Data Mining Project

- Your task is to analyze a dataset using some of the techniques that we have learned in class in order to solve a problem, answer a question, or simply explain the relationships within a dataset for some business or organization.

- Consider what data set you would want to use for your Data Mining project. You can use almost any dataset, with my approval. You have been provided with several links and we have a lab devoted to the task of identifying an appropriate dataset. A common mistake in this course is to assume that finding an appropriate dataset is a trivial task. *Do not underestimate the time that it will take to find, examine and review datasets for your proposal!!*

  Many datasets come with some type of metadata file. Be sure to download that as well, as it is crucial to your understanding of the data. Some of the datasets (such as the Breast Cancer datasets) might be valuable when considered alone or together with other datasets in the same folder. (The Breast Cancer data folder has a file on diagnosis and one on prognosis.) Note that the full version of our sample "Iris" dataset is posted there as well.

  Some may be beyond the scope of this course. For instance, Forest Fires uses some sophisticated Regression techniques. If you are comfortable with this, go for it. Otherwise, another dataset may be better for you. Alternatively, you can read the description of the work that has been done, and try both to duplicate it and to take a fresh look, using your own techniques.

1. What is the problem presenting to the business/organization?

2. What is the dataset available that you think may help to solve this problem?

3. Data Exploration: Describe the dataset:

   - The target (label, class) attribute, if the goal is classification. (For most of these datasets, it is.)
   - The regular attributes
   - Anything interesting in the descriptive statistics?
   - Missing values? If so, how did you handle them? (Try more than one way!)
   - Anything else that made this problem more interesting? More understandable? More challenging?

4. Data Visualization: In describing the dataset, did you attempt any graphs, charts, etc., to make the data more understandable? You will also probably want to use some graphs and charts for your results later.

   ➔ You must use at least two different visualization techniques. Explain why you chose them, and interpret the output. Possible techniques might be scatter plots, density

diagrams, histograms, etc. You can save those plots as .jpg and include them in your final project report and presentation.

5. Data Preparation:

    - Did you perform any attribute reduction? (Some of the datasets have a lot of attributes.)
    - Did you perform normalization? (Why? What models were you considering that might perform better with normalized data?)
    - Data Types: Discretization? Binning? Aggregation?
    - Missing values, as listed above.
    - Outliers? How did you choose to treat those?
    - Note that some classification approaches require certain data set characteristics (for example, they may only work on nominal/categorical data)

6. Model Selection:

    - What types of models did you consider? Why?

    - Identify at least two classification models that can be applied to this data, and build those models after applying the data preparation that is appropriate to each model. (Note that you may have to create different models, as each may require different data preparation.) Suggested approaches include one decision tree approach and one rule-based approach, such as n-nearest neighbor.

    - For each model that you select, explain why you thought that model was appropriate to use, given the data and the type of problem.

    - Evaluate the models. Explain the metric (accuracy?). If there is a confusion table, explain that as well.

    - If appropriate, apply one or more visualization techniques on the preprocessed and classified data (in addition to doing that on the original data in step 1).

7. Submit RapidMiner process flows (File→Save Process), as well as your output. Incorporate these into your report and your presentation, and submit a report and your PPT presentation.

- Many of the datasets do not have separate training and test data sets. Therefore, you can try several different techniques to split the data into training and test (scoring) sets, as discussed in class and demonstrated in lab. For some datasets, it is difficult to see what they're trying to test from the documentation given. For some, such as Heart Disease, or Glass Identification, or Contraceptive Method Choice, it's very clear.

- If you start early enough, you can take a risk. You are not alone—I am here to help you, and I'd love to help you do something a little beyond your comfort level!!
- As you are working, save your discarded versions, or your intermediate versions. Data Mining is a process. I would like to see your thought process in adding models, or determining that a model was not appropriate. I would like to see the visuals that were helpful and those that didn't add anything to your knowledge. You may not want to include everything in your presentation to the class, but do include your process in your report
- MOST IMPORTANT!! Do not assume that you will find a dataset in 5-10 minutes, and then you can focus on running some DM models. That's not going to happen. Count on spending some quality time experimenting with the dataset, to see what types of data conversion you need, or what types of models [don't] work well on those data.