

## CS699 A2 – Fall 2022 Project Assignment

### Important:

- **You must not use a project that you used or you are using for some other course as a project for this course. It is a violation of BU Academic Conduct Code.**
- **You must not use the datasets from UCI Machine Learning Repository.**
- **You must not use the following datasets:**
  - NBA 5 year
  - Employee attrition
  - Accident (unless new data)
  - Airport customer satisfaction survey
- **Try to find a "fresh" dataset.**
- **You may use a dataset from Kaggle only if the dataset has been downloaded less than 500 times.**
- **When you submit deliverables, always include names of your team members as part of file names. For example, if two members are John Smith and Becky Hill, then the file names should be (notice that names are sorted by the last names):**

**Hill\_Becky-Smith\_John-Proposal.pdf**  
**Hill\_Becky-Smith\_John-Presentation.pdf**  
**Hill\_Becky-Smith\_John-Report.pdf**

The project must be performed by a team of two students. Every team must present their project.

The first thing you need to do is to form a team. On the Blackboard course site, there is *Emails* link on the left pane. You can communicate with other students using the link.

Once you form a team, you must include team members in your proposal. **Only one student** needs to submit a proposal, report, and slides.

The goal of the assignment is to give students an opportunity to perform a **classification** data mining task.

### Project Outline

- Choose a dataset, define a data mining goal (classification), and submit a proposal.
- Proposal approved.
- Perform appropriate preprocessing as needed.
- Select 5 classification algorithms.

- For each classification algorithm, build a classification model from the preprocessed dataset and test it using 10-fold cross-validation, and collect and keep the performance result.
- Split the preprocessed dataset into a training dataset and a test dataset.
- Repeat the following 5 times:
  - Choose an attribute selection method and create a reduced training dataset and a reduced test dataset.
  - From the reduced training dataset, build 5 classification models (using the same 5 classification algorithms you selected above) and test them on the reduced test datasets. You must use the same 5 classification algorithms in all five iterations.
- Now, you have built and tested 25 classification models.
- Compare them and select one model as your best model.
- Compare the performance of your best model with the performance of the model that was built using the same classification algorithm from the dataset with all attributes (the dataset after preprocessing).

Refer to the diagram in the next page for an overall process.

Detailed requirements are given below.

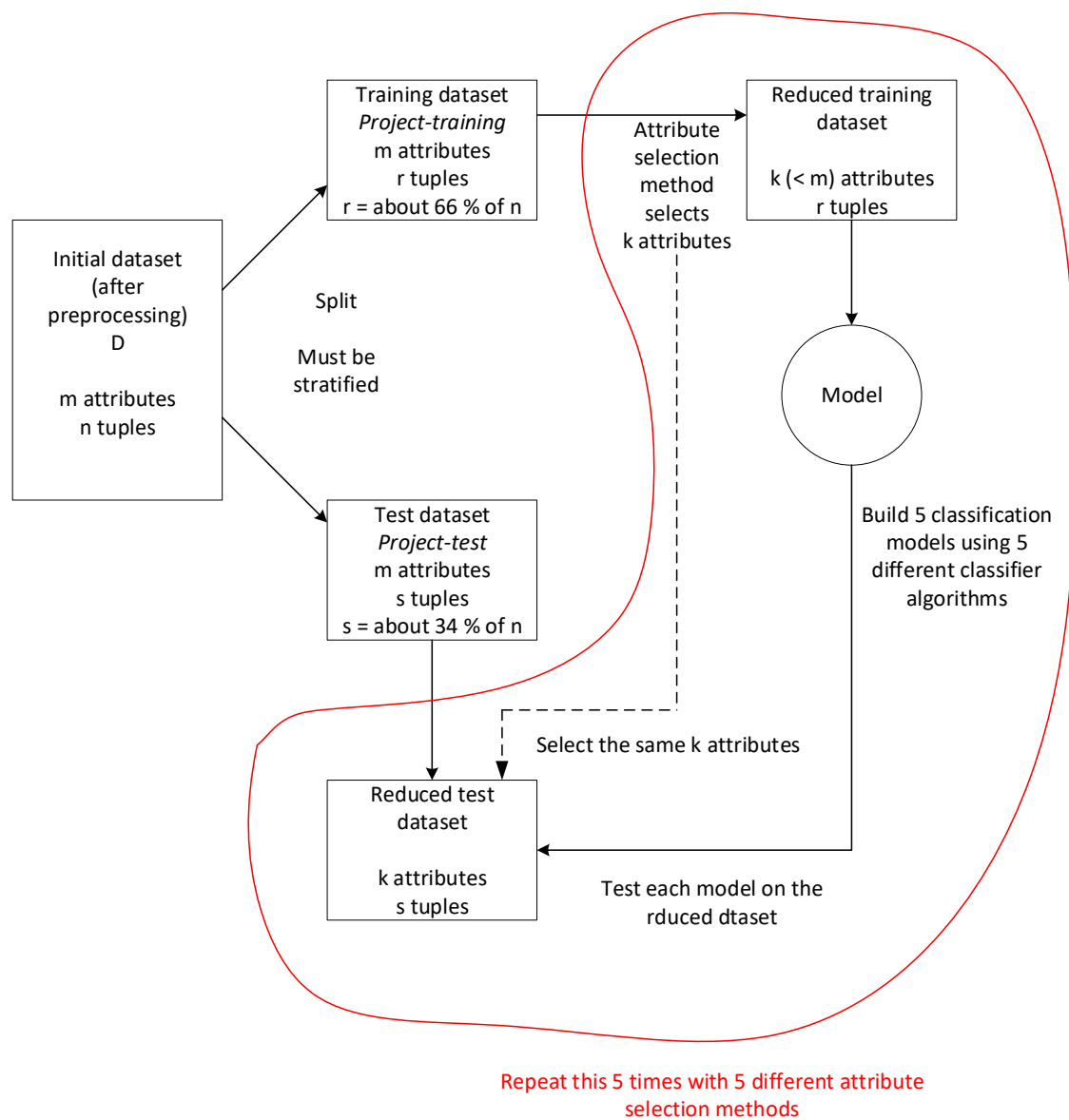
You choose a real world dataset, define your own data mining goal (a **classification**), and perform necessary data mining tasks to achieve the goal. It is strongly suggested that you choose a data mining goal that has a potential for practical use. It is also strongly suggested that you find a “fresh” dataset, which, to the best of your knowledge, was rarely used by other people. 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 website that has been used by many people. You may want to check government (federal, state, or municipal) websites.

Once you build data mining models, you must evaluate the data mining result using appropriate performance measures.

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.**”
- Dataset minimum requirements
  - At least 20 attributes
  - At least 500 tuples
- Data mining minimum requirements
  - You need to consider at least five attribute selection methods.
  - You need to build classifier models using at least five different classifier algorithms for each chosen set of attributes. So, you need to build and test total at least 25 classifier models.

- You may try any data preprocessing/preparation/transformation to increase the performance of your classifier models.
- Model building and 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 (**not on the training dataset**).
  - Refer to the diagram above.



- Performance comparison
  - Compare performance of all 25 classifier models you built using appropriate performance measures. You must decide which measure to use when comparing models, and you must explain why you used those measures.
  - Choose one model that you think is the best for your data mining goal. You need to justify why you chose that model.
  - Compare the performance of your best model with the performance of the model that was built using the same classification algorithm from the dataset with all attributes (the dataset after preprocessing).

### Schedule and Deliverables

**(Only one member of each team needs to submit deliverables)**

1. Proposal
  - a. Due: 9/22
  - b. Include the names of your team.
  - c. Dataset description: You must include the source of your dataset and detailed description of it. 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. Final project report due: 11/3  
 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 the deadline.
3. 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) Brief description of data mining tool(s) you used
    - (5) Brief description of classification algorithms you used.
    - (6) Brief description of attribute selection methods you used.
    - (7) The set of attributes selected by each attribute selection method.
    - (8) Detailed description of data mining procedure (the procedure you actually followed) including all data preprocessing you performed.
    - (9) Data mining result and evaluation:
      - a. Performance measures of all 25 models, including all 25 confusion matrices. Do not include the screenshots of confusion matrices. Instead, you must create your own confusion matrices in your report.

- b. You must present your result using tables, graphs, charts, or in other visual format so that readers of your report can easily and effectively understand your result.
  - c. Justification for your selection of the best model.
- (10) 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).
- 4. When you submit your project report, you also need to submit **the following datasets**, including
  - a. Initial dataset
  - b. The dataset after preprocessing
  - c. Initial training dataset and test dataset
  - d. The training dataset and the test dataset that were used for your best model
  - e. Other intermediate dataset(s) if needed
- 5. Other deliverables may be required based on the nature of your individual project, which will be determined after I have more information about your project.
- 6. Presentation:
  - a. Each team will have about 25 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), 5 points will be deducted for each missed presentation.

#### Software Tools and Performance Measures

- You may use any software tool(s) to perform data preprocessing.
- You may use any software tool(s) for attribute selection.
- You may use any software tool(s) to build and test classifications models.
- You must submit all screenshots, programs/scripts, and relevant files that illustrate all steps of the project you performed.
- Regardless of which tool(s) you use, you must collect and include the following performance measures for each of 25 classification models:
  - Confusion matrix
  - **For each class:** TP rate, FP rate, precision, recall, F-measure, ROC area, and MCC.
  - Weighted averages of the above performance measures.
  - An example is shown below:

```

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
          1.000    0.000    1.000     1.000    1.000     1.000    1.000    1.000    Iris-setosa
          0.960    0.040    0.923     0.960    0.941     0.911    0.992    0.983    Iris-versicolor
          0.920    0.020    0.958     0.920    0.939     0.910    0.992    0.986    Iris-virginica
Weighted Avg.   0.960    0.020    0.960     0.960    0.960     0.940    0.994    0.989

=== Confusion Matrix ===

  a  b  c  <-- classified as
50  0  0 | a = Iris-setosa
 0 48  2 | b = Iris-versicolor
 0  4 46 | c = Iris-virginica

```

- The above is an output from Weka. It also includes PRC Area. However, the PRC Area is not required (if you use other tools).

### Grading

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

### Project overall and report (70)

- Project report is due 11/3. 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 final report. Otherwise, up to 15 points will be deducted.
- Organization of your documentation. If your documentation is poorly organized, up to 10 points will be deducted.
- If your results are not effectively presented using tables, graphs, and/or charts, up to 10 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 11/17, 12/1, and 12/8.
- The order of presentation will be determined randomly.
- Presentation slides are due as follows:
  - Teams presenting on 11/17: 11/14
  - Teams presenting on 12/1: 11/28
  - Teams presenting on 12/8: 12/5
  - 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 3 points will be deducted.
- Whether presentation material is well organized in describing what you did. Otherwise, up to 3 points will be deducted.
- Whether graphs and/or tables were effectively utilized to present the result. Otherwise, up to 3 points will be deducted.
- Whether questions are properly answered. Otherwise, up to 3 points will be deducted.

### Participation (10)

- If a student misses a presentation, 53 points will be deducted for each missed presentation.

### **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**.