Predicting Heart Disease using Machine Learning Algorithm

Anagha Sridhar[1], Anagha S Kapardhi[2]

[1,2]Student, Dept. of Information Science and Engineering, NIE, Mysore, India

Abstract - *Machine Learning is used across many ranges around the world. The healthcare industry is no exclusion. Machine Learning can play an essential role in predicting presence/absence of locomotors disorders, Heart diseases and more. Such information, if predicted well in advance, can provide important intuitions to doctors who can then adapt their diagnosis and dealing per patient basis. In this paper, we'll discuss a project where we worked on predicting possible Heart Diseases in people using Machine Learning algorithms. The algorithms are Naïve Bayes and Decision Tree Classifier. The dataset has been taken from Kaggle.*

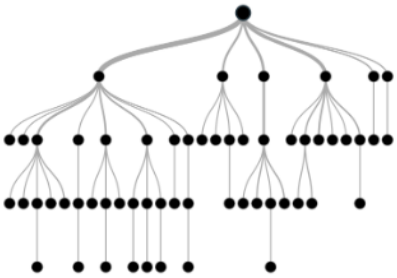
Key Words: Heart Disease, Naïve Bayes, Decision Tree, Datasets

1. INTRODUCTION

Heart is a vital organ of the humanoid body. It pumps blood to every part of our anatomy. If it miscarries to function correctly, then the brain and various other organs will stop functioning, and within few minutes, the person will die. Change in lifestyle, work related stress and wrong food habits add to the increase in rate of several heart related illnesses. Heart diseases have occurred as one of the most prominent cause of death all around the world. According to World Health Organization, heart associated 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 top cause of death. 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 outlay on health care and also reduce the efficiency 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 cardiovascular diseases. Thus, reasonable and accurate prediction of heart related diseases is very important. Medical organizations, all around the world, collect data on various health related issues. These data can be oppressed using various machine learning techniques to gain useful understandings. But the data collected is very massive and, many a times, this data can be very noisy. These datasets, which are too devastating 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 ailments accurately.

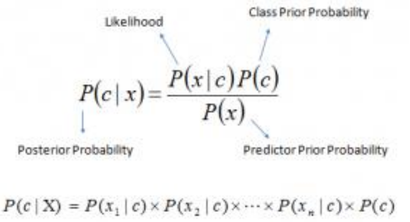
1.1 Decision Tree

Decision tree is controlled method used for the prediction of unconditional as well as numerical value. It represents the data occurrences along with their class label in the form of a tree. A set of rules can be construed from the tree which can be used to order the unknown data record to its output value. A test on an attribute is performed on the core node. The result of the test is depicted by the branch of tree and class label are present at the leaf node. In this technique the whole data set or the whole group of sample points in split into two or more homogenous classes. The split is recognized from the parameter or the factor which is dogged to be the best splitter or differentiator.



1.2 Naïve Bayes

Rather than a single classifier it actually is a grouping of multiple classifiers all working on the basic Naïve Bayes principle of autonomous features. Hence each feature is assumed to be independent and autonomous paying individually to the training data point's likelihood of belonging to a particular class. As per the Bayes theorem,



$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$$

2. LITERATURE SURVEY

Python is the language used. Python files are saved with "**.py**" extension. Same is also applied on Jupyter notebook whose extension is **.ipynb**. You need to install python as well as Jupyter for using this project. If you want to envision the result using only python, you should install **Pycharm** where Jupyter notebook is used to visualize the result on the web browser.

Machine Learning is used for numerous purpose such as colour based division, forecasting diseases, image processing applications such as object recognition, image classification and transfer learning. When deep learning came the computation power has upgraded to such level that now it is possible for the machines to effort like humans. Companies like Interest, Google, Facebook and Amazon is consuming this technology to so large extent that their incomes have increased dramatically. In this theory we have tried to use decision tree to perform multi-class classification for heart disease.

2.1 Datasets

Datasets in a perfect world is a flawlessly curated group of observations with no missing values or irregularities. However, this is not true. It can be disordered, which means it needs to be clean and wrangles. Data cleaning is a essential part in data science problems. Machine learning models learn from data. It is crucial; however, that the data you feed them is precisely pre-processed and refined for the problem you want to solve. This includes data cleaning, pre-processing, feature engineering, and so on.

2.2Train and Test Model

Model selection is the process of joining data and prior information to select among a group of arithmetical models. The first pass models, few can contend with logistic regression. This linear model is fast to train, easy to know, and classically does pretty well “out of the box”. The Scikit Learn logistic regression model works well combined with Pipeline and Grid Search CV pre-processing tools. This will help to streamline the process of model training and hyper parameter optimization.

Sklearn’s pipeline functionality makes it easier to repeat commonly occurring steps in your modelling process.

2.3 Model Accuracy

When you train your classification predictive model, you will want to assess how good it is. Interestingly, there are many different ways of evaluating the performance. Most data scientists that use Python for predictive modeling use the Python package called scikit-learn. Scikit-learn contains many built-in functions for analyzing the performance of models.

3. RESULT AND DISCUSSION

In this system, training data set is tested in predicting the heart illness. 13 attributes were taken for the prediction system of the disease. Two algorithms were selected for assessment – Decision tree & Naïve Bayes. The most exact & effective system was tested among the two. Reading found that Decision tree was more precise in its calculation of heart disease. Various test cases included in the paper demonstrate the above said fact in this regard.

3.1 Flow Chart and Table

We performed computer simulation on one dataset. Dataset is a Heart dataset. The dataset is available in UCI Machine Learning Repository.

Table -1: Heart Disease Attributes

Name	Type	Description
Age	Continuous	age: age in years
Sex	Discrete	sex: sex (1 = male; 0 = female)
Cp	Discrete	chest pain location (1 = substernal; 0 = otherwise)
trestbps	Continuous	resting blood pressure (in mm Hg on admission to the hospital)
Chol	Continuous	serum cholestoral in mg/dl
fbs	Discrete	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
Restecg	Discrete	resting electrocardiographic results (0,1,2)
Thalach	Continuous	maximum heart rate achieved
exang	Continuous	exercise induced angina (1 = yes; 0 = no)
oldpeak	Discrete	ST depression induced by exercise relative to rest
slope	Continuous	the slope of the peak exercise ST segment -- Value 1: upsloping
Ca	Continuous	number of major vessels (0-3) colored by flourosopy
Thal	Discrete	3 = normal; 6 = fixed defect; 7 = reversable defect
Num	Discrete	diagnosis of heart disease (angiographic disease status)

3.2 Main Flow Diagram

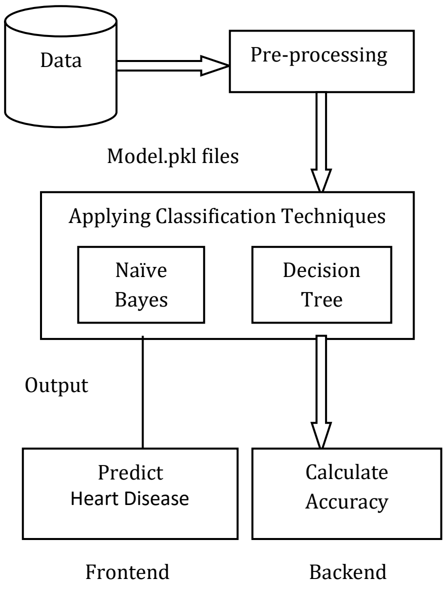


Chart -1: Main Flow diagram

3.3 Sub Flow Diagram

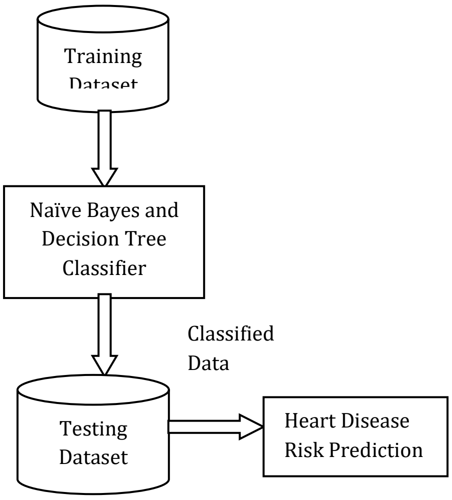

Training Dataset
Naïve Bayes and Decision Tree Classifier
Classified Data
Testing Dataset
Heart Disease Risk Prediction

Chart -2: Sub Flow diagram

4. CONCLUSION

In this paper various suggestion and classification methods are executed on the heart datasets to predict the heart diseases. Classification algorithms are used to predict small set of relations between attributes in the databases to build an correct classifier. The main contribution of the present study to attain high calculation accuracy for early diagnoses of heart diseases. The proposed hybrid associative classification is implemented on scipy environment. Finally an skilled system is developed for the end user to check the risk of heart diseases on the basis of assumed parameters and the best associative classification method. The experimental results show that large number of the rules support to the better determines of heart diseases that even support the heart professional in their diagnosis in decisions.

REFERENCES

1. https://www.researchgate.net/publication/319393368_Heart_Disease_Diagnosis_and_Prediction_Using_Machine_Learning_and_Data_Mining_Techniques_A_ReviewJ

2. https://dzone.com/articles/a-tutorial-on-using-the-big-data-stack-and-machine

3. https://pythonhow.com/html-templates-in-flask/

4. Intelligent Heart Disease Prediction System Using Data Mining Techniques-Sellappan Palaniappan, Rafiah Awang 978-1-4244-1968-5/08/ ©2008 IEEE.

5. Intelligent Heart Disease Prediction System Using Data Mining Techniques-Sellappan Palaniappan, Rafiah Awang 978-1-4244-1968-5/08/ ©2008 IEEE

6. Blake, C.L., Mertz, C.J.: "UCI Machine Learning Databases", http://mlearn.ics.uci.edu/databases/heartdisease/, 2004

7. Chapman, P., Clinton, J., Kerber, R. Khabeza, T., Reinartz, T., Shearer, C., Wirth, R.: "CRISP-DM 1.0: Step by step data mining guide", SPSS, 1-78, 2000.