**CS453 Data Mining Fall 2022**

**Course Project**

***(100 points in total)***

Your Course project consists of the following steps/components:

1. Find a dataset (choose any available on-line. Do not choose the ones given for the Data Mining textbook). You may modify an existing one and create your own. Your dataset should have reasonable number of attributes and reasonable number of instances. Your dataset should have at least 3 nominal attributes and at least 3 numerical attributes. You should understand the dataset and its attributes to be able to ask meaningful Data Mining questions. - **Due: November 1.**
2. Formulate at least 3 data mining problems that makes sense for your dataset (Classification? Association? Clustering? Numerical Prediction?). The number of Data Mining questions in your project must be not less than 3\* (the number of team members)! - **Due: November 10.**
3. Choose models for knowledge representation (Rules, Decision Trees, ...) for each of the questions/problems asked. - **Due: November 10.**
4. Chose a few (at least 3 algorithms) algorithms/methods available/implemented in WEKA for your problem/knowledge representation. – **Due: November 10.**
5. Prepare your dataset for mining:
6. Explore your dataset for irrelevant and/or duplicate attributes. Decide what attributes are related to the problems/questions should be considered and what attributes should be dropped. If you are discarding some of them, explain your decision!
7. Are there missing attribute values? If yes, fix the problem. Describe the method chosen for it. Explain your rational.
8. Explore your dataset for errors. If errors are detected fix them
9. Write your dataset in .arff format if it is not already. – You need to use WEKA for Data Mining problems.
10. Explore any other issues that your dataset might have and fix them before doing any Data Mining processing.

**Due: November 15.**

1. Use R to do interesting and meaningful visualization for your dataset. – Your project must have at least 4 data visualizations for your dataset per a team member! - **Due: November 25.**
2. Use R to calculate interesting and meaningful statistics for your dataset: Analyze numerical attribute dependency using correlation measure. Use correlation map to visualize the attribute correlation values. Analyze linear dependency between pairs of numerical attributes. Use statistical summary for numerical attributes as well as frequency distribution for nominal attributes. - Your project must have at least 3 different statistical analysis results per a team member. - **Due: November 25**
3. Run your Data Mining algorithms. Use different Test options. Evaluate your results for efficiency/accuracy. Fr each Data Mining question you ask/problem you solve use at least 3 different algorithms with at least 4 testing options. Compare results by efficiency. **Due: November 25.**
4. Write an abstract for your project (one paragraph describing what was done) - **Due: November 30.** (**Please email it to me with the names of the team members**) and include it into your Final report. Your project abstract should have the following structure:

**Authors:** *names for the members of you project team*

**Title:** *your project title*

**Abstract:** *one paragraph describing in general what you have done and what are the results*

1. Create a slide presentation for your project – **Due on your presentation dates: Course projects presentations will run 2 – 3 last classes based on the number of projects/presentations. Email me your project presentation slides – Due: Thursday, December 8, till 5:00pm.**
2. Present your project in class (15 - 20 minutes) – **Due on your presentation dates: Course projects presentations will run 2 – 3 last classes based on the number of projects/presentations. All course projects have to be presented in class. If a course project is not presented in class, you are losing 10 points of your course project grade.**
3. **Be prepared to participate at the “Celebration of Scholarship and Creativity” Day, WSU, Wednesday, middle of April, 2023 – poster presentations.**
4. Write a final report for your project - **due on December 15, till 11:00pm, Word file - Blackboard submission.**

**The Final Report should have the following structure:**

1. Project Abstract

**Authors:** *names for the members of you project team*

**Title:** your project title

**Abstract:** *one paragraph describing in general what you have done and what are the results*

1. Describe your dataset: size, attributes, the attribute type and possible values;
2. Describe any problems with the dataset: too many attributes, missing values, …;
3. Describe solutions you made dealing with that (C.)
4. Include visualizations done for the dataset and discuss them;
5. Include calculated statistics for the dataset and discuss them;
6. Define/Describe your Data Mining problems (Define/describe your Data Mining questions/problems such that they are understandable by a general population)
7. Describe your Knowledge representation model/s used for each of your Data Mining problems.
8. Describe what algorithms you are choosing/running
9. Describe the choice of your training/testing dataset sizes (testing options used)
10. Report efficiency results for each run. Report discoveries, if there are any.
11. Analyze efficiency of the results for each model constructed. – Is there anything interesting? Is there anything strange? Try to explain, if it is.
12. Describe possible application/s of your project.
13. Describe future work: what questions you would like to answer/solve if you continue working on the dataset in the future?
14. Reference:
    1. Reference to any literature/courses used
    2. Reference to any tools used
    3. Programming languages used in the project

**Final Report is due on December 15, till 11:00pm, Word file - Blackboard submission.**

**You may work individually or as a team. A team should consist of not more than three students. If you need any of my help/advice about any of the project steps/phases, please let me know!**

**Here are some links to available, interesting datasets:**

1. <http://www.kdnuggets.com/2011/02/free-public-datasets.html?c2b>
2. <http://www.kdnuggets.com/datasets/index.html>
3. <http://www.cs.waikato.ac.nz/ml/weka/index_datasets.html>
4. data.world