Biomedical Data Science & AI

Exercise sheet 1 - Introduction

SoSe2025

Prof. Dr. Holger Fröhlich, Jannis Guski, Sebastian Schwick, Diego Valderrama

Due date: Apr 15th

Questions

Exercise 1 - Descriptive Statistics & Data Visualization (total: 11 points)

1. Load the Iris dataset into your notebook from Scikit-Learn. (2 points)
2. Report the descriptive statistics of the features of the iris dataset. (3 points)
   a. Mean, Median, Mode
   b. Variance, MAD, Standard deviation
   c. Quantiles, IQR
3. Plot a density plot for each of the variables. Interpret the plots. (2 points)
4. Create a violin plot for the sepal width feature for each class. What can be seen from the plots? (2 points)
5. Combine your violin plot with boxplots as shown in the lecture. (2 points)

Exercise 2 - Data Pre-processing (total: 11 points)

1. Load the *banknote authentication* dataset from the given *data_banknote_authentication.csv* file. How many rows and columns does the dataset contain? (2 points)
2. Mention the different types of variables. Which types does your dataset contain? (2 point)
3. Count the number of duplicate rows in the dataset. How can you remove the duplicate rows? (2 points)
4. Count the number of missing values in the dataset. (1 points)
5. How can you deal with missing values in your dataset? Implement one of the possible methods (2 points)
6. Based on the dataset you get after dealing with missing values, create a parallel coordinates plot. Color the lines based on class assignment. (2 points)

Exercise 3 – Scatterplot (total: 3 points)

1. Load the dataset from the given *dataset.csv* file. (1 points)
2. Plot the scatterplot matrix for the given dataset. What can be seen in the scatterplot matrix? (2 points)