|  |
| --- |
|  |

Fundamentals of Data Science - project

## The aim of the project is to familiarize you with practical algorithms and problems of modern data analysis and tools used to implement selected task. The project is **individual**.

## The report (~10 pages) along with any attachments (source code) should be submitted in electronic form using Microsoft Teams. Special assignment section will be prepared there.

## The report should not contain too much theoretical information - only the most important practical aspects of the problem under consideration. Particularly appreciated will be the ability to critically analyze and formulate valuable technical conclusions.

## The deadline for submitting the report is **June 9th 2022**. After this date, the final evaluation will take into account the delay in the implementation of the task. On this day we have classes and each of you will have 5 minutes to present your results.

## Project meetings are not obligatory, but rather consultations for those interested.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | Thu. 19.5.2022 | 8:50 - 11:05 | | Thu. 26.5.2021 | 8:50 - 11:05 | | Tue. 2.6.2022 | 8:50 - 11:05 | |  |  | |  |  | |
|  |

## Of course, you can also ask questions via e-mail/Microsoft Teams.

**The task:**

Please analyze the dataset containing descriptions of football matches

<https://www.kaggle.com/datasets/hugomathien/soccer>

You can choose a league of your preference. We focus mainly on the match table, however, you may use other tables as well.

Your task is to predict the results of the match (Win, Draw, Lost by the home team), having, as the input data, all the matches which happened before. Please use 80% of the matches as the training set, and 20% as the testing set. The assignment of matches to both sets should be done randomly. The choice of the classifier is up to you.

Notes:

1. Please perform a statistical summary of the data (max, min, mode, avg, median) and discuss it.
2. You have to decide how many previous matches should be considered for your model, and if it should only use matches of the given team (or even matches between those specific teams).
3. Please report the accuracy of your model for both training and testing sets, and report the confusion matrix.
4. When it works? The home team wins about 46% of the time so if your classifier is better than that in terms of accuracy you have achieved your goal.

In your report, please present your approach and obtained results, code should be sent as an attachment or a link to a repository.

Using SQLite files:

<https://datacarpentry.org/python-ecology-lesson/09-working-with-sql/index.html>