CS699 – Spring 2019
Project Assignment

The goal of the project assignment is to give students an opportunity to perform a classification data mining task. The project must be performed by a team of two students. Every team must present their project.

You choose a real world dataset, define your own data mining goal, and perform necessary data mining tasks to achieve the goal. It is strongly suggested that you choose a data mining goal that has a potential for practical use. Once you build data mining models, you must evaluate the data mining result using appropriate performance measures. You should not select a synthetically generated dataset and you should not use a dataset from UCI Machine Learning Repository. You must also avoid using a dataset on the Kaggle web site that has been used by many people. You may want to check government (federal, state, or municipal) web sites.

The following specifies minimum requirements. You can choose a larger dataset and you can perform additional tasks not mentioned in the requirements if you want.

- The project must be "classification." If you are interested in other types of data mining, indicate it in your proposal. I will review it and may approve it.

- Dataset minimum requirements
    - At least 20 attributes
    - At least 300 tuples

  If you are interested in a certain dataset but it does not meet the above requirements, then indicate that in your proposal. I will review it and may approve it.

- Data mining minimum requirements
    - You need to consider at least four attribute selection algorithms (which are implemented on Weka) plus a set of attributes chosen by yourself.
    - You need to build classifier models using at least four different classifier algorithms for each chosen set of attributes. So, you need to build and test total 20 classifier models.
    - You may try any data preprocessing/preparation/transformation to increase the performance of your classifier models.
- Model testing
    - Once you complete data preprocessing, you must split your dataset into a training dataset and a test dataset. You must make sure that the class distribution is preserved in both datasets.
    - You build your models from the training dataset and you test your models on the test dataset.

- Performance comparison
    - Compare performance of all 20 classifier models you built using the following performance measures: accuracy, TP rate, FP rate, ROC curve (or area under curve), and other measures if you want.
    - Choose one model that you think is the best for your data mining goal. You need to justify why you chose that model.
    - Since you need to use at least five attribute selection algorithms and at least four classifier algorithms, you need to compare at least twenty classifier models.

Schedule and Deliverables

(Only one member of each team needs to submit proposal, intermediate report, and final project report.)

1. Proposal
    a. Due: **2/6**
    b. Include the names of your team.
    c. Include detailed description of your dataset. Your description must include the names and meanings of all attributes as well as the number of tuples and the number of attributes.
    d. Clearly state your data mining goal (e.g., I want to predict whether a new customer will buy a computer or not).
    e. **Clearly indicate which attribute is the class attribute**.
    f. You also need to submit your dataset.
2. Intermediate report
    a. Due: **3/6**
    b. Include detailed description of what you did so far.
    c. You need to submit your training dataset and test dataset. If you changed your dataset, then you must resubmit your initial dataset.
    d. If not sufficient progress has been made, up to 10 points will be deducted.
3. Project (and the report) due: **3/27**
   You must submit all project documentation as described below. This is a hard deadline and there will be a 10% late penalty per day after that.
4. Project report:
    a. A project report should include:
        (1) Cover page
        (2) Statement of your data mining goal
        (3) Detailed description of the dataset
        (4) Detailed description of data mining tool(s) or algorithm(s) you used
        (5) Detailed description of data mining procedure (the procedure you actually followed) including all data preprocessing you performed
        (6) Data mining result and evaluation:
            a. You must include all performance measures, including confusion matrices, from Weka’s output window for all 20 models.
            b. Justification for your selection of the best model

(7) Discussion and conclusion, including what you learned from this project.

   b. In your report, you must clearly state what each team member did for this project.
   c. Your report must be at least 10 pages long (with 12pt font and single spaced).
5. You also need to submit **all datasets**, including the initial dataset and the final dataset plus intermediate datasets if necessary.
6. Other deliverables may be required based on the nature of your individual project, which will be determined later after I have more information about your project.
7. Presentation:
   a. Each team will have 15 minutes for presentation.
   b. All students must be present in the class during the presentations. If you do not attend a presentation (when other teams are presenting), 3% will be deducted for each unattended presentation.

Grading

- Project overall and project report: 70%
- Presentation: 20%
- Participation: 10%

Project overall and report (70)

- The intermediate report is due 3/6. If not sufficient progress has been made by that time, up to 10 points will be deducted.
- Project report is due 3/27. There is no grace period and there will be a late penalty of 10 points per day if you submit late.
- Whether the data mining result is practically usable. If your dataset was not used by other people (to the best of your and my knowledge), your project has potential for some practical use, and the performance of your model is reasonably good, then you may get an extra credit up to 10 points.
- Technical soundness of your approach. Otherwise, up to 10 points will be deducted.
- The performance of your best classification model. Note that there is no performance threshold which is used to grade your project. This is because different datasets and different data mining goals can result in different performance. I will use my own judgement considering your dataset and your data mining goal. If the performance of your models is very low (e.g., 60% or lower accuracy), then you must try to increase the performance and/or try to explain why it is so low. If you do not address such a low performance in one way or another, up to 10 points will be deducted.
- Whether all necessary components are included in the documentation. Otherwise, up to 15 points will be deducted.
- Organization of your documentation. If your documentation is poorly organized, up to 5 points will be deducted.

- Whether your discussion and conclusion is substantive and technically and logically sound. Otherwise, up to 10 points will be deducted.

Presentation (20)

- Presentations will be done on 4/10, 4/24, and 5/1.
- The order of presentation will be determined alphabetically by the last name of your team members.
- Presentation slides are due on:
    - 4/7 for 4/10 teams
    - 4/21 for 4/24 teams
    - 4/28 for 5/1 teams
    - If you submit late, there will be 1 point late penalty per day.

Your presentation will be graded based on the following criteria.

- Whether the presentation accurately represents what you did. Otherwise, up to 2 points will be deducted.
- Whether presentation material is well organized in describing what you did. Otherwise, up to 2 points will be deducted.
- Whether graphs and/or tables were well utilized to present the result. Otherwise, up to 2 points will be deducted.
- Whether questions are properly answered. Otherwise, up to 2 points will be deducted.

Participation (10)

- If a student misses any presentation day, 3 points will be deducted per one missing day.

Important

It is very important that I should be able to reproduce your data mining model and data mining result based on your documentation. So, the description of your data mining procedure, including all preprocessing you performed, must be detailed and accurate. If I cannot reproduce your model and result, you will lose up to **40 points**.