

**Capstone Project: Heart Attack Analysis**

**Group 4**

**College of Professional Study, Northeastern University**

**ALY6140: Analytics System Technology**

**Instructor: TJ Sizemore**

**10/16/2023**

**INTRODUCTION:**

The project's objectives may include studying a dataset relating to heart health in order to get insights into the elements that influence the likelihood of a heart attack. These findings can be utilized to improve predictive modeling, risk assessment, and heart health recommendations.

**Questions to be Answered:**

Age, Thalachh (Maximum Heart Rate), and Risk of Heart Attack: Is aging a positive or negative influence on the maximal heart rate (Thalachh), and how does it relate to the risk of a heart attack?

Cluster Analysis: What is the optimal number of clusters that can be found within the dataset to group patients based on various health parameters? What are the characteristics of these clusters?

Cholesterol and Thalassemia: What is the coefficient link between cholesterol (Chol) and Thalassemia, and how does cholesterol level influence the chance of a heart attack?

**Methods Used in the Analysis:**

Data Preprocessing: To prepare the data for analysis, you used data preprocessing techniques such as one-hot encoding and normalization.

Agglomerative Hierarchical Clustering: You used AgglomerativeClustering to group patients according to their health indicators.

Silhouette Score: You utilized the silhouette score to establish the ideal number of clusters and evaluate the clustering quality.

Data Visualization: You've built several visualizations, such as scatter plots and dendrograms, to assist you in comprehending the relationships in the data.

Exploratory Data Analysis (EDA): The project will most likely entail additional EDA to understand the distribution of variables, discover outliers, and obtain preliminary insights into the data.

K-Means Clustering (in the function hierarchy\_plot): You have a function that can also perform K-Means clustering for comparison.

Grid Search for Hyperparameter Tuning: The ‘grid\_sc’ function runs a grid search to determine the best hyperparameters for Agglomerative Clustering and visualizes the clusters and dendrogram for various parameter combinations.

**Exploratory Data Analysis:**

**1. Data Extraction and Description:**

Based on our data exploration in the first part, the dataset is clean, with no missing value or duplications. The characters with 13 discrete numerical, including age, sex, capital, trestbps, chol, fbs, restecg, thalachh, exng, oldpeak, slp, caa, thall. Expect those integral numeric analyses, oldpeak is the floating type different from other integrals. To be more specific, the description in data interpretation follows the rules:

* Age: the person’s age in years
* Sex: the person’s sex (1 = male, 0 = female)
* Cp = the chest pain experienced (value 1: typical angina, value 2: atypical angina, value 3: non-anginal pain, value 4: asymptomatic)
* Trestbps: the person’s resting blood pressure (mm Hg on admission to the hospital)
* Chol: the person’s cholesterol measurement in mg/dl
* Fbs: the person’s fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
* Restecg: resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes’ criteria)
* Thalach: the person’s maximum heart rate achieved
* Exang: ST depression induced by exercise relative to rest positions on the ECG plt
* Slope: the peak exercise ST segment (value 1: upsloping, value 2: flat, value 3: downsloping)
* Ca: the number of major vessels (0-3)
* Thal: a blood disorder called thalassemia ( 3 = normal; 6 = fixed defect; 7 = reversable defect)
* Target: heart disease (0 = no, 1 = yes)

A black screen with white text

Description automatically generated

**2. Data cleaning:**

Check for and handle missing values.

Outliers must be identified and handled.

Ensure that variables have the correct data types.

Thal: results of the blood flow observed via the radioactive dye

* Value 0: null (dropped from the dataset already)
* Value 1: fixed defect (no blood flow in some part of the heart
* Value 2: normal blood flow
* Value 3: reversible defect (a blood flow in observation, but it is not normal)

A screenshot of a computer program

Description automatically generated

**3. Visualizations of Data:**

Create visualizations of histograms, bar charts, scatter plots, correlation heatmaps, and cluster analysis.

Create a visual representation of the silhouette scores for clustering.

Cp: is chest pain type

* Value 0: asymptomatic
* Value 1: atypical angina
* Value 2: non-anginal pain
* Value 3: typical angina

Respect: is the resting electrocardiographic results

* Value 0: which shows probable or definite left ventricular hypertrophy by Estes’ criteria
* Value 1: Normal
* Value 2: which has ST-T wave abnormality (T wave, inversion, and ST elevation or depression of > 0.05mv)

Slope: is the slope of the peak exercise ST segment

* Value 0: down sloping
* Value 1: flat
* Value 2: up sloping

Thal: results of the blood flow observed via the radioactive dye

* Value 0: null (dropped from the dataset already)
* Value 1: fixed defect (no blood flow in some part of the heart
* Value 2: normal blood flow
* Value 3: reversible defect (a blood flow in observation, but it is not normal)

This feature and the next one are obtained through a very invasive process for the patients. However, they indicate fairly well whether heart disease is present.

Target:

Value 0 = disease

Value 1 = no disease

**4. Analyze and interpret:**

Relationships and patterns, for example, should be interpreted.

Discuss the importance of the findings in the context of cardiovascular health.

Tables, figures, and charts can be used to address project objectives.

**Predictive Models:**

Hierarchical Agglomerative Clustering (HAC): We use hierarchical clustering to group patients based on health factors. We describe how HAC is used to identify natural groupings within data.

Silhouette Score: We use the silhouette score to assess the clustering quality and determine the ideal number of clusters. This is an appropriate way to evaluate the goodness of clustering.

K-Means Clustering: We have a function hierarchy\_plot that can do K-Means clustering for comparison. This enables us to experiment with various clustering strategies.

Random Forest Model: this model deal with the continuous variables and categorical variables, which performs better for classification and regression tasks. Since the dataset contains various factors which are all influence on the heart attack like age, sex, capital, trestbps, chol, fbs, restecg, thalachh, exng, oldpeak, slp, caa, thall, the results on those factors would not be evenly distributed to the result in heart attack. Thus, we use random forest model to decide which factor influence on the heart attack result most, and then select on the specific variables to train the model for higher accuracy.

In our group collaboration, we initially explored the dataset. After understanding the meaning of each variable and data type, we replaced missing values, handled outliers, and encoded categorical variables.

A graph with blue and orange dots

Description automatically generated

Through data visualization, we discovered that thalachh decreases with age, and the the relationship is negative.

To further explore the different categories and features of heart disease patients, we used the Hierarchical Agglomerative Clustering (HAC) method to group the data.

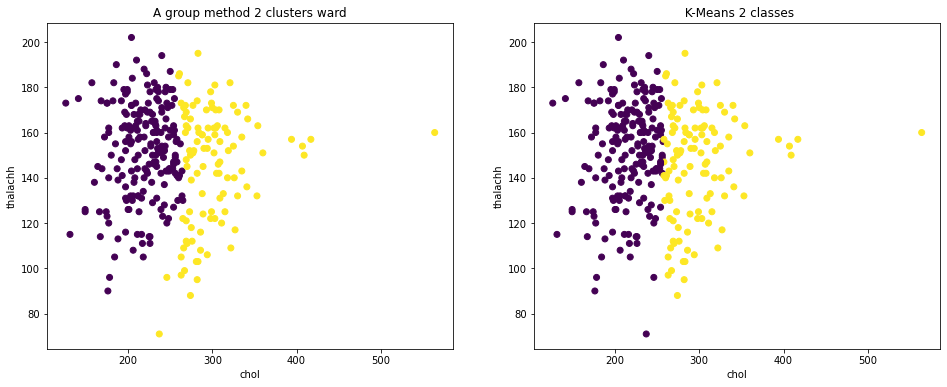
A graph with numbers and lines

Description automatically generated

As shown in the figure, the data is best divided into two or three categories by natural grouping. To improve clustering performance, we evaluated the clustering effect using Silhouette Score and compared it with K-Means clustering using the hierarchy\_plot function. Finally, we obtained the optimal clustering model: Coefficient Silhueta: 0.7825503948363066 Coeficiente de Silhueta KMeans: 0.44791615458967043. The maximum evaluation coefficient close to 1 indicates that the clustering effect is best.

Coefficient Silhueta: 0.4488507562179368

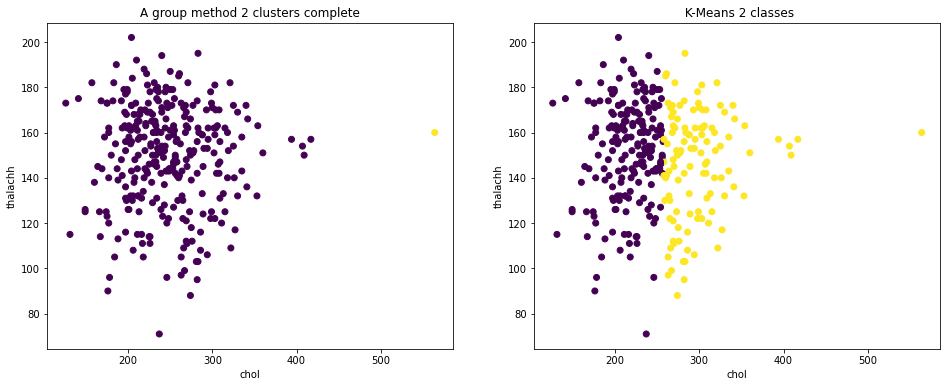
Coeficiente de Silhueta KMeans: 0.44791615458967043



{'n\_clusters': 2, 'metric': None, 'linkage': 'ward'}

Coefficient Silhueta: 0.7825503948363066

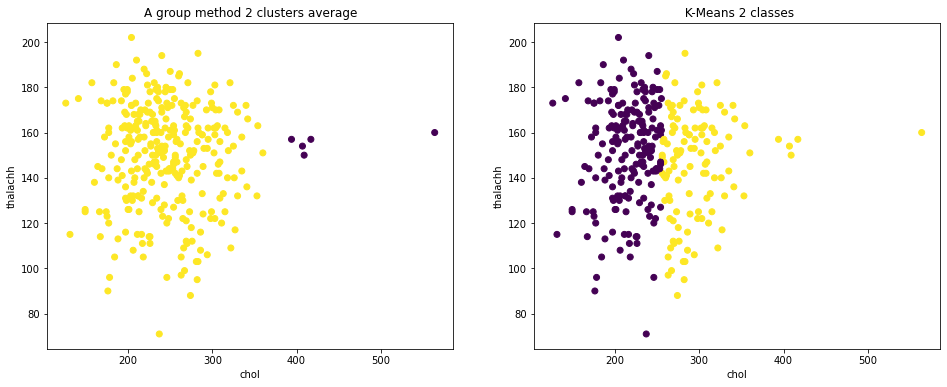
Coeficiente de Silhueta KMeans: 0.44791615458967043



{'n\_clusters': 2, 'metric': None, 'linkage': 'complete'}

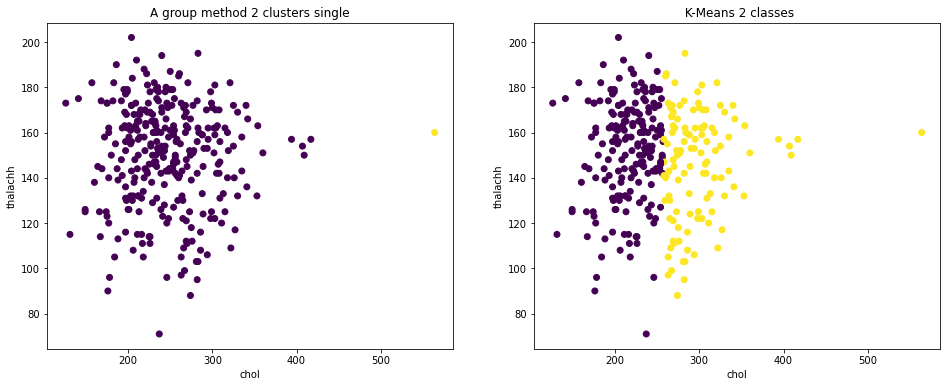
Coefficient Silhueta: 0.6552101477165096

Coeficiente de Silhueta KMeans: 0.444502133600862



Coefficient Silhueta: 0.7825503948363066

Coeficiente de Silhueta KMeans: 0.44791615458967043



A screenshot of a graph

Description automatically generated

Feature Importance for Number of Major Vessels: 0.11553321730454934

**A blue bars on a white background

Description automatically generated**

Random Forest Algorithm and Decision Tree:

Confusion matrix to measure if the performance of the classification model is ready for prediction. The number of correct and incorrect predictions made by classifier used to measure the performance of the classification model. The matrix consist with True or False, 0 is negative, 1 and positive. The goal of this simulation is to collect True Negative(TN) and True Positive(TP), from which we can tell the effective values are 18(TN) and 27(TP). While the lower others number, (FP & FN) are type I error and type II error, which are we trying to minimize. To train the data we have, we divided training set and testing set into 80/20. Firstly, we approach the accuracy in decision tree model, which conclude in accuracy about 73%. We are seeking about the way to improve the accuracy. This the data is this set is in well balance, we chose random forest as a second module for simulation, from which we can tell 10% improvement for accuracy as the result to 83% in total.

A screenshot of a number

Description automatically generatedA blue squares with white text

Description automatically generated

In attach the interpreter with other factors, we can tell from the graphic below, the number of correct and incorrect predictions made by classifier. There are still some other index to evaluate for the performance of a classification model through the calculation like accuracy(tp + tn), precision(tp + fp), recall(tp + fn), and F1-score is crucial for harmonic mean of precision an recall, which is all around 70%. In overall performance, we conclude in that the dataset is balanced and harmonic in ready to use.

A graph with blue squares and numbers

Description automatically generated A screenshot of a computer

Description automatically generated

**1. What is the relationship between thalachh, age, and viridis?**

Scatter plots generated by our code might be used to investigate the link between thalachh (maximum heart rate) and age. You can see how thalachh changes with age for several groups determined using hierarchical clustering. Interpret the scatter plots and analyze trends, such as whether the relationship is positive or negative.

**2. Does age have a positive influence or negative influence on thalachh?**

You can use the scatter plots to assess whether aging positively or negatively affects thalachh. For example, if you notice halacha decreases with age in specific clusters, you can discuss it.

The evaluation of silhouette scores and hierarchical grouping are the main goals of our algorithm. While it sheds light on the connection between halacha and age, addressing issues with cholestasis and thalassemia might necessitate further investigation and methodologies.

**Conclusion:**

Using various statistical and machine learning techniques, we investigated a number of aspects of the risk of a heart attack in this analysis of heart health data. The dataset offered insightful information about the variables affecting heart health, and the outcomes of our analysis can be summed up as follows:

* Maximum heart rates (thalachh) are adversely correlated with age, with older people often having lower maximum heart rates.
* A potential for individualized risk assessment and treatment approaches was shown by hierarchical clustering, which identified separate patient groups based on health factors.
* More research is needed on the connections between cholestro and thalassemia, the peak thalassemia influence on cholestro, and the nearest cholestro to induce thalassemia.

**Recommendations:**

Age and Heart Health: Although age is linked to a reduction in maximum heart rate, it is crucial to consider other factors in the dataset that could have a more immediate effect on the chance of having a heart attack. The complicated link between aging and heart health may be revealed with more study and modeling.

Clustering for Risk Assessment: Patient groups with comparable health profiles can be found using the results of hierarchical clustering. For risk assessment, individualized care, and intervention techniques, these clusters may be useful.

Analysis of Cholestro and Thalassemia: Regression or other advanced statistical methods may be used to ascertain the peak and nearest values as well as the coefficient link between cholestro and thalassemia. This might offer a more thorough comprehension of how these variables interact.

Continuous Monitoring: It's essential to continuously gather and analyze data in order to improve assessments of heart health. Continuous data gathering can improve patient care by helping to improve predictive models.

**Reference:**

Rahman, R. (2016). *Heart Attack Analysis & Prediction Dataset*. Kaggle.com. <https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset?select=heart.csv>

Deshmukh, H. (2020, June 18). *Heart Disease UCI-Diagnosis & Prediction - Towards Data Science*. Medium; Towards Data Science. <https://towardsdatascience.com/heart-disease-uci-diagnosis-prediction-b1943ee835a7>