Project 2

Information

This part of the project consists of three exercises:

- In **exercise 1** you retrieve demographic and economic data for countries from the Worldbank API.
- In **exercise 2** you combine the data from the previous exercise with data on the performance of countries at the Olympic Games 2012 and prepare the combined data for exercise 3.
- In **exercise 3** you train a linear regression model, to predict the number of Gold medals of countries at Olympic Games based on demographic and economic features.

Note: Since machine learning is not a focus topic of this course, you will not need to optimize the model. Just demonstrate that you are able to apply the steps we discussed in the course and correctly interpret the results.

Exercise 1

Write a Python function that can be used to query data from the Worldbank Indicator API. Your function should:

- take the following input parameters: `indicators`, `countries`, and `years`.
- return a Pandas DataFrame of the queried data
- have a docstring that explains what the function does, what the input parameters are, and what the output is
- minimize the number of API calls necessary to retrieve the data

Demonstrate that your function works by querying the following data (codes are provided in parentheses):

a) The total population (SP.POP.TOTL) of Germany (DE) and France (FR) between 2015 and 2020.
b) The total population (SP.POP.TOTL), GDP in current US$ (NY.GDP.MKTP.CD), and life expectancy in years at birth (SP.DYN.LE00.IN) of all countries (all) in 2012. Print the shape of the resulting DataFrame and display its first 10 rows.
c) State how many API calls your function makes for a) and b) respectively.

Notes:

- To solve the exercises study the documentation of the basic call structures. Most of the information you need is provided there.
- If needed, additional information about the API is available here. For instance, you will find links to the list of available indicators and countries, and explanations on error codes.
- Note that by default the API returns only the first 50 items. Check the documentation on how to retrieve more items.

Exercise 2

The file `medal_table_2012.csv` contains information about the number of medals won by each country at the Olympic Games 2012. (It probably looks similar to the medal table that you calculated in the first part of the project. Small differences are possible, but the overall structure should be the same.)

a) Preprocess both the medal table data and the Worldbank data retrieved in exercise 1 b) and combine the two datasets suitably into one tidy dataset. The final dataset should be such that it allows you to answer the following exercises. Explain your actions and decisions in a few sentences.
b) Create an alternative medal table for the 2012 Olympic Games by calculating the number of Gold, Silver, and Bronze medals won per 10 million inhabitants. Display the 10 most successful countries according to this alternative medal table.

Note: If there are missing values in the Worldbank data set (e.g. if no population data is available for Germany), then you do NOT need to impute these values.

Exercise 3

Carry out a simple supervised machine learning experiment, in which you train a model to predict the number of Gold medals a country wins at the Olympic Games 2012 based on demographic and economic features. **Note**: Since machine learning is not a focus topic of this course, you do not need to optimize the model. Just demonstrate that you are able to apply the steps we discussed in the course and correctly interpret the results.

a) Train and evaluate a linear regression model: 1. Split your data into a training and a test set. 2. Train a linear regression model using population, life expectancy and the GDP per capita of a country as features. 3. Evaluate the model using the root mean squared error as the performance metric.
b) Discuss your results: How do you judge the performance? What are possible reasons for this performance? How could the model be improved?
c) Due to an unfortunate "data error", the country Netherlands was not included in the Olympic Games data and is therefore not present in medal table. Use your trained machine learning model to predict the number of Gold medals the Netherlands has won in 2012, just based on their demographic and economic characteristics.