MIS 545 – Data Mining for Business Intelligence

Course Project Instructions

The course project aims to simulate a business problem and apply the entire data mining process to solve this problem. Each person will choose a dataset and analyze it thoroughly, employing the techniques we have covered in the course. The focus of the project is to provide business intelligence analysis for making practical decisions.

Steps Involved:

1. Select datasets

Find the dataset that you will use for the project. There are numerous free datasets and data repositories that you can access online and use for this project. The important thing is finding an interesting one, which can lead to good research questions. Your dataset should contain a minimum of 5,000 records after preprocessing. If you have doubts about your chosen dataset, please check with the instructor to see if it is applicable for this project.

2. Define the business problem

Define the business problem you want to investigate using the dataset; for example, use customers' transaction data to predict the approval of a loan.

3. Analyze data

Use R language to analyze the dataset and solve the problem you defined earlier. Utilize at least **three data mining methods** to solve the problem. Two of them should be predictive methods (e.g., decision tree, naïve Bayes, SVM, neural network), and the other should be a descriptive method (cluster analysis, association rule mining).

The steps involved are:

- Pre-processing of the datasets
- Descriptive statistics and exploratory analysis
- Classification and prediction
- Cluster analysis or association rule mining
- Model evaluation

4. Analyze and interpret results

- Provide insightful observations and comments on the results of the experiments.
- Present results and recommendations in the video presentation and the final report.

Project assignment

● Team Building Form (optional)

Students have the option to work individually or in a two-person project team by submitting a *Team Building Form* before the end of week 1 (July 11th). Each student in a group will receive the same grade based on the assessment of the project description, project presentation, and project final report. Each group should be responsible for its own interactions and individual participation.

● Project Description

Each group or individual should submit a one-page project description summarizing the proposed project by August 3rd. The description should include:

- a brief introduction with the overall idea of your project, the goals, and business value (who cares, why it is an important problem)
- the data source(s) you would like to use
- at least three (3) different data mining techniques that will be used to solve your problem
    - two predictive methods (e.g., decision tree, naïve Bayes, SVM, neural network)
        - briefly explain why you choose each method.
    - AND one descriptive method (cluster analysis, association rule mining)

● Project Presentation Video

Please record and upload your project presentation video to the Panopto folder (*Tools-> Panopto-> Project Presentation*) by August 22nd. The video name should contain the *group name* and the *project name*. There is no time requirement for the presentation. Ideal duration will be under 15 mins. Presenter(s) also need to answer questions on the discussion board. The presentation grade will be determined by the peer reviews. Please see *Project Presentation Rubric* for detail. The video should include the content shown below:

- **Problem description and the current state of the domain:** Define the business problem(s) you want to investigate using the dataset. Also, briefly describe existing work or what others have done in this domain.
- **Dataset description: origin, data points, variables:** Your data set should contain minimally 5,000 data points after preprocessing. Describe its origin, the number of variables (at least 5), names or description of variables, and descriptive statistics for your data (*e.g.*, mean, standard deviation, min, max, *etc.*)
- **Data preprocessing activities and results:** Describe your data preprocessing activities. Describe the data transformations made, the rationale behind them, and the results with descriptive statistics. You need to show descriptive data that shows WHY you will preprocess your data as claimed. Then briefly describe the transformations you made, the rationale behind them, and the transformation results with descriptive statistics.

- **Algorithms used and rationale:** Describe your data mining approach. Which algorithms did you use, which variables, and why? Show details of the algorithm, such as the parameters used and the model illustrations (equations or plots). **OR Intended algorithms to be used and rationale:** Describe your data mining approach. Which algorithms do you plan to use, with which variables, and why? (The algorithms can be improved in the final report.)
- **Preliminary analysis and results:** Show the details of your preliminary analysis and the results. The analysis is based on a smaller dataset or a subset of variables, or both. The goal is to show that your project has a good chance of being successful.
- (**Optional) Results and Interpretation:** Show the results of your analysis. What are the explanations for these results? What conclusion, suggestion, or recommendation do you want to make?
- **(Optional) Evaluation:** Describe your evaluation of your project. This should be more comprehensive than one simple measure (such as accuracy). You should provide a detailed evaluation of strengths and weaknesses for the datasets, subsets of data, different algorithms, etc. The entire project and its outcomes need to be evaluated rather than simply an individual algorithm.

● Project Peer Review

Students receive a grade for completing peer reviews. Each person will evaluate 3 other teams/individuals and provide feedback. Please see *Project Presentation Rubric* for detail.

● Project Final Report

The final project report will be due on the last day of the semester, August 29th. The report should include the following items:

1. Executive summary of results and findings
2. A section about the business problem, the importance of the problem and the implications of solving the problem, etc.
3. A section about the dataset: What the data is about, what the records and attributes are, what kind of preprocessing it required, etc.
4. A section for summary statistics of data.
5. One section for each model: Two models are predictive methods, one is descriptive. You need to explain each model and justify the parameters that you use.
6. Results from model executions.
7. Model evaluation and a recommendation of a better model.
8. Implications and conclusion
9. Additional screenshots, images, etc., can be provided in appendices.

There is no page limit or minimum page requirement for the report. However, 10 pages should be long enough to contain all necessary content.

Common mistakes to avoid

While every project is different, there are some fundamental data analysis mistakes to avoid. A few projects inevitably contain these errors every semester.

- Results do not answer the question.
- Using clustering for a classification problem, and vice versa.
- Using continuous *dependent variables* in classification
- Using discrete variables in clustering
- Failing to report evaluation criteria for a classification model (e.g., accuracy, precision, recall).
- Misunderstanding summary statistics as models.

Project Schedule

	Availability Period	
Project Deliverable (Sorted by Due Date)	**Open & Available (12:30am AZ Time)**	**Close & Due Date (11:59pm AZ Time)**
Team Building Form	Monday, 07/05/21	Sunday, 07/11/21
Project Description	Monday, 07/19/21	Tuesday, 08/03/21
Project Presentation Video	Monday, 08/9/21	Sunday, 08/22/21
Project Peer Review	Monday, 08/16/21	Wednesday, 08/25/21
Project Final Report	Monday, 08/16/21	Sunday, 08/29/21

Project Grades

Project	**15%**
Project Description	2%
Project Presentation Video	5%
Project Peer Review	1%
Project Final Report	7%