CST 8390 Final Project (Mandatory)

(Report due: April 14, Presentations: April 3 –14)
(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 can find 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.sandiego.gov/datasets/?category=public-safety
https://data.sfgov.org/browse?category=Public+Safety

A dataset cannot be used by more than 5 teams. You have to decide on 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 on 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 a 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. You can combine datasets if possible. Don't select Covid datasets as they all are summarized ones.

You should follow CRISP-DM while working on the data. Data Understanding and Preparation will take some time. As part of your data understanding, you should describe and explore data and verify its quality. Once you understand your data, you can select, clean, construct, integrate, and format data. You can then apply different methods for classification, clustering, outlier detection, associations, and regression methods. All of these steps must be reported in a professional style.

Write a report on how you found your dataset and the initial guesses regarding trends and patterns within the data before any analysis. The link to the dataset should be provided in the report and 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 the analysis afterward, which either confirms or denies your original guess. Include screenshots and graphs to justify your results. Also, when you build some prediction model, 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 in your analysis. This question should be written at 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 Understanding (with the source link), Data Preparation, Modeling, Discussion of results, and conclusion. The 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, the question that you want to answer by your analysis, various data understanding and preparation steps, etc. Describe your analysis and the results by mentioning the algorithms. Briefly explain whether the analysis confirmed or denied your expectations and explain any surprises that you found. Also include an analysis of the accuracy of your results and their importance. You need to discuss how the results are helpful to future work or society. You should have 5 main titles (along with related slides) –Data Understanding (with the source link), Data Preparation, Modeling, Discussion of results, and conclusion.

Submission:

The presentations will be during the last two weeks of school, either in the lecture or in the lab. **This should be from the perspective of you providing a report to a company or job interview where they aren't sure what data science is about** (Just creating some tables and pictures is 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 the project is mandatory for this course.