Analysis of Heart Disease Data

Presentation on the 07.12.2022

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Eric Jacquomé, Finn Hülsbusch, Lasse Lemke, Thilo Dieing, Timo Gumpp

Why did we choose the Topic?

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Heart diseases are one of the highest causes of mortality on earth
- The healthcare sector is information rich but knowledge poor
- Applicable in the real world
- Goal: predict if patient has a heart disease or not (Classification)

Dataset Description
Data Origin

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Published by the University of California (UCI) in 1988
- Consists of 4 unsorted subsets:
  - Switzerland
  - Hungary
  - USA Cleveland
  - USA Long Beach
- Total of 495 positive & 404 negative = 899 measurements

Dataset Description
Features

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Total of 76 attributes (33 Numeric 42 Categorical 1 Constant)
  - Patient data
  - (Exercise) Electrocardiogram
  - Cardiac uoroscopy
- 21245 (31%) missing values

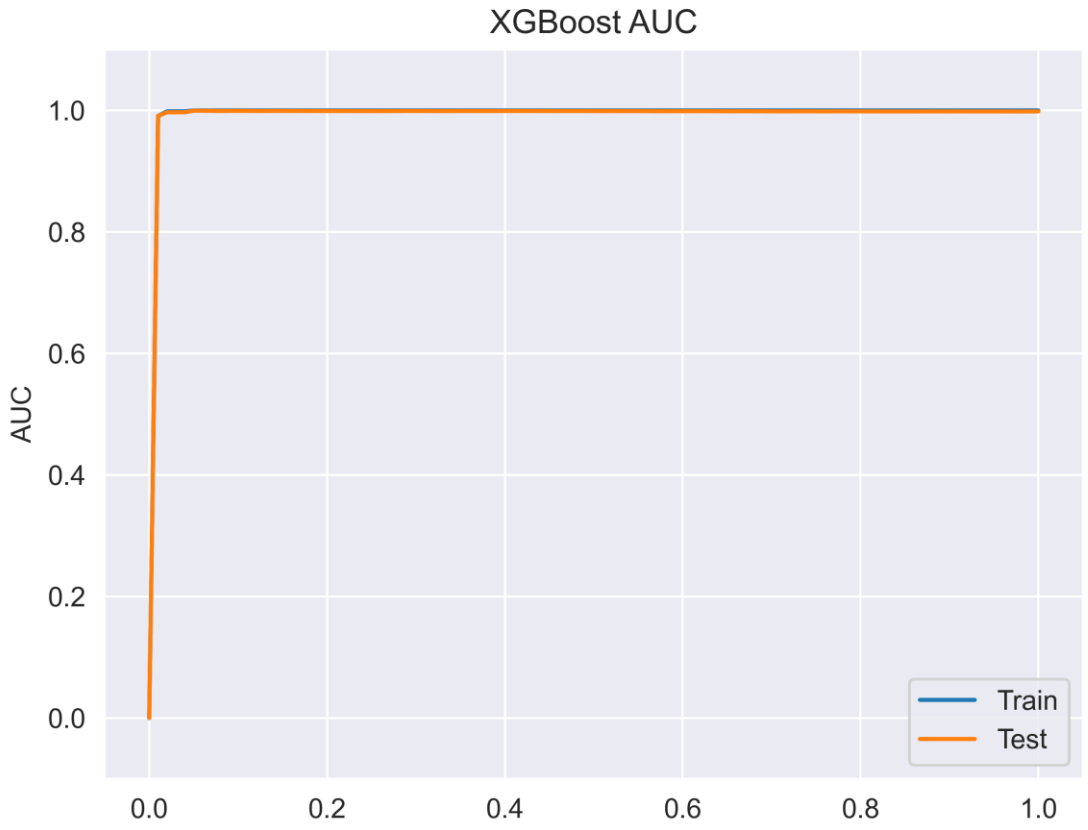
Data Understanding
Dropped Features

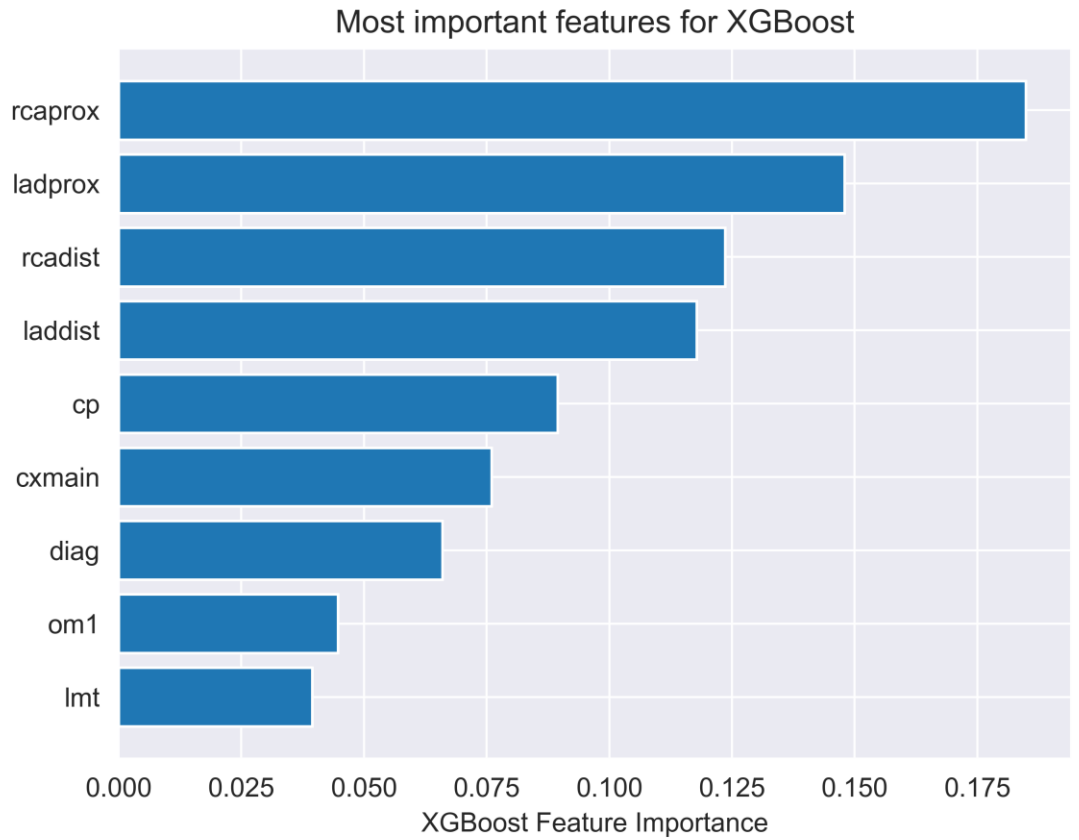
UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- IDs
- Dates of measurements
- Names
- Constants
- Irrelevant columns according to UCI
- Unspecified columns

Data Understanding
False Predictors

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics





Data Understanding
Dropped Features

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- IDs
- Dates of measurements
- Names
- Constants
- Irrelevant columns according to UCI
- Unspecified columns

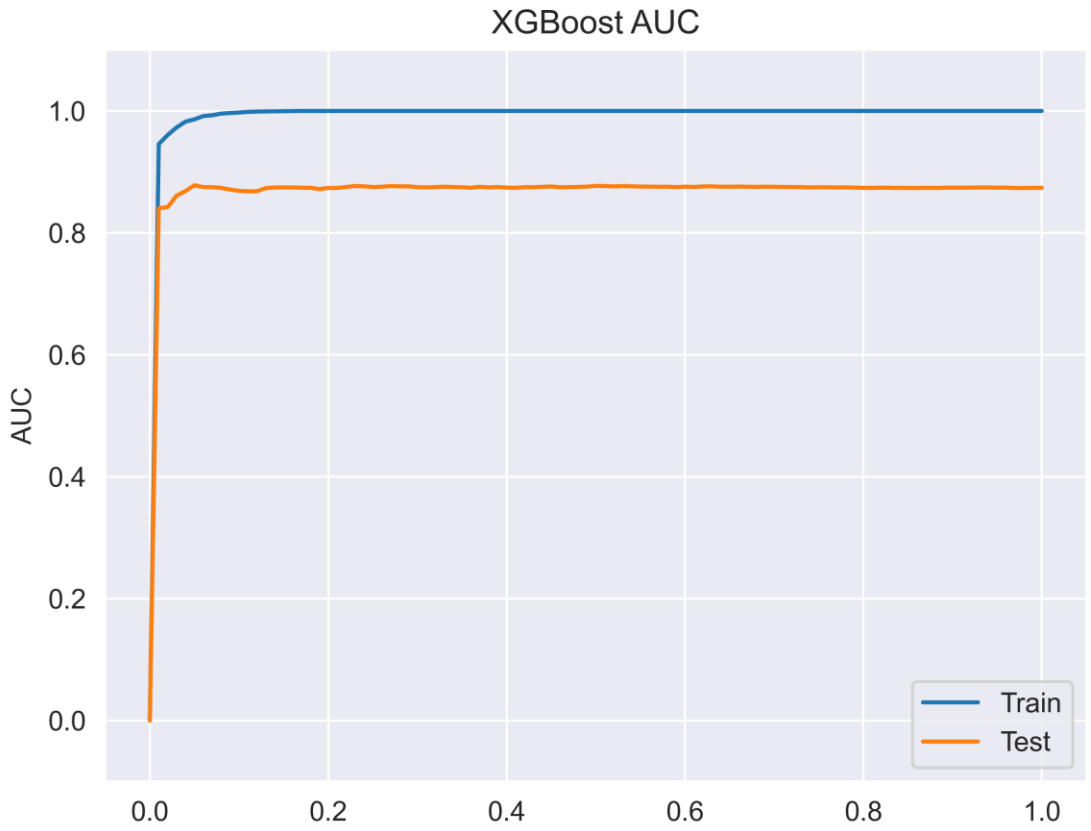
Data Understanding
Dropped Features

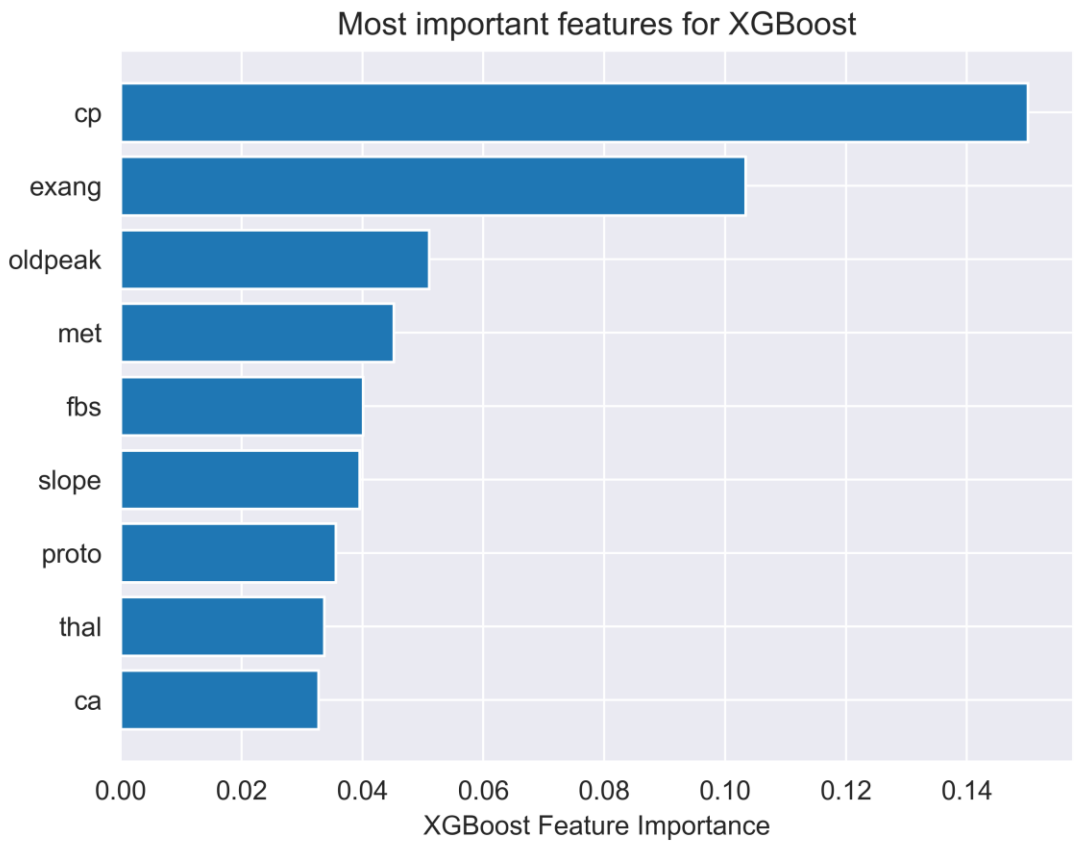
UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- IDs
- Dates of measurements
- Names
- Constants
- Irrelevant columns according to UCI
- Unspecified columns
- Coronary arteries

Data Understanding
False Predictors

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics





Data Understanding

Missing Features / Feature Subset Selection

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Pipeline Hyperparameters

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Name	Factor
minimum_percentage_to_be_dropped	8

Data Preparation

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- 4 Datasets -> Different encodings -> harmonised
- Validity tests
  - Cholesterin = 0
  - Blood pressure= 0

Data Preparation

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- 4 Datasets -> Different encodings -> harmonised
- Validity tests
- Outliers
  - Values outside the specified value range of UCI
  - No consideration of medical outliers

Data Preparation

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- 4 Datasets -> Different encodings -> harmonised
- Validity tests
- Outliers
- Check for inconsistencies
  - Maximum Heartrate < resting Heartrate
  - TimepointOfPeakHeartRateWhileExercising < ExerciseDuaration
  - isSmoking, NumberOfYearsSmoking, AvgNumberOfCigarettes

Data Preparation

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- 4 Datasets -> Different encodings -> harmonised
- Validity tests
- Outliers
- Check for inconsistencies
- Generate delta between rest and peak measurements
  - heart rate (resting against peak)
  - ECG (resting against exercise)

Data Preparation

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- 4 Datasets -> Different encodings -> harmonised
- Validity tests
- Outliers
- Check for inconsistencies
- Generate delta between rest and peak measurements
- OneHotEncoding applied to 5 features

Data Preparation

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- 4 Datasets -> Different encodings -> harmonised
- Validity tests
- Outliers
- Check for inconsistencies
- Generate delta between rest and peak measurements
- OneHotEncoding applied to 5 features
- Binning for age

Pipeline Hyperparameters

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Name	Factor
minimum_percentage_to_be_dropped	8
age_binning	3

Scaler and Estimator Selection

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- We had no idea. So why don‘t we take everything?
  - Let‘s go…

Pipeline Hyperparameters

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Name	Factor
minimum_percentage_to_be_dropped	8
age_binning	3
estimators	10
scaler	7

Pipeline Hyperparameters

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Name	Factor
minimum_percentage_to_be_dropped	8
age_binning	3
estimators	10
scaler	7
sampler	3

Pipeline Hyperparameters

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Name	Factor
minimum_percentage_to_be_dropped	8
age_binning	3
estimators	10
scaler	7
sampler	3
Imputer	3

Evaluation

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Low number of samples
  - Nested stratified Cross-Validation with 10 folds
  - Saving classification report for outer loop

Pipeline Hyperparameters

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Name	Factor
minimum_percentage_to_be_dropped	8
age_binning	3
estimators	10
scaler	7
sampler	3
imputer	3
nestedCV	100

Pipeline Hyperparameters

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Name	Factor
minimum_percentage_to_be_dropped	8
age_binning	3
estimators	10
scaler	7
sampler	3
imputer	3
nestedCV	100
hyperparameters (model)	~30

Total number of fits

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Name	Factor
minimum_percentage_to_be_dropped	8
age_binning	3
estimators	10
scaler	7
sampler	3
imputer	3
nestedCV	100
hyperparameters (model)	~30
Total	**~45360000**

$$45360000 \cdot 0{,}01s = 453600s = 126h$$

Results

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Baseline: Every patient has a heart disease
  - Accuracy: 495/899 = 0,55
- Best configuration:
  - Full grown DecisionTree with gini as criterion
  - Minimum percentage to be dropped: 0
  - 2 bins for age (<=52,>52)
  - No scaling and sampling
  - Accuracy: 0,77

Results KNN
Confusion Matrix

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Leave-One-Out Cross-Validation

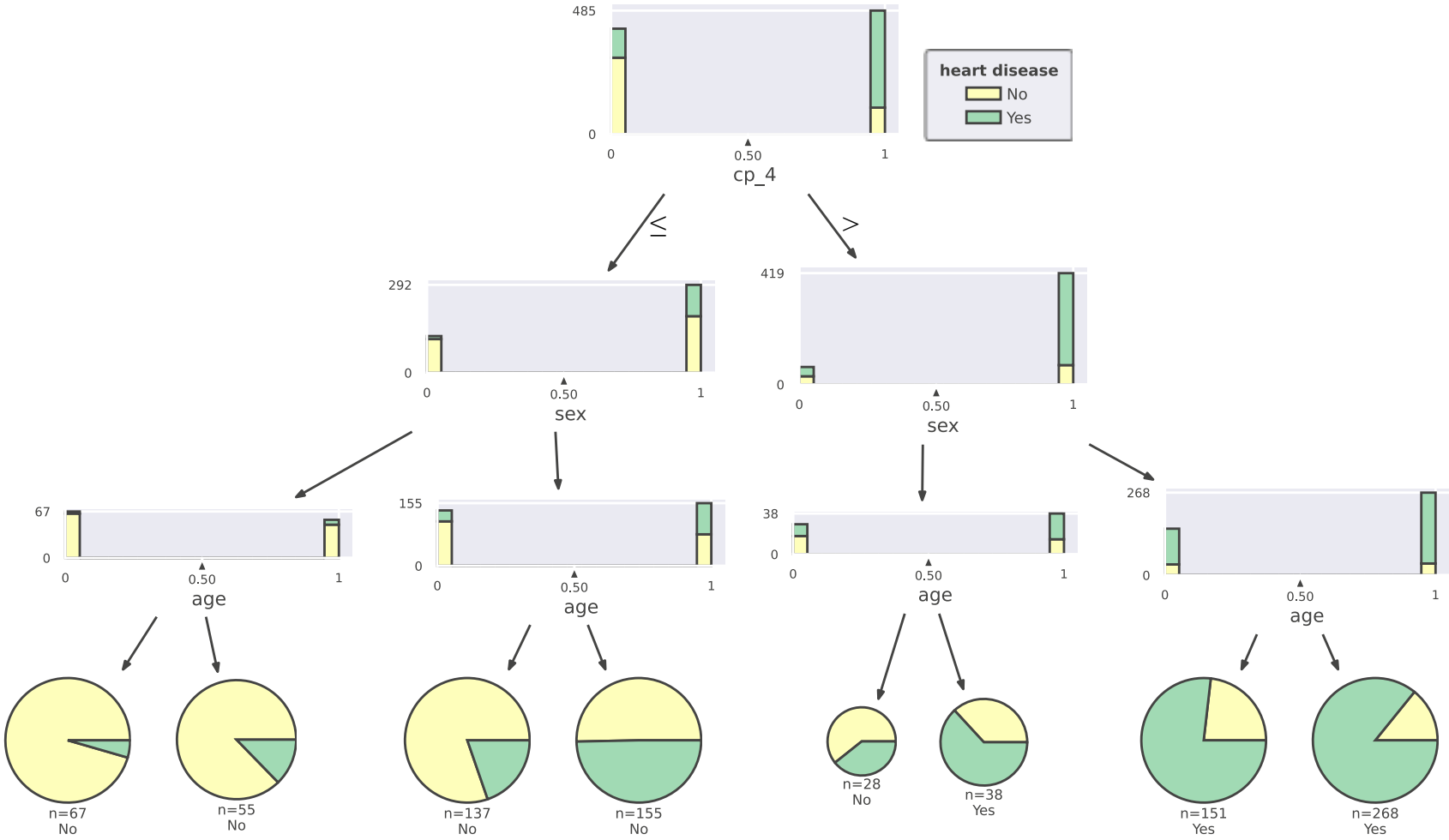
	Heart disease	No heart disease
Heart disease	431	64
No heart disease	134	270

Leave-One-Group-Out Cross-Validation

	Heart disease	No heart disease
Heart disease	419	76
No heart disease	146	258

Visual result

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics



Question?

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Results

Classification Report

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Leave-One-Out Cross-Validation

	Precision	Recall	F1-score	support
No disease	0,81	0,67	0,73	404
Disease	0,76	0,87	0,81	495
Accuracy			0,78	899
Macro avg	0,79	0,77	0,77	899
Weighted avg	0,78	0,78	0,78	899

Leave-One-Group-Out Cross-Validation

	Precision	Recall	F1-score	support
No disease	0,77	0,64	0,70	404
Disease	0,74	0,85	0,79	495
Accuracy			0,75	899
Macro avg	0,76	0,74	0,74	899
Weighted avg	0,76	0,75	0,75	899

Learning 1

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

```
df[df['column'] > 1] = 1
df.loc[df['column'] > 1, 'column'] = 1
```

Learning 2

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

```
one_hot_encoded_features = ['cp', 'restecg', 'slope', 'ca', 'restwm']
discretizeParameters = {
    'columnTransformer__discretize': [
        'passthrough',
        KBinsDiscretizer(2, encode='ordinal', strategy='uniform'),
        KBinsDiscretizer(5, encode='ordinal', strategy='uniform')]
}
columnTransformer = ColumnTransformer(
    transformers=[
        ('discretize', KBinsDiscretizer(), ['age']),
        ('oneHotEncoder', OneHotEncoder(handle_unknown='ignore'), lambda X: [value for value in one_hot_encoded_features if value in X.columns]),
    ], remainder='passthrough')
```