IAI 5101: Foundations of ML for Engineers & Scientist

Winter 2022

Assignment 2

Submission Deadline: 20th March 2022 on Brightspace.

Today, heart disease is one of the biggest causes of morbidity and mortality. The ability to predict cardiovascular disease has become an important subject area for clinical data analysis. The process of identifying the presence of a heart disease can be difficult due to several contributing factors, e.g., cholesterol, pulse rate, blood pressure, etc. Machine learning is now extensively used for early diagnosis to increase the chances of survival. You are hereby presented with a sample of heart disease dataset containing a collection of demographic and clinical characteristics from 303 patients. Below are the attribute description. Using a train (70%) and test (30%) dataset split, complete the following:

Age: *age of the patient [years]*
Sex: *sex of the patient [M: Male, F: Female]*
ChestPainType: *chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]*
RestingBP: *resting blood pressure [mm Hg]*
Cholesterol: *serum cholesterol [mm/dl]*
FastingBS: *fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]*
RestingECG: *resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]*
MaxHR: *maximum heart rate achieved [Numeric value between 60 and 202]*
ExerciseAngina: *exercise-induced angina [Y: Yes, N: No]*
Oldpeak: *oldpeak = ST [Numeric value measured in depression]*
ST_Slope: *the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]*
HeartDisease: *output class [1: heart disease, 0: Normal]*

A. **EDA (20 marks):**

- Univariate analysis:
  - Using a histogram, plot a distribution of the numerical values
- Bivariate analysis:
  - Plot a histogram showing the age against the target variable (positive vs. negative cases)
  - Compare the median age for male and female using a boxplot
- Multivariate Analysis:
  - Use a heatmap to check for correlation between predictor variables

B. **Feature Engineering (20 marks):**

- Check for duplicates & missing values. Drop, if present
- There are some outliers in the dataset, (e.g., 0 cholesterol, negative oldpeak) handle them before building the model
- Check for class imbalance and handle, if necessary
- Convert categorical data into numerical data using one-hot encoding or any other label encoding approach
- Scale the data using a standard scaler

C. **Model Development I (20 marks):**

Ensemble Method:

- Use a majority voting approach to predict class label using KNN (k=5), SVM (kernel = rbf), DT (ensure you find optimal tree), and XGboost classifiers based on a soft voting (i.e., weighted average)

approach. Note: In majority voting, the predicted class label for a particular sample is the class label that represents the majority of the class labels predicted by each individual classifier.

D. **Model Development II (20 marks):**

Deep Learning:

- Train a deep neural network using Keras with 3 dense layers
- Try changing the activation function or dropout rate. What effects does any of these have on the result?

E. **Model Comparison, Evaluation (20 marks):**

- Compare the results of the ensemble with the deep neural network model in terms of the following criteria: precision, recall, accuracy, F-measure.
- Identify the model that performed best and worst according to each criterion.