains many features which are irrelevant to the class problem and they won't give further information than the currently selected features.

5.1. Symmetrical Uncertainty

Symmetrical uncertainty is an entropy based filter; it measures the feature to class correlation using entropy. It computes the weights of every feature and provides a rank to every feature. If the weight of features is less than zero then those are ignored or irrelevant to classification task. Symmetrical uncertainty is balanced by normalization in the range [0-1] when bias occurs towards more weights.

5.2. Correlation-based Feature Selection

Correlation-based Feature Selection (CFS) algorithm is a simple filter method. Given a full set, it finds an optimal subset that contains features that are highly correlated with class label and uncorrelated with each other. The 'class label' field in the training set is the target value of that particular instance of the training set. CFS evaluates correlation of the feature subset on the basis of this hypothesis: "A good feature subset contains features highly correlated with (predictive of) the classification, yet uncorrelated with (not predictive of) each other".

5.3. Consistency-based Subset Evaluation

CSE evaluates the feature subsets and finds an optimal subset of relevant features that are consistent to each other. To determine the consistency of a subset, the combination of feature values representing a class are given a pattern label. All instances of a given pattern should thus represent the same class. A pattern is inconsistent if there exist at least two instances such that their patterns are same but they differ in their class labels.

Table 1: CHD dataset

Test Sample	Feature	Test Sample	Feature
P1	Gender	P9	Diabetes
P2	Age	P10	TotChol
P3	Education	P11	Glucose
P4	CurrentSmoker	P12	DiaBP
P5	CigsPerDay	P13	BMI
P6	BPMeds	P14	HeartRate
P7	PrevelentStroke	P15	SysBP
P8	PrevelentHyp	P16	TenYesCHD

6. ANALYSIS OF RESULTS

In this work the main objective of this work is to accurately detect those patients are having CHD disease. i.e. need to reduce the false negatives, false negatives and need to deal with class imbalance problem. All the experiments are implemented using "R-language" and executed on Intel i3 4-core machine with 4GB RAM PC.

6.1. Without feature selection:

The CHD dataset is normalized and k-fold cross-validation is performed on the data set, where k=10 in our work. We mainly focus on the evaluation of the Classifier with four metrics presented in Table 2.

Table 2: All Features

Algorithm	Precision	Recall	F1	Accuracy
NN	85.46	99.52	91.95	85.24
LDA	85.87	98.36	91.69	84.88
SVM	84.82	100	91.79	84.83
RF	84.96	99.65	91.72	84.75
XGBoost	85.99	96.19	90.81	83.49

The AUC obtained through XGBoost, NN, SVM, RF and LDA are 0.6547, 0.5053, 0.5056, 0.5024 and 0.509 respectively. XGB is 0.6547 which is higher than the AUC obtained through other methods with all features.

6.2. With feature selection

In this section we discuss the Symmetric Uncertainty (SU) method, Consistency based Subset Evalution (CSE) and Correlation based features Selection (CFS). It was observed that with the feature selection methods there is no much loss in accuracy. The detailed results with other metrics in

combination with three feature selection methods are presented in Table 5 and the Figure(1).

Table 5: Average accuracy of 10-fold cross-validation of filtering methods

Filter Methods	Classifier	Precision	Recall	F1	Accuracy
SU	NN	85.79	98.87	91.87	85.16
	LDA	85.67	98.32	91.56	84.64
	SVM	85.14	99.74	91.86	85.02
	RF	85.23	99.03	91.62	84.64
	XGB	86.19	96.23	90.93	83.73
CFS	NN	85.61	99.03	91.84	85.08
	LDA	85.67	98.32	91.56	84.64
	SVM	85.05	99.61	91.76	84.83
	RF	85.15	99.45	91.74	84.83
	XGB	85.96	96.39	90.88	83.60
CSE	NN	85.32	99.68	91.64	85.18
	LDA	85.57	98.65	91.64	84.75
	SVM	84.96	99.81	91.79	84.86
	RF	84.85	99.87	91.75	84.77
	XGB	85.82	96.42	90.81	83.46

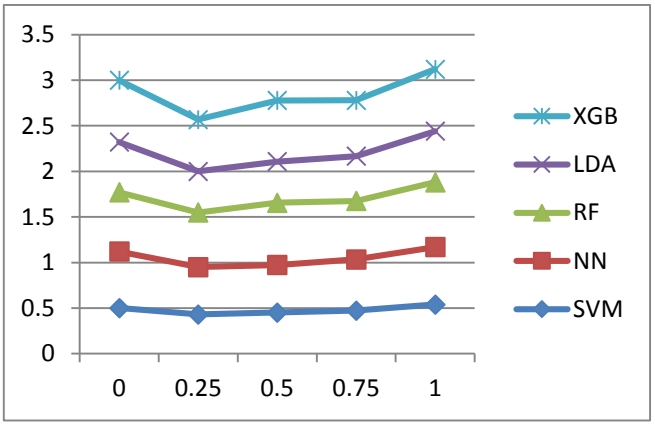

Figure (1): AUC values of different Classification Methods

7. CONCLUSION

In this paper we present detection of CHD using Classification prediction methods. Our work was carried out on Framingham heart dataset by explaining NN, LDA, SVM, RF and XGBoost (EXtreme Gradient Boosting). We obtained average accuracies of XGB in without feature selection. We also examined the importance of features using feature selection method for Classification. In this feature selection method also XGB performs very well compared to all other classification methods. Finally, we conclude that class imbalance is a critical problem in the medical field, which is demolished by Extreme Gradient Boosting Technique.

REFERENCES

[1]. Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M., Haussler, D., 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proceedings of the National Academy of Sciences 97 (1), 262–267.

[2]. Chakraborty, K., Mehrotra, K., Mohan, C. K., Ranka, S., 1992. Forecasting the behavior of multivariate time series using neural networks. Neural networks 5 (6), 961–970.

[3].Chaurasia, V., 2017. Early prediction of heart diseases using data mining techniques.

[4]. Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM, pp. 785–794.

[5]. Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., Lawler, J. J., 2007. Random forests for classification in ecology. Ecology 88 (11), 2783–2792.

[6]. Dash, M., Liu, H., 2003. Consistency-based search in feature selection. Artificial intelligence 151 (1), 155–176.

[7]. Enke, D., Thawornwong, S., 2005. The use of data mining and neural networks for forecasting stock market returns. Expert Systems with applications 29 (4), 927–940.

[8].Friedman, J. H., 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189–1232.

[9].Gagliardi, A. C., Miname, M. H., Santos, R. D., 2009. Uric acid: a marker of increased cardiovascular risk. Atherosclerosis 202 (1), 11–17.

[10].Gislason, P. O., Benediktsson, J. A., Sveinsson, J. R., 2006. Random forests for land cover classification. Pattern Recognition Letters 27 (4), 294–300.

[11]. Hall, M. A., 1999. Correlation-based feature selection for machine learning. Ph.D. thesis, The University of Waikato.

[12]. Hsu, C.-W., Chang, C.-C., Lin, C.-J., et al., 2003. A practical guide to support vector classification.

[13]. Hua, S., Sun, Z., 2001. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach1. Journal of molecular biology 308 (2), 397–407.

[14]. Huang, C.-L., Chen, M.-C., Wang, C.-J., 2007a. Credit scoring with a data mining approach based on support vector machines. Expert systems with applications 33 (4), 847–856.

[15]. Huang, M.-J., Chen, M.-Y., Lee, S.-C., 2007b. Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis. Expert systems with applications 32 (3), 856–867.

[16]. Iskandar, D. A., Ujir, H., 2015. Spatio-temporal semantic representation of cardiac mri in heart attack patients. In: IT in Asia (CITA), 2015 9th International Conference on. IEEE, pp. 1–5.

[17]. James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An introduction to statistical learning. Vol. 112. Springer.

[18]. Kyatam, S., Alhayajneh, A., Hayajneh, T., 2017. Heartbleed attacks implementation and vulnerability. In: Systems, Applications and Technology Conference (LISAT), 2017 IEEE Long Island. IEEE, pp. 1–6.

[19]. Lewoniewski, W., We˛cel, K., Abramowicz, W., 2016. Quality and importance of wikipedia articles in diff erent languages. In: International Conference on Information and Software Technologies. Springer, pp. 613–624.
[20]. Maglogiannis, I., Loukis, E., Zafiropoulos, E., Stasis, A., 2009. Support vectors machine-based identification of heart valve diseases using heart sounds. Computer methods and programs in biomedicine 95 (1), 47–61.
[21]. Masethe, H. D., Masethe, M. A., 2014. Prediction of heart disease using classification algorithms. In: Proceedings of the world congress on Engineering and Computer Science. Vol. 2. pp. 22–24.
[22]. Merlet, P., Valette, H., Dubois-Randé, J.-L., Moyse, D., Duboc, D., Dove, P., Bourguignon, M. H., Benvenuti, C., Duval, A. M., Agostini, D., et al., 1992. Prognostic value of cardiac metaiodobenzylguanidine imaging in patients with heart failure. Journal of Nuclear Medicine 33 (4), 471–477.
[23]. Nag, P., Mondal, S., Ahmed, F., More, A., Raihan, M., 2017. A simple acute myocardial infarction (heart attack) prediction system using clinical data and data mining techniques. In: Computer and Information Technology (ICCIT), 2017 20th International Conference of. IEEE.pp. 1–6.
[24]. Nielsen, D., 2016. Tree boosting with xgboost-why does xgboost win" every" machine learning competition? Master's thesis, NTNU.
[25]. Parascandolo, G., Heittola, T., Huttunen, H., Virtanen, T., et al., 2017. Convolutional recurrent neural networks for polyphonic sound event detection. IEEE/ACM Transactions on Audio, Speech, and Language Processing 25 (6), 1291–1303.

BIOGRAPHIES

K.Venkata Nagendra, working as Assistant Professor in the Department of Computer Science Engineering at Geethanjali Institute of Science and Technology, Nellore, Andhra Pradesh, India. He has 8 years of experience in the field of teaching. He is a research scholar in Vikrama Simhapuri University. He did his M.Tech in ANU, Guntur. His areas of interests are Data warehousing and Data Mining and Cloud Computing.

Dr.Maligela Ussenaiah, working as Assistant Professor in the Department of Computer Science in Vikrama Simhapuri University, Nellore, Andhra Pradesh, India. He is having 9 years of teaching experience. He did his PhD in Computer Science from SriKrishna Devaraya Univeristy, Ananthapur, Andhra Pradesh. His areas of interests are Networks, Mobile Wireless Networks, Data warehousing and Data Mining and Image processing.