CST 8390 Final Project (Mandatory)

(Report due: Aug 14, Presentations: Aug 2 - 12)
(20 marks)

Data Science is still a relatively new field and not many people know how to apply it. Moving forward, you will have to explain to others what you are able to calculate from data and how it can be used to help them. The goal of this final project is to think about how to apply the algorithms we learned in this course to a real dataset. You need to work on this project **with a partner.** For the presentation, both students must speak an equal amount of time, a total of around 10 minutes.

Part 1- Project Report **(10 marks)**

Find a data set from the following online portals:

https://data.brla.gov/
https://opendata.maryland.gov/

A dataset cannot be used by more than 5 teams. You have to decide your dataset in the coming days. Every team should have a list of 3 datasets ready (along with the source link, prioritized) and we will decide the dataset based on the availability. Datasets from other sources will not be accepted. You must use the Project Dataset Selection Form to submit your dataset choices. When you select dataset, make sure that there are at least 10 relevant attributes (including the ones that you can extract or create) and 300 instances in it.

Make sure that each instance corresponds to one event/person/activity/situation etc. If each instance is summarized or an aggregate, then you must avoid such datasets. Don't select Covid datasets as they all are summarized ones.

You should clean the data, remove outliers and then run the algorithms you have learned on the data. For numeric attributes, you can calculate min, max, average, standard deviation, and co-variance/correlation. If the attributes are nominal, then you can calculate the frequency of each label. If the attributes are of mixed (numeric and nominal) type, then you can convert numeric to nominal using filters or using some other meaningful translations (for example, convert numbers to ranges like "low", "medium", and "high"). You can then apply different methods for classification, clustering, outlier detection, data mining and regression methods.

Write a report on how you found your dataset, and the initial guesses regrading trends and patterns within the data before any analysis. The link of the dataset should be provided in the report and in the presentation. Then describe which of the algorithms you want to use to find whether your assumptions are correct. Lastly, describe what you found in analysis afterwards, which either confirms or denies your original guess. Include screenshots and graphs to justify your results. Also, when you build some prediction model or equation, give a detailed screenshot of the results. Describe the accuracy of your prediction by presenting confusion matrices, R^2 values, etc.

You should frame a question that you want to answer by your analysis. This question should be written on the bottom of the cover page. This question cannot be easily answered using Excel (which means your question should be dependent on more than 3 factors).

You should have 5 main sections – Data collection (with the source link), Preprocessing, Data Analysis, Results, and Conclusion. Report should have a cover page (with names of both students and student numbers), table of contents, tables, pictures etc., introduction, conclusion and references.

It should be written in a professional report style.
Font: Times New Roman size 12 with 1.5 line spacing, justified

Part 2 - Project Presentation **(10 marks)**

Give a short 10-minute presentation (use PowerPoint slides) which summarizes the steps in your report from part 1. Briefly describe your data set and how it was collected, the question that you want to answer by your analysis, various data preparation steps etc. Describe your analysis by explaining the algorithms and the results you found. Briefly explain whether the analysis confirmed or denied your expectations and explain if any surprises that you found. Also include an analysis of the accuracy of your results and its importance. You need to discuss how the results are helpful to the future work or to the society. You should have 5 main titles (along with related slides) – data collection (with the source link), preprocessing, analysis, results, and conclusion.

Submission:

The presentations will be during the last two weeks of school, either in the lecture or in lab. **This should be from the perspective of you giving a report to a company or job interview where they aren't sure what data science is about** (Just creating some tables with numbers are not enough). You should also submit your report (along with final arff files and model files) and presentation through Brightspace as a zipped folder named lastnameFirstStudent_firstnameFirstStudent_lastnameSecondStudent_firstnameSecondStudent.zip.

To get grades, BOTH submission (report & slides before due date) and presentation are required. Successful completion of project is mandatory for this course.