Heart Disease Clustering

Karolina Mączka, Tymoteusz Urban

Business purpose

- Anonymized data of patients who have been diagnosed with heart disease
- Patients with similar characteristics might respond to the same treatments
- Help doctors understand which treatments might work with their patients

Dataset:
https://www.kaggle.com/datasets/kingabzpro/heart-disease-patients

Original:
https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Dataset

1. age - age in years (male risk > 55, female risk > 65)

2. sex - sex (1 = male; 0 = female)

3. cp - chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic)

4. trestbps - resting blood pressure (in mm Hg on admission to the hospital, 120 norm, 130< overpressure)

5. chol - serum cholesterol (in mg/dl, normal: <200 mg/dL, borderline high: 200 to 239 mg/dL, high: >240 mg/dL)

6. fbs - fasting blood sugar > 120 mg/dl (1 = true; 0 = false, diabetes)

7. restecg - resting electrocardiographic results (0 = normal; 1 = having ST-T; 2 = hypertrophy)

8. thalach - maximum heart rate achieved

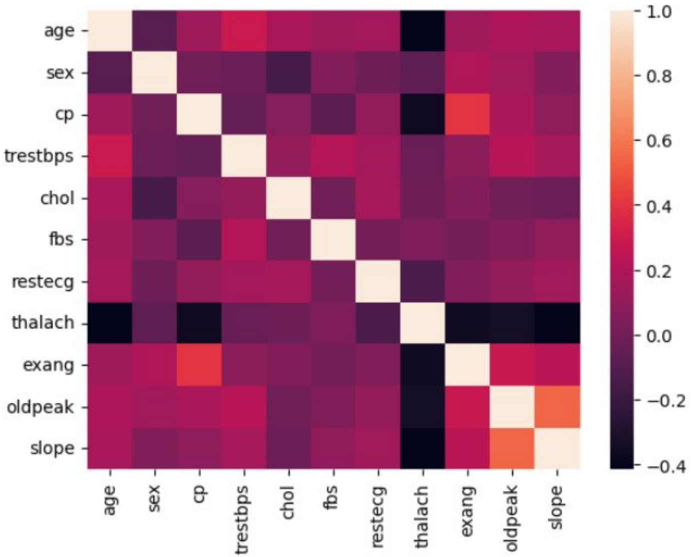
9. exang - exercise induced angina (1 = yes; 0 = no)

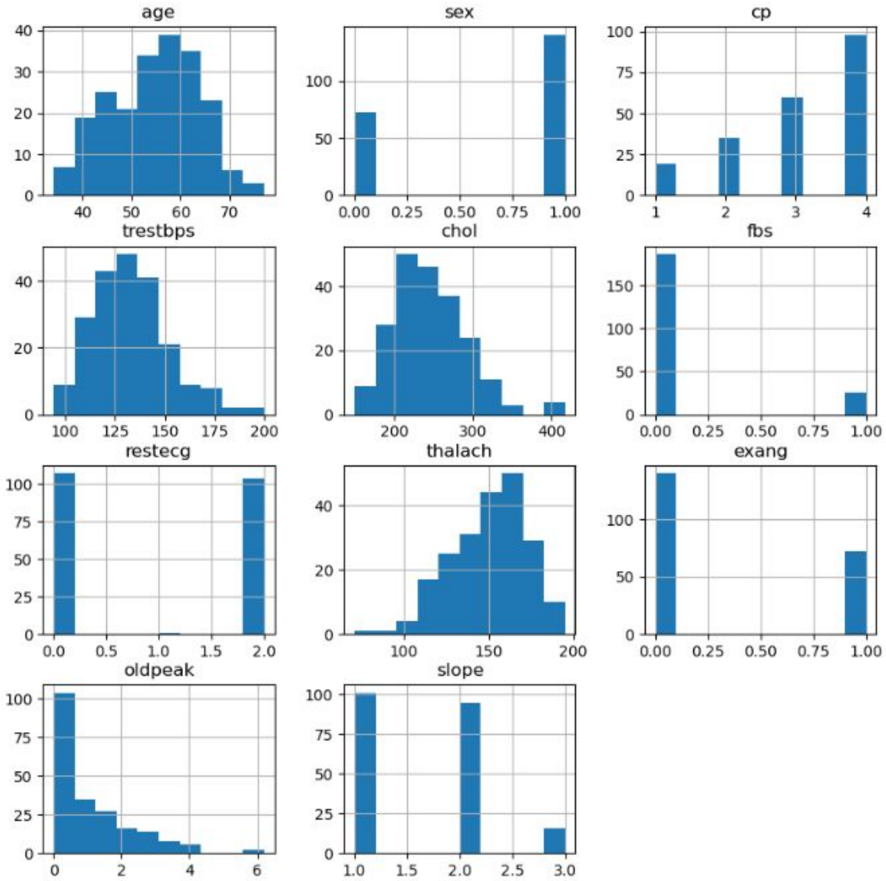
10. oldpeak - ST depression induced by exercise relative to rest

11. slope - the slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = downsloping, flat is good)

We did a thorough study of the data and asked medical expert to help us understand it

Data Exploration





Preprocessing

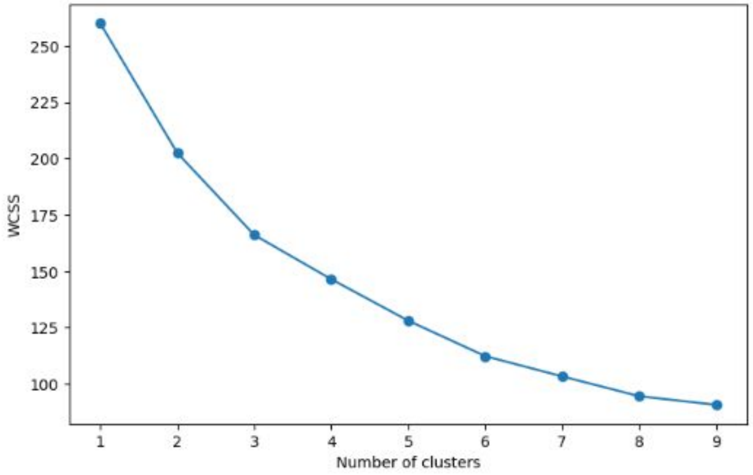
- Removing unnecessary columns (ID)
- Handling NaNs (replacing with medians)
- Replacing outliers
- Encoding chest pain type column
- Applying MinMax scaler
- Reducing dimensionality with PCA

Preprocessing

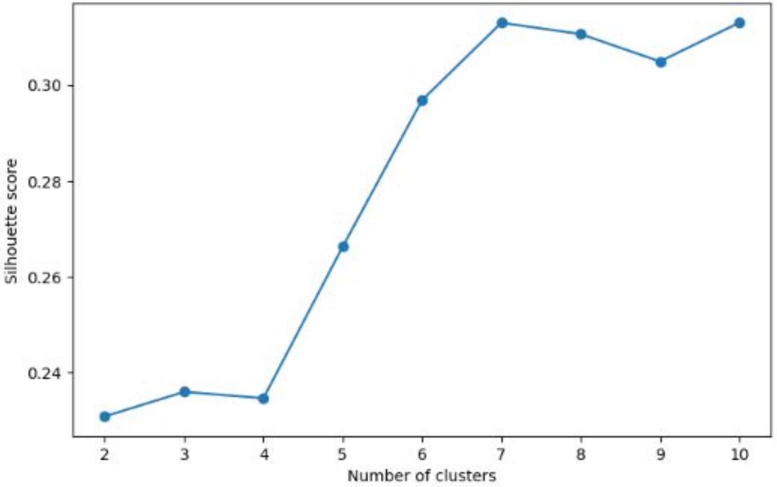
- Removing unnecessary columns (ID)
- Handling NaNs (replacing with medians)
- ~~Replacing outliers~~ - outliers are important in medical data
- ~~Encoding chest pain type column~~ - clusters were based only on cp variable
- Applying MinMax scaler
- Reducing dimensionality with PCA

Number of clusters

Elbow method

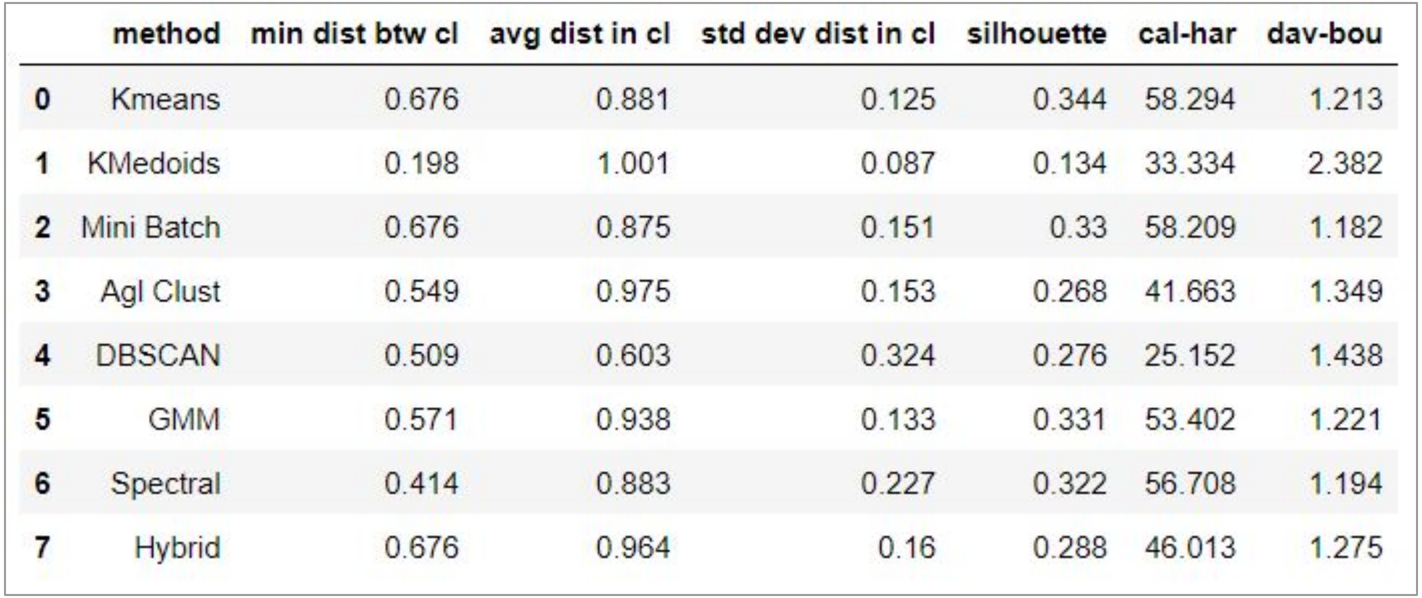


Silhouette method



It was hard to choose optimal number of clusters from elbow method. However in silhouette the best results were obtained by 7 and 10 clusters. We chose 7 as 10 is too much.

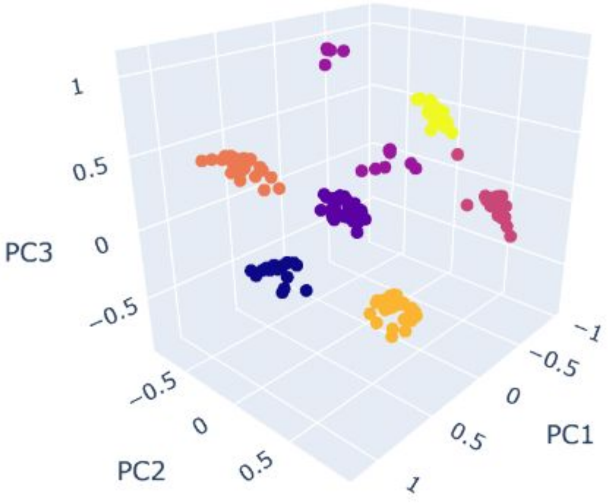
Clustering algorithms

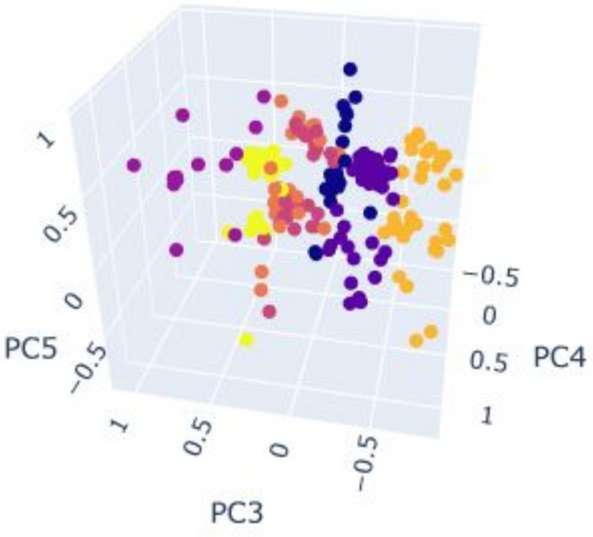


	method	min dist btw cl	avg dist in cl	std dev dist in cl	silhouette	cal-har	dav-bou
0	Kmeans	0.676	0.881	0.125	0.344	58.294	1.213
1	KMedoids	0.198	1.001	0.087	0.134	33.334	2.382
2	Mini Batch	0.676	0.875	0.151	0.33	58.209	1.182
3	Agl Clust	0.549	0.975	0.153	0.268	41.663	1.349
4	DBSCAN	0.509	0.603	0.324	0.276	25.152	1.438
5	GMM	0.571	0.938	0.133	0.331	53.402	1.221
6	Spectral	0.414	0.883	0.227	0.322	56.708	1.194
7	Hybrid	0.676	0.964	0.16	0.288	46.013	1.275

KMeans has the best performance (1st in minimal distance between clusters, 3rd in average distance in clusters, 2nd in standard deviation in clusters)

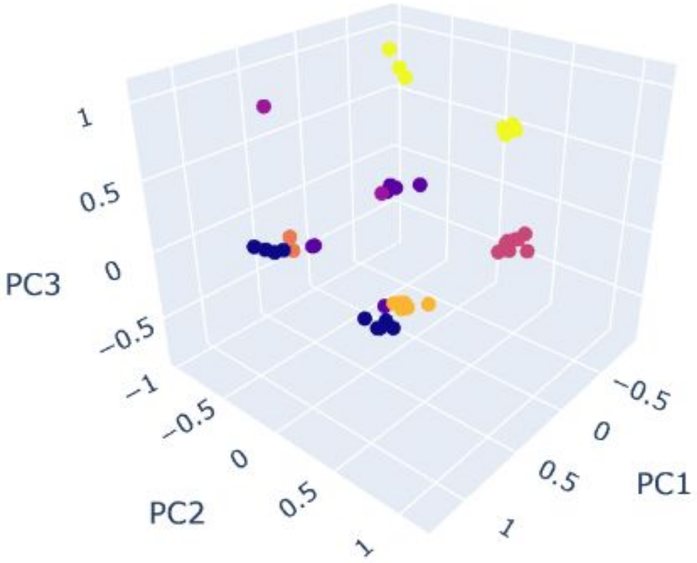
KMeans and principal components





On the visualizations we can see that our algorithm has accurately divided data into clusters

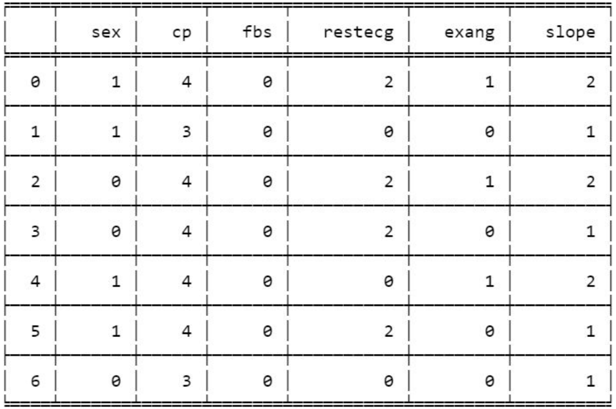
Testing



Min distance between clusters: **0.192**
Avg dist. between points in cluster: **1.163**
Std Dev of distance between points in cluster: **0.14**
Silhouette score: **0.14**
Calinski-Harabasz index: **7.179**
Davies Bouldin index: **1.728**

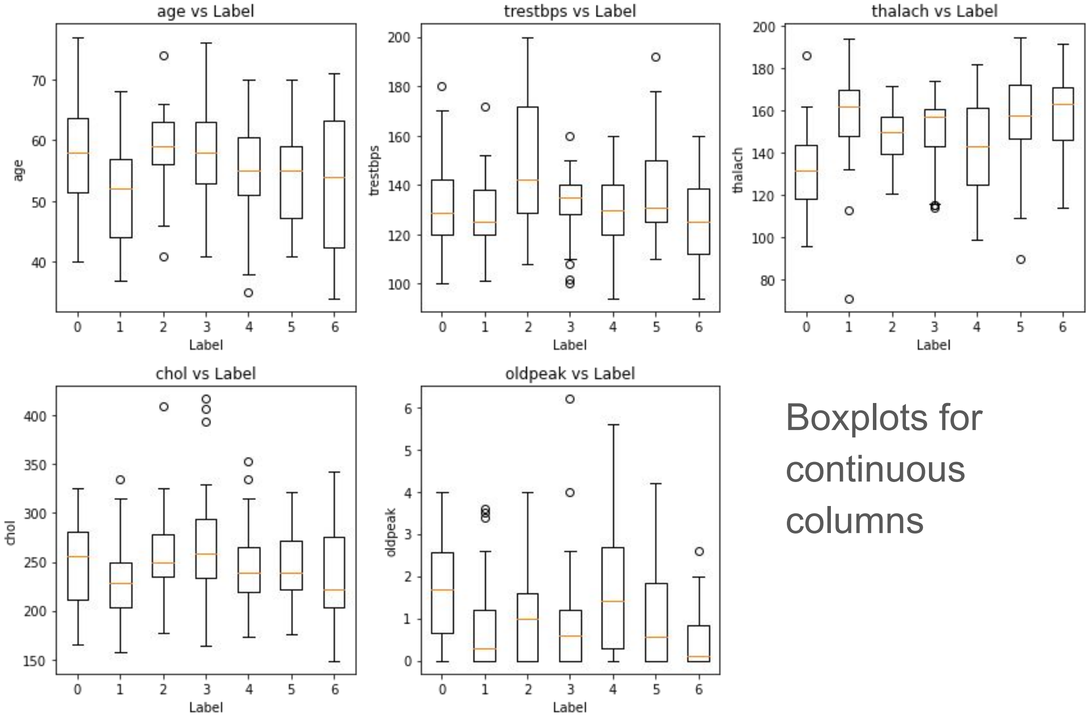
We can see that test data is also well clustered. Metrics are worse, but this is because this dataset is much smaller, so it cannot be directly compared.

Results



	sex	cp	fbs	restecg	exang	slope
0	1	4	0	2	1	2
1	1	3	0	0	0	1
2	0	4	0	2	1	2
3	0	4	0	2	0	1
4	1	4	0	0	1	2
5	1	4	0	2	0	1
6	0	3	0	0	0	1

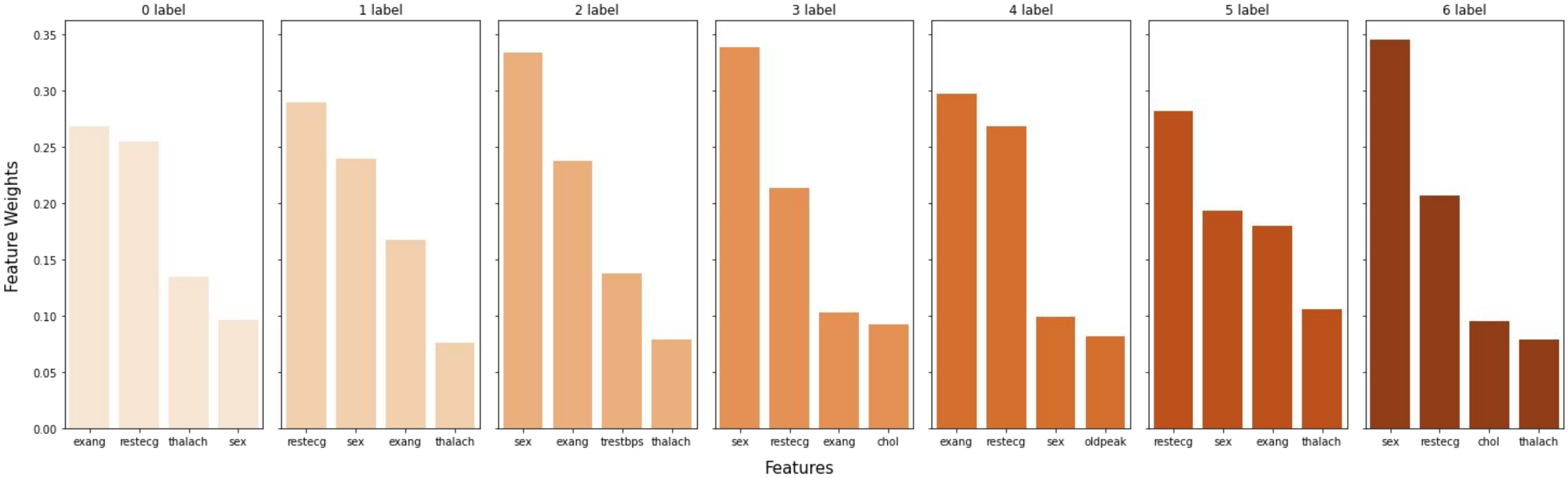
Modes for non-continuous columns



Boxplots for continuous columns

Characteristics of clusters vary, each cluster is distinguishable and has some specific features

Feature importance



In each case labels were binary encoded and fed to Random Forest classifier, which was classifying one class versus many. We extracted most important features for each class. We can see that sex, restecg, exang, thalach and chol are most significant variables.

Interpretability

By analyzing means, medians, boxplot charts and feature importance we created descriptions of specific clusters:

MEN:

- **cluster 0** - around 60 years old, low maximum heart rate, hypertrophy, exercise induced angina, asymptomatic chest pain
- **cluster 1** - around 50 years old, high maximum heart rate, low resting blood sugar, normal ECG, no exercise induced angina
- **cluster 4** - normal ECG, exercise induced angina, high oldpeak
- **cluster 5** - hypertrophy, high maximum heart rate, no exercise induced angina

WOMEN:

- **cluster 2** - around 60 years old, high resting blood pressure, high cholesterol, exercise induced angina
- **cluster 3** - around 60 years old, high cholesterol, hypertrophy, no exercise induced angina
- **cluster 6** - no diabetes, normal ECG, no exercise induced angina, small oldpeak

Validation - main concerns

- No introduction and train test split
  - Both things were added
- Replacing outliers
  - We abandoned this idea
- Functions overwriting data
  - Fixed functions and technical issues