Featured Engineering on Heart disease dataset using ML

Gopal Murlidhar Kholade #1, Adarsh Vishnu Tayde #2, Pramod T. Talole#3

[1]Student,Information Technology,Anuradha Engineering College, Chikhli - 443201, Maharashtra, India
[2]Student,Information Technology,Anuradha Engineering College, Chikhli - 443201, Maharashtra, India
[3]Assistant Professor,Information Technology,Anuradha Engineering College, Chikhli - 443201, Maharashtra, India

ABSTRACT

Featured engineering is the first and must step in any machine learning project. It can be simplified as a process of filtration of raw unusable, Uneven, hard to understand data into a well understandable format. Here when we receive data it is most;y scattered because it is formatted and entered by humans hence prone to error. so, This process will help in increasing our model impact and will increase the accuracy of results while evaluation.

Keywords- *Machine Learning(ML), Cardiovascular Disease Prediction, Feature Recommendation Engineering, Data Analysis, Data Filtering.*

1. INTRODUCTION

Heart disease is a disease in which a people dies due to an abnormality and malfunctioning in their heart. Under the banner of heart disease, there are many types of diseases that cause dysfunction of different parts of the heart. By the World Health Organisation, approximately 17.9 million people died from cardiovascular disease in 2019, which is for 32% of world wide deaths. However, it would be useful to be able to predict heart disease before it becomes fatal and to seek medical attention in a timely manner. This is where the machine learning comes to mind for understanding and solving the issue. Now, the capabilities of human technology are increasing and we can see the difference between 15 years ago and today. This has drastically/significantly changed the world around us. phones in our pockets, getting cheaper , easy to carry and the processing power of these small and portable devices has significantly improved and is expected to continue to increase in the time ahead. So things like machine learning combined with the Internet will change an era in human history.

2. DATA

The data sets used in the DATA process are from IEEE. The dataset consists of numerous information bundled in rows and columns, where rows direct us to the number of individual data is from and columns will direct us to the attributes based on which we have classified and noted the data. Data can be aggregated in the CSV(comma separated file) or Exel document. For the model, we prefered using CSV.

The attributes used in the following are given below:

- Age: Signifies the age of the patient.
- Sex: It displays the gender of patients in the following format :
  0 = Male
  1= Female
- Chest-pain type: It displays the chest pain in the following format :

  1 = typical angina
  2 = atypical angina

3 = non —anginal pain
4 = asymptotic

- Resting Blood Pressure: It is the resting blood pressure value of an patient in mmHg (units).
- Cholesterol: It is cholesterolof blood in mg/dl (unit)
- Fasting Blood Sugar: It is the checking of the fasting blood sugar value of an patient with 120mg/dl:
  if FBS > 120mg/dl =>1 (TRUE)
  else => 0 (FALSE)
- Resting ECG: It is resting electrocardiographic results:
  0 - normal
  1 - having ST to T wave abnormality
  2 - left ventricular hypertrophy
- Max heart rate: This means the max heart rate achieved/reached by the patient's heart.
- Old Peak: ST depression included exercise relative to rest.
- ST slope: Slope of the peak exer. of ST segment.
- Target: It is the result that denotes whether the individual is diseased or not and will be used while training and testing
  0= Not diseased
  1=for diseased

3. Related Surveys:

3.1 Feature engineering is the task of improving predictive modeling performance on a dataset by transforming its feature space. Existing approaches to automate this process depend on either transformed feature space exploration through evaluation-guided search, or explicit expansion of datasets with all transformed features followed by feature selection. Such approaches incur high computational costs in runtime and/or memory. We present a unique technique, called Learning Feature Engineering (LFE), for automating feature engineering in classification tasks. Our empirical results show that LFE outperforms other feature engineering approaches for an overwhelming majority (89%) of the datasets from various sources while incurring a substantially lower computational cost.

3.2 The identification of such fluctuations and, so, of the presence of individuals within the churches has been applied through three different methods. the primary is an unsupervised clustering algorithm here termed density peak, the second may be a supervised deep learning model supported by a standard convolutional neural network (CNN) and therefore the third may be a novel unplanned engineering feature approach unexpected mixing ratio (UMR) peak.

4. FEATURE ENGINEERING

Feature engineering consists of the various processes

4.1 Feature Creation: This can be adding or removing some features. For high precision.

4.2 Transformation: This is a function that transforms a function from one representation to another. The goal here is to plot and visualize data with plots and features. If you have problems with a new feature, you can remove it and try the new feature with another feature, speed up training, or improve the accuracy of a particular model.

4.3 Feature Extraction: The process of extracting useful data from a source. It compresses the amount of data into a manageable amount that algorithms can process without distorting or removing original relationships or sensitive information.

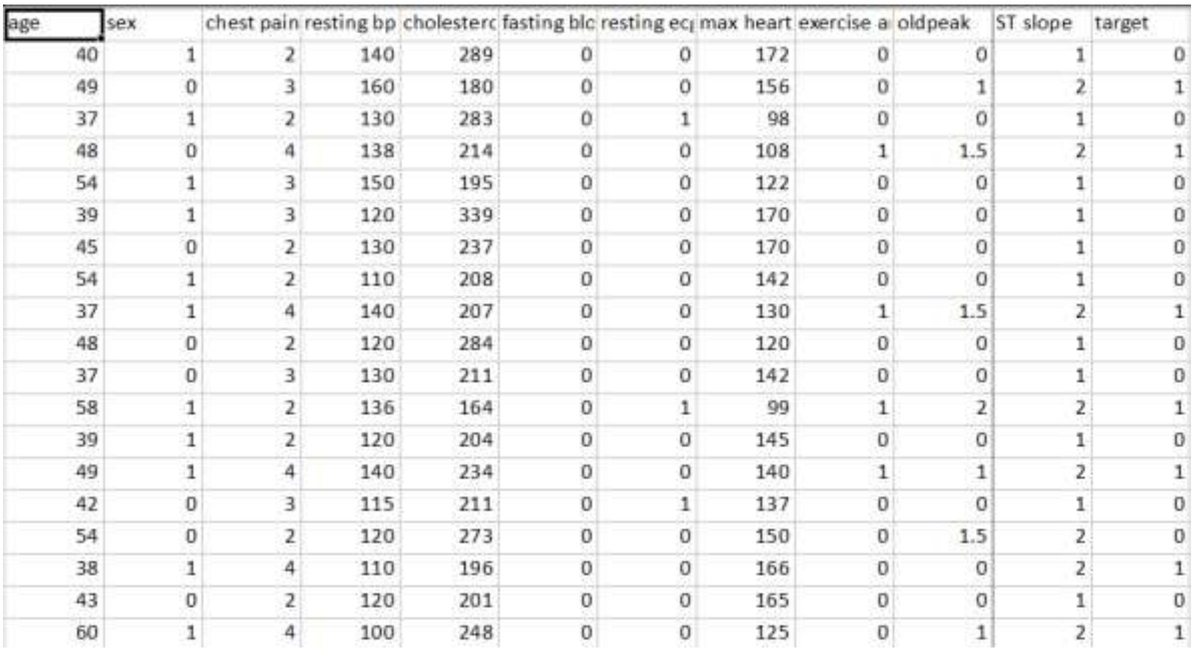
4.4 Exploratory Data Analysis: Exploratory Data Analysis: EDA is a simple and powerful tool that you can use to improve your understanding of the data in your data set by examining its properties.The technique is useful when a goal is created like, Heart Disease Prediction Model then it finds patterns and links between the data. often used on huge amounts of quantitative or qualitative data that have not been analyzed before.

4.5 Benchmark: It is the most useful and transparent technique used in feature engineering. It uses various algorithms to check the efficiency of data and accuracy in the model. It is different than that of the actual model but useful for feature engineering.

Our work of feature engineering on a dataset is shown in the jupyter notebook file given in the link:

https://drive.google.com/drive/folders/10uTNcyegepxYtyycHbj_2dLH6VLC-qwq?usp=sharing

In the end, we get the dataset in the following format,



age	sex	chest pain	resting bp	cholesterc	fasting blc	resting ec	max heart	exercise a	oldpeak	ST slope	target
40	1	2	140	289	0	0	172	0	0	1	0
49	0	3	160	180	0	0	156	0	1	2	1
37	1	2	130	283	0	1	98	0	0	1	0
48	0	4	138	214	0	0	108	1	1.5	2	1
54	1	3	150	195	0	0	122	0	0	1	0
39	1	3	120	339	0	0	170	0	0	1	0
45	0	2	130	237	0	0	170	0	0	1	0
54	1	2	110	208	0	0	142	0	0	1	0
37	1	4	140	207	0	0	130	1	1.5	2	1
48	0	2	120	284	0	0	120	0	0	1	0
37	0	3	130	211	0	0	142	0	0	1	0
58	1	2	136	164	0	1	99	1	2	2	1
39	1	2	120	204	0	0	145	0	0	1	0
49	1	4	140	234	0	0	140	1	1	2	1
42	0	3	115	211	0	1	137	0	0	1	0
54	0	2	120	273	0	0	150	0	1.5	2	0
38	1	4	110	196	0	0	166	0	0	2	1
43	0	2	120	201	0	0	165	0	0	1	0
60	1	4	100	248	0	0	125	0	1	2	1

fig-1: Dataset representation

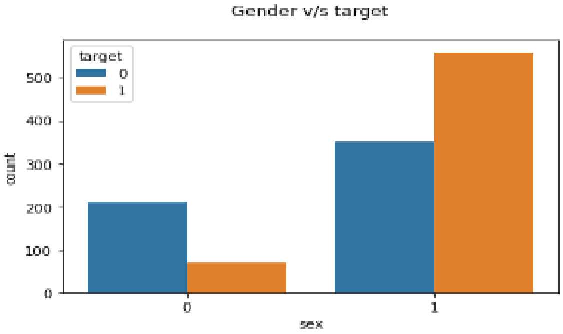


fig-2: Graph- Sex vs Target

Here, from the dataset, it is clear that human males are more susceptible to get Heart Disease than females. Men are more suceptible to heart disease more than women. Sudden Heart Attacks are experienced by men between 70% to 89%. A woman may experience a heart attack with no chest pressure at all, they usually experience nausea or vomiting which are often confused with acid reflux or the flu.

5.Conclusion

From the above illustration of data, it is clear that feature engineering is the most important process for machine learning. And the accuracy we have got is good enough.
Accuracy of Logistic Regression Model = 81.30 % ,
Accuracy of Decision Tree Model = 74.68 %,
Accuracy of Support Vector Machine = 84.13 %,
Accuracy of k-NN Model = 81.93 %,
Accuracy of Naive Bayes Model = 80.67 %,
Accuracy of Random forest classifier = 88.23 %.

6. References

[1]. Related surveys:

- Learning Feature Engineering for Classification (ijcai.org)
- https://www.academia.edu/77970375/Machine_learning_and_engineering_feature_approaches_to_detect_events_perturbing_the_indoor_microclimate_in_Ringebu_and_Heddal_stave_churches_Norway_

[2]. Data:
the dataset is collected from :
https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive#files