The First Part of the Assignment of IDS 2019-2020

Introduction

The assignment guides you through the analysis of a data set using the techniques and tools provided in the course. This part of the assignment tests the understanding of the material in lectures 1-9. It is necessary to follow the assignment in the given order since the result of some questions might depend on answers to previous steps. The questions are detailed in the provided jupyter notebook.

The Dataset

Dataset: Use the provided data set "**population_density.csv**". The data set has the following attributes:

- holiday: US national holiday
- temperature: Average temperature in kelvin
- rain_1h: Amount in mm of rain that occurred in the hour
- snow_1h: Amount in mm of snow that occurred in the hour
- clouds_ percentage: Percentage of cloud cover
- weather_type: Short textual description of the current weather
- weather_type_details: Longer textual description of the current weather
- date_time: DateTime Hour of the data collected in local CST time
- population_density: Number of people in center of the city

You should pass some steps before starting the assignment as preprocessing steps. The details of preprocessing steps are mentioned in the jupyter notebook file. After passing these preprocessing steps, export your final dataset as 'population_density_categorical.csv' dataset and use that for the next stages of the assignment. Make sure that you submit your extracted dataset with your results in Moodle.

Submission and Deliverables

The deadline for the assignment is 30/11/2019 23:59. You will need to hand in your submission via Moodle. Note that there is no extension for the deadline and also late submissions will not be considered.

The assignment should be done in groups of 2-3 and only one of the group members should upload the submission. Make sure to include all group members names and student ids in that submission!

Your submission should include a **jupyter notebook**, which presents your results and also contains the python code used to obtain the results. Next to this jupyter notebook, upload a zip-file that contains all requested data sets, including the **extracted dataset** (.csv) based on your student number (see the jupyter notebook).

Report requirements:

- You are allowed to upload 3 separate items via Moodle.
    1. Jupyter notebook.
        - Use the provided jupyter notebook to present results and code.
        - Make sure that the name and student id of **all the members** are in the jupyter notebook.
    2. datasets.zip including all the requested data sets.
    3. In the cases that the result of an algorithm is pdf, jpg, etc, you should attach the result to this notebook file and refer to that in the text.

Grading

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

The grade of the assignment counts 40% towards the final grade. In this first part of the assignment, 100 points are obtainable, 90 points for the seven main sections and 10 points related to your report style:

1. Preprocessing of the Data set – 5 points
2. Insight into the Data – 20 points
3. Decision Trees – 15 points
4. Regression – 10 points
5. SVM – 10 points
6. Neural Networks – 20 points
7. Evaluation – 10 points

- As a data scientist, adequately presenting your results is just as important as what you have done, therefore, 10 points are obtainable for report style.

Please note that correctness of your code, its result and also the accuracy of your explanation are important.