Heart attack analysis & prediction

Joose Tikkanen

TIES4450 — Data mining and machine learning

Structure (I)

1. Knowledge mining process in short
2. The data selection, domain analysis and setting the goals
3. Data mining
4. Interpretation and evaluation

Structure (II)

Data mining steps:

- Preprocessing the data
- Data transformation and visualization for EDA
- Selecting the methods
- Mining
- Visualizing the results

The data, the domain & the goals

- Heart attack analysis & prediction dataset
- Subset of 14 attributes out of 76 from 303 patients
- Utilizing the first 13 features to predict the last target variable (heart disease or not)
- 9 classifying and 5 continuous features

The data, the domain & the goals

age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1

- sex – 0 = female, 1 = male
- cp – Chest pain type
  - 0 - Typical angina: Chest pain related to decreased blood supply to the heart
  - 1 – Atypical angina: Chest pain not heart related
  - 2 – Non-anginal pain: Typically esophageal spasms (non heart related)
  - 3 – Asymptomatic: Chest pain not showing signs of disease
- trtbps – Resting blood pressure (in mm Hg on admission to the hospital)
  - > 130-140 is typically a cause of concern
- chol – Serum cholesterol in mg/dl
  - > 200 is a cause for concern
- fbs – Fasting blood sugar > 120 mg/dl (1 = true, 0 = false)
  - > 126 mg/dl signals diabetes
- restecg – Resting electrocardiographic results
  - 0 – Nothing to note
  - 1 – ST-T Wave abnormality, non-normal heartbeat, can be mild symptoms or severe problems
  - 2 – Possible or definite left ventricular hypertrophy, enlarged heart's main pumping chamber
- thalachh – Maximum heart rate achieved
- exng – Exercise induced angina (1=yes, 0=no)

The data, the domain & the goals

age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1

- oldpeak – ST depression induced by exercise relative to rest
  - Stress of heart during exercise, unhealthy heart will stress more
- slp – The slope of the peak exercise ST segment
  - 0 – Upsloping: Better heart rate with exercise (uncommon)
  - 1 – Flatsloping: Minimal change (typically healthy heart)
  - 2 – Downsloping: Signs of unhealthy heart
- caa – Number of major vessels (0-3) colored by fluoroscopy
  - Signs of better blood movement (no clots), the doctor can see the blood moving
- thall – Thallium stress result, can show if the heart is damaged
  - 1 – Normal
  - 2 – "Fixed defect": Sign of damaged heart muscle
  - 3 – Reversible defect: No proper blood movement to the heart
- output – The predicted attribute: Is the patient diagnosed with heart disease? (1 = yes, 0 = no)

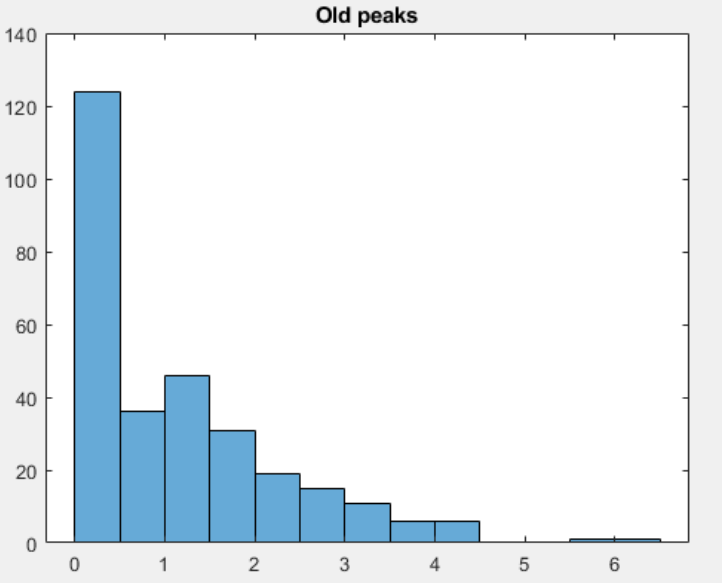
The data, the domain & the goals

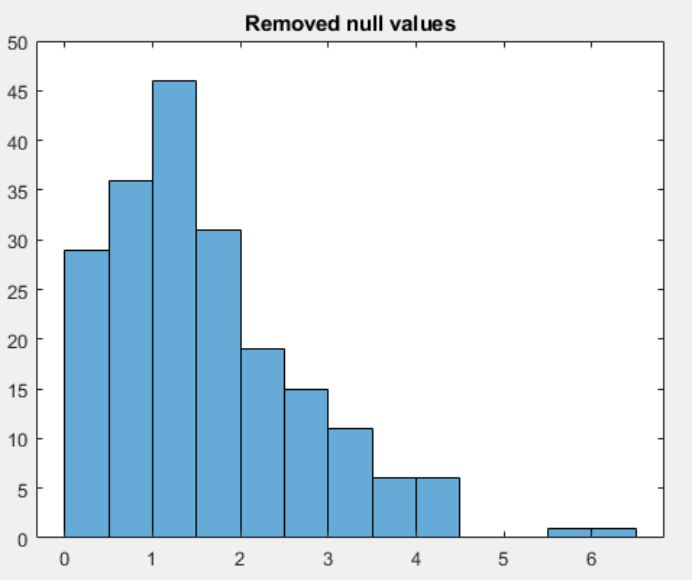
- Limitations of the project
  - Little prior knowledge about the domain
  - Poor explanations about the data
  - The quantity of data is somewhat low
- Goals for the KM process
  - Get understanding how different variables may affect the chance of heart disease
  - Try to get a model to predict if a person is prone to having a heart attack
  - Learn to use different DM techniques learned during the course

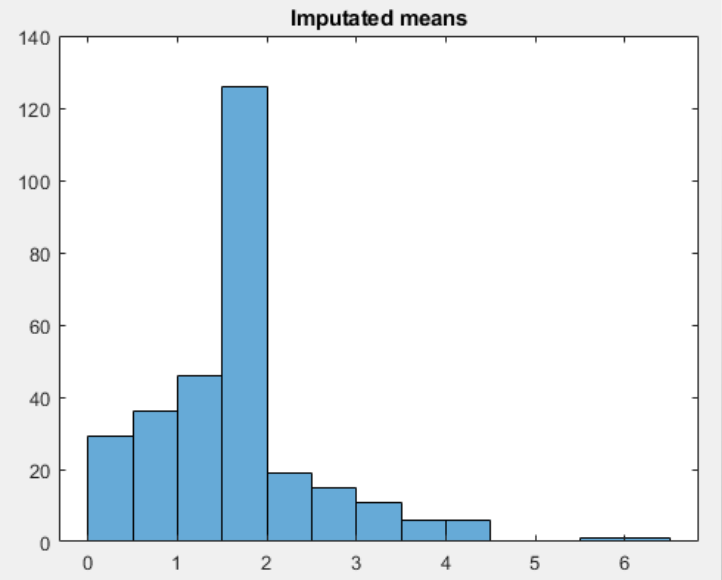
Data mining: Preprocessing

- Omitted separate o2Saturation data set – no explanation about the relation to the main data set
- Removed observations with null values for thall – 2 rows removed
- Removed observations with null values for caa – 5 rows removed
- Problem with oldpeak variable – 95 values of zero (~32 % of the observations)
  - Relevant or not measured?

Data mining: Preprocessing







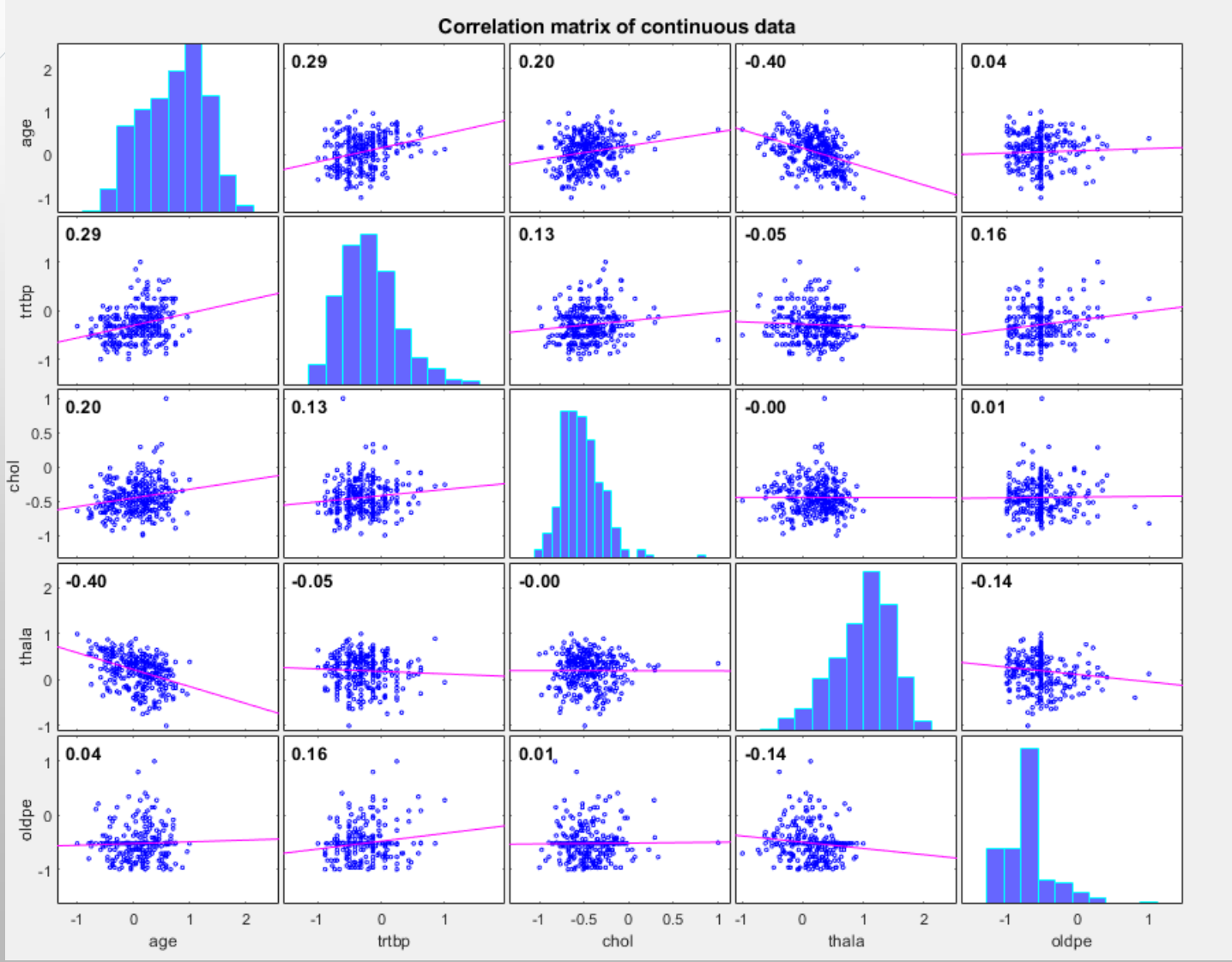
Missing values handled with mean imputations

- Better ways to do imputation?

Data mining: Preprocessing

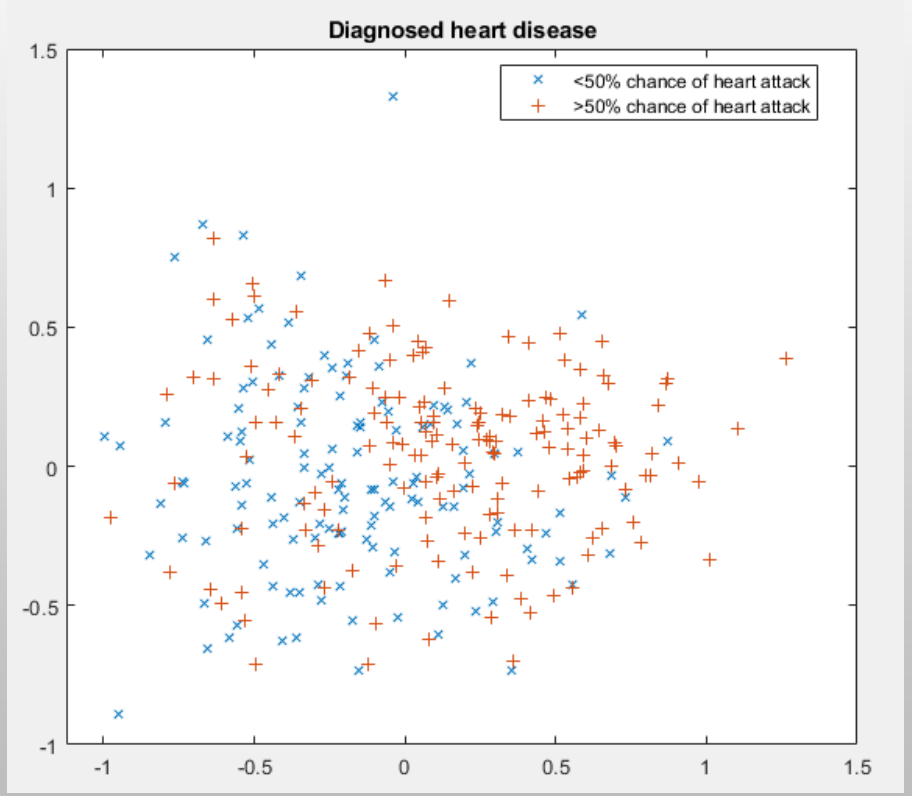
- Separated the target variable (last column) from the other predictor features
- Scaled the whole data to the range of [-1, 1]
- Separated the continuous and categorical variables for EDA
  - 5 continuous and 9 categorical variables, how to visualize?

Data mining: Data transformation and EDA

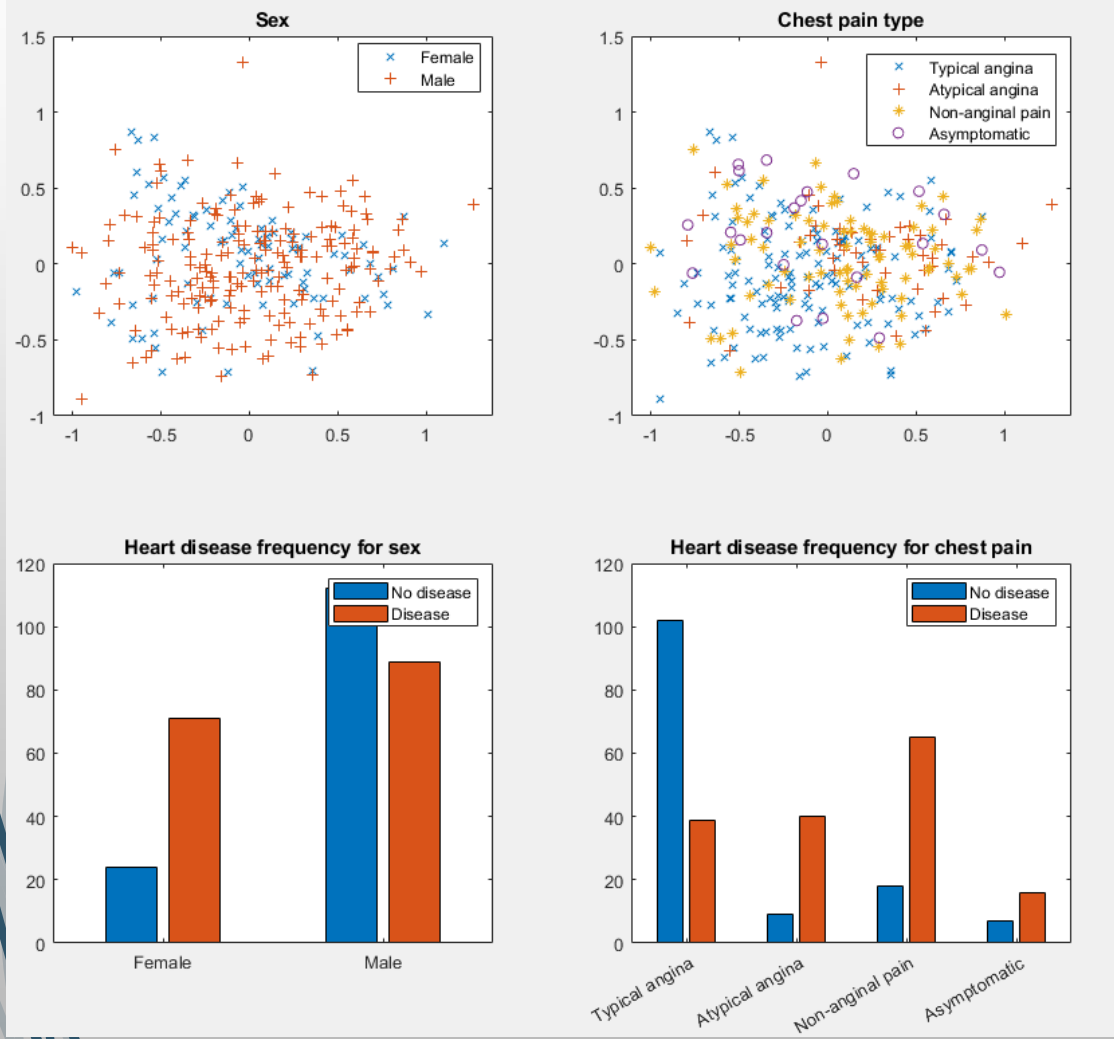


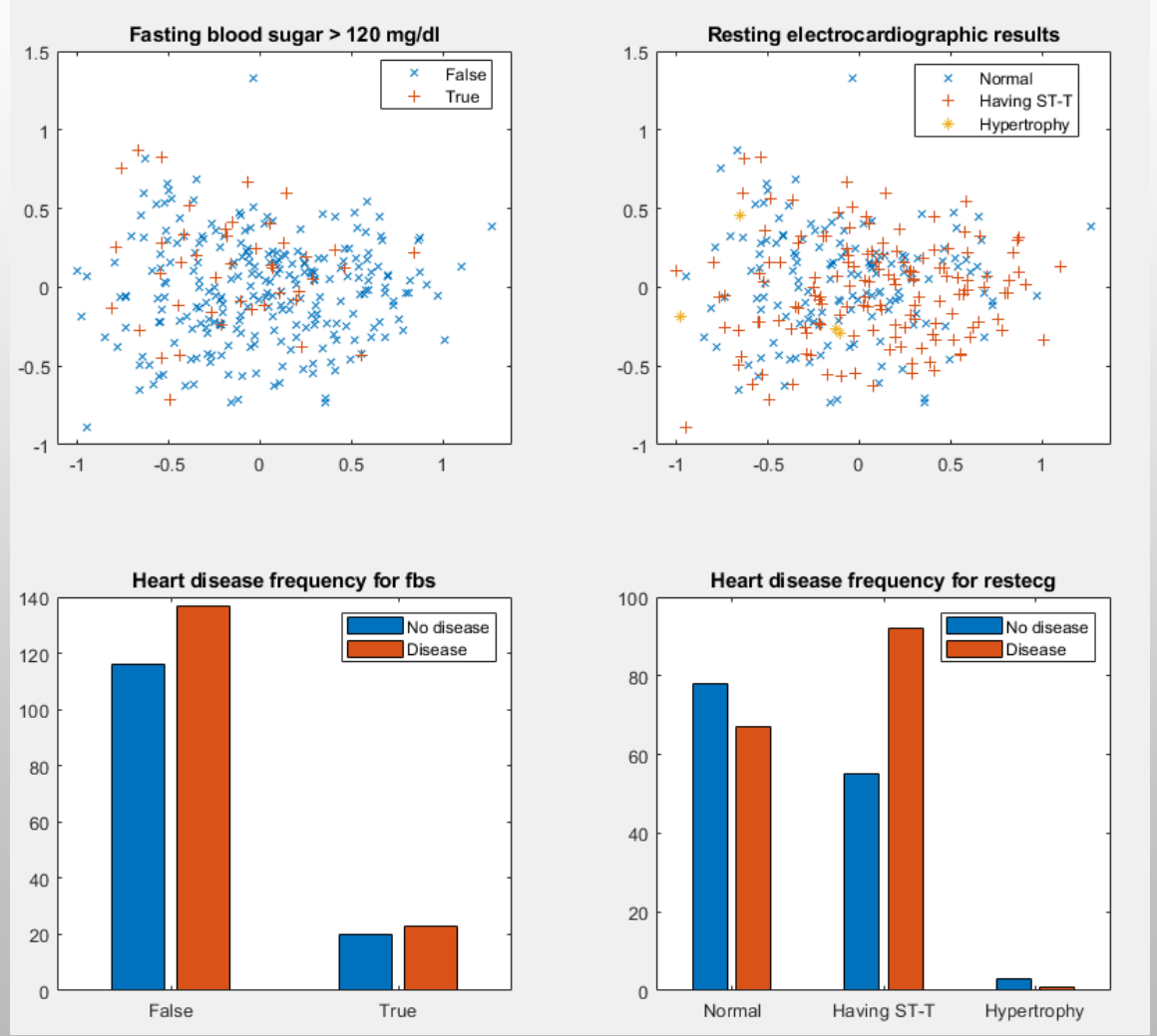
Data mining: Data transformation and EDA

- Solution: Visualize all of the continuous data in 2D with PCA
- PCA: Determine new (2D) coordinate system that represents the original distances of the data points as accurately as possible
- Reflect the categorical data to the data points to explore their correlations visually

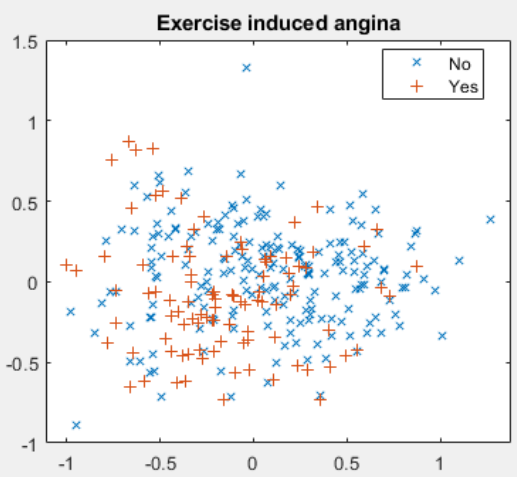


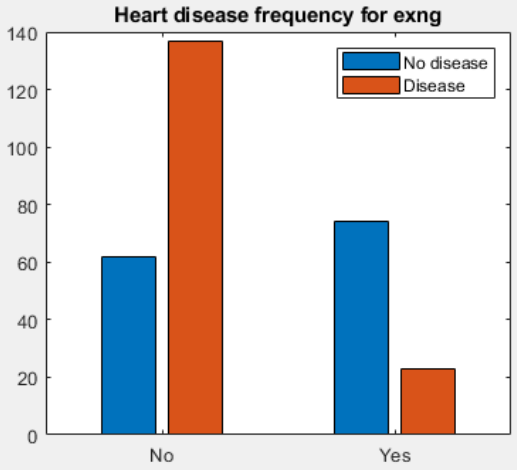
Data mining: Data transformation and EDA

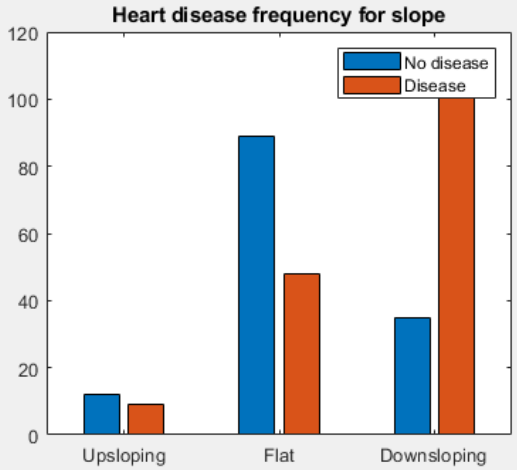


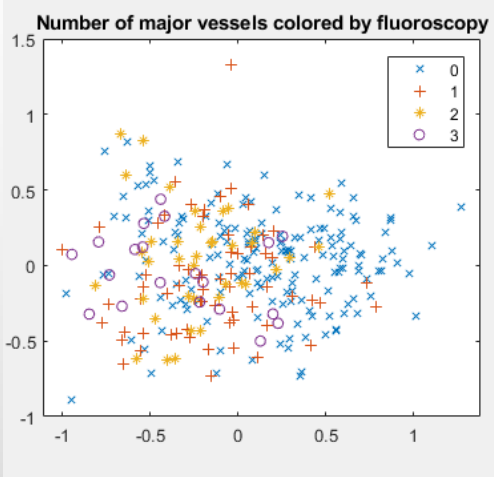


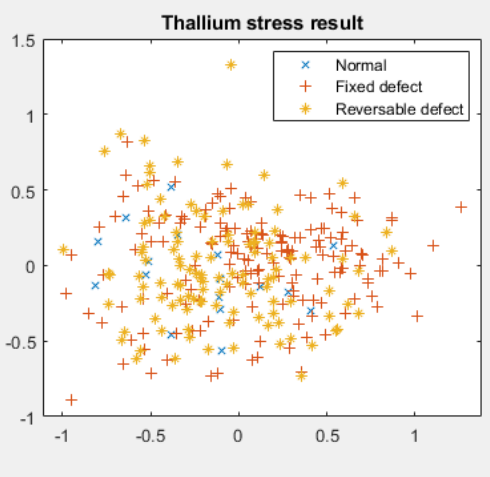
Data mining: Data transformation and EDA

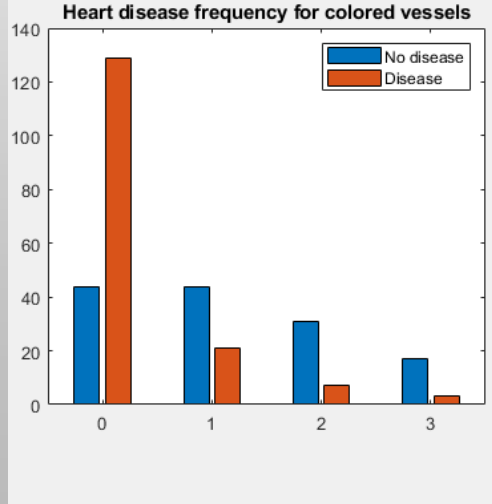


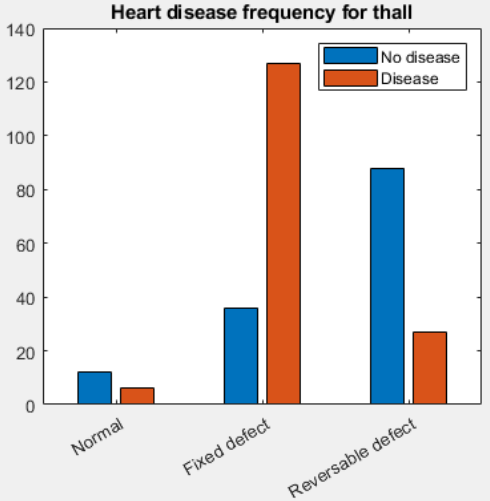








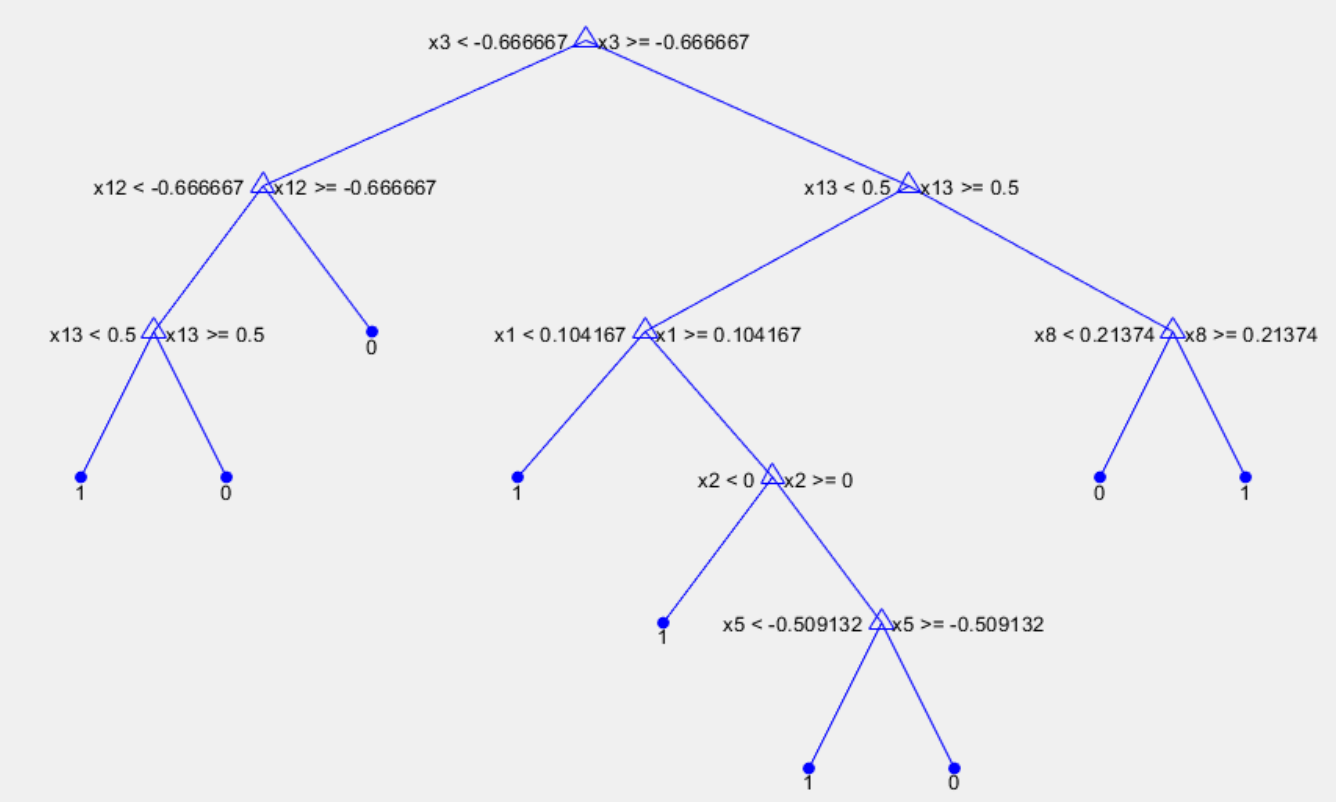


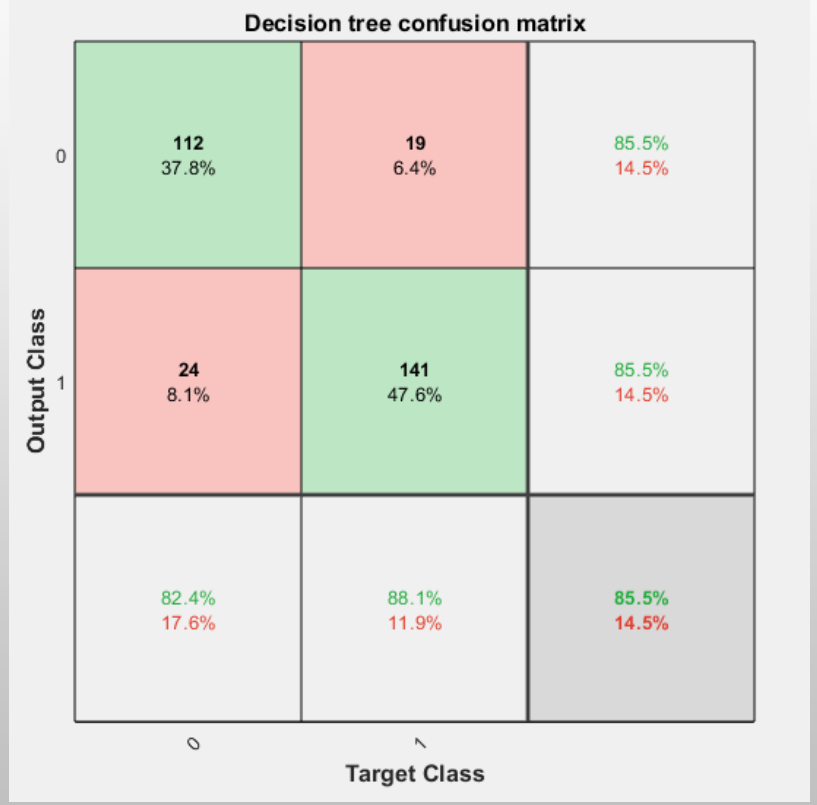


Data mining: Method selection & Mining

- Predictive modelling with classification
- Using whole dataset (13 input variables, 1 output)
- Comparing different classifiers, which is the most accurate?
  - Decision tree
    - What are the best attributes that can predict heart disease?
  - k-Nearest Neighbors
    - 10-fold cross validation to determine k
  - Naive Bayes
    - Normal density estimation
  - LDA
    - Maximally separating hyperplane between classes

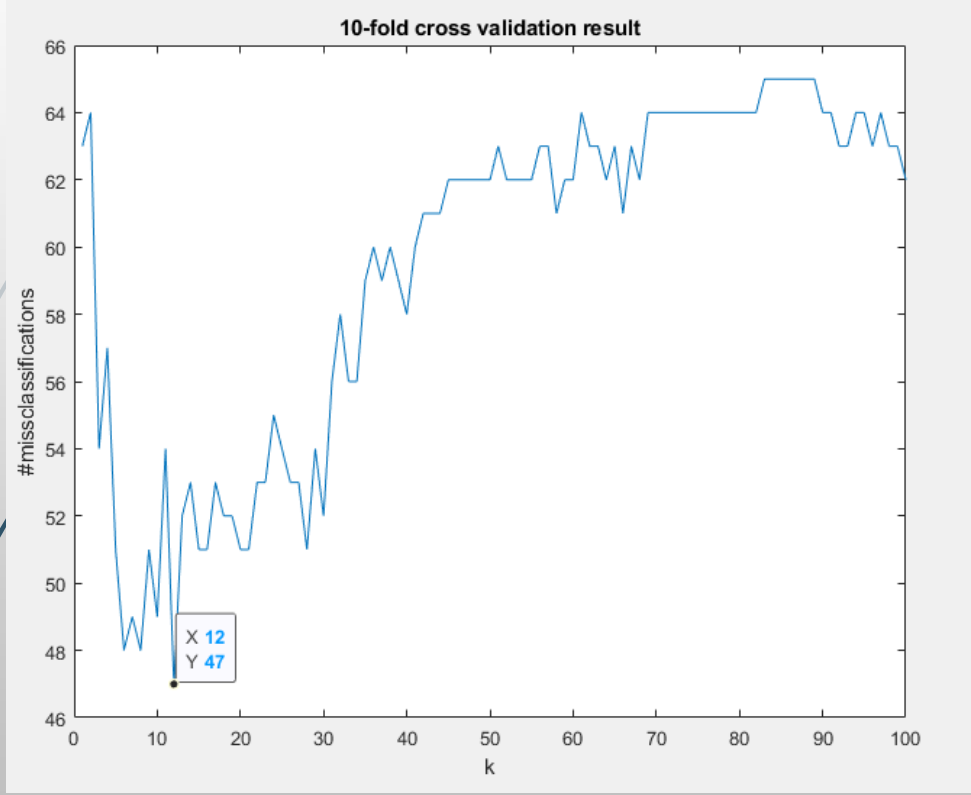
Data mining: Decision tree classifier

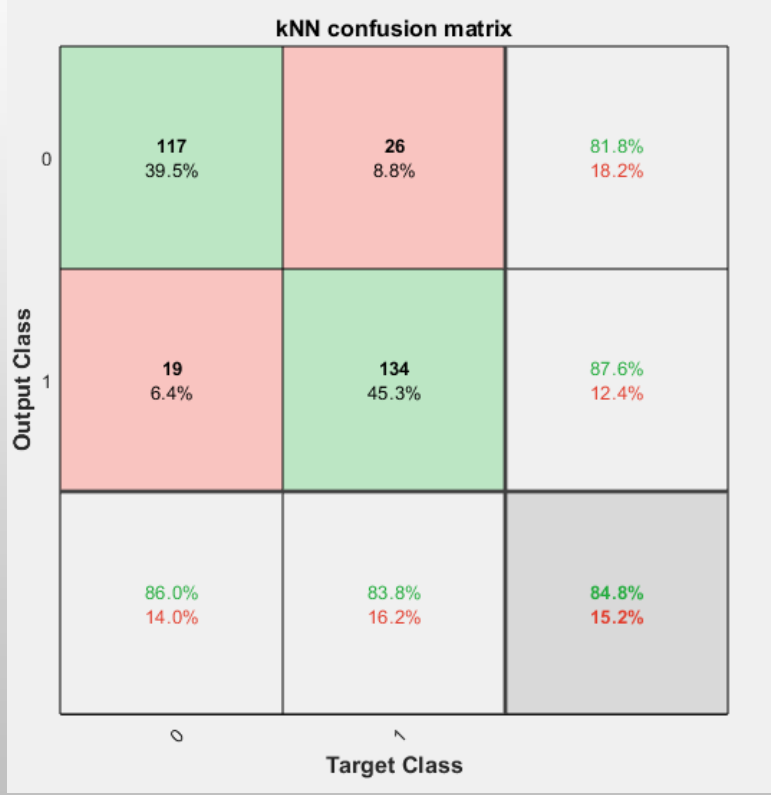




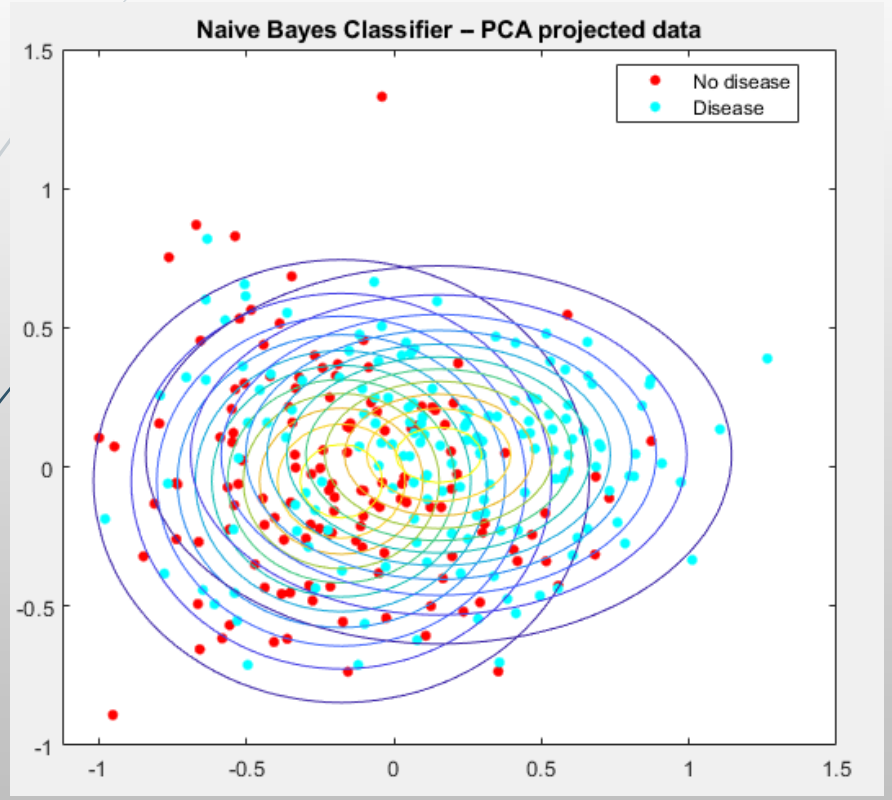
Most significant predictors according to the tree are chest pain type, thallium stress result and number of major vessels colored by fluoroscopy, followed by age and maximum heart rate

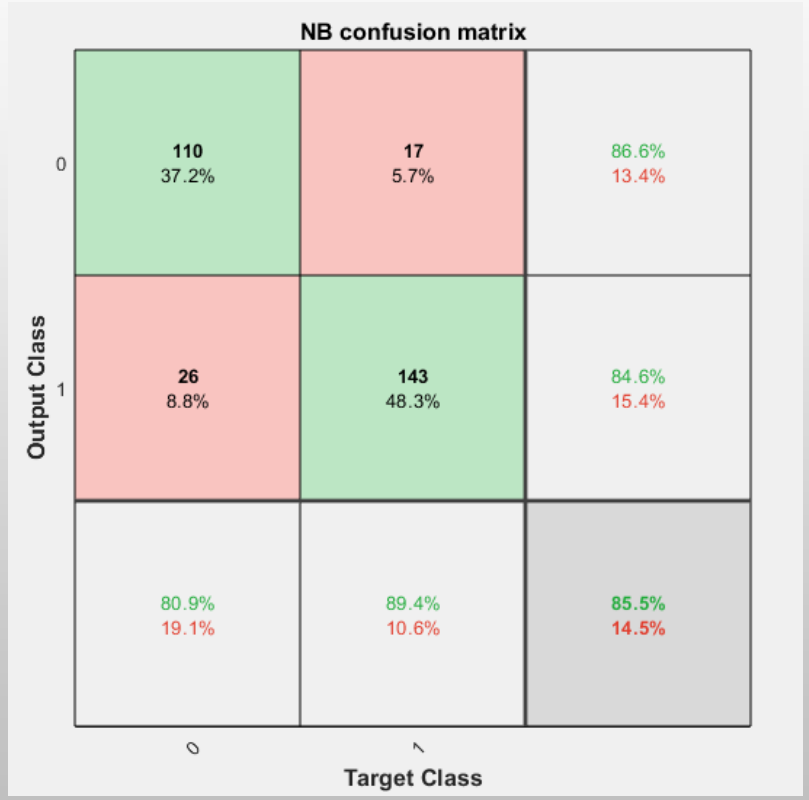
Data mining: k-NN classifier



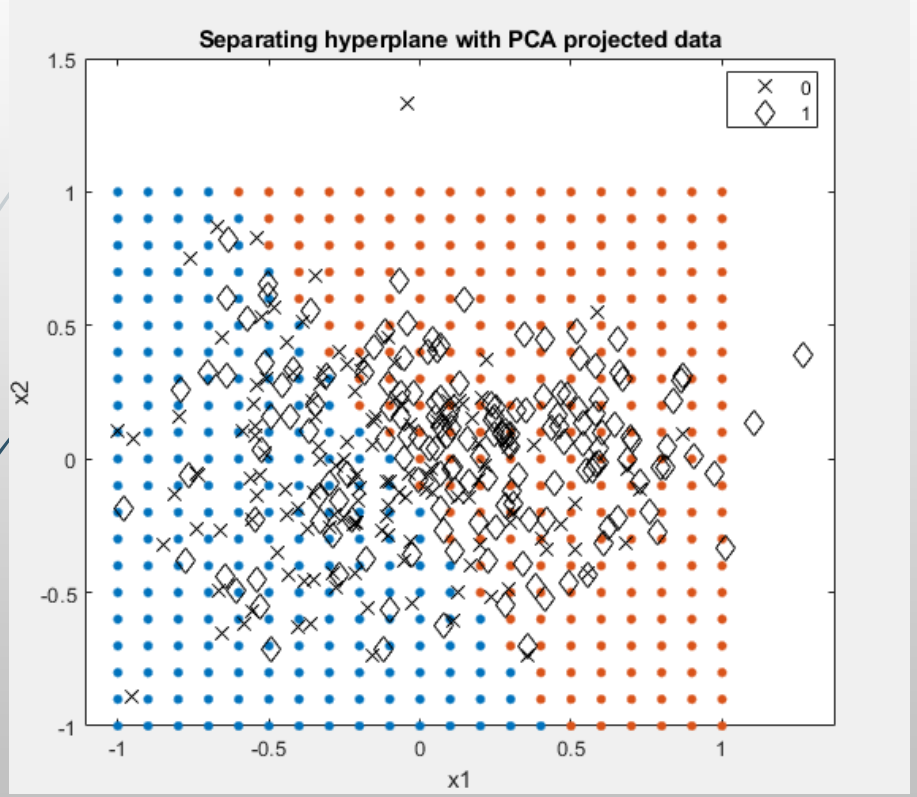


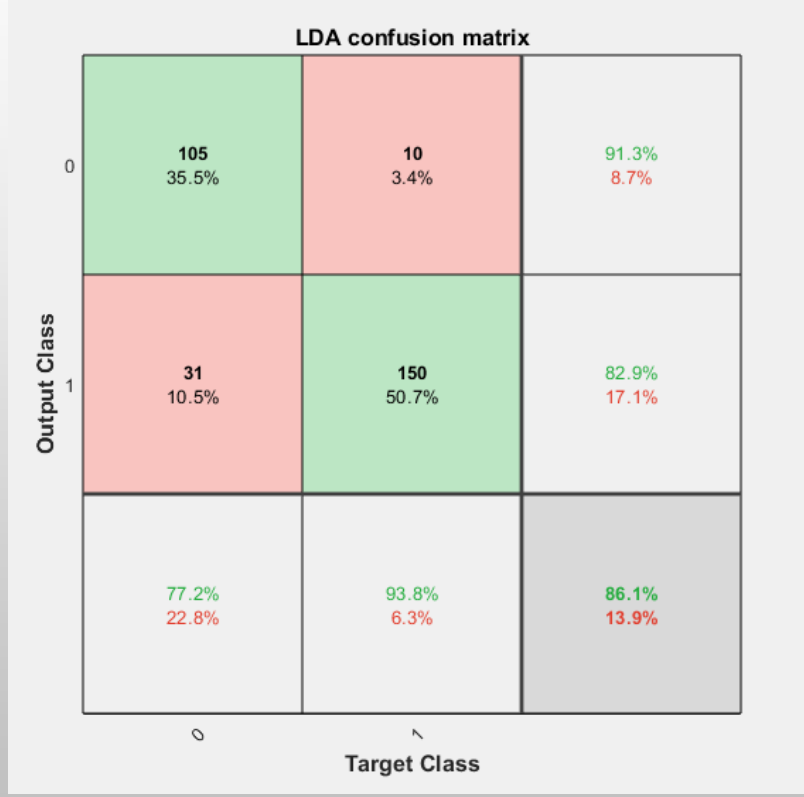
Data mining: Naive Bayes classifier





Data mining: LDA classifier





Interpretation and evaluation

- None of the classification methods reached 90 % accuracy
  - Classification models might need more tuning
  - Quantity of the data might be too low
- The most accurate classifier was LDA, decision tree was close
- Classifiers tend to give more false positives (with exception of kNN)

Interpretation and evaluation

- Some of the results were expected
  - Increase in age, resting blood pressure, serum cholesterol, more stressing heart and max heart rate tend to increase chances of having a heart disease
  - If no major vessels can be seen with fluoroscopy -> very likely to have a heart disease
  - Signs of damaged heart muscle increases chances of a heart disease
- Some unexpected findings
  - Heart related chest pain indicated no heart disease
  - Reversible defect from thallium stress result (no proper blood movement to heart) indicated no heart disease
  - Age < 55 indicated of having a heart disease

References

- Data set: https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset
- Zaki, M. J., Meira Jr, W., & Meira, W. (2014). Data mining and analysis: fundamental concepts and algorithms. Cambridge University Press.