The first part of the Assignment, IDS 2020-2021

Introduction

The assignment guides you through analysis of a dataset using techniques and tools provided in the course. The first part of the assignment tests your understanding of the material discussed in the lectures 1-9. The assignment questions are given in the Jupyter notebook. It is necessary to follow the assignment in the given order because certain questions might require an output obtained in the previous steps. Please note, it is important to use **the Python environment provided for this course** to answer the questions.

The dataset

The dataset "**dataset.csv**" contains entries about environmental factors of the different habitats and different types of amphibians living in those habitats. The dataset has several features:

- ID: identification number of each data row
- SurfaceR: the overall surface of the water reservoirs in the habitat (m^2)
- NumberR: the number of reservoirs in the habitat
- TypeR: the type of the reservoirs
    - A: natural
    - B: newly formed natural
    - C: settling ponds
    - D: close to houses
    - E: technological
    - F: garden pond
    - G: trenches
    - H: wet meadows, flood plains, marshes
    - I: river valleys
    - J: streams, creeks
- VegetationR: the vegetation density in the reservoirs (low to high)

- SurroundingR1, SurroundingR2, SurroundingR3: the main, second and third type of surrounding area
    - A: forest
    - B: meadows
    - C: garden
    - D: parks
    - E: dense buildings
    - F: sparse buildings
    - G: river
    - H: roads
    - I: agricultural
- UseR: intensity of use by humans (low to high)
- FishingR: intensity of fishing (low to high)
- AcessR: percentage of shore allowing access to land habitat (no buildings/roads)
- RoadDistance: km to the next road
- BuildingR: km to the next settlement
- PollutionR: intensity of pollution (low to high)
- ShoreR: type of shore
- The different amphibians: '1' if this type is present in the habitat, '0' otherwise

Before answering questions, perform the data preprocessing as explained in the Jupyter notebook. Export the resulting dataset '**sampled_data.csv**' and submit it with your results.

Submission and deliverables

The deadline for the assignment is 23/12/2020 23:59. You will need to hand in your submission via the Moodle. Please note, that a deadline extension is not possible and the late submissions will not be considered.

The assignment should be done in groups of 2-3 students. Make sure to include all group member names and their ids.

Your submission should include a **Jupyter notebook** with your results and your Python code that indicates how you have obtained the results. In addition, please upload a **zip-file** with all requested datasets and other outputs, in particular, the 'sampled_data.csv' created in the preprocessing step. An additional report is **NOT** required and will not be considered for the grading.

Submission summary:

You have to submit <u>two</u> items to the Moodle:

1. **A Jupyter notebook.**
    - Use provided Jupyter notebook to present your results and code.

- Make sure that names and student ids of all group members are provided in the notebook.

2. A '**datasets.zip**' with all requested datasets and other outputs, such as pdf, jpg, etc.. Please do not forget to discuss the outputs in the notebook.

Grading

Successful participation in the assignment, i.e. scoring at least 50%, is one of the prerequisites for taking the written exam. The results of the assignment are valid only in the current semester and will expire afterwards. The assignment can only be redone in the next academic year.

The assignment counts as 40% of the final grade (20% each part). In the first part of the assignment, 100 points are assigned: 90 points for eight main sections and 10 points related to your reporting style:

1. Preprocessing of the Data set – 5 points
2. Insight into the Data – 15 points
3. Decision Trees – 15 points
4. Regression – 14 points
5. SVM – 8 points
6. Neural Networks – 15 points
7. Evaluation – 10 points
8. Clustering – 8 points

- For a data scientist, a sufficient and proper presentation of the results is as much important as analysis and code. 10 points are therefore given for your reporting style. A few useful tips how to present your work well:
    - Add comments to your code.
    - Do not mix code and answers to the questions. Please use markdown cells to explain your solutions and provide answers to the stated questions in a sufficient and coherent manner.
    - Do not change the structure of the notebook, except adding blocks for the better readability.
    - In general, make your answers readable and understandable. Please, check the spelling.

Please note, that correct and full results, sound code and sufficient explanations are highly important.