Diagnosis of Heart Disease Proposal

CP610 Data Analysis Spring 2021

Dee Wu
Physics and Computer Science (MAC)
Wilfrid Laurier University
Waterloo, Canada
wuxx2586@mylaurier.ca

1. Abstract

Diagnosis of Heart Disease Project analyzes Heart Disease Data to predict the diagnosis of heart disease, by utilizing data analysis methodologies, with the evaluation and comparison of different algorithms and the optimizations in necessity, which also applies current data visualization libraries in data visualization, as well as develops a Machine Learning API using Flask Microservice.

This project will also practise the knowledge learnt from CP610 Data Analysis and will seek the opportunities to apply the algorithms, theorems, methods in the course learned, and investigate the frontline data analysis application.

Task ID	Task
1	Extract from DB
2	Exploring Processing Data
3	Building and Evaluating Models
4	Prediction and Evaluation
5	Model Visualization
6	Machine Learning API using Flask

Table 1.1 Task List

2. Description of Applied Problem

The goal is to implement the fundamental machine learning and data analysis methodologies including supervised and unsupervised learnings, in which Diagnosis of Heart Disease is a good candidate to apply the algorithms for clustering to predict the diagnostic result.

In the project the following algorithms have been applied, Dummy Classifier, Logistic Regression Model, K-NN Classifier, Gradient Boosting Classifier, Bagging Classifier, and Area Under the ROC Curve (AUG).

There will be an evaluation session for the algorithms applied and visualizations with Matplotlib and other libraries. To access prediction result as a RESTful request dispatching the project will also create a Python API by using Flask microframework.

3. Description of Available Data

This directory contains 4 databases concerning heart disease diagnosis. All attributes are numeric-valued. The data was collected from the four following locations:

- Cleveland Clinic Foundation (cleveland.data)
- Hungarian Institute of Cardiology, Budapest (hungarian.data)
- V.A. Medical Center, Long Beach, CA (long-beach-va.data)
- University Hospital, Zurich, Switzerland (switzerland.data)

Each database has the same instance format. While the databases have 76 raw attributes, only 14 of them are used (Table 2.1).

Missing Attribute Values: Several. Distinguished with value -9.0.

No.	Attribute	Description
1.	age	age in years
2.	sex	sex (1 = male; 0 = female)
3.	cp	chest pain type -- Value 1: typical angina -- Value 2: atypical angina -- Value 3: non-anginal pain -- Value 4: asymptomatic
4.	trestbps	resting blood pressure (in mm Hg on admission to the hospital)
5.	chol	serum cholestoral in mg/dl
6.	fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7.	restecg	resting electrocardiographic results -- Value 0: normal -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8.	chalach	maximum heart rate achieved
9.	exang	exercise induced angina (1 = yes; 0 = no)
10.	oldpeak	ST depression induced by exercise relative to rest
11.	slope	the slope of the peak exercise ST segment -- Value 1: upsloping -- Value 2: flat -- Value 3: downsloping
12.	ca	number of major vessels (0-3) colored by flourosopy
13.	thal	3 = normal; 6 = fixed defect; 7 = reversable defect
14.	num	diagnosis of heart disease (angiographic disease status) -- Value 0: < 50% diameter narrowing -- Value 1: > 50% diameter narrowing (in any major vessel: attributes 59 through 68 are vessels)

Table 2.1: Attribute Information

4. Analysis and Visualization Techniques

4.1. Preprocessing

Data preprocessing tasks include exploring data, data processing, building prediction models, training the models, the evaluations, and data visualizations.

4.2. Analysis

4.2.1. Regression and Classification

- Implement Linear Regression, Logistic Regression models, and other models.
- Handle overfitting, underfitting and the regularizations.
- Apply the algorithms and comment on the purpose, functionality of these algorithms.

4.2.2. Model evaluation

Import sklearn and other Python libraries to evaluate the models in comparison.

- Implement the model evaluation metrics, such as accuracy, precision.
- Describe generalization error in scope for the learning algorithms.

4.3. Visualization

Matplotlib is a de-facto method to present the mining results and other data visualization findings. Due to its versatility and the accountability qualifications, it is a good fit in plotting the Diagnosis of Heart Disease Project.

5. References

[1] Chaurasia, V. and Pal, S. (2013) Data Mining Approach to Detect Heart Disease. International Journal of Advanced Computer Science and Information Technology (IJACSIT), 2, 56-66. [Citation Time(s):1]

[2] UCI Machine Learning Repository, Heart Disease Data Set, http://archive.ics.uci.edu/ml/datasets/Heart+Disease