Prediction of Heart Disease using Ensemble Learning Methods

[1]Lakshmi T N,[2]D R Nagamani

[1]Student, M.Tech, [2]Assistant Professor
[1]Department of CSE,
[1]Bangalore Institute of Technology,Bangalore, India

Abstract:Heart disease is one of the major cause of death that occurs across the world. Data involved regarding heart disease can be used to diagnose the type of disease so that can assist the specialists in order to identify disease during an early stage. Many developed software is to provide assistance for doctors in making decisions of heart disease. This can be used to assist the specialist in predicting the status of Heart disease presence of the clinical data of various patients. In this paper, the Heart disease prediction using classification algorithms such as Decision trees, Random Forest, Gradient descent and AdaBoosting is used. The main idea of this project is to extract only the patterns present in it by applying algorithms to heart diseases and to predict the heart disease patients where this presence is valued from no presence to likely presence. The experimental results demonstrate that this approach improves classification accuracy, and the presented model can help health care professional in predicting heart disease.

***IndexTerms*-Heart disease, Classification Techniques, Feature Selection, Decision Tree, Random Forest, Gradient descent, Adabooting.**

I. INTRODUCTION

Healthcare industry today generates large amounts of complex data about patients, hospital's resources, disease diagnosis, electronic patient records, medical devices, etc [1]. The large amounts of data are a key resource to be processed and analyzed for knowledge extraction that enables support for cost-savings and decision making. Heart disease is also called coronary heart disease (CHD) where it is a condition that affects the heart. This is a leading cause of death worldwide. Physicians generally make decisions by evaluating current test results of the patients. Some of the decisions are taken previously for other patients with the same conditions are also examined. So diagnosing heart disease requires experience and highly skilled physicians. Heart disease diagnosis is important to diagnose in the early stage yet it is a complicated task. Today many hospitals collect patient data to managing the health care of patients. This information is in different format like numbers, charts, text, and images. But this database contains rich information but poorly used for clinical decision making.

Data mining is one of the explorations for large datasets to extract some of the hidden and previously unknown patterns, relationships, and knowledge that are difficult to detect with traditional statistical methods. The implementation of work is done on Cleveland heart diseases dataset from the University of California Irvine (UCI) machine learning repository to test on different data mining techniques. This automatically infers some of the diagnostic rules and help specialists to make a diagnosis. The main objective is to find out the suitable machine learning technique that is computationally efficient as well as accurate for the prediction of heart disease.

II. LITERATURE SURVEY

The number of systems for the prediction of different diseases are proposed and implemented by using different techniques and methods. George et al. have proposed a decision support system for dementia patients using support vector machines to define and detect agitation transition. In this system, two new SVM architectures are presented, which were applied to the detection of agitation and agitation transition. This approach gives the accuracy of 91.4%, which is higher as compared with 90.9% for the traditional SVM [3]. A system for diagnosis of heart disease that is based on support vector machine along with a sequential minimal optimization algorithm is presented. In this system, the network structure of Radial Basis Function is also used and it is trained using Orthogonal Least Square algorithm and applied to the dataset based on Indian patients. The result shows that the SVM is equivalently as good as compared to Radial bias function in the detection of heart disease with an accuracy of 86.42% [4]. Ankita Dewan and Meghna Sharma have discussed various kinds of techniques for developing a heart disease prediction system and proposed using Backpropagation Algorithm as the best classification technique for the targeted system. They also have proposed using Genetic Algorithm as optimizer against the Backpropagation Algorithm drawback of being stuck in local minima. The proposed methodology was intended for implementing in future with an accuracy of nearly 100% or with minimal error. [2]

Panahiazar et al. applied decision trees, Random Forests, Adaboost, SVM and logistic regression to a dataset extracted from the EHR of the Mayo Clinic. The dataset included 5044 HF patients admitted to the Mayo Clinic from 1993 to 2013. For each patient, 43 predictor variables, expressing demographic data, vital measurements, lab results, medication, and co-morbidities, were recorded. The class variable corresponded to mortality status, consequently, three versions of the dataset were created, each one corresponding to survival period (1-year, 2-year, 5-year). 1560 instances out of 5044 were used for training and the rest 3484 instances for testing. The authors observed that logistic regression and Random Forests were more accurate models compared to others, also, among the scenarios, the best prediction accuracy was 87.12%.[5]

III. METHODOLOGY

In the heart disease prediction system, there are input variables, which are disease risk factors which are obtained from the dataset,
and the output variables categorized as “disease absence” and “disease presence”. Prediction of heart disease is called supervised learning problem. Because of having output variables are in category type, the prediction heart disease is “classification type of supervised learning”.

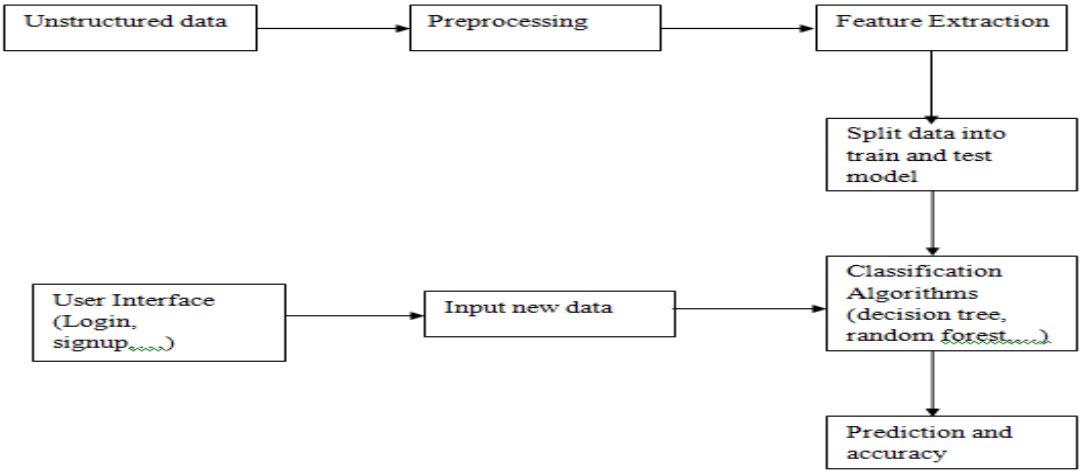

Figure 1: Architecture Diagram

The heart disease datasets are collected from source as given in [10]: The UCI machine learning is the most widely used repository which contains different datasets from various locations. These data sets are used for data mining and machine learning purposes. As for the heart disease prediction, data is collected from Cleveland, Switzerland, Long Beach VA, and Hungary. The numbers collected in csv represents the values of attributes which are an indication to either presence or absence of heart disease in the patient through another attribute called class. The range of this attribute is from 0 (no presence) to 4. Most of the experiments associated with Cleveland database are focused on absence ("class" value 0) and presence ("class" values from 1 to 4). For experimentation, considered only 2 classes for prediction, that is 0 being absent and 1 being present. Due to personal security, the personal identification information of the patients is replaced with dummy values. The directory in repository contains a dataset related to heart disease diagnosis. The database contains a total 76 raw attributes, but in these experiments, only 14 attributes of them are used. The dataset used in this experiment contains different important parameters like ECR, cholesterol, chest pain, fasting sugar, MHR (maximum heart rate) and many more. The detailed information about these attributes and their domain range are as follows: [11]

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	0 = female 1 = male
Cp	Discrete	Chest pain type: 1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 =asymptom
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar>120 mg/dl: 1true 0=False
Exang	continuous	Exercise induced angina (1 = yes; 0 = no)
Thalach	Continuous	Maximum heart rate achieved
Old peak ST	Continuous	Depression induced by exercise relative to rest
Slope	Discrete	The slope of the peak exercise segment : 1 = up sloping 2 = flat 3 = down sloping
Ca	Continuous	Number of major vessels colored by fluoroscopy that ranged between 0 and3
Thal	Discrete	3 = normal 6 = fixed defect 7= reversible defect
Class	Discrete	Diagnosis classes: 0 = No Presence 1=Least likely to have heart disease 2= >1 3= >2 4=More likely have heart disease

Table 1: Attribute Information

The different kind of data mining classification techniques, i.e. Decision Trees, Random forest, Gradient descent and Ada boosting are used to analyze the dataset in the healthcare industry.

Decision Tree

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from the root to leaf represent classification rules. In decision analysis, a decision tree and the closely related influence diagramare used as a visual and analytical decision support tool, where the expected values(or expected utility) of competing alternatives are calculated. Decision trees are commonly used in operations researchand operations management. If in practice, decisions have to be taken online with no recall under incomplete knowledge, a decision tree should be paralleled by a probabilitymodel as the best choice model or online selection model algorithm. Another use of decision trees is as a descriptive means for calculating conditional probabilities.

Random Forest

Random Forest is predominantly an ensemble of unpruned classification trees. It provides remarkable performance on a number of practical problems, such as health care prediction problems as it is not sensitive to noise in the data set, and it is not subjected to over fitting. It is built by combining the predictions of several trees, each of them is trained separately. It works fast and usually exhibits a significant performance improvement over many other tree-based algorithms such as decision tree.
There are three main choices to be made when constructing a random tree:
1. The method for splitting the leaves.
2. The type of predictor to use in each leaf.
3. The method for injecting randomness into the trees.

Ada Boosting Classifier

Ada Boosting classifier combines weak classifier algorithm to form strong classifier. A single algorithm may classify the objects poorly. But if it combines multiple classifiers with the selection of training set at every iteration and assigning right amount of weight in the final voting, it results in good accuracy score for an overall classifier. After training a classifier at any level, ada-boost assigns a weight to each training item. Misclassified item is assigned higher weight so that it appears in the training subset of next classifier with higher probability. After each classifier is trained, the weight is assigned to the classifier as well based on accuracy. The more accurate classifier is assigned higher weight so that it will have more impact in the final outcome.

Gradient Descent Boosting

The idea of boosting came out of the idea of whether a weak learner can be modified to become better. The statistical framework cast boosting as a numerical optimization problem where the objective is to minimize the loss of the model by adding weak learners using a gradient descent like procedure. Gradient boosting involves three elements:

A loss function to be optimized., a weak learner to make predictions, and an additive model to add weak learners to minimize the loss function. The loss function used depends on the type of problem being solved. It must be differentiable, but many standard loss functions are supported, and it can define its own. Decision trees are used as the weak learner in gradient boosting. Specifically, regression trees are used that output real values for splits and whose output can be added together, allowing subsequent models outputs to be added and "correct" the residuals in the predictions. Trees are constructed in a greedy manner, choosing the best split points based on purity scores like Gini or to minimize the loss. Trees are added one at a time, and existing trees in the model are not changed. A gradient descent procedure is used to minimize the loss when adding trees. Traditionally, gradient descent is used to minimize a set of parameters, such as the coefficients in a regression equation or weights in a neural network. After calculating error or loss, the weights are updated to minimize that error.

IV. EXPERIMENTAL RESULTS

In this section, the results of different classification models are presented that used in this approach. Accuracy, precision, and recall have used as performance measures. All these measures are calculated by using confusion matrix because it contains all TP, TN, FP and FN assessments. Results are obtained by different classifiers on heart disease patient dataset are given in Table 1. It uses all ensemble learning classification models including Decision Tree, Random Forest, Gradient descent and AdaBoosting. It is clear from figure 2 that Gradient Descent has the highest accuracy 84.669811% respectively than any other classification model for the dataset of heart disease patients.

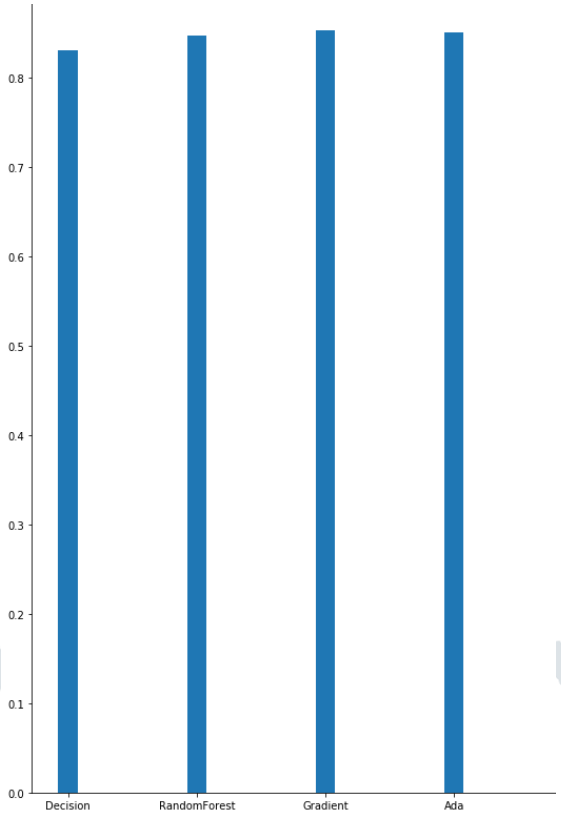

Figure 2 shows the experimental results of Algorithms

The resulted performance score for the heart disease dataset are represented as follows by confusion Matrix for the classification algorithm to predict the data values for the presence of heart disease

Table 1 Confusion matrix values predicted

predicted	false	true	__all__
false	302	5	307
true	63	2	65
__all__	365	7	372

The dataset is distinguished among classes are shown below for Class Distribution:

Table 2 class distribution of datavalues

database:	0	1	2	3	4	total
cleveland:	164	55	36	35	13	303
hungarian:	188	37	26	28	15	294
switzerland:	8	48	32	30	5	123
long beach va:	51	56	41	42	10	200

V.CONCLUSION

As the healthcare data being generated from time to time in the medical field, it can be processed faster for predicting diseases with none overhead. Here, a scalable solution is proposed for predicting heart disease attributes and validated its accuracy. It is implemented with ensemble learning algorithms for predicting heart disease, and it is shown that with as small as dataset records for Gradient descent algorithm to achieve 84.669811% accuracy. The proposed ensemble model is further aimed to magnify and augment in the future by assimilating more machine learning methods and additional medical attributes.

REFERENCES

[1] B.L DeekshatuluaPriti Chandra "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm ", M.Akhiljabbar International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013

[2] A. Dewan and M. Sharma, "Prediction of heart disease using a hybrid technique in data mining classification," 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 704-706, 2015

[3] G. E. Sakr, I. Elhajj, and H. Huijer, "Support vector machines to define and detect agitation transition", IEEE Transactions On Affective Computing, vol. 1, pp. 98–108, December 2010

[4] S. Ghumbre, C. Patil, and A. Ghatol, "Heart disease diagnosis using support vector machine" ,In proceedings of International Conference on Computer Science and Information Technology (ICCSIT), pp. 84–88, December 2011

[5] M. Panahiazar, V. Taslimitehrani, N. Pereira and J. Pathak, "Using EHRs and machine learning for heart failure survival analysis." Stud HealthTechnolInform,vol.216,pp.40–44,2015

[6] Purushottam, Prof. (Dr.) Kanak Saxena, Richa Sharma "Efficient Heart Disease Prediction System using Decision Tree" International Conference on Computing, Communication and Automation, ICCCA, 2015

[7] W. Ouwerkerk, A.V. Adriaan and A.H. Zwinderman, "Factors influencing the predictive power of models for predicting mortality and/or heart failure hospitalization in patients with heart failure," JACC: Heart Fail, vol. 2, pp. 429-436, April 15, 2014

[8] Y. Xing, J. Wang, Z. Zhao, and Y.Gao, "Combination data mining methods with new medical data to predicting outcome of coronary heart disease," Proc. IEEE, Convergence Information Technology, pp. 868-872, 2007

[9] V. Taslimitehrani, and G. Dong, "A new CPXR based logistic regression method and clinical prognostic modeling results using the method on traumatic brain injury," in: Proc. of IEEE International Conference on Bioinformatics and Bioengineering, pp. 283–290, November 2014

[10] https://archive.ics.uci.edu/ml/datasets/Heart+Disease.

[11] V. Chaurasia and S. Pal, "Early prediction of heart diseases using data mining techniques" Caribbean Journal of Science and Technology, vol.1, 208-217, 2013

[12] A.K. Pandey, P. Pandey, K.L. Jaiswal, and A.K. Sen, "A heart disease prediction model using decision tree", IOSR Journal of Computer Engineering (IOSR-JCE), vol. 12, Issue 6, PP 83-86, Aug. 2013

[13] M. Shouman, T. Turner and R. Stocker, "Integrating decision tree and kmeans clustering with different initial centroid selection methods in the diagnosis of heart disease patients", Proceedings of the International Conference on Data Mining, 2012

[14] Aljaaf, A.J., Al-Jumeily, D., Hussain, A.J., Dawson, T., Fergus, P. and Al-Jumaily, M., 2015, April. Predicting the likelihood of heart failure with a multi level risk assessment using decision tree. In Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE), 2015 Third International Conference on (pp. 101-106). IEEE

[15] Bourjeily C., Badr G., Hajjam Al Hassani A., Andres E., Heart Failure Prediction with CPT+, In 9th International Conference on e-health,