Mandatory Assignment 1

Datasets & Data Processing

Markus Kinn

**Classification:**

**The dataset:**

The dataset I used for the classification task is: The Titanic Disaster Dataset (<https://www.kaggle.com/competitions/titanic/data>). The goal of this dataset is to predict whether a person, given some attributes, was going to survive the Titanic Disaster. The given attributes are: PassengerId, Pclass, Name, Age, SibSp, Parch, Ticket, Fare, Cabin and Embarked. In this notebook I process the data and look for correlations between the independent variables and the dependent variable.

**Data processing:**

In notebook

**Exploratory Data Analysis:**

In notebook

**Regression:**

**The dataset:**

The dataset I used for the regression task is: Medical Cost Dataset. (<https://www.kaggle.com/datasets/mirichoi0218/insurance>). The goal of this dataset is to predict how much people would most likely spend on treatment at a hospital based on the independent variables: age, sex, bmi, children, smoker and region. In this notebook I process the data and look for correlations between the independent variables and the dependent variable.

**Data processing:**

In notebook

**Exploratory Data Analysis:**

In notebook

**Clustering:**

**The dataset:**

The dataset I’m using for this clustering task is: Airline Passenger Satisfaction (https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction). This dataset is usually used for classification, with the goal of predicting whether a customer was satisfied with its flight experience. For this task I’ve removed the dependent variable/ the ground truth to convert it to a clustering problem. The rest of the dataset contains scores for different aspects of the flight experience.

**Data processing:**

In notebook

**Exploratory Data Analysis:**

In notebook