Data mining Project Guidance:

The goal of the project is for you to show that you can do something substantial, and practice the data mining process on some actual data sets. The goal of your reports are to provide *evidence of learning*. Show me you learned and now understand something you did not before.

Project Components (Five Parts):

1. **Project Proposa**l

   One to two pages including:

   a. Project title
   b. What you want to work on data mining
   c. What are the challenges involved
   d. Give a brief and precise description or definition of the problem or question
   e. Where are you getting the data, with appropriate references
   f. How will you evaluate your method? How will you measure performance or success of your method?
   g. Why did you select this project?

2. **Project Research Paper**

   Ten pages, 10 point font, double spaced one inch margins, with figures and with a final page of references.

   a. Project title
   b. No title page.
   c. Who has done this before?
   d. What challenges did they face?
   e. Are there any ethical considerations?
   f. Is there a business case for this work?
   g. What issues are there with the data?
   h. Where did you get the data?
   i. What can you say about the data?
   j. How clean is the data?
   k. What methods and algorithms were used in the past?
   l. Other issues related to the project?
   m. What distance metrics are used for comparing data records?
   n. How do you plan to solve the problem?

   Example:

   (a) We will create a dataset of all the Allstate data. We will clean the data and then break it into equivalent groups of training, testing and validation using Z-Score. We will also break each of these groups into the A-G groups, and have equivalent Training, testing and validation groups. We will use Principal components analysis to determine what columns/attributes are important.

   (b) Our hypothesis is that we may need to predict the Allstate groupings in 2 different ways: individually and all together as a group (ie. Group A-yes, Group B-No, Group C-Yes, Group D- Yes, etc.) as well as trying to predict each group individually.

   (c) We will implement a back propagation algorithm, J48, Naïve Byes and DBSCAN for clustering. We will also use a Support Vector Machine.

   o. How will you validate your results?
   p. What algorithms from our course do you plan to use?

   I expect you to try to use at least five standard algorithms (minimum) as a basis for comparison. Use more if you can.

3. **Mid-Project Progress Presentation/Report**
   A five minute presentation describing:
   a. The background of the project – because others will not know
   b. What you are trying to do
   c. How you are doing on it
   d. What data cleaning you are using
   e. What algorithms you used
   f. What you found
   g. What distance metrics you used
   h. What is interesting

4. **Final Project Report – probably due before the presentations**
   This is another eight to ten pages *added* to your earlier research paper. You will re-submit the entire report, with this added. This discusses what you did on the project, with graphs and results visualized.
   a. Which algorithms did you finally use?
   b. What went wrong, or what challenges did you face?
   c. What was interesting about this?
   d. Which algorithm worked best?
   e. What else would you like to share about the project?
   f. What did you learn about data mining by doing this project?
   g. An overview of the data, how you cleaned it and how you made up your test/train/validate sets
   h. Your results from all of the algorithms including confusion matrix (if applicable)
   i. A detailed description of each algorithm used and what parameters you used

5. **Project Presentation**
   A ten to twelve minute presentation describing your project. Topics might include:
   a. The background of the project. What would others want to know?
   b. What data cleaning did you use?
   c. What attributes and/or features were most important to you?
   d. Did you do any data reduction?
   e. Did you use principle components analysis to rotate your data?
   f. Did you modify your data?
   g. Did you discretize your data?
   h. What algorithms did you use?
   i. Did you do any data visualization?
   j. What did you find out?
   k. What distance metrics you used?
   l. What was interesting?
   m. What surprised you?
   n. Did everything go as planned?
   o. Etcetera…

Project Types to Select From:

1. **Traditional Data Project Investigation:**
   Explore one of the competitions (links to some of them follow).

   Re-create the CRISP process of data analysis, data cleaning…

   Re-create some of the findings from these groups.
   These projects use a significant amount of real-world data.

   Your task would be to describe the data, pull the data into one of the standard toolkits, build a classifier, and re-create the results as best you can.

2. **Alternative Presentation:**
   I am open to other ideas related to data mining that push your limits and expand your knowledge. The project must generate measurable or demonstrable results.

   Go get your own data from: CDC, NHSTA, NIH, USDA – United States Department of Agriculture, (Agricultural Research Service, USDA Branded Food Products Database), weather data from data underground, traffic flow data, highway construction data… Facebook, or Twitter.

   Note: getting your own data is different than creating your own data. Do not try to do your own study. Do not use data on: any sports teams, video games, or computer games.

   Examples:

   a. Cloud computing using Amazon Web Services – how to do it, and use it for data mining.

   b. Learning how to use graph theory and adjacency matrices to do social network analysis.

   c. Learning and incorporating text data mining – twitter data or facebook data, using an API.

   d. New trends in data mining.

   e. How to use Google TensorFusion on a project.

   f. How to incorporate data into Google Earth for visualization using an API.

   g. Displaying and demonstrating data that is geo-visualized.

   h. Something else I haven't thought of.

APPENDIX:
Googling for: "data mining data sets" generates many results, including the following. Most of these are pages that link to other pages…

http://www.kdnuggets.com/datasets/index.html

Including: **https://www.kaggle.com/kaggle/us-baby-names**

http://archive.ics.uci.edu/ml/

http://www.rdatamining.com/resources/data

http://www.abbottanalytics.com/data-mining-resources-sets.php

http://www.r-bloggers.com/datasets-to-practice-your-data-mining/ (some links are bad)

https://www.quora.com/Where-can-I-find-large-datasets-open-to-the-public

http://www.augmentedintel.com/wordpress/index.php/augmented-intel-free-online-analytics-applications-for-corporate-intelligence/searchable-list-of-public-data-mining-datasets/

http://web.stanford.edu/class/cs341/data.html

http://www.cs.cmu.edu/~awm/10701/project/data.html

http://www.cs.waikato.ac.nz/ml/weka/datasets.html

Old Links to KDD PROJECTS IN THE PAST: (Possibly bad links.)

General KDD website:
http://www.sigkdd.org/

Data mining research project on real data from KDD competitions
http://www.sigkdd.org/kddcup/index.php
KDD Cup is the annual Data Mining and Knowledge Discovery competition organized by ACM Special Interest Group on Knowledge Discovery and Data Mining. Below are links to the descriptions of all past tasks.

KDD Cup 2010: Student performance evaluation

KDD Cup 2009: Customer relationship prediction

KDD Cup 2008: Breast cancer

KDD Cup 2007: Consumer recommendations

KDD Cup 2006: Pulmonary embolisms detection from image data

KDD Cup 2005: Internet user search query categorization