No code lab

Home Assignment

Your home assignment is to find a dataset that is related to your research discipline (even if only tangentially) and perform three operations on the dataset. You may use the following three websites to search for datasets (of course, you can use the dataset that you already have):

- https://www.kaggle.com/search
- https://archive.ics.uci.edu/datasets
- https://catalog.data.gov/dataset/

Your first task is to visualize the dataset. Choose 1-2 visualization techniques that can tell an interesting story about the dataset. Make screenshots and write 1-2 paragraphs describing the visualization, in particular, say what insights or hypotheses are visible in the data.

Next, use an unsupervised learning technique to cluster your data. You may choose any algorithm you want (k-means, DBSCAN, hierarchical clustering). Again, make a screenshot and add your commentary to the results. Do you agree with the way data is clustered? Is clustering showing some interesting property of the data that you were not aware of before? Try to explore the data by modifying the parameters (e.g., the number of clusters).

Finally, use your data to build a classifier. Use any algorithm you want (with the exception of majority voting, obviously). If possible (i.e., if you are using an algorithm which produces some structure such as a set of rules or a tree), visualize the classifier. Test the classifier, using either cross validation or random sampling. Are you satisfied with the performance of the classifier? Write a paragraph describing your experiences and your impression of the model.

Put your assignment into a *.pdf document and send it directly to Mikolaj.Morzy@put.poznan.pl The deadline is September 30, 2024. If you struggle with clustering or classification, you may omit this step, but at least you have to visualize the data and describe what you see in the visualizations.