Assignment- Mining Over Datasets

In this exercise, you are called to apply data mining practices to existing datasets. You need to provide:

- one report containing individual sections per dataset, which will describe the methods, tools, and results of the analysis, based on the (per-dataset) questions and requirements, elaborated below.
- code file(s) with all the required information to reproduce the analysis

(Use the e-class functionality to upload a single -possibly compressed- file containing the above)

For the evaluation of the assignment, the following aspects will be taken into account:

- the clarity of writing and coherence of the report
- the completeness of the application parameters of the method(s) used
- the explanation of why the selected methods were appropriate
- the reproducibility of the process

Dataset 1: Airlines Dataset Inspired in the regression dataset from Elena Ikonomovska

Data: https://users.iit.demokritos.gr/~izavits/datasets/Airlines.arff.zip

The dataset is in "arff" format. You will easily understand the features and their possible values by inspecting the file.

You may apply whatever transformation you like to transform the dataset in a more convenient form for your task.

Tasks:

- Provide an overview of the dataset size, features, and distribution of feature values.
- Describe the average delays per airport/airline.
- Identify and report the most prominent rules of association between delays and point of origin AND/OR point of arrival.
- Try to predict the delay given all other features and report the appropriate performance on cross-validation.
- Identify patterns/rules regarding delays and try to explain when delays should be expected, based on these patterns.

Dataset 2: Religion data

Description:
https://github.com/aaronpenne/data_visualization/blob/master/religion/data/1952.txt

Data: https://github.com/aaronpenne/data_visualization/blob/master/religion/data/1952.xls

You are free to consult or use the code here if you want to provide a summary using a map:

https://github.com/aaronpenne/data_visualization#mapping-religion

Tasks:

- Get to know your data. Provide a small description of the attributes that will help you in the following tasks (in bullet points).
- Summarize the data to help understand the overall picture of religious groups over the US.
- Which are the counties with the highest per-person ratio of Orthodox Christian members?
- Find the 3 most extreme counties with respect to the distribution of their churches across religions?
- Where would you create a cross-religion center of discussion between religions to maximize its impact? Support the proposal based on data analysis results.