CS699 – Spring 2025

Project Assignment

The project is a practice of building and testing classification models using a real-world dataset. The goal of the project is to let students explore different methods to build high-performance classification models. The project ***must*** be performed by a team of two students.

Dataset

The project dataset is a part of 2023 American Community Survey (https://www.census.gov/programs-surveys/acs). The original dataset was downloaded from https://www2.census.gov/programs-surveys/acs/data/pums/2023/1-Year/ and modified for this assignment.

The given dataset *project_data.csv* has 4318 tuples and 117 variables.

In the dataset, each tuple is a person and *Class* is the class attribute. The class attribute value represents whether a person has difficulty living independently.

Goal

The goal of the project is, using the given dataset, to build multiple classification models, compare their performances, and select the "best" model.

Requirements

The requirements of the project are divided into three parts – process requirement, performance requirement, and deliverable requirement.

Process Requirement

- In general, you may use any preprocessing method and you may use any classification algorithm to find a model with high performance.
- You must use R for the whole project, including preprocessing.
- You must use at least two methods to create balanced datasets.
- You must use at least three attribute selection methods (after preprocessing).
- You must use at least six different classification algorithms.
- Refer to a simplified overall process diagram attached at the end of this assignment.

Performance Requirement: Refer to the "Grading Guideline" section.

Deliverable Requirement: Refer to the "Final Report" section.

Project Schedule

- Intermediate Report
  - Due date: 2/18
  - You are expected to finish preprocessing by the due date. However, you can revisit and redo preprocessing if needed later (so this preprocessing may not be your final preprocessing).
  - You may want to perform data exploration to better understand the data.
  - You must submit an intermediate report that includes ***detailed*** description of all preprocessing you did. You also need to submit your R code.
  - Your intermediate report must include a cover page with the names of team members.
  - Late submission penalty: 2 points will be deducted for each late day.
- Final Report
  - Due date: 4/7
  - Your final report must include:
    - Cover page (including the names of team members)
    - Brief description of data mining tool(s) you used.
    - Brief description of all classification algorithms you used.
    - ***Detailed*** description of data mining procedure (the procedure you actually followed) including all data preprocessing you performed. This must be a step-by-step description which I may have to rely on to reproduce your results.
    - Data mining result and evaluation:
      - Performance measures of all models you built.
        For each model, you must include the following performance measures:
        - **Confusion matrix**
        - **Performance measure table:** Show, for each class, TP rate, FP rate, precision, recall, F-measure, ROC area, MCC, and kappa statistic. You must also show the weighted average of each measure (see an example format below).
        - Any other additional measures if you want.

      Do not include the screenshots of confusion matrices and performance measures in your report. Instead, you must create confusion matrices and performance measure tables in your report yourself.

      Example format of performance measure table:

	TPR	FPR	Precision	Recall	F-measure	ROC	MCC	Kappa
Class No								
Class Yes								
Wt. Average								

      - 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.
      - All parameters of your final best model (the best among 36 models).
      - Discussion and conclusion.
      - Division of work between team members.
    - You must also submit the following datasets along with the final report:
      - The dataset after all your preprocessing. Name this *preprocessed_data.csv*.
      - The initial training and the test datasets. Name these files *initial_train.csv* and *initial_test.csv*, respectively.
      - If you created an intermediate file(s) that were used to build models, you must submit those file(s), if needed.
  - Your final report must have at least 20 pages and at most 40 pages.
  - Late submission penalty: 5 points will be deducted for each day.
- Presentation
  - Presentations are scheduled on 4/23 and 4/28.
  - All teams must submit presentation slide by 4/14.
  - Late submission penalty is 2 points per day.
  - All students must attend both presentations. If a student misses a presentation, 5 points will be deducted (for each missed presentation).

Grading Guideline

- Points distribution
  - Intermediate report: 10%
  - Presentation: 10%
  - Final report and overall project quality: 80%
- Performance criteria
  - Minimum requirement
    - class Yes TPR >= 81%   AND class No TPR >= 79%
  - If the minimum requirement is not met, 10% will be deducted.
  - If performance satisfies the following criterion, extra credit of 20% will be given.
    - class Yes TPR >= 84% AND class No TPR >= 82%
  - If the performances of models cannot be reproduced:
    - At least 20% will be deducted.
    - If the team received any extra credit, the extra credit is revoked (in addition to 20% deduction)
- Intermediate report
  - If preprocessing has not been finished by the intermediate report due date, up to 6 points will be deducted.
  - If the intermediate report is not detailed or is not substantive, up to 6 points will be deducted.

- Other grading criteria
  - You must do your best to achieve high performance, such as:
    - try different methods to create balanced training datasets
    - try different attribute selection methods
    - try different classification algorithms
    - try parameter tuning
    - etc.
  - If there is no clear evidence showing that a team did their best to achieve high performance, up to 20 points will be deducted.
  - If your performance results are not reliable, up to 20 points will be deducted.
  - If the final report is not well organized, up to 10 points will be deducted.
  - If the description of all data mining procedure is not detailed enough, up to 10 points will be deducted.
  - Your project results must be presented in your report effectively using tables and graphs so that readers of your report may understand the results easily and clearly. Otherwise, up to 10 points will be deducted.
  - If your discussion/conclusion is not substantive up to 5 points will be deducted.
  - If the final report does not include all required components, up to 10 points will be deducted.
  - If all required datasets are not submitted, 5 points will be deducted.

Important:

- I should be able to reproduce your performance by running your R code and following your description of data mining procedure (see the above criteria). So, it is very important that the description of your data mining procedure must be detailed and precise.
- You must submit a single R code file. This R code must include all the steps you have taken for the project. When I grade your project, I will run this R code to try to reproduce your results. It must include all preprocessing you performed and all model building and testing.
- In your R program, you must include sufficient comments so that I can easily understand what each code segment does.
- **Additional requirements on R code submission is posted on Blackboard separately.**

Deliverable and file naming convention

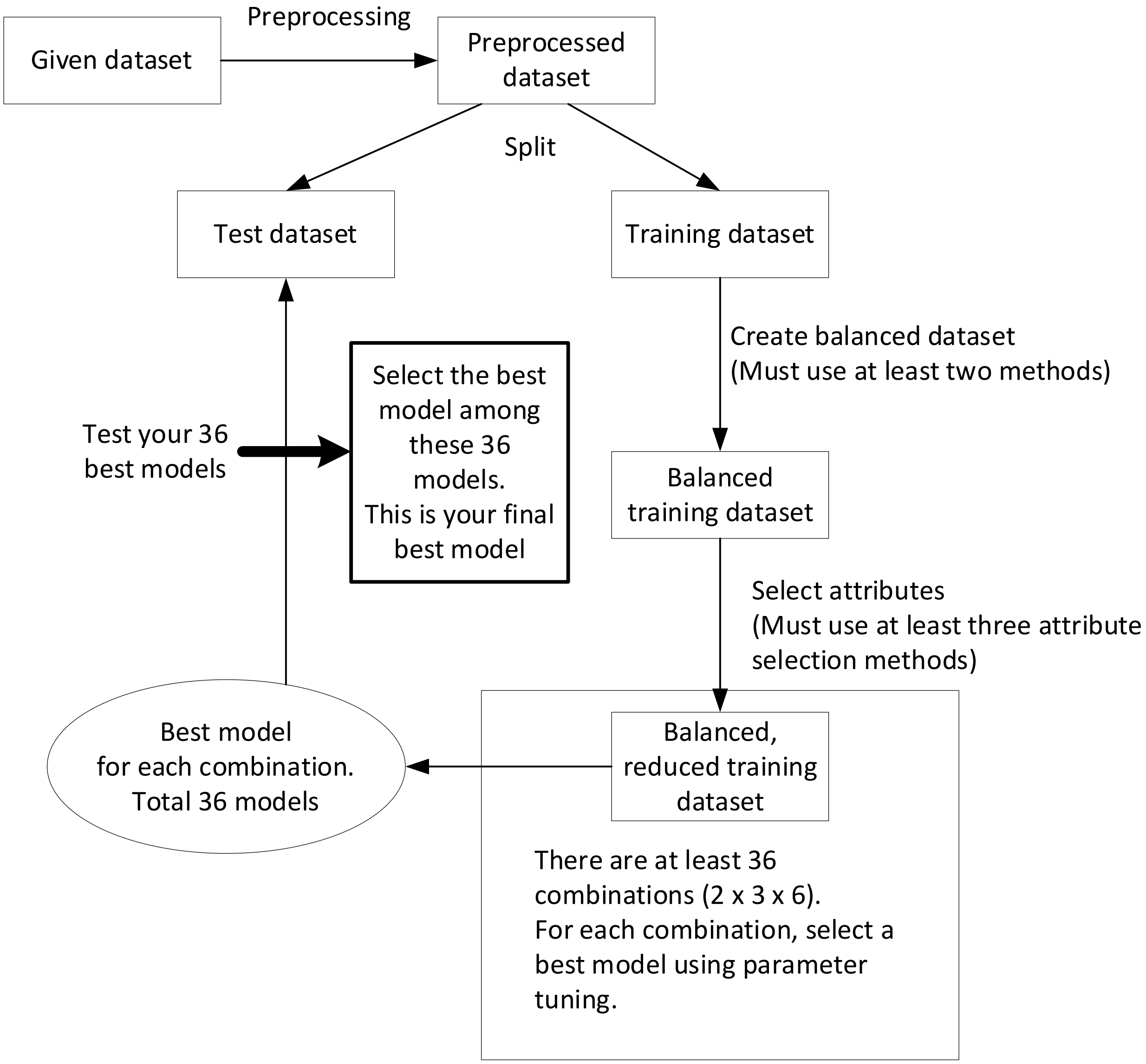
- Final report: *LastName_FirstName_Report.pdf* or *LastName_FirstName_Report.docx*. If your team has two members, then include names of both members in the fine name.
- Intermediate report: Use the same naming convention, except that you use *IntermediateReport* instead of *Report*.
- Dataset after preprocessing: *preprocessed_data.csv*
- Initial train and test datasets: *initial_train.csv* and *initial_test.csv*.

- Include all R programs in a single file and name it *project_code.R*.

- When submitting the final report, include the final report, datasets, R program file, and any other files you may have in ***a single archive file*** and name it *LastName_FirstName.<ext>*, where *<ext>* is an appropriate file extension, such as z*ip* or *rar*. Again, include names of both team members.

- **Only one team member (not both) must submit all deliverables.**

Overall Project Process (simplified)


Given dataset
Preprocessing
Preprocessed dataset
Split
Test dataset
Training dataset
Create balanced dataset
(Must use at least two methods)
Balanced training dataset
Select attributes
(Must use at least three attribute selection methods)
Balanced, reduced training dataset
There are at least 36 combinations (2 x 3 x 6).
For each combination, select a best model using parameter tuning.
Best model for each combination. Total 36 models
Test your 36 best models
Select the best model among these 36 models. This is your final best model