# CSC 535/635 – Data Mining Project Guide

The goal of the course project is to apply data mining techniques to real-world problems. It is recommended to use scikit-learn for the project. You can use any built-in IPython or Python functions for the project. The use of other libraries may be allowed upon approval from the instructor.

The following guidelines will be used to evaluate projects:

* 1. Whether the motivation of the project is well described.
  2. Whether the dataset(s) is properly explained.
  3. Whether the choice of data mining techniques is well justified.
  4. How well the experimental results are discussed.
     1. Why does a method produce better results than others? (if applicable)
     2. Why do not the methods give significantly different performance? (if applicable)
  5. Whether the method is properly applied to realistic data set(s) that may be used in practice.
  6. How well the experimental settings are described.
     1. What preprocessing was done to the data? (if applicable)
     2. Was feature selection considered?
     3. What methods do you consider?
     4. How do you tune the optimal parameter(s)?
     5. What are the evaluation strategies (e.g., k-fold cross validation) for the performance comparison?
     6. How many times are the experiments repeated? (if applicable)
  7. How well the experimental results are discussed.
     1. How can the experimental results be interpreted?
     2. Are the results and data set properly visualized? (if applicable)
  8. Conclusion
  9. Misc.
     1. What obstacles have you overcome?
     2. Any issues that you want to share?
     3. Any future plans for related research?

For the project proposal, find a dataset that you may be interested in. You can refer to homework 1 for possible places to look for datasets. Kaggle.com and UCI Machine Learning Repository contain a lot of interesting datasets and well-defined problems. If you have any specific related dataset, you can suggest it too.