Heart Disease Prediction using Machine Learning Algorithms

Devara Sandhya[1], Dr. Kamalraj R[2]

[1]PG STUDENT, Department of Computer Application, JAIN(Deemed-To-Be) University Banglore, Karnataka, India
[2]Assistant Professor, Department of CS and IT, JAIN(Deemed-To-Be) University, Karnataka, India

---

Abstract -In This Project Cardiovascular diseases are the most common cause of death worldwide over the last few decades in the developed as well as underdeveloped and developing countries. Early detection of cardiac diseases and continuous supervision of clinicians can reduce the mortality rate. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time and expertise. In this project, we have developed and researched about models for heart disease prediction through the various heart attributes of patient and detect impending heart disease using Machine learning techniques like backward elimination algorithm, logistic regression and REFCV on the dataset available publicly in Kaggle Website, further evaluating the results using confusion matrix and cross validation. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine.

Keywords: Machine Learning, Logistic regression, Cross-Validation, Backward Elimination, REFCV, Cardiovascular Diseases.

1 Introduction

According to the World Health Organization, every year 12 million deaths occur worldwide due to Heart Disease. The load of cardiovascular disease is rapidly increasing all over the world from the past few years. Many researches have been conducted in attempt to pinpoint the mostinfluential factors of heart disease as well as accurately predict the overall risk. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications. This project aims to predict future Heart Disease by analyzing data of patients which classifies whether they have heart disease or not using machine-learning algorithms.

2 Problem Statement

The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but either they are expensive or are not efficient to calculate chance of heart disease in human. Early detection of cardiac diseases can decrease the mortality rate and overall complications. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time and expertise. Since we have a good amount of data in today's world, we can use various machine learning algorithms to analyze the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data.

3 Proposed Solution

section depicts the overview of the proposed system and illustrates all of the components, techniques and tools are used for developing the entire system. To develop an intelligent and user-friendly heart disease prediction system, an efficient software tool is needed in order to train huge datasets and compare multiple machine learning algorithms. After choosing the robust algorithm with best accuracy and performance measures, it will be implemented on the development of the smart phone-based application for detecting and predicting heart disease risk level. Hardware components like Arduino/Raspberry Pi, different biomedical sensors, display monitor, buzzer etc. are needed to build the continuous patient monitoring system.

The below figure shows the process flow diagram or proposed work. First we collected the Cleveland Heart Disease Database from UCI website then pre-processed the dataset and select 16 important features.

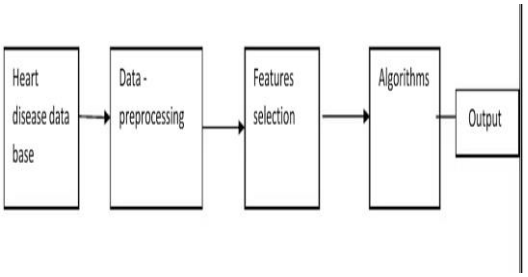

Fig : System Architecture

For feature selection we used Recursive feature Elimination Algorithm using Chi2 methodand get

16 top features. After that applied ANN and Logistic algorithm individually and compute the accuracy. Finally, we used proposed Ensemble Voting method and compute best method for diagnosis of heart disease.

Data Pre-processing:

This file contains all the pre-processing functions needed to process all input documents andtexts. First we read the train, test and validation data files then performed some preprocessing like tokenizing, stemming etc. There are some exploratory data analysis is performed like response variable distribution and data quality checks like null or missing values etc.

Feature:

Extraction In this file we have performed feature extraction and selection methods from sci-kit learn python libraries. For feature selection, we have used methods like simple bag-of-words and n-grams and then term frequency like tf-tdf weighting. We have also used word2vec and POS tagging to extract the features, though POS tagging and word2vec has not been used at this point in the project.

Classification:

Here we have built all the classifiers for the breast cancer diseases detection. The extracted features are fed into different classifiers. We have used Naive-bayes, Logistic Regression, Linear SVM, Stochastic gradient decent and Random forest classifiers from sklearn. Each of the extracted features was used in all of the classifiers. Once fitting the model, we compared the f1 score and checked the confusion matrix. After fitting all the classifiers, 2 best performing models were selected as candidatemodels for heart diseases classification. We have performed parameter tuning byimplementing GridSearchCV methods on these candidate models and chosen best performing parameters for these classifier. Finally selected model was used for heart disease detection with the probability oftruth. In Addition to this, we have also extracted the top 50 features from our term-frequencytfidf Vectorizer to see what words are most and important in each of the classes. We have also used Precision-Recall and learning curves to see how training and testset performs when we increase the amount of data in our classifiers.

Prediction:

Our finally selected and best performing classifier was algorithm which was then saved on disk with name final_model.sav. Once you close this repository, this model will bec opied to user's machine and will be used by prediction.py file to classify the Heart diseases. It takes a news article as input from user then model is used for final classification output that is shown to user along with probability of truth.

4 Literature Study

Machine Learning techniques are used to analyze and predict the medical data information resources. Diagnosis of heart disease is a significant and tedious task in medicine. The term Heart disease encompasses the various diseases that affect the heart. The exposure of heart disease from various factors or symptom is an issue which is not complimentary from false presumptions often accompanied by unpredictable effects. The data classification is based on Supervised Machine Learning algorithm which results in better accuracy. Here we are using the Random Forest as the training algorithm to train the heart disease dataset and to predict the heart disease. The results showed that the medicinal prescription and designed prediction system is capable of prophesying the heart attack successfully.

Machine Learning techniques are used to indicate the early mortality by analyzing the heart disease patients and their clinical records .have brought about the two Machine Learning techniques, k-nearest neighbor model and existing multi line arregression to predict the stroke severity index of the patients. Their study show that k-nearest neighbor performed better than Multi Linear Regression model. Have suggested various Machine Learning techniques such as support vector machine, penalized logistic regression to predict the heart stroke. Their results show that Support vector machine produced the best performance in prediction when compared to othermodels. Boshra Brahmi et al, developed different Machine Learning techniques to evaluate the prediction and diagnosis of heart disease. The main objective is to evaluate the different classification techniques, Decision Tree, KNN and Naïve Bayes. After this, evaluating some performance in measures of accuracy, precision, sensitivity, specificity are evaluated.

5 Future Scope

As illustrated before the system can be used as a clinical assistant for any clinicians. The disease prediction through the risk factors can be hosted online and hence any internet users can access the system through a web browser and understand the risk of heart disease. The proposed model can be implemented for any real time application .Using the proposed model other type of heart disease also can be determined. Different heart diseases asrheumatic heart disease, hypertensive heart disease, ischemic heart disease,

cardiovasculardisease and inflammatory heart disease can be identified. Other health care systems can be formulated using this proposed model in order to identify the diseases in the early stage. The proposed model requires an efficient processor with good memory configuration to implement it in real time. The proposed model has wide area of application like grid computing, cloud computing, robotic modeling, etc. To increase the performance of our classifier in future, we will work on ensembling two algorithms called Random Forest and Adaboost. By ensembling these two algorithms we will achieve high performance.

6 Conclusion

This project, we introduce about the heart disease prediction system with different classifier techniques for the prediction of heart disease. The techniques are Random Forest and Logistic Regression: we have analyzed that the Random Forest has better accuracy as compared to Logistic Regression. Our purpose is to improve the performance of the Random Forest by removing unnecessary and irrelevant attributes from the dataset and only picking those that are most informative for the classification task.

References

[1] P.K. Anooj, --Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules‖; Journal of King Saud University – Computer and Information Sciences (2012) 24, 27 – 40. Computer Science & Information Technology (CS & IT) 59

[2] Nidhi Bhatla, Kiran Jyoti"An Analysis of Heart Disease Prediction using Different DataMining Techniques". International Journal of Engineering Research & Technology

[3]Jyoti Soni Ujma Ansari Dipesh Sharma, Sunita Soni. "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction".

[4] Chaitrali S. Dangare Sulabha S. Apte, Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques" International Journal of Computer Applications (0975 – 888)

[5] Dane Bertram, Amy Voida, Saul Greenberg, Robert Walker, "Communication, Collaboration, and Bugs: The Social Nature of Issue Tracking in Small, Collocated Teams".

[6]M.Anbarasi, E. Anupriya, N.Ch.S.N.Iyengar, ―Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm‖;

International Journal of EngineeringScience and Technology, Vol. 2(10), 2010.

[7] Ankita Dewan, Meghna Sharma," Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification", 2nd International Conference on Computing for Sustainable Global Development IEEE 2015 pp 704-706. [2].

[8] R. Alizadehsani, J. Habibi, B. Bahadorian, H. Mashayekhi, A. Ghandeharioun, R.Boghrati, et al., "Diagnosis of coronary arteries stenosis using data mining," J Med SignalsSens, vol. 2, pp. 153-9, Jul 2012.

[9] M Akhil Jabbar, BL Deekshatulu, Priti Chandra," Heart disease classification using nearest neighbor classifier with feature subset selection", Anale. Seria Informatica, 11,2013

[10]Shadab Adam Pattekari and Asma Parveen," PREDICTION SYSTEM FOR HEARTDISEASE USING NAIVE BAYES", International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624, Vol 3, Issue 3, 2012, pp 290-294.

[11] C. Kalaiselvi, PhD, "Diagnosis of Heart Disease Using K -Nearest Neighbor Algorithmof Data Mining", IEEE, 2016

[12] Keerthana T. K., "Heart Disease Prediction System using Data Mining Method",International Journal of Engineering Trends and Technology", May 2017.

[13] Data Mining Concepts and Techniques, Jiawei Han and Micheline Kamber, ELSEVIER. Animesh Hazra, Arkomita Mukherjee, Amit Gupta, Prediction Using Machine Learning and Data Mining July 2017, pp.2137-2159.