Data Mining, Fall 2021
Final Exam

This exam is take-home, but is to be done INDIVIDUALLY, not in groups and not discussing with each other. You are free to use your book, the web, and your past assignments as you see fit.

Each of the following is a short-essay answer. Limit your answers to at most N words for each question, where N is specified for each question. Please type your answers directly into this word file and use the sample 12pt font size and double line spacing. Feel free to add tables/graphs as you see fit if that helps improve your answer. You do not need to run any code/experiments to answer these questions, but, if for some reason you want to you can.

Question 0:

I, Krista Miller, do declare that all work on this exam is my own and I have not discussed this with anyone.

Question 1 (10 points, maximum 1 page)

Why is it important to clean and normalize/modify your data for data mining? (In addition to the obvious dealing with missing and NaN values). Consider each of the five main algorithms we focused on: {Association Rules, Decision Tree Classification, Naive Bayes Classification, K-Means Clustering, DBscan Clustering}.

Question 2 (20 points, maximum 2 pages)

Consider the five algorithms we focused on: Association Rules, Decision Tree Classification, Naive Bayes Classification, K-Means Clustering, DBscan Clustering. When and why would you choose to use on over the other?

Question 3 (10 points, maximum 1 page)

For supervised learning we used the quality evaluation metrics of accuracy, precision, and recall. Why consider all three? Under what circumstances are different metrics better than the others? Provide a small example that illustrates your point.

Question 4 (10 points, maximum 1 page)

When is a datacube better to use than one of {association rules, classification, or clustering}? When are one of {association rules algorithms, classification, clustering} better than datacubes?