Machine Learning SS2025

Homework 1 – kNN

This homework consists of two parts worth 4 points in total:

- **Coding part** (use **both R and Python** for E1-E3, each part is worth **1 point**, non-runnable or uncommented code is worth 0 points)
- **Non-coding part (2 points)**

For this homework you'll use the provided *"Fisher's Iris Dataset"* (**iris.csv**).

Part 1 – Coding *(2 points)*

Load the data set and solve the following tasks:

- **E1**: Load the iris dataset and select only entries of the classes "**iris virginica"** or "**iris versicolor"** (so we have a binary classification problem).

- **E2**: Use the kNN-classes of sklearn in Python and the caret package in R with K=5 and a train-test-split of 70-30 for an initial classification and calculate the accuracy using the test set. Explain any discrepancies between the results from Python and R.

- **E3**: Use the extensive search approach to identify a good k, visualize the accuracy for all k you tried via a plot, and explain your choice of a "good k". Additionally, discuss the possible impact of different distance metrics on the classification performance.

Part 2 – Non-Coding *(2 points)*

- **E4**: Elaborate on the strengths and weaknesses of the kNN-classifier and relate them to your results in E1-E3. Provide examples of real-world scenarios where kNN would be an appropriate or inappropriate choice, and justify your reasoning.
- **E5**: Discuss potential biases in the iris dataset and how they might affect the classification results.