CS699 A2 – Spring 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.

There are three options. You must choose one of these options. I strongly suggest that you choose Option 1.

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.

Option 1

The goal of option 1 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 five 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 (details will be discussed later).
- Split the preprocessed dataset into a training dataset and a test dataset.
- Choose five attribution selection methods and prepare five reduced training datasets.
- From each reduced training dataset, build five classification models (using the five classification algorithms you selected above) and test them on the corresponding reduced test dataset.
- 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).

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.

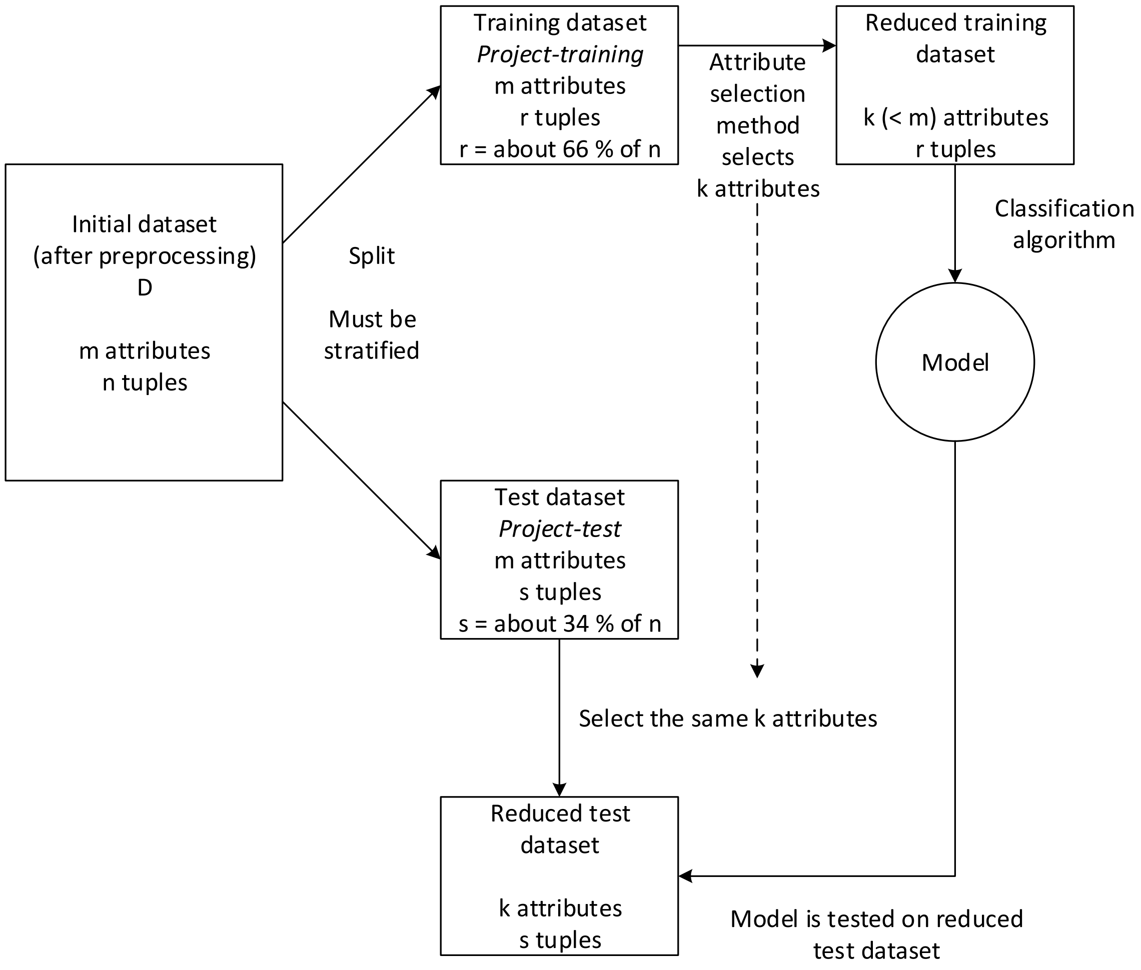
- The project must be "**classification**."

- Dataset minimum requirements
  - At least 20 attributes
  - At least 500 tuples

  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 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 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**.

The following diagram is a simplified illustration of an overall process:



1. 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.

Schedule and Deliverables

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

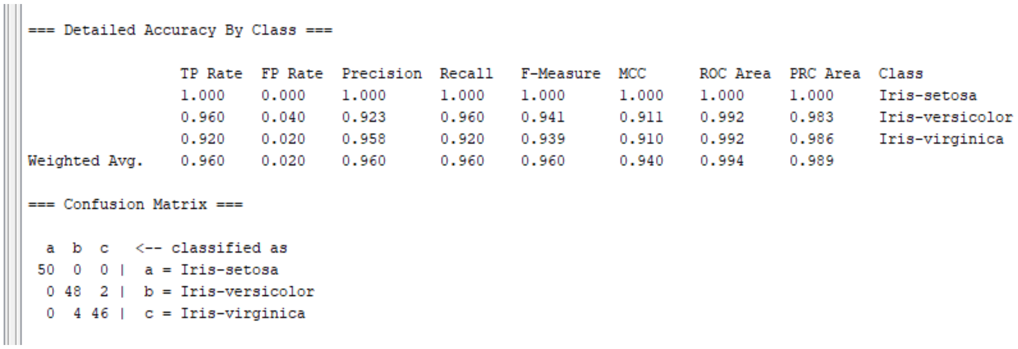
1. Proposal
    a. Due: 2/9
    b. Include the names of your team.

   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: 4/6
   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.
         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 **all 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 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 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 4/6. 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 4/13, 4/27, and 5/4.
- The order of presentation will be determined randomly.
- Presentation slides are due as follows:
  - Teams presenting on 4/13: 4/10
  - Teams presenting on 4/27: 4/24
  - Teams presenting on 5/4: 5/1
  - If you submit late, there will be 1point 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, 3 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**.

Option 2

Option 2 is an experiment to determine whether a bagging method and a boosting method increase the performance of classifier models. Follow the instruction given below.

- Select 20 datasets for classification.
- Select 5 classification algorithms.
- For each dataset *D* and each classifier algorithm *A*, perform the following:
  - Run *A* on *D*, with 10-fold cross-validation chosen as the test method, and collect the following performance measures: *TPR*, *FPR*, *F-measure*, and *AUC*.
  - Run *Bagging* with *A* on *D*, with 10-fold cross-validation chosen as the test method, and collect the following performance measures: *TPR*, *FPR*, *F-measure*, and *AUC* for each class.
  - Run *AdaBoostM1* with *A* on *D*, with 10-fold cross-validation chosen as the test method, and collect the following performance measures: *TPR*, *FPR*, *F-measure*, and *AUC* for each class.
- You must repeat the above 100 times (20 datasets x 5 classifier algorithms).
- Then, organize your result, present your result (as a table, graph, or any other format), and draw your conclusion. Try to be creative when you present your result so that your result may be effectively conveyed to readers of your report. Remember that your goal is to determine whether those ensemble methods increase classifier performance.

Schedule and Deliverables

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

1. Proposal
   - Due: 2/9
   - Submit all datasets you chose.
   - Description of all datasets:
     For each dataset, you must include:
     - The name of the dataset
     - The number of tuples and the number of attributes
     - Names and meanings of all attributes
     - Name of the class attribute and class distribution
     - Source of the dataset
   - Names of the classification algorithms you chose

2. Project report
    - Due: 4/6
    - Your project must include:
        - Cover page
        - If you performed any preprocessing on any dataset, you need to describe in detail the preprocessing you performed and you also need to submit the final dataset that was created after the preprocessing.
        - Result of the experiment: You need to 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.
        - Discussion and conclusion

Grading

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

Project overall and report (70)

- Project report is due 4/6. There is no grace period and there will be a late penalty of 10 points per day if you submit late.
- If the whole or part of the experiment is not technically sound/correct, up to 20 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 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.
- If the presentation of your result is considered "excellent" you will get extra 10 points.

Presentation (20):

- Schedule: Same as Option 1

- 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 well 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, 3 points will be deducted for each missed presentation.

Option 3

This option is an experiment to compare an undersampling method and an oversampling method to handle unbalanced datasets. Follow the instruction given below.

- Select at least 10 unbalanced datasets for classification. Make sure that the class attribute is a binary attribute and the fraction of the minority class is no more than 20%.
- Select 5 classification algorithms.
- For each dataset *D* and each classifier algorithm *A*, perform the following:
  - Split D into a training dataset D_{tr} and a test dataset D_{ts}. Use about 2/3 as the training dataset and 1/3 as the test dataset. Make sure that the class distribution is preserved.
  - Build a classifier model using the algorithm *A* from the training dataset D_{tr}, and test the model on the test dataset D_{ts}, and collect the following performance measures: *TPR*, *FPR*, *F-measure*, and *AUC* for each class.
  - From the training dataset D_{tr}, create an undersampled dataset $D_{tr\text{-}us}$.
  - Build a classifier model using the algorithm *A* from the undersampled training dataset $D_{tr\text{-}us}$, and test the model on the test dataset D_{ts}, and collect the following performance measures: *TPR*, *FPR*, *F-measure*, and *AUC* for each class.
  - From the training dataset D_{tr}, create an oversampled dataset $D_{tr\text{-}os}$.
  - Build a classifier model using the algorithm *A* from the overrsampled training dataset $D_{tr\text{-}os}$, and test the model on the test dataset D_{ts}, and collect the following performance measures: *TPR*, *FPR*, *F-measure*, and *AUC* for each class.
- You must repeat the above at least 50 times (at least 10 datasets x 5 classifier algorithms).
- Then, organize your result, present your result (as a table, graph, or any other format), and draw your conclusion. Try to be creative when you present your result so that your result may be effectively conveyed to readers of your report. Remember that your goal is to determine whether undersampling is better or oversampling is better for unbalanced dataset.

- You may try other methods to address the issue of unbalanced dataset for classification.

Schedule and Deliverables

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

1. Proposal
   - Due: 2/9
   - Submit all datasets you chose.
   - Description of all datasets:
     For each dataset, you must include:
     - The name of the dataset
     - The number of tuples and the number of attributes
     - Names and meanings of all attributes
     - Name of the class attribute
     - Show which class is the minority class and which class is the majority class, and also show the ratio of the two.
     - Source of the dataset
   - Names of the classification algorithms you chose

2. Project report
   - Due: 4/6
   - Your project must include:
     - Cover page
     - If you performed any preprocessing on any dataset, you need to describe in detail the preprocessing you performed and you also need to submit the final dataset that was created after the preprocessing.
     - Result of the experiment: You need to 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.

     Discussion and conclusion

Grading

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

Project overall and report (70)

- Project report is due 4/6. There is no grace period and there will be a late penalty of 10 points per day if you submit late.
- If the whole or part of the experiment is not technically sound/correct, up to 20 points will be deducted.
- Whether all necessary components are included in the project 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.
- Whether your discussion and conclusion is substantive and technically and logically sound. Otherwise, up to 10 points will be deducted.
- If the presentation of your result is considered “excellent” you will get extra 10 points.

Presentation (20)

- Schedule: Same as Option 1

- 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 well 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, 3 points will be deducted for each missed presentation.