Student name: Leung Chun kit

Student ID: 21018713D

COMP4433 Data mining and data warehousing

Individual Project report

Analysis and prediction on Heart Attack Dataset

Objective

Analysis of the heart attack analysis and prediction dataset using PCA, LDA, and association rule. And build a model to predict whether the client has heart disease or not by using given information (e.g. age, number of major vessels …)

Data description

* Age: Age of the patient
* Sex: Sex of the patient
* cp: Chest Pain type chest pain type
  + Value 1: typical angina
  + Value 2: atypical angina
  + Value 3: non-anginal pain
  + Value 4: asymptomatic
* trtbps: resting blood pressure (in mm Hg)
* chol: cholesterol in mg/dl fetched via BMI sensor
* fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
* 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
* thalachh : maximum heart rate achieved
* exng: exercise induced angina (1 = yes; 0 = no)
* oldpeak: Previous peak
* slp: Slope of ST
* caa: number of major vessels (0-3)
* thall: Thal rate
* output:
  + 0: less chance to have heart disease
  + 1: more chance to have heart disease

There is total of 303 observations and 14 features.

Data analysis

Exploratory data analysis (EDA)

We are trying to find out some relationship between the chance of getting heart disease and client information.

Correlation matrix for numeric variables and the output

A screenshot of a computer

Description automatically generated with medium confidence

We found out that that ‘thalachh’ (maximum heart rate achieved

) and ‘slp’ (slope of ST) has a weak positive linear relationship with the chance of getting heart disease.

And ‘oldpeak’ (Previous peak) had a weak negative linear relationship with the chance of getting heart disease.

Scatterplot of thalachh against output

Histogram

Description automatically generated with medium confidence

The lower bound of the thalachh of the client with output 0 is less than the lower bound of the thalachh bound of the client with output 1.

Scatterplot of oldpeak against output

Chart

Description automatically generated with medium confidence

The client with output 0 tends to have a lower oldpeak than the client with output 1.

Principal component analysis (PCA)

PCA can be used to reduce the dimensions of data. It retains most of the variations in the data.

Chart, scatter chart

Description automatically generated

The PCA plot with the first two PC explains 33% of the variation in the dataset.

The PCA separates two types of clients. The clients with a higher chance of having heart disease to the left side and the client with a lower chance of having heart disease to the right slide. But in the middle part of the graph, there are some overlaps parts with the two types of clients.

The first 2 PC only explain 33 % of the variation of the data. Therefore, we may consider using more PC to analyze the data.

Linear Discriminant Analysis (LDA)

Graphical user interface

Description automatically generated with low confidence

LDA with separate features based on their label.

In the LDA plot, the client with less chance of getting heart disease tends to have a lower value, and the client with a higher chance of getting heart disease tends to have a higher value. But we cannot use the LDA value to classify the output because there is some overlap part on the LDA value for both types of clients.

Association rule mining

We are going to apply the apriori algorithm to find the association rule from the dataset.

First, we convert the features in the database to labels, based on their characteristic.

For example, we separate the age into three groups.

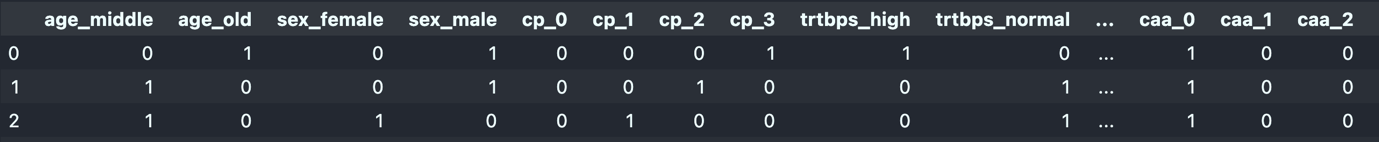
If the client's age is less than or equal to 25, we label it as 'young'.

If the client's age is less than or equal to 50, we label it as 'middle'.

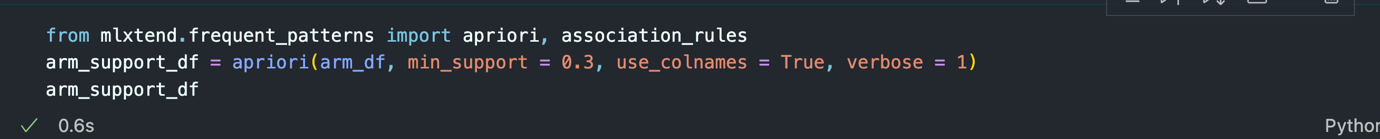
If the client's age is greater than or equal to 50, we label it as 'old'.

Then, we use the pd.dummies() function to apply one hot encoding.

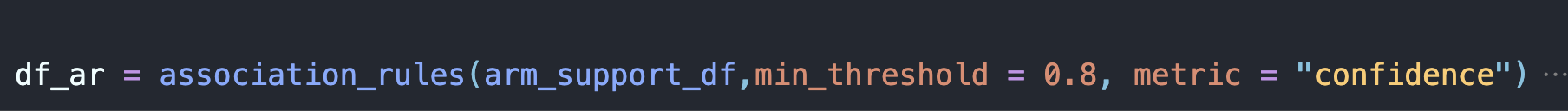
Example output:



We use the apriori function to find the frequent items.



And use the association rule function to find the association rule



We set the minimum support to 30% and the minimum confidence to 80%.

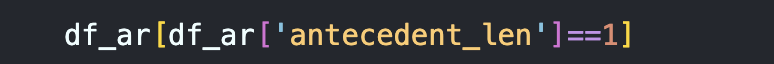
Some of the association rules mined



There are a total of 266 rules mined with confidence greater than 80% and support greater than 30%

The rule with the highest confidence is (chol: high, fbs: False, slp: 1) -> (thalachh : normal ) ( support 0.31, confidence 0.93)

Some rules with only one antecedent:



A picture containing graphical user interface

Description automatically generated

Some meaningful rules mined:

Client with false resting electrocardiographic results and age greater than 50 years old tends to have high serum cholesterol levels. (restecg: false, age: old) -> (chol: high) (support: 0.33, confidence: 0.9)

Clients with a high chance of getting heart disease tend to have high serum cholesterol levels.

(Output: disease)-> (chol: high) (support: 0.44, confidence: 0.81)

we find an interesting rule to point out that clients with a low chance of getting heart disease tend to be male. (Output: no disease) -> (sex: male) (support 0.564356, confidence 0.826087)

Before we conclude that people with no heart disease are more likely to be male in general. We need to investigate more on the real condition.

Graphical user interface, text

Description automatically generated

We find out that the number of males and females in the dataset is not equally distributed. There is the more male client in our dataset. Therefore the rule (output: no\_disease ) -> (sex: male) is not true in general.

Rules mined related to output: 1 (higher chance of getting heart disease)

Graphical user interface

Description automatically generated with low confidence

We check for the distribution of clients with a high chance of getting heart disease and a low chance of heart disease.

A screenshot of a computer

Description automatically generated with medium confidence

The normal rate of higher chance of getting heart disease is 54%.

But in the following condition, the rate of the client having a high chance of getting heart disease is greater than normal (> 80%).

A picture containing text

Description automatically generated

In clients with fixed defects and 0 major vessels colored by flourosopy, there are 89% of the client have a high chance of getting heart disease. (thall : fixed\_defect, caa : 0) -> (output: disease) (support 0.33, confidence 0.89)

In clients with no exercise-induced angina and 0 major vessels colored by flourosopy, there are 85% of the client have a high chance of getting heart disease. (exng: false, caa : 0) -> (output: disease) (support 0.37, confidence 0.85)

Model selection to predict the chance of client having heart disease

First, we split the raw data into a training dataset and a testing dataset. The training dataset is used to train the model and select the useful model and the testing data validates the performance of our model chosen when facing unseen data.

We select 30% data as testing data and the remaining 70% data as training data.

The number of observations of training data is 212 and the number of observations of testing data is 91.

I will select the best model from Random forest, logistic regression, LDA classifier, SVM, Decision tree, Naive bayes classifier, KNN, and Multilayer perceptron.

I am going to use 10-fold cross-validation to calculate the mean accuracy rate among difference iterations, and choose the models with the highest mean accuracy rate.

For the models with hyperparameters, I will use 10 cross-validations to find out the best parameter, then use the best parameter to build the model.

Model parameter choosing example:

For the KNN model, we need to choose the parameter K (the number of the neighborhood) to train the model. To find out the K with the best performance. We will perform a 10-fold validation on turning the K parameter.

We will choose the K with the highest accuracy rate in the 10-fold cross-validation.

Mean Accuracy rate on difference K- value

Chart, scatter chart

Description automatically generated

According to the plot, we can find out that when K =25, the model performs the best in the 10-fold cross-validation. Therefore, we choose k =25 as our parameter for training the KNN model.

Result of different models’ performance on 10-fold cross-validation (training data)

i) score and false negative table ii) Accuracy rate of difference models

Graphical user interface, application

Description automatically generated Chart, bar chart, histogram

Description automatically generated

iii) False negative rate of difference models

Chart, bar chart, histogram

Description automatically generated

If based on the cross-validation accuracy rate, we should select Random Forest since it has the highest accuracy rate.

But our task is to predict whether a client has or has not had heart disease. So that we care about the false negative rate (type II error). Because if we cannot classify the patient, they may not get help on time and worsening the condition of the disease. The cost of getting a false negative result is much larger than getting a false positive result. Therefore, we need to consider the false negative rate in choosing the model.

Since the false Negative rate of LDA is lower than RF, and the accuracy rate is just slightly lower than the Random Forest model, we should choose the LDA model to use for predicting client heart disease.

After building the model of LDA, we are going to check whether LDA is the best model when facing unseen data.

We are going to use the testing dataset to validate the performance of LDA compared to the others.

Result of different models’ performance on simulation of unseen client (testing data)

i) score and false negative table ii) Accuracy rate of difference models

A screenshot of a computer

Description automatically generated with medium confidence Chart, bar chart, histogram

Description automatically generated

iii) False negative rate of difference models

Chart, histogram

Description automatically generated

We find out that the LDA model has the high accuracy rate among all models. And has the lowest false negative rate. Therefore, LDA is the ideal model for predicting a client's chance of getting heart disease.