Computer Science Department

Final Project: Heart Disease Prediction using Machine Learning
Course: Introduction to AI
Course Instructor: Yulia Newton, Ph.D.

Yosha Mundhra

Introduction/Background

Heart related diseases or Cardiovascular Diseases (CVDs) are the main reason for a huge number of death in the world over the last few decades and has emerged as the most life-threatening disease**. Heart disease** is the leading cause of death in the United States, causing about 1 in 4 deaths. The term “heart disease” refers to several types of heart conditions. In the United States, the most common type of heart disease is coronary artery disease (CAD), which can lead to heart attack.

As per the recent study by WHO, heart related diseases are increasing. 17.9 million People die every-year due to this. With growing population, it gets further difficult to diagnose and start treatment at early stage. Early detection and treatment of several heart diseases is very complex because of the lack of diagnostic centers, qualified doctors and other resources that affect the accurate prognosis of heart disease.

Dataset description

Benchmark dataset of UCI Heart disease is used for prediction in the model. Models based on supervised learning algorithms such as Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Logistic Regression, Decision Trees (DT), Random Forest (RF) and ensemble models will be used for the development of model. This model will be helpful to the medical practitioners at their clinic as decision support system.

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient.

Independent variables consists of both Numeric and Categorical values.
The data does not consists of free texts.

Independent Variables:
age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal

Dependent Variable: target

Hyper-parameters: Penalty, C

Methodology

ML Algorithm used:

Models based on supervised learning algorithms such as Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Logistic Regression, Decision Trees (DT), Random Forest (RF) are used for the development of model. Hyper-parameters are used for regularization.

Figure 1: The source for this image is google.com

Methodology:

The objective is to build a ML model for heart disease prediction based on the related parameters.

After reading the data, a correlation matrix is plotted to determine the correlation between the independent variables.
Subplots between different independent variables shows the chances of having a heart disease in an Individual based on the value of the independent variable for improved prediction.

Data modelling is started with data cleaning where outliner detection is performed followed by transformation of some input variable to categorical data. Data is then split into test and training set followed by training of model to get predictions and accuracy results.

Finally a confusion matrix is plotted followed by Hyper-parameter tuning to improve accuracy.

Application to the project

This model will be helpful to the medical practitioners at their clinic as decision support system.
It will help in early detection and treatment of several heart diseases with accurate prognosis of heart disease.

Analysis and Results

Independent Variable Analysis

After reading the data, a correlation matrix is plotted to determine the correlation between the independent variables and it is observed that there is not much of correlation.

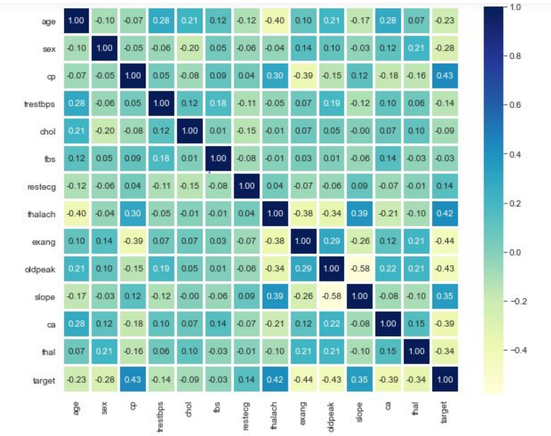

Figure 2: Correlation Matrix

Model Training

For model training the data is split in 70/30 ratio with 30% test data and 70% training data with random state 100 to get same examples every time to reproduce the same results.
The model is trained using the training data.

Test results

Predictions and accuracy scores are calculated using different models to determine the performance of different models and find the model with ML algorithm that gives more accuracy in less time i.e. most efficient.

Logistic Regression	: 0.8681318681318682
Decision Tree Classification	: 0.7802197802197802
Random Forest Classification	: 0.8241758241758241
Gradient Boosting Classification	:0.7802197802197802
Ada Boosting Classification	: 0.8241758241758241
Extra Tree Classification	: 0.8461538461538461
K-Neighbors Classification	: 0.6703296703296703
Support Vector Classification	: 0.6043956043956044

After Hyper-parameter tuning, Model with Logistic Regression gives 90% accuracy i.e. accuracy increases by 3%.

Classification report shows about the precision, recall, f1 score and support. In most of the cases f1 score is the most important, but in some cases we prioritize precision or recall over f1 score. For instance, in our case and generally in the medical community, a false negative is usually more disastrous than a false positive for preliminary diagnoses.

Confusion Matrix and Accuracy without Hyper-parameter tuning:

```
[[37  9]
 [ 3 42]]
              precision    recall  f1-score   support

           0       0.93      0.80      0.86        46
           1       0.82      0.93      0.87        45

    accuracy                           0.87        91
   macro avg       0.87      0.87      0.87        91
weighted avg       0.87      0.87      0.87        91

0.8681318681318682
```

Confusion Matrix and Accuracy with Hyper-parameter tuning:

```
[[38  8]
 [ 1 44]]
              precision    recall  f1-score   support

           0       0.97      0.83      0.89        46
           1       0.85      0.98      0.91        45

    accuracy                           0.90        91
   macro avg       0.91      0.90      0.90        91
weighted avg       0.91      0.90      0.90        91

0.9010989010989011
```

Summary/Conclusions

Started with data analysis and visualizations followed by correlation and outlier detection among input variables to remove unwanted data if present followed by modelling and hyper-parameter tuning to obtain the model with best accuracy.

Given more time, Ensemble models can be added or techniques like Stacking of different models can be used to improve predictions. In addition we can generate more features from existing features to improve model generalizability.

Key References

https://www.kaggle.com/ronitf/heart-disease-uci

Acknowledgements

I would like to acknowledge and convey my sincere gratitude to my course instructor : **Professor Yulia Newton** for her ever obliging and motivational attitude towards me.