STS 3130 Data Mining I – Final Project Introduction

***Description*:**

The final project involves exploring a technique appropriate for a regression or classification problem that we have not covered in class (or a previous course in the Data Analytics and Statistics program). The purpose of this project is to help students practice the skill of independently acquiring knowledge of statistical methods and to practice communicating that knowledge to others. You must choose a method and find an appropriate dataset for implementing the method.

***Important Dates*:**

* Milestone 1: Via email, by Saturday, November 20th at 11:59 PM, submit a brief description (one paragraph) of your intended technique and your proposed dataset. I must approve your technique and dataset before you move forward. ***It is recommended that you make a submission sooner***.
* Milestone 2: Via email, by Wednesday, December 3rd at 11:59 PM, submit your dataset any evidence of progress since Milestone 1. This could include your code, basic data exploration, part of your paper, etc.  ***It is recommended that you make a submission sooner***.
* Paper: This is due by the end of your final exam time period on Wednesday, December 15th from 12 - 3 PM. NOTE: This deadline is the only deadline of the semester for which we do not have an 11:59 PM deadline.

When picking a project, you should pick a topic that is connected to either the regression or classification style of problem. You may pick any topic that we have not covered in a Data Analytics course as long as it is appropriate for either regression, classification, or both. Do NOT pick CART or random forest since we will cover those topics in the remaining lectures.

Ultimately, you will implement the method and compare the results to at least one of the other methods that we have used. For example, if your topic is random forest (it cannot be because we covered it in class, but I am just making a point here), you could compare your results to that of logistic regression.

If you are looking for a topic, you can start by looking at our textbook. There are many topics in there that we have not covered. You could also search for “machine learning techniques for classification” or “machine linear techniques for regression” or “machine learning techniques for supervised learning.”

Grading details follow.

***Grading***:

I will provide details on how to earn a B (85%) on the project. If you wish to earn a higher grade, show me that you are willing to go above and beyond.

* Deadlines - 10 Points:
  + 5 Points - Meet Milestone 1 Deadline (Due 11/20/21 - Submit a brief description (one paragraph) of your intended technique and your proposed dataset.)
  + 5 Points - Meet Milestone 2 Deadline (Due 12/3/21 - Submit your dataset any evidence of progress since Milestone 1. This could include your code, basic data exploration, part of your paper, etc.)
  + 0 Points - Meet Milestone 3 Deadline (Due 12/15/21 - Submit your final paper.) There are no points for meeting this deadline, but there will be a deduction if you miss it.
* Paper - 75 Points
  + Abstract (5 Points) - one paragraph providing an overview of the paper. This should a brief description of your data, the method you will employ, and planned comparisons (i.e., if you are doing classification, what methods will you compare your method to).
  + Exploration of Data (10 Points) - you should adequately explore your dataset. Discuss types of variables, amount of missing data, basic summary statistics, plots, etc. This should be a thorough exploration and write up. The write up should include paragraphs and a discussion beyond “The summary statistics are shown in Table 1” and “Here are some plots.”
  + Background (15 Points) - provide relevant information about the technique you have chosen, including when technique is used, types of data for which it is applicable, competitors, etc.
  + Implementation (15 Points) - talk about how to implement the method (this is basically a discussion of the code)
  + Results (20 Points)
    - Implement the method on your dataset
    - Implement a competing method (any of those that we discussed in class)
    - Compare the results
  + References (5 Points) - You should include a minimum of three references
    - At least one of the references can be to a website/tutorial that you found on the method
    - At least one of the references should be more an academic reference (journal, book, etc.)
  + Professionalism of paper (5 Points)
    - Does the paper look professional?
    - Are figures labeled and properly referenced?
    - Did you only include relevant figures (i.e., the ones discussed in the paper)
    - Are spelling and capitalization correct?
    - Are paragraphs used and well written?
    - You should include a description of how you went above and beyond.
  + Above and Beyond - Include a section in which you describe how you went above and beyond.
* Above and beyond - 15 Points - Let your creativity shine.
  + There are many ways that you can go above and beyond, but here are a few examples:
    - Talk about when the method fails to work well and/or other competing methods
    - Did your analysis include other ideas that we have not included (such as extensive data cleaning, analysis of text data, etc.)
    - Many other ways - you’re smart, think about it
  + NOTE: The amount of additional points that you earn for any of the above tasks depends on both the amount of effort and the quality of the work.