Data Mining
Fall 2024

Lab Assignment #2

Objective

In this assignment, you will learn how to build an accurate classification model, evaluate the model performance, and analyze the important features of the given dataset.

Data Description

The whole dataset contains **a training set (with class labels)** and **a testing set (without class labels)**. The number of the training set and the testing set is 44,939 and 19,260, respectively.

This dataset, including demographics, medical examination results, and so on, consists of patients' information on hospital admission. In this assignment, you need to predict the survivability of admitted patients (who died or did not die). In other words, you must build a model from the **training set** to predict the binary target class **has_died** in the **testing set.**

The dataset will be available for **downloading on the Kaggle competition** (please check the Grading part).

*The reference for this dataset will be announced only after the deadline of this assignment to keep the source blind.

Steps

1. Model Training:
   You may use any languages like Python, Java, and C++, as well as open-sourced libraries/tools like scikit-learn, Weka, etc., for building the classification model. Multiple methods can be used/integrated to build your classification model. However, the result file must be in **CSV** format for the Kaggle submission.
2. Model Internal Validation:
   Do the **internal KFold cross-validation** for your model to give you a reliable measure of model quality, which will help you know the performance of the trained model before submitting your prediction results. Please refer to this reference to learn how to do that. Besides, the number of folds must be **5** (**at least)**. Finally, you must count the **average** AUROC and macro F1 score with all folds. Please describe the whole splitting process clearly in your report.
3. Competition:
   Try your best to get higher performance for classification in terms of the metrics given in the Evaluation part. In addition, You must participate in the Kaggle competition to compare the model performance with your classmates.
4. Explainable experiment:
   Analyze the importance of features and give the **Top 20 most important features**.

Evaluation

請注意，每一項均有分配分數；助教會根據下方準則為標準，並依據撰寫的品質來給分（非有寫就滿分）。

Your project will be evaluated in 2 parts:

1. A detailed report including the following parts:
    - Data pre-preprocessing and any other data-centric procedures you conducted:
        - Procedures **must** involve, but are not limited to:
            - (1) Dataset analysis; (2) Data cleaning; (3) Data transformation; (4) Data imputation; (5) Data imbalance handling.
        - You **must** (1) describe what you do in those processes, (2) discuss problems encountered and (2) explain how you deal with them clearly and completely.
        - You **must** do the visualization analysis on this dataset.
        - Some hints: do summary statistics, box plot, and histogram…, and detect if there exist any outliers or anomalies in the dataset.
    - Classification Methods:
        - Describe and introduce clearly the machine learning algorithms (give References to the used algorithms) and related packages you used.
            - Note: the algorithm you selected should has explainable property.
        - Describe how to reproduce the results with your source code files.
    - Results & Analysis:
        - Count Average AUROC and macro F1-Score with **cross-validation method**.
            - F1-Score:

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}$$

$where$:

$$precision = \frac{TP}{TP+FP}$$

$$recall = \frac{TP}{TP+FN}$$

            - AUROC: Area Under ROC curve.
            - Screenshot your results (i.e., Average AUROC and macro F1-Score).
        - Explainable experiment:
            - Show the **Top 20** most important features.
            - Give some analysis and insights about the high-importance features.
            - You can utilize other methods/tools to complete this experiment (e.g., SHAP value).

2. Classification performances on the Testing Set (in Kaggle competition):
    - Your classification model will be evaluated based on **Macro F1-Score**.
        - F1-Score:

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}$$

$where$:

$$precision = \frac{TP}{TP+FP}$$

$$recall = \frac{TP}{TP+FN}$$

    - Please check the Grading part (Performance scoring).

Grading

1. Report (75%)
    - Please refer to the "Evaluation" part for the sections you must include in the report.
    - Note: Make all descriptions as clear and complete as possible. Your score will be deducted if the report is poorly organized (including the clarity and completeness).
2. Performance Scoring (25%)
    - Participate in Kaggle competition (2024 NYCU Data Mining Lab Assignment 2):
        - https://www.kaggle.com/t/60acea88f2e247cca12e03e68c7eac74
        - After reading this requirement, please carefully check the Kaggle competition information. (IMPORTANT: *Overview*, *Data*, and *Rules*)
        - End Date: **11/27 (Wed) 23:00:00**
    - Pass baseline score (Macro F1-Score) on the testing set (15%)
        - You must pass a baseline score (benchmark) shown on the Kaggle leaderboard.
    - Ranking (10%)
        - Your ranking on the Kaggle leaderboard will determine the scores you can get. The top 3 places will get all scores (i.e., 10); the top 40% (round up 無條件進位) of the remaining places will get 7 points, and others will get 4 points.
        - Note: Before ranking, **you must pass the baseline score mentioned above**. Otherwise, you'll get 0 points for this ranking part.
    - **Note: No cheating. We will check the code you submitted to verify the correctness of the algorithms. If you use a mining tool (e.g., weka), please describe the running steps in your report clearly.**

What to Turn in

Zip the following two files together with the name "**DM_HW2_{your student id}_{your name}.zip**". (e.g., DM_HW2_310XXX_王大明.zip)

1. A report with the name "**HW2_Report.pdf**"
2. A submission file with the name "**testing_result.csv**"
    - Just output the result of your **model (with the best place on the Kaggle leaderboard)**.
    - Do remember to submit this file on Kaggle competition.
    - Data format example (same as submission on Kaggle):

patient_id	pred
19566	0
69039	0
77670	1
93534	1
104990	0
127397	1

3. Put all source code files in a folder named **"Code"**
    - You **do not need** to submit these if you use mining tools (e.g., weka), but you must clearly describe the running steps in your report.

Important Date

- Deadline (Your report & source files): **11/27 (Wed) 23:59:59**

Penalty

- Format error
  - The report is not in pdf. (-5%)
- Kaggle competition
  - Violate the Team Name rule (-5%)
  - Violate the "One account per participant" rule (**zero score** in this assignment)
  - Violate the "Team Limit" rule (**zero score** in this assignment)
- Classification Method
  - Failed to provide sufficient information of algorithm, packages or references (-10%)
- Reproduce problem
  - Cannot reproduce the same results with your source files (**zero score** in this assignment)
- Late submission
  - If you submit your work within one day after the deadline, you will get **80%** of the original score.
  - If you submit your work within two days after the deadline, you will get **50%** of the original score.
  - If you submit your work more than two days after the deadline, you will get **zero score** on this homework.