Explainable Artificial Intelligence **Paderborn University**
Summer Semester 2025 Dr. Stefan Heindorf

Exercise Sheet 0

Due date April 14, 2025

The main goal of this exercise is to ensure that you have a working Python environment and are familiar with the fundamental tools (e.g., `NumPy`, `Pandas`, `Matplotlib`, etc.) that we will use in this course. If you find working with some of the libraries challenging, we highly recommend that you take the time to learn more about them during the course. If you have any questions regarding any of the tasks, please feel free to ask them in the `Forum` section of our course on Panda platform.

1 Task: Environment Setup

a. What are the benefits of using Python virtual environments?

b. What are the advantages of using a `conda` environment over a `venv` environment?

c. Suppose you want to work on two different projects, one requiring Python `3.12` and the other requiring Python `3.13`. Which Python environment management tool (`venv` or `conda`) would be the best choice in this case? Explain why.

d. Create a `conda` environment called `xai` with Python `3.12`.

e. A file named `requirements.txt` is provided with the necessary libraries for this course. How can you install all the libraries listed in the file in the `xai` environment with a single command?

f. You also need to install the `torch` library in your environment. For CPU-only support, you can use the command: `pip install torch`. However, if you want to install the GPU version, you can follow the guide provided on the official website[1].

2 Task: Working with NumPy Arrays

a. The `.npy` format is a file format for storing NumPy arrays[2] on disk. There is a file named `upb.npy` in the `data` folder. Load the file into a NumPy array and display its shape.

[1] `https://pytorch.org/get-started/locally/`

[2] You can find an introduction to NumPy in the `https://www.datacamp.com/tutorial/python-numpy-tutorial`.

b. The data that you loaded in the previous step is a 3D NumPy array representing an RGB image. The shape of the array is `(height, width, channels)`. Display the image using the `matplotlib` library[3].

c. What is your interpretation of the `.size` attribute of this array?

d. Display a figure with three subplots (vertically stacked) showing the red, green, and blue channels of the image separately[4]

e. Convert the image to grayscale by taking the average of the three channels. Display the grayscale image.

f. Convert the image to a binary image by setting all pixel values less than 128 to 0 and all pixel values greater than or equal to 128 to 255. Display the binary image.

g. Suppose we want to crop the image to a specific region. Crop the image to the region between the coordinates `(100, 100)` and `(550, 700)`. Display the cropped image.

3 Task: Data Analysis & Wrangling

a. A dataset of hotel customers' reservation information[5] is provided in the `data` folder. Load the dataset in a `Pandas`[6] `DataFrame`. How many rows does the dataset have? Display the first 5 rows of the dataset.

b. Drop the column `Booking_ID` and remove those rows where the column `no_of_adults` is less than 1.

c. Map the values of the column `booking_status` to the following: `Not_Canceled` $\rightarrow$ `0`, `Canceled` $\rightarrow$ `1`.

d. Answer the following questions based on the dataset[7]:

   a) What percentage of the bookings were canceled?

   b) Which month is the busiest for hotels?

   c) On average, do solo travelers book a room for more nights[8] than families?

   d) What is the cancellation rate for each market segment?

---

[3] You can use the `imshow` function from `matplotlib.pyplot` to display the image. For more information, you can refer to the `https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.imshow.html`.

[4] Choose the corresponding color map (`cmap`) for each channel.

[5] The dataset is also available at `https://www.kaggle.com/datasets/ahsan81/hotel-reservations-classification-dataset`

[6] `https://pandas.pydata.org/docs/user_guide/index.html`

[7] Try to use `Pandas` built-in functions like `groupby`, `value_counts`, etc.

[8] Sum of both weekend nights and weekday nights

e. A histogram shows the frequency distribution of a single variable. Plot a histogram of lead time for the bookings separated by the booking status (shown in different colors)[9]. What can you infer from the plot?

f. As you know, most machine learning algorithms support only numerical data. However, our dataset contains some columns with categorical data (e.g. `type_of_meal_plan`, `room_type_reserved`, etc.). To handle this, we need to encode these categorical columns into numerical values. There are several ways to do this[10], but for this task, we will use one-hot encoding[11]. Please find all the categorical columns in the dataset and encode them using one-hot encoding. For each categorical column with n unique values, create $n - 1$ binary columns (by removing the first value)[12].

g. You can see that the range of values in each column is different. In other words, the values are not on the same scale. This can be problematic for some machine learning algorithms. To address this issue, we need to normalize the data. There are different normalization techniques available in the `scikit-learn` library. For this task, use the `MinMaxScaler` to normalize the dataset[13]. This scaler transforms each feature to a given range (by default, between 0 and 1).

h. A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables which can range from -1 to 1. A value of 0 indicates no correlation, while a value of 1 indicates a perfect positive correlation. Similarly, a value of -1 indicates a perfect negative correlation[14]. Plot the correlation matrix of the dataset[15]. What can you infer from the plot, especially in terms of the target variable (`booking_status`)?

---

[9] In this case, the horizontal axis should represent the lead time, and the vertical axis should represent the count of bookings. For more information, you can refer to the `https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.hist.html`. Note that you need to show two histograms in the same plot, one for each booking status.

[10] Please get yourself familiar with different encoding techniques available in the `scikit-learn` library, as we will use them frequently in the ML tasks. You can learn more about encoding categorical variables in the `https://www.geeksforgeeks.org/encoding-categorical-data-in-sklearn/`.

[11] A good explanation of one-hot encoding and its implementation via `Pandas` and `Scikit-learn` can be found in the `https://www.datacamp.com/tutorial/one-hot-encoding-python-tutorial`.

[12] You can use the `get_dummies` function from Pandas and set the parameter `drop_first` to `True`. For more information, you can refer to the `https://pandas.pydata.org/docs/reference/api/pandas.get_dummies.html`. It would be truly beneficial to research more about why we used one less column for encoding.

[13] You can use the `MinMaxScaler` from `sklearn.preprocessing`. For more information, you can refer to the `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html`.

[14] For more information, you can refer to the `https://builtin.com/data-science/correlation-matrix`.

[15] You can use the `corr` function from Pandas to calculate the correlation matrix and the `heatmap` function from `seaborn` to plot the heatmap. For more information, you can refer to the `https://seaborn.pydata.org/generated/seaborn.heatmap.html`