**Data Science**

**Spring semester**

**term project additional specification**

1. **Dataset**

Must contain at least a few categorical features and a few numerical features.

Must contain a reasonable amount of dirty data (missing data and/or wrong data) -- that is, you should not use mostly clean dataset, because you will not get to practice data cleaning.

The dataset size should be over 1,000 records.

1. **Data preprocessing**

You must show statistical summary of your dataset, explore the dataset (using plotting), and perform (and explain and show) a variety of data preprocessing that you learned in this course.

This is a major part of the term project.

1. **Algorithms**

Must use data scaling (standardization/normalization)

Must use 2 of the following 3 types of algorithm

regression (linear, polynomial, multiple),

classification (decision tree, k-nearest neighbor, random forest),

clustering (k-means, HAC)

Use of ensemble learning (which you will learn in 2 weeks in class) is optional (but encouraged)

You may use just one algorithm that was not taught in this course. (However, you will need to demonstrate (in your own way) that you really understand it; that is, you did not just copy code from some source without even knowing what it is.

1. **Evaluation**

Must use k-fold cross validation for testing classification models

May use confusion matrix or ROC (which you will learn in class in 2 weeks), or heat map, Matplotlib plots