Mandatory Assignment 2

Classification & Clustering

Markus Kinn

**Classification:**

**Dataset 1:**

For the first dataset I chose to continue with the titanic dataset from the first mandatory assignment. The goal is the same as before; predict whether a person was going to survive the titanic disaster or not.

The final dataset (after preprocessing) contains 15 columns:

Survived(the ground truth), Sex, FamilyMembersCount, Pclass\_1, Pclass\_2, Pclass\_3, Embarked\_C, Embarked\_Q, Embarked\_S, AgeGroup\_0, AgeGroup\_1, AgeGroup\_2, AgeGroup\_3, AgeGroup\_4, AgeGroup\_5.

**Dataset 2:**

The second classification dataset I used was: Heart Attack Analysis & Prediction Dataset. (<https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset>). This dataset is used to predict if a person is likely to have a heart attack given current metrics.

There are 10 columns (including ground truth):

Age: age of patient, Exng: exercised induced angina, Ca: number of major vessels, Cp: Chest Pain type, Trtbps: resting blood pressure, Chol: cholestoral, Fbs: fasting blood sugar, Rest\_ecg: resting electrocardiographic results, Thalach: max heart rate achieved, Target: 0 = lower chance of heart attack and 1 = higher chance of heart attack

**Clustering:**

**Dataset 1:**

For the clustering part of the assignment I chose to use the same clustering dataset form the first mandatory assignment: Airline Passenger Satisfaction (<https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction>). This dataset is usually used as a classification dataset with the ground truth: Satisfaction, but for this task the ground truth has been removed.

This dataset contained a total of 104.000 rows and 25 columns. This lead me to removing 90.000 rows for both the Birch algorithm and Agglomerative.