CSCI - 4146/6409 - The Process of Data Science - Summer 2022

Assignment 1

The submission must be done through Brightspace. Due date and time as shown on Brightspace under Assignments.

- To prepare your assignment solution use the assignment template notebook available on Brightspace.
- The detailed requirements for your writing and code can be found in the evaluation rubric document on Brightspace.
- Questions will be marked individually with a letter grade. Their weights are shown in parentheses after the question.
- Assignments can be done by a pair of students, or individually. If the submission is by a pair of students, only one of the students should submit the assignment on Brightspace.
- We will use plagiarism tools to detect any type of cheating and copying (your code and PDF).
- Your submission is a single Jupyter notebook and a PDF (With the compiled results generated by your Jupyter notebook). File names should be:
    - **A1-<your_name1>-<your_name2>.ipynb**
    - **A1-<your_name1>-<your_name2>.pdf**
- **Forgetting to submit both files results in 0 markings for both students.**

In this assignment, you will need to predict the Australian wildfire.
Link for the dataset
https://www.kaggle.com/datasets/brsdincer/australia-and-investigative-special-wildfires-data

- The size of the data is quite large. It may take a lot of time to train on a (Google) Colab instance or on a local machine depending upon your machine configuration. So, you should use subsampling during experimentation. Use the contents from the lectures about sampling and justify your choice of sampling.
- The reduced dataset should be able to capture a similar distribution as complete data and the number of subsampled data instances should be at least 150k.

1. Data Exploration and preprocessing (0.1)

1. Data quality report.
    a. Generate data quality reports for the continuous and the categorical features of the data set
    b. Plot the heatmap correlation among the variables in the data
    c. Identify data quality issues and build the data quality plan
2. Preprocess your data according to the data quality plan
3. Answer the following questions:
    a. What are the dates on which bushfires present the high number of incidents?
    b. Based on the data quality report, which attributes do you think are useful to predict the confidence of an incident? Explain why you think that the selected attribute is important.

2. Spatio-temporal data (0.2)

For all the parts below, describe how you generated the plot and analyze any findings that you observe in it.

1. Plot a geographical heat map of FRP for the Aqua area. Use data from the year 2017 and any aggregation method (e.g. mean, summation, maximum, or something else) of your choice. Justify your choice of aggregation method.
2. Mark the "Fire activity" (lat, long = -35.6,149.12) on the city map.
3. Mark the regions with the highest recorded fire radiation in a day for measurements where "acq_date = 2020-01-08".
4. Find a visualization to plot the progress of the fire activity across all points from Nov 1, 2019, to Jan 31, 2020.

3. Build a model for spatial prediction of wildfire (0.7)

1. In between 'Brightness temperature I-5' and 'Fire Radiative Power' choose either as the target feature.
2. Explain what the task you're solving is (e.g., supervised x unsupervised, classification x regression x clustering or similarity matching x, etc)
3. Use a feature selection method to select the features to build a model.
4. Select one or more evaluation metrics. Justify your choice.
5. Build a baseline model
    1. Perform hyperparameter tuning if applicable.
    2. Train and evaluate your model

   3. How do you make sure not to overfit?
   4. Visualize the learning process and/or the knowledge acquired by your model. Explain why you think this is the right visualization.
   5. Analyze the results
6. Build a candidate final model (you are encouraged to experiment with more than one model but only include and discuss your “best” model)
   1. Perform hyperparameter tuning if applicable.
   2. Train and evaluate your model
   3. How do you make sure not to overfit?
   4. Visualize the learning process and/or the knowledge acquired by your model
   5. Analyze the results
7. Compare the two models with a statistical significance test. Use a box plot to visualize your comparison.