Lab: Dataset Exploration

Goal: To identify three possible datasets and the possible DM problems that can reasonably be addressed by these datasets.

Problem: In many data mining applications, you are presented with data and with a business (or other domain) applications problem that must be addressed. However, for this project, you are required to identify a problem, and use a dataset that is appropriate for that problem. Often, the process is reversed: you will have to explore quite a few datasets, look at the format and the types of variables, and determine from that what types of data mining applications might be appropriate when using that dataset. This is "backwards", but it is valuable to look at these datasets and determine what types of data mining algorithms can be used and what types of problems these datasets can help to explore or to clarify. It is also "forwards" because you hopefully have some idea of the type of problem that you want to explore.

Please do not make the mistake of thinking that it is non-trivial to simply "find" or "get" a dataset. **This is the single most serious error that students have made in this course!!** You do not want to find yourself two weeks before the project is due, exploring a dataset that simply doesn't have the data that you need to fulfill the requirements of the project. Unless you already have a dataset in mind for a problem that you want to work on (perhaps from a research project, a previous class, or a work-related problem), you will have to spend some quality time with your team mates exploring some datasets. In fact, even if you already have a dataset in mind, you will have to determine that the dataset is appropriate for the type of project that you want to do in this course, and you are still required to look at alternative datasets and project ideas. This is because it has been my experience that once students start to work on a predetermined project, there are often changes in direction and scope that render the initial dataset no longer appropriate.

Process:

1. Identify three possible topics that your team is interested in exploring. This can be a problem of classification and/or clustering in practically any domain of interest. (For this project, you must use some kind of classification algorithm. You may also use clustering.) We have looked at some business applications, text mining, community-based, scientific and medical applications. Any of these, or any others that may interest you, are appropriate for the project.
2. Explore datasets: You may want to begin by looking at some of the links that I included in the document DatasetLinks. But you are by no means limited to those datasets!! If you have a specific scientific or other domain of interest, please explore those datasets as well. For each dataset:
    - **a.** *FORMAT:* Preview a subset of the dataset. It may be in a format that is compatible with RapidMiner. (With the newer rules of the community license, many types of data are now available for import.) Or, it may be in a format that is easily converted to a usable format. **What is the format of this dataset, and is it usable?**
    - **b.** *EXAMPLES/OBSERVATIONS/ROWS:* Look at the attributes, as well as any metadata (descriptive information) of the dataset. **What type of data do these**

examples/observations represent? In other words, what does each row (example) represent? You are not bound by the existing format. But: **Do these data have the potential to be converted into examples/observations that can be used by a set of DM processes in RM?**

c. *ATTRIBUTES/COLUMNS:* Do these attributes lend themselves to the types of analyses that we have discussed in class? If not, can processing be done that will render them usable? Are there numeric attributes that perhaps could be discretized? Is there something that looks like it could be a label/target attribute? Does it lend itself to classification? In other words, does this dataset have the potential to be used for this project? . Are there attributes that are highly correlated with each other, or even redundant? For instance, different levels of "location" data stored in different attributes might lead you to believe that there are quite a few independent attributes, when they are really semantically the same thing. **How many truly semantically distinct attributes are in this dataset? Try to state what type of problem(s) this dataset could address, given the attributes included in the dataset. Be as specific as you can, referencing the specific attributes that make this possible.**

d. *DATASET SIZE:* **Look at the examples: How many are there? Will you have to sample** (an extremely large sample can slow down your algorithms. So you may want to sample, at least initially.) **Do you have enough examples for both a training and a test set.**

e. *DATA PREP:* Look at the data: **Do you have a lot of missing values? Other issues that require cleaning? Normalization required? Changes of data types? What kinds of preprocessing do you think might be appropriate on this dataset,** especially considering the type of problem that you are considering? You want a dataset that will allow for some transformations, so that you can demonstrate that in your project. But you don't want a dataset that can't be transformed into something you can work with.

f. *Overall judgment:* Very briefly, do you feel that this dataset has the potential to be used for the project for this course.

3. Repeat step 2 for each dataset, and identify at least one dataset for each topic of interest. So that means a minimum of 3 datasets. Answer the questions that are in **bold** in step 2.
4. Given your examination of the three datasets, select one to explore further. For the selected dataset, run some preliminary RM processes on the dataset:
    a. Import into RM
    b. Look at some of the descriptive statistics
    c. Identify a target (label) attribute, even for a trivial classification application. This will force you to review on a practical level your earlier conclusions. It will also force you to focus on the types of classification that can actually be done with this dataset.
    d. Run a preliminary classification.

Submit: A document with clear bullet points for each dataset, and within each dataset, your answers for each of the questions in bold. For step 4, include a screenshot of the classification that you ran, and a one-paragraph description of what you learned about the dataset. One document per team.