Technical Elective: Data Mining

Project: Heart disease classification

Data Description

The heart disease dataset consisted of 303 consecutive patients referred for coronary angiography at the Cleveland Clinic between May 1981 and September 1984. No patient had a history or electrocardiographic evidence of prior myocardial infarction or known valvular or cardiomyopathic disease. All 303 patients provided a history and underwent physical examination, electrocardiogram at rest, serum cholesterol determination and fasting blood sugar determination as part of their routine evaluation. Historical data were recorded and coded without knowledge of noninvasive or angiographic test data. In addition, after giving informed consent, the patients underwent 3 noninvasive tests as part of a research protocol. The results of these tests (exercise electrocardiogram, thallium scintigraphy and cardiac fluoroscopy) were not interpreted until after the coronary angiograms had been read. These tests were analyzed and the results recorded without knowledge of the historical or angiographic results. Work-up bias was therefore not present. Angiograms were interpreted by a cardiologist without knowledge of other test data. Further details of this data collection are described elsewhere.

This database contains 76 variables, but all published experiments refer to using a subset of 14 of them. Hence, we only consider these 14 variables in this project. The details of these 14 variables as shown in Table 1.

Table 1 Variables description

Variable Name	**Role**	**Type**	**Description**	**Units**
age	Feature	Integer		years
sex	Feature	Categorical		
cp	Feature	Categorical		
trestbps	Feature	Integer	resting blood pressure (on admission to the hospital)	mm Hg
chol	Feature	Integer	serum cholestoral	mg/dl
fbs	Feature	Categorical	fasting blood sugar > 120 mg/dl	
restecg	Feature	Categorical		
thalach	Feature	Integer	maximum heart rate achieved	
exang	Feature	Categorical	exercise induced angina	
oldpeak	Feature	Integer	ST depression induced by exercise relative to rest	
slope	Feature	Categorical		
ca	Feature	Integer	number of major vessels (0-3) colored by flourosopy	
thal	Feature	Categorical		
num	Target	Integer	diagnosis of heart disease	

Question

Question 1

Split the training and testing set. The first 50 samples in the table are the testing set, while the rest are the training set. Please explore the data preliminarily. For example, calculate the statistical values of all integer features in the training set, including mean, median, and quartile.

Question 2

The raw dataset is of low quality, such as the appearance of missing values and outliers—Preprocess raw data to obtain a high-quality format dataset.

Question 3

Use the training set to train classifiers (at least three different classifiers, such as decision trees, support vector machines, etc.) and evaluate the model's performance using Accuracy, Recall, Precision, F1-score, and AUC value. It is worth noting that you need to deal with the imbalanced dataset. In addition, strategies such as grid search should be used to optimize the hyperparameters of the classifier. Please perform sensitivity analysis on the hyperparameters of the classifier.

Question 4

Redo Question 3, requiring feature selection algorithms (at least two different algorithms, such as forward selection) to select the best set of features for training the classifier.

Question 5

Redo Question 3, requiring unsupervised dimensionality reduction algorithms (at least two different algorithms, such as PCA) to reduce the dimensionality of the dataset and use it for training the classifier.

Question 6

Please compare the performance of the best model in Question 3, Question 4, and Question 5, and write a summary.

Question 7

Based on the data in this project, please raise a valuable question and resolve it.

There are also some noteworthy points:

- ✓ **Please use programming software (such as Python) and the given dataset to solve the questions. All code needs to be submitted.**
- ✓ **Data visualization can provide a more intuitive understanding and analysis of data. In this project, rich data visualization is required. For example, the performance of different models in Question 3 can be presented more intuitively through data visualization.**
- ✓ **Each group needs to submit a report (more than 5 pages). In the report, including but not limited to the principles of the methods and algorithms used and analyzing experimental results.**
- ✓ **Each group needs to give a presentation (about 20 minutes).**

Data Reference

Janosi, Andras, Steinbrunn, William, Pfisterer, Matthias, and Detrano, Robert. (1988). Heart Disease. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X.