

A ESSENTIAL CARDIAC DISEASE DETECTION PREDICTION USING ENSEMBLE MACHINE LEARNING

M.M.BALAKRISHNA¹, T.JOICE SWAPNA²

¹Asst.Professor [M.Tech] , Department of CSE, St.Mary's Group of Institution, Guntur, Ap.
balakrishna2508@gmail.com

²PG Scholar, Dept of CSE, St.Mary's Group of Institution, Guntur, Ap, India.
thadimallajoiceswapna@gmail.com

ABSTRACT: Machine Learning is used across many ranges around the world. The healthcare industry is not an exception. Machine Learning can play an essential role in predicting presence/absence of loco motors' disorders, heart diseases and more. Such information, if predicted well in advance, can provide important intuitions to doctors who can then adapt their diagnosis as per patients need. We work on predicting possible heart diseases in people using Machine Learning algorithms. In this project we perform the comparative analysis of classifiers like Decision Tree, SVM, KNN. Based on the analysis, we propose an Ensemble classifier which performs hybrid classification by taking strong and weak classifiers. Since it can have multiple samples for training and validating the data, we perform the hybrid analysis of both classifiers Neural Networks and XG-boost that can produce better accuracy and predictive analysis.

1. INTRODUCTION

Heart is an important organ of the human body. It pumps blood to every part of our anatomy. If it fails to function correctly, then the brain and various other organs will stop working, and within few minutes, the person will die. Change in lifestyle, work related stress and bad food habits contribute to the increase in rate of several heart related diseases. Heart diseases have emerged as one of the most prominent cause of death all around the world. According to World Health Organization, heart related diseases are responsible for the taking 17.7 million lives every year, 31% of all global deaths. In India too, heart related diseases have become the leading cause of mortality. Heart diseases have killed 1.7 million Indians in 2016, according to the 2016 Global Burden of Disease Report, released on September 15, 2017. Heart

related diseases increase the spending on health care and also reduce the productivity of an individual. Estimates made by the World Health Organization (WHO), suggest that India have lost up to \$237 billion, from 2005-2015, due to heart related or Cardio Vascular diseases.

Thus, feasible and accurate prediction of heart diseases is very important. Medical organizations, all around the world, collect data on various health related issues. These data can be exploited using various Machine Learning techniques to gain useful insights. But, the data collected is very massive and many a times, this data can be very noisy. These datasets, which are to over whelming for human minds to comprehend can be easily explored using various machine learning techniques. Thus, these algorithms have become very useful, in recent times, to predict the presence or

absence of heart related diseases accurately.

2. LITERATURE SURVEY

“Prediction and Analysis of the occurrence of Heart Disease Using Data Mining Techniques”

The main objective is to predict the occurrence of heart disease for early automatic diagnosis of the disease within result in a short time. The proposed methodology is also critical in a healthcare organization with experts that have no more knowledge and skill. It uses different medical attributes such as blood sugar and heart rate, age, sex are some of the attributes are included to identify if the person has heart disease or not. Analyses of the dataset are computed using WEKA software.

“Implemented Hybrid Machine Learning for Heart Disease Prediction”

The data set used is Cleveland data set. The first step is data pre-processing step. In this the tuples are removed from the data set which has missed the values. Attributes age and sex from data set are also not used as the authors think that it's personal information and has no impact on predication. The remaining 11 attributes are considered important as they contain vital clinical records. They have proposed their own Hybrid Random Forest Linear Method (HRFLM) which is the combination of Random Forest (RF) and linear method (LM). In the HRFLM algorithm, the authors have used four algorithms. First algorithm deals with partitioning the input dataset. It is based on a decision tree which is executed for each sample of the dataset. After identifying the feature space, the dataset is split into the leaf nodes. Output of first algorithm is Partition of dataset. After that in second algorithm they apply rules to the data set

and output here is the classification of data with those rules.

This algorithm deals with finding the minimum and maximum error rate from the classifier. Output of this algorithm is the features with classified attributes. In forth algorithm apply Classifier which is hybrid method based on the error rate on the Extracted Features. Finally, they have compared the results obtained after applying HRFLM with other classification algorithms such a decision tree and support vector machine. In result as RF and LM are giving better results than other, both the algorithms are put together and new unique algorithm HRFLM is created. The authors suggest further improvement in accuracy by using combination of various machine learning algorithms.

3 EXISTING SYSTEM

Heart disease is even being highlighted as a silent killer which leads to the death of a person without obvious symptoms. The nature of the disease is the cause of growing anxiety about the disease & its consequences. Hence continued efforts are being made to predict the possibility of this deadly disease in prior. So various tools & techniques are regularly being experimented with to suit the present-day health needs. Machine Learning techniques can be a boon in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting the data from various sources, classifying them under suitable headings & finally analysing to extract the desired data we can conclude. This technique can be very well adapted to the do the prediction of heart disease. As the well-known quote says, “Prevention is better than cure”, early

Prediction & its control can be helpful to prevent & decrease the death rates due to heart disease.

4. PROPOSED SYSTEM

The working of the system starts with the collection of data and selecting the important attributes. Then the required data is pre-processed into the required format. The data is then divided into two parts training and testing data. The algorithms are applied and the model is trained using the training data. The accuracy of the system is obtained by testing the system using the testing data.

Methodology

Dataset collection is collecting data which contains patient details. Attributes selection process selects the useful attributes for the prediction of heart disease. After identifying the available data resources, they are further selected, cleaned, made into the desired form. Different classification techniques as stated will be applied on preprocessed data to predict the accuracy of heart disease. Accuracy measure compares the accuracy of different classifiers.

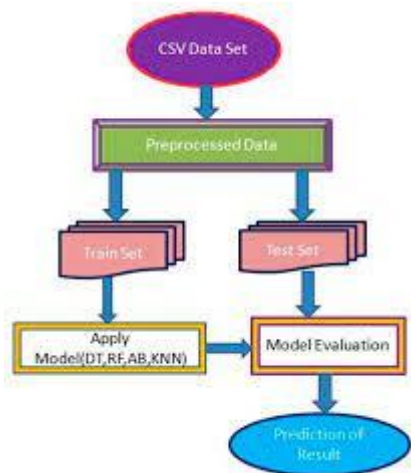


Fig: System Architecture

5. IMPLEMENTATION

1. DATA COLLECTION :

Initially, we collected a dataset for our heart disease prediction system. After the

collection of the dataset, we split the dataset into training data and testing data. The training dataset is used for prediction model learning and testing data is used for evaluating the prediction model. For this project, 70% of training data is used and 30% of data is used for testing. The dataset used for this project is heart.csv from Kaggle. The dataset consists of 76 attributes; out of which, 14 attributes are used for the system.

2. SELECTION OF ATTRIBUTES:

Attribute or Feature selection includes the selection of appropriate attributes for the prediction system. This is used to increase the efficiency of the system. Various attributes of the patient like gender, chest pain type, fasting blood pressure, serum cholesterol, exang, etc. Are selected for the prediction. The Correlation matrix is used for attribute selection for this model

3. DATA PRE-PROCESSING

Data pre-processing is an important step for the creation of a machine learning model.

Initially, data may not be clean or in the required format for the model which can cause misleading outcomes. In pre-processing of data, we transform data into our required format. It is used to deal with noises, duplicates, and missing values of the dataset.

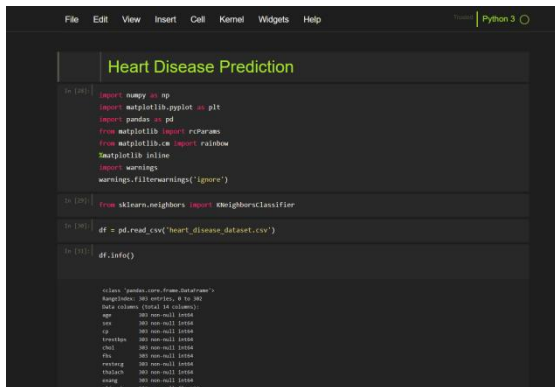
Data pre-processing has the activities like importing datasets, splitting datasets, attribute scaling, etc. Preprocessing of data is required for improving the accuracy of the model

4. PREDICTION OF DISEASE:

Now we should train the model on the training dataset and test dataset. Thus, it is chosen ML Models. Various machine learning algorithms like SVM, Naive Bayes, Decision Tree, Random Tree,

Logistic Regression, K-NN, Xg-boost are used for classification. Comparative analysis is performed among algorithms and the algorithm that gives the highest accuracy is used for heart disease prediction.

6.RESULT



```

File Edit View Insert Cell Kernel Widgets Help Python 3
Heart Disease Prediction
In [10]: import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from matplotlib import rcParams
from matplotlib.cm import rainbow
import matplotlib
import warnings
warnings.filterwarnings('ignore')

In [11]: from sklearn.neighbors import KNeighborsClassifier

In [12]: df = pd.read_csv('heart_disease_dataset.csv')

In [13]: df.info()

Out[13]: <class 'pandas.core.frame.DataFrame'>
Int64Index: 300 entries, 0 to 299
Data columns (total 14 columns):
age          300 non-null int64
sex         300 non-null int64
trestrest   300 non-null int64
trestrest   300 non-null int64
trestrest   300 non-null int64
trestrest   300 non-null int64
trestrest   300 non-null int64
trestrest   300 non-null int64
trestrest   300 non-null int64
trestrest   300 non-null int64
trestrest   300 non-null int64
trestrest   300 non-null int64
trestrest   300 non-null int64
trestrest   300 non-null int64
trestrest   300 non-null int64
  
```

7. CONCLUSION:

Heart diseases are a major killer in India and throughout the world, application of promising technology like machine learning to the initial prediction of heart diseases will have a profound impact on society. The early prognosis of heart disease can aid in making decisions on life style changes in high-risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine. The number of people facing heart diseases is on the rise each year. This prompts its early diagnosis and treatment. The utilization of suitable technological support in this regard can prove to be highly beneficial to the medical fraternity and patients. In this paper, the seven different machine learning algorithms used to measure the performance are SVM, Decision Tree, Random Forest, Naïve Bayes, Logistic Regression, K-Nearest Neighbors, Extreme Gradient Boosting and Neural Networks applied on the dataset.

The expected attributes leading to heart disease in patients are available in the

dataset which contains 76 features and 14 important features that are useful to evaluate the system are selected among them. If all the features are taken into the consideration, then the efficiency of the system the author gets is less. To increase efficiency, attribute selection is done. In this n features must be selected for evaluating the model which gives more accuracy. The correlation of some features in the dataset is almost equal and so they are removed. If all the attributes present in the dataset are taken into account, then the efficiency decreases considerably.

All the seven machine learning methods accuracies are compared based on which one prediction model is generated. Hence, the aim is to use various evaluation metrics like confusion matrix, accuracy, precision, recall, and f1-score which predicts the disease efficiently. Comparing all seven the Random Forest classifier gives the highest accuracy of 95.08%.

FUTURESCOPE:

The expected attributes leading to heart disease in patients are available in the dataset which contains 76 features and 14 important features that are useful to evaluate the system are selected among them. If all the features taken into the consideration, then the efficiency of the system the author gets is less. To increase efficiency, attribute selection is done. In the future, the work could be improved by creating a web application premised on the logistic regression algorithm and by using a larger dataset than the one used in this study, which would help to provide better outcomes and aid health professionals in predicting heart disease efficiently and effectively.

REFERENCES:

1. Kumar Arun, GargIshan, Kaur Sanmeet, —Loan Approval

- Prediction based on Machine Learning Approach||, IOSR Journal of Computer Engineering (IOSR-JCE), Vol. 18, Issue 3, pp. 79-81, Ver. I (May-Jun. 2016).
2. AboobydaJafar Hamid and Tarig Mohammed Ahmed, —Developing Prediction Model of Loan Risk in Banks using Data Mining||, Machine Learning and Applications: An International Journal (MLAIJ), Vol.3, No.1, pp. 1-9, March 2016.
 3. S. Vimala, K.C. Sharmili, Prediction of Loan Risk using NB and Support Vector Machine||, International Conference on Advancements in Computing Technologies (ICACT 2018), vol. 4, no. 2, pp. 110-113, 2018.
 4. PidikitiSupriya, MyneediPavani, NagarapuSaisushma, NamburiVimalaKumari, kVikash, "Loan Prediction by using Machine Learning Models", International Journal of Engineering and Techniques. Volume 5 Issue 2, Mar-Apr 2019
 5. Nikhil Madane, Siddharth Nanda," Loan Prediction using Decision tree", Journal of the Gujrat Research History, Volume 21 Issue 14s, December 2019.
 6. Wei Li, Shuai Ding, Yi Chen, and Shanlin Yang, Heterogeneous Ensemble for Default Prediction of Peer-to-Peer Lending in China, Key Laboratory of Process Optimization and Intelligent Decision-making, Ministry of Education, HefeiUniversity of Technology, Hefei 2009,
 7. Short-term prediction of Mortgage default using ensembled machine learning models, Jesse C.Sealand on July 20, 2018.
 8. Clustering Loan Applicants based on Risk Percentage using K-Means Clustering Techniques, Dr. K. Kavitha, International Journal of Advanced Research in Computer Science and Software Engineering.
 9. Toby Segaran, "Programming Collective Intelligence: Building Smart Web 2.0 Applications." O'Reilly Media.
 10. Drew Conway and John Myles White," Machine Learning for Hackers: Case Studies and Algorithms to Get you Started," O'Reilly Media.
 11. Trevor Hastie, Robert Tibshirani, and Jerome Friedman,"The Elements of Statistical Learning: Data Mining, Inference, and Prediction," Springer ,Kindle
 12. PhilHyoJin Do Ho-Jin Choi, "Sentiment analysis of real-life situations using location, people and time as contextual features," International Conference on Big Data and Smart Computing(BIGCOMP), pp. 39–42. IEEE, 2015.
 13. Bing Liu, "Sentiment Analysis and Opinion Mining," Morgan &Claypool Publishers, May2012.
 14. Bing Liu, "Sentiment Analysis: Mining Opinions, Sentiments, and Emotions," CambridgeUniversity Press, ISBN:978-1-107-01789-4.
 15. Shiyang Liao, Junbo Wang, RuiyunYu, KoichiSato, and Zixue Cheng, "CNN for situations understanding based on sentiment analysis of twitter data," Procedia

- computer science, 111:376–381,
2017.CrossRef.
16. K I Rahmani, M.A. Ansari, Amit
Kumar Goel, "An Efficient
Indexing Algorithm for
CBIR,"IEEE- International
Conference on Computational
Intelligence & Communication
Technology ,13-14 Feb 2015.