Biomedical Data Science & AI

Exercise sheet 1 - Introduction

SoSe2021

Prof. Dr. Holger Fröhlich, Mohamed Aborageh, Vinay Bharadhwaj, Yasamin Salimi

Due date: Apr 27th

Questions

Exercise 1 - Descriptive Statistics & Data Visualization (total: 9 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)

Exercise 2 - Data Pre-processing (total: 9 points)

1. Load the heart dataset from the given *heart.csv* file. How many rows and columns does the dataset contain? (2 points)
2. How many unique values does each column contain? (1 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 (3 points)

Exercise 3 - Correlation (total: 7 points)

1. Load the dataset from the given *dataset.tsv* file. (1 points)
2. Plot the scatterplot matrix for the given dataset. What can be seen in the scatterplot matrix? (2 points)
3. Which correlation would suit the comparison of feature_1 and feature_3? Calculate the relevant correlation coefficient for the 2 features. (2 points)
4. Plot the correlation heatmap of the entire dataset. (2 points)