CS699

Project Assignment

The project is a practice of building and testing classification models using a real-world dataset. The project must be performed by a team of at most two students.

Dataset

The project dataset is a 2020 Behavioral Risk Factor Surveillance System (BRFSS) Survey Data, which was downloaded from https://www.cdc.gov/brfss/annual_data/annual_2020.html and modified for this assignment.

Some preprocessing was already performed and the dataset, which is given to the students, *project_dataset-5K.csv*, has 5000 tuples and 276 attributes. Students are expected to perform further preprocessing as needed.

In the dataset, each tuple is a person and *Class* is the class attribute. The class attribute value Y means a person has experienced a depressive disorder and N means a person has not experienced a depressive disorder.

Goal

The goal of the project is, using the given dataset, to build multiple classification models, which would predict a person with a depressive disorder, compare their performance, and select the "best" model.

Requirements

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

Process Requirements

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

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

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

Project Schedule

- Intermediate Report
  - Due date: 2/14
  - As mentioned earlier, some preprocessing was already done. However, further processing would be necessary and you need to finish preprocessing by the due date.
  - You may want to do data exploration to better understand the data.
  - You must submit an intermediate report that includes ***detailed*** description of all work you did.
  - Late submission penalty: 2 points will be deducted for each late day.
- Final Report
  - Due date: 4/3
  - Your final report must include:
    - Cover page
    - Brief description of data mining tool(s) you used.
    - Brief description of all preprocessing methods/work you performed.
    - 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.
    - 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 measures 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 measures tables in your report document yourself.

        Example format of performance measures table:

	TPR	FPR	Precision	Recall	F-measure	ROC	MCC	Kappa
Class 0								
Class 1								
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 best model.
      - 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) too.
  - Your final report must have at least 10 pages.
  - Late submission penalty: 5 points will be deducted for each day.

- Presentation Slides
  - Teams presenting on 4/24: Submit slides by 4/20
  - Teams presenting on 5/1: Submit slides by 4/27
  - Late submission penalty: 2 points will be deducted for each day.

<u>Grading Guideline</u>

- Points distribution
  - Intermediate report: 10%
  - Presentation: 10%
  - Final report and overall project quality: 80%
- Performance criteria
  - Minimum requirement
    - class Y TPR >= 60% AND class N TPR >= 70%
  - 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 Y TPR >= 65% AND class 1N TPR >= 75%
  - Among those teams who satisfy the minimum requirement but do not satisfy the 20% extra credit requirement, top two teams will receive an extra credit of 10%. The top two teams will be determined by the weighted average of the TPR's of the two classes.
  - If the performance of the model of a team cannot be replicated (on an independent test dataset):
    - 20% will be deducted.
    - If the team received any extra credit, the extra credit is revoked (in addition to 20% deduction)
- Intermediate report
  - If not all preprocessing has been performed by the intermediate report due date, up to 10 points will be deducted.
  - If the intermediate report is not detailed or is not substantive, up to 10 points will be deducted.

- Presentation grading criteria
    - If presentation does not accurately reflect the project report, students do not answer questions properly, or presentation is not substantive, then up to 10 points will be deducted.
    - If a student is absent on the presentation day, 5 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 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.
    - You 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 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:

- **It is very important that I should be able to reproduce your performance on an independent dataset 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.**

<u>Deliverable and file naming convention</u>

- 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*.

- Presentation slides: Use the same naming convention, except that you use *Presentation* 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 *zip* or *rar*. Again, include names of both team members if there are two students in a team.**

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

Overall Project Process (simplified)