**CS483 - 01 Big Data Analytics - Capstone Spring 2020**

**Course Project**

***(100 points in total)***

**Please form a team of 1 – 2- 3 students (from 1 up to 3 students).** Please do the following for your course project:

1. Find a dataset of big size. You do not have to use the entire dataset for your project, but it should represent a Big Data problem. Use a reasonable portion of the dataset for your project. You might use different datasets for the project (data comes from different resources/files) - **Due: April 8.**

**Please e-mail me the names for your team members and a link to your dataset by April 8, 11:00pm.**

1. Look at the dataset/s and understand its nature. – **Due: by April 10.**
2. Analyze the dataset asking meaningful and interesting questions and answering them using a variety of tools/languages/techniques. Your course project must have:
   1. Data Visualization part (meaningful data visualization that helps to answer good questions – question/s and tools are up to you) – Need to have at least 3 (per team member!) meaningful and interesting visualizations. The number of visualizations has to be at least 3\*# of team members.
   2. Statistical Analysis of the data (Asks meaningful and interesting statistical questions – answer them with appropriate tools/techniques) – Need to have at least 3 (per team member) statistical questions answered. The questions should be meaningful and interesting. The number of statistical questions should be at least 3\*# of team members.
   3. Ask at least 3 (per team member) Data Mining/Machine Learning type questions: Classification, Numerical Prediction, Clustering, … (meaningful and interesting, not obvious) and answer them using an appropriate tool. The number of Data Mining/Machine Learning questions should be at least 3\*# of team members. **Do not use WEKA!**
   4. Use R, Python and/or any other tool/software/language.
   5. Please use at least three different languages/software/tools.

1. Write a final report for your project – all your work should be described, it should be organized well and written clearly. I suggest the following structure:
   1. **[10 points] Introduction:** Description of the initial set, where it comes from, what kind of information represents, its size (all parameters that might represent its size), …. Say what size subset of the initial dataset is used in your project. Discuss any challenges with the dataset (Any data cleaning, missing attribute values, duplicate attributes, irellevant attributes – adding/deleting attributes form the initial dataset…). Describe how you have resolved them.
   2. **[15 points] Section 1 – Data Visualization:** What you visualize, what questions you try to answer using the visualization, what tool/language is used, visualization results/graphs (should be easy to understand and really to help to answer the question/s you asking). Your answers/conclusions/hypothesis based on the visualization. Discuss challenges/problems if there are any. If you have more than one choice for a tool/language used, discuss your choice.
   3. **[15 points] Section 2 – Statistical Analysis:** Tell what statistical measures you like to calculate (the statistical questions should make sense and be interesting). What are the calculated measures/statistical analysis results. Discuss what the results tell you. Say what tool/package/language are used to calculate these statistical measures. Discuss challenges/problems if there are any. If you have more than one choice for a tool/language used, discuss your choice.
   4. **[30 points] Section 3** – (*Your Data Mining/Machine Learning Questions*)**:** Formulate your questions. Describe what model is used for knowledge representation (Rules, Tree, Regression, …), what algorithms are used (you may apply a few and compare results), which tool is used (set of implemented algorithms will depend on a tool). Report the results, compare them (results for the same problem/question obtained using different algorithms/models, use different test-train options). Discuss challenges if there are. Discuss your choices.
   5. **[5 points] Section 5** – **Future Work:** Based on what you have done what are the other questions that you would like to ask working on the dataset in the future.
   6. **[5 points] Section 6** – **References:** List everything that was used for the project.
   7. **[5 points] Appendix** – **Code:** Include your code (if any) for each of the sections. – Attach all your files to a zip file for your course project submission.
2. **[15 points] You need to present your course project in class, on Thursday, April 30 and Tuesday, May 5. – Not presenting in class you will lose 15 points from your Course project grade! - Your course projects must be ready by April 30/May 5 (depending when you present them – we will randomly select dates)**
3. **Final course project report and slides presentation have to be submitted by Tuesday, May 12, 2020, 11:00pm.** Please submit both files (report and slides) using “Course Project” Blackboard submission link (both files in the same submission).