UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Data and Web Science Group
B6, 26 – B1.15
68159 Mannheim

Data Mining- Python

Exercise 1: Simple Preprocessing and Visualization

1.1. Python Installation

Download Anaconda and install the software on your laptop. Additional information on how to install Anaconda can be found here: https://docs.anaconda.com/anaconda/install/

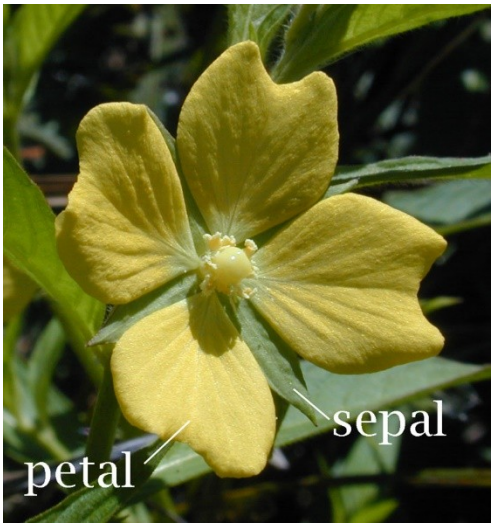
Optionally use Google Colab.

1.2. Load and Preprocess the Students Dataset

Import the *students* data set into Python with pandas. The *students* data set is provided in ILIAS as an Excel file. Use the read_excel function.

1. What is the most common mark that has been given in FSS2010? To find the answer filter the examples and draw a histogram afterwards. Use the plot functionality of pandas.
2. Is there a correlation between the mark and the number of attended classes? Find the answer using a scatter plot.
3. Does this correlation hold for all students? Find the answer by aggregating the examples by student and use a scatter plot afterwards.

1.3. Visual Exploration of the Iris Dataset



The data set describes three types of Iris flowers:

- Setosa
- Virginica
- Versicolour

There are four (non-class) attributes

- Sepal width and length
- Petal width and length

Retrieve the Iris data set from the ILIAS group and put it in the folder where you started the notebook. Use different plotters to visualize and explore the data set.

1. Which attribute combination and (approximate) value ranges determine the type of Iris flower?

Answer:

Type of Iris Flower	Attribute combination and value ranges
Setosa	
Virginica	
Versicolour	

ChatGPT Bonus Exercises

Reminder: Do not take the answers of ChatGPT at face value! Always cross-check with lecture slides, literature and/or the teaching staff!

C.1. Discuss the Data Mining Process, Data Preparation and Data Visualization

- Use ChatGPT to ask about the data mining process and its main steps. What makes the data mining process an iterative process?
- An important part of data preparation is the identification of problems such as missing values, outliers and duplicate records. Ask ChatGPT for some methods on handling these 3 encountered problems. Are there more problems that can arise during the data preparation step?
- Ask ChatGPT about some visualization techniques and when each of them can be used. What visualization technique would be best to spot outliers in a given dataset?

C.2. Learn about profiling and visualizing datasets

- Ask ChatGPT about some important features of DataFrames that can be used for data profiling. Ask if it can generate you some code for profiling the “Iris” dataset that was used in the last exercise. What information did the profiling provide about the dataset?
- Given the “Iris” dataset, ask ChatGPT which visualization technique can help you distinguish different groups of flowers by using different dataset features. Ask again if it can generate the code for the answer using the matplotlib package and try the code.

C.3. Self-Assessment

- Ask ChatGPT to generate some multiple-choice questions regarding the data mining process for graduate students and try to answer them.
    - You can ask for hard questions if you would want additional more difficult questions to test your knowledge.