Asalah Mohammed Zarban

Asalah063@gmail.com

Comparative study of ML model to predict heart disease

Project Motivation

In my project, we will analyze the most relevant/risk factors of heart disease as well as predict the overall risk using logistic regression. Since the early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes to reduce complications and save the life of high-risk patients. Where the World Health Organization has estimated 12 million deaths occur worldwide, every year due to heart diseases.

Question/need:

- Whose is more susceptible to heart disease male or female?
- Is the Total cholesterol effect to increase the probability of susceptible CHD?
- Which is the most algorithm predict accurately?

Data Description:

Dataset source:

https://www.kaggle.com/dileep070/heart-disease-prediction-using-logistic-regression

Dataset info:

Our data was taken from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. And is contain over 4,000 observations of 15 features.

Description of the Dataset:

Demographic:

1. Sex: male or female (Nominal)

2. Age: Age of the patient; (Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)

Behavioral

3. Current Smoker: whether or not the patient is a current smoker (Nominal)
4. Cigs Per Day: the number of cigarettes that the person smoked on average in one day. (can be considered continuous as one can have any number of cigarettes, even half a cigarette.)

Medical(history):

5. BP Meds: whether or not the patient was on blood pressure medication (Nominal)
6. Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal)
7. Prevalent Hyp: whether or not the patient was hypertensive (Nominal)
8. Diabetes: whether or not the patient had diabetes (Nominal)

Medical(current):

9. Tot Chol: total cholesterol level (Continuous)
10. Sys BP: systolic blood pressure (Continuous)
11. Dia BP: diastolic blood pressure (Continuous)
12. BMI: Body Mass Index (Continuous)
13. Heart Rate: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)
14. Glucose: glucose level (Continuous)

Predict variable (desired target):

15. TenYearCHD: 10 year risk of coronary heart disease CHD (binary: “1”, means “Yes”, “0” means “No”)

Algorithms:

Resampling imbalance dataset

P-value approach

I take different machine learning algorithm as Logistic Regression, Naive Bayes, Random Forest Classifier, Extreme Gradient Boost, K-Nearest Neighbour, Decision Tree, Support Vector Machine and try find algorithm which predict accurately.

Tools:

-Pandas, numpy, seaborn for loading dataset

- sklearn.utils for resampling dataset

-sklearn.metrics, sklearn.linear_model, sklearn.naive_bayes, sklearn.ensemble, xgboost, sklearn.neighbors, sklearn.tree, sklearn.svm for modeling dataset

-matplotlib.pyplot for visualisation

-sklearn.model_selection, statsmodels.api, scipy.stats, collections, statsmodels.tools for splitting dataset

-sklearn.preprocessing for preprocessing dataset

-mlxtend.classifier for ensembling dataset

MVP Goal:

- Basic Data Exploration
  Analyze the dataset
  Cleaning data
  Fix imbalance data
  Comparative study of ML model