UDACITY MACHINE LEARNING NANODEGREE – 2020

CAPSTONE PROJECT PROPOSAL

PREDICTING THE PRESENCE OF HEART DISEASE IN PATIENTS

HAVISHA KALWAD

MAY, 2020

DOMAIN BACKGROUND

Heart disease describes a range of conditions that affect your heart. Diseases under the heart disease umbrella include blood vessel diseases, such as coronary artery disease; heart rhythm problems (arrhythmias); and heart defects you're born with (congenital heart defects), among others. The term "heart disease" is often used interchangeably with the term "cardiovascular disease." Cardiovascular disease generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain (angina) or stroke. Other heart conditions, such as those that affect your heart's muscle, valves or rhythm, also are considered forms of heart disease.

Heart disease statistics in United States of America:

- Heart disease is the leading cause of death for men, women, and people of most racial and ethnic groups in the United States.
- One person dies every 37 seconds in the United States from cardiovascular disease.
- About 647,000 Americans die from heart disease each year—that’s 1 in every 4 deaths.
- Heart disease costs the United States about $219 billion each year from 2014 to 2015. This includes the cost of health care services, medicines, and lost productivity due to death.

PROBLEM STATEMENT

Of all the applications of machine-learning, diagnosing any serious disease using a black box is always going to be a hard sell. If the output from a model is the particular course of treatment (potentially with side-effects), or surgery, or the absence of treatment, people are going to want to know why. This dataset gives a number of variables along with a target condition of having or not having heart disease. Using this dataset I wish to build a model that helps predict the presence of heart disease in any new patient information which is input into this model.

The goal of this project is: Given clinical parameters about a patient, can we predict whether or not they have heart disease?

DATASET AND INPUTS

The dataset for this project shall be obtained from the below mentioned links:

1. https://archive.ics.uci.edu/ml/datasets/Heart+Disease
2. https://www.kaggle.com/ronitf/heart-disease-uci

Both the above mentioned link denote the same dataset. The Kaggle version of the dataset was obtained from the UCI dataset.
This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. The names and social security numbers of the patients were removed from the database.
The features/columns used in this dataset are as follows:

1. age - age in years
2. sex - (1 = male; 0 = female)
3. cp - chest pain type
    - 0: Typical angina: chest pain related decrease blood supply to the heart
    - 1: Atypical angina: chest pain not related to heart
    - 2: Non-anginal pain: typically, esophageal spasms (non heart related)
    - 3: Asymptomatic: chest pain not showing signs of disease
4. trestbps - resting blood pressure (in mm Hg on admission to the hospital) anything above 130-140 is typically cause for concern
5. chol - serum cholesterol in mg/dl
    - serum = LDL + HDL + .2 * triglycerides

    - above 200 is cause for concern

6. fbs - (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
    - '>126' mg/dL signals diabetes
7. restecg - resting electrocardiographic results
    - 0: Nothing to note
    - 1: ST-T Wave abnormality
        - can range from mild symptoms to severe problems
        - signals non-normal heart beat
    - 2: Possible or definite left ventricular hypertrophy
        - Enlarged heart's main pumping chamber
8. thalach - maximum heart rate achieved
9. exang - exercise induced angina (1 = yes; 0 = no)
10. oldpeak - ST depression induced by exercise relative to rest looks at stress of heart during exercise unhealthy heart will stress more
11. slope - the slope of the peak exercise ST segment
    - 0: Upsloping: better heart rate with exercise (uncommon)
    - 1: Flatsloping: minimal change (typical healthy heart)
    - 2: Downslopins: signs of unhealthy heart
12. ca - number of major vessels (0-3) colored by fluoroscopy
    - colored vessel means the doctor can see the blood passing through
    - the more blood movement the better (no clots)
13. thal - thallium stress result
    - 1,3: normal
    - 6: fixed defect: used to be defect but ok now
    - 7: reversable defect: no proper blood movement when exercising
14. target - have disease or not (1=yes, 0=no) (= the predicted attribute)

SOLUTION STATEMENT

I will prepare the data by first splitting the feature and target columns. The check for the quality of the given data and clean the data. Check for missing values and impute them, if possible, or delete them.

I shall then perform some data visualization to help me understand the distribution of the data. I shall also look into the correlation matrix to see which feature is dependent on what other features.

There are several non-numeric features which need to be categorically encoded. I shall also try to perform feature engineering to further add more details to the model which will help predict better.

I shall then split the data into training and test sets (80% training set – 20% test set). I shall then normalize and scale the data to try various iterations on the chosen models to see if there are any differences in the performance with each model. I also with to apply hyperparameter tuning to the machine learning models I would be building.

I wish to try the following set of algorithms on my dataset to decide which would be the best fit:

- Logistic Regression
- K-Nearest Neighbors
- Support Vector Machine
- Decision Tree Classifier
- Random Forest
- XGBoost Classifier

BENCHMARK MODEL

I Shall choose XGBoost as my benchmark and try to beat the benchmark with hyperparameter tuning.

EVALUATION METRICS

The evaluation matrix that I would use are: Precision, Recall, Accuracy and F1-score.

PROJECT DESIGN

The workflow would be as follows:

1. Exploring the data:
   - Loading the Libraries and the dataset
   - Looking into the dataset
   - Understanding the statistics of the dataset
2. Data Pre-processing and Cleaning:
   - Identify features column and target column
   - Categorical encoding for non-numeric columns
   - Handling missing values
   - Split the data into training and test sets
   - Feature Scaling – scaling and normalizing data
3. Build the model:
   - Build the model
   - Make predictions on the test set
   - Feature selection
4. Hyperparameter tuning for the models to improve results
5. Conclusion with the summary of models built