Course Project Instructions

# Data Set

* Do not choose the data sets that we have analyzed in class, such as the Titanic data, the zoo data, etc.
* It should not be a small or made-up dataset. For this semester, “small” is defined as fewer than 50examples in the data set.
* Choose a data set that does not require excessive data preprocessing.

# Experiment Design

* Define a problem on the dataset and describe it in terms of its real-world organization or business application. The complexity level of the problem should be comparable to HW4 or HW7 assignment.
* The problem may use one or more of the types of data mining algorithms that we have studied this semester: Classification, Clustering and Association Rules, in an investigation of the solution to the problem.
* This investigation must include some aspects of experimental comparison: depending on the problem, you may choose to experiment with different types of algorithms, e.g. different types of classifiers, and some experiments with tuning parameters of the algorithms. Alternatively, if your problem is suitable, you may use more than one of the algorithms (Clustering + Classification, e.g.). If there are a larger number of attributes, you can try some type of feature selection to reduce the number of attributes. You may use summary statistics and visualization techniques to help you explain your findings.

# Project Idea Proposal – DUE WEEK 7

* Submit a one-pager to describe the data mining problem, the dataset, and your initial strategies for data analysis 48 hours prior to the Week 7Live Session ("Naïve Bayes").

# Project Progress Report– DUE WEEK 9

* Each student should submit a project progress report 48 hours prior to the Week 9 Live Session. You are expected to have finished all experiments and a draft project report. This round of feedback is expected to help you fix major problems in your research design and result analysis. Your report will be commented but not graded.

# Live Presentation – DUE WEEK 10

* Introduce - what is your area (briefly!) and what are your goals.
* Discuss 5 key analyses, their results and their visualizations. You will have more than 5 analyses in your report, but will choose the most interesting 5 to present.
* Offer clear and anyone-can-understand conclusions.
  + Include nothing about the data cleaning, etc.

# Final Project Report– DUE AFTER WEEK 10

* To complete this project, write a final report that conforms to general academic paper format. Make sure to cite relevant work appropriately**. Your report should be within 8 pages, 1-inch margin on all sides, and at least 12 point Arial or Times New Roman.** Submit your final report 48 hours after the Week 10 Live Session.

# Team Meetings – Every Friday at 7:00 PM EST

|  |  |  |
| --- | --- | --- |
| Meeting / Deliverable Due | Date | Comments |
| Meeting 1 | 7/17/2020 | Choose a dataset |
| Meeting 2 | 7/22/2020 | Dataset chosen. Need clarifications of project deliverables due dates. |
| Meeting 3 | 7/25/2020 |  |
| Meeting 4 | 8/7/2020 |  |
| Proposal | 8/12/2020 |  |
| Meeting 5 | 8/14/2020 |  |
| Meeting 6 | 8/21/2020 |  |
| Progress Report | 8/26/2020 |  |
| Meeting 7 | 8/28/2020 |  |
| Final Report & Presentation | 9/2/2020 |  |

# Required Methods

* Full cleaning, prep, EDA and visual EDA
* **Association Rule Mining** (unsupervised) Offer the top 10 rules for the highest sup, the top 10 for conf, and the top 10 for lift. All rules must have at least one element on the left and one on the right. Also choose to set the left as a given value and show the top 10 (based on your dataset and determinations).
* **Clustering** (k means and one other – your choice) (unsupervised) Include visualization. Run k means for three choices for k and choose the best.
* **Decision Tree** (supervised). Include at least three different trees and their visualizations.
* **Naïve Bayes**
* **SVM** – use at least three different kernels, and for each kernel, look at and discussion 3 different costs.
* **Text Mining** \*\*\* OPTIONAL\*\*\*\*

# Notes

* You must have at least two different vis types that support each method.
* Each method above will have its own subsection(s) in the Analysis and in the Results.
* Place the visualizations where they belong in the your project – not at the end.

# Required Project Headings

* Introduction - (3-4 paragraphs all about the area of interest. Do NOT discuss the data in the intro). The intro is non-technical and offers background, objectives, history, value, goals, etc. It must be readable by a person who is 10 or 11.
  + Grade Notes:
    - The Intro discusses the data or dataset. (-10 points)
    - The Intro discusses data science in general or models and methods. (-10 points)
    - The Intro is shorter than 3 paragraphs (a paragraph is 7 sentences). (-10 points)
    - The Intro is not clearly or well written and/or does not offer a good basis for what the project is about. (up to -10)
    - The Intro is written below the graduate level, such as by saying things like “In my project, I …”, or “This Project is about….” Consider this project to be an academic paper. It’s never about you. Avoid all use of “I”, “we”, “you”, …
* Analysis
  + **Subsection 1: The Data**
    - This is where you talk all about the data, the variables, the cleaning, measures of cleanliness, etc.
    - Even if your data is clean, you must write code to clean it (pretend you cannot see that it is clean). Minimally: NAs, changing things to factors as needed, outliers, discretization, incorrect values, such as .23 for age, etc.
    - **Have a sub-subsection for each variable** and show that you looked at all the criteria noted above. Show AND MEASURE the before and after. For example, if you find that a variable has 10 missing values and you update these with the mean, then the before is the mean before, and the after is the mean after. In this case, you should also include the variance before and after. The measure is \*very\* dependent on what you clean and how you clean it. So this will be for you to think about.
  + **Subsection 2: EDA: statistical and visual.** 
    - Create a vis and/or table for each variable in the dataset.
  + **Subsections 3 - n:**
    - Here you run all methods learned in this class. One subsection for each. Include tuning and/or different options as applicable - such as different kernels and C for SVM, different k for kmeans, etc.
    - Explain as you go. Pretend you are writing a tutorial paper.
  + Grade Notes:
    - If there is not a clear subsection (and sub-subsections) that are all about the data, dataset, data cleaning, variables, EDA, visual EDA, prep, etc. (up to -15)
    - If there is not a subsection dedicated to each required method noted above. (-10 for each missing method).
    - If there is not at least one vis for each required method in the Analysis area and within the subsection for that method. (-5 per missing vis)
    - If the model, model parameters (such as kernel for SVM), etc. are not well explained points can be lost.
    - If a model or method is not correctly applied or gives results that should not occur, points can be lost.
* Results
  + **Subsections 1 – n**
    - You will have and will discuss results, issues, and limitations for all the analysis. You will note which ones worked well, and why, which ones did not, and why, etc.
  + Grade Notes:
    - If any method is missing. (-10 per missing method)
    - At least one results-supporting vis for each method. (Missing a vis – 5). **These are not the same as the visualizations required in Analysis.**
    - The Results section will have a subsection for each method. Each method subsection will be different in the sense that each method is different, has different parameters, etc. For example, when you do kmeans, you will talk about the different values of k, how each performed, what each revealed, and which was selected and why. Results talk about what happened in each method, but technically.
* Conclusions
  + 3-4 paragraphs - NON-TECHNICAL. What was the outcome - what did you find, discover, predict, classify? Why does it matter to humans?
  + Grade Notes:
    - Less than 3 paragraphs (-10)
    - Any mention of technical results (which belong in the Results above) -10
    - A lack of flow or clarity about what actual findings were, why they matter, who they matter to, how they can improve things, etc…