# Data Mining

## Lab Assignment #2

## Objective

In this assignment, you will use the classification methods to build an accurate classification model as well as analyzing the important features from the given dataset.

## Data Description

The whole dataset contains **a training set (with class labels)** and **a testing set (without class labels)**. The total number of the dataset is 64,199.

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

The dataset will be available for downloading on E3.
*The reference for this dataset will be announced only after the deadline of this assignment to keep the source blind.

## Steps

1. 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 together to build your classification model.
2. Do the internal training and validation for your models. Specifically, you need to divide the training set we provided into two sets first: Internal training set and Internal validation set (proportion: 70% and 30%, respectively) by yourself. After the splitting, you should train the models with internal training set, validate the models with internal validation set, and finally predict the results on testing set we provided. So you can check your model's performance on the internal validation set by yourself, while your model's final performance on the testing set will be evaluated by TAs based on your predicted results on the testing set. Please describe the whole splitting process in your report clearly.
3. Try your best to get a highest performance for classification in terms of the metrics given in the part of **Evaluation**.
4. For the model you build on the training set and the classification results on the validation set, analyze the importance of features and give the **Top 20 features with highest importance**.

## Evaluation

Your project will be evaluated in 2 parts:

1. Classification performances on the Testing Set:
    - Your classification model will be evaluated based on two metrics, **F1-Score** and **AUROC**.
    - 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 the ROC Curve

2. A detailed report including the following parts:
   - Data pre-preprocessing and any other data-centric procedures you conducted:
     - Procedures may involve, but are not limited to:
       - (1) Data Cleaning; (2) Data Transformation; (3) Data Reduction; (4) Data imputation, etc.
     - You must (1) discuss problems encountered, and (2) explain how you deal with them.
     - Some hints: Do summary statistics, box plot, and histogram…, and detect if there exist any outliers or anomalies in the dataset.
   - Classification Methods:
     - Describe clearly the algorithms your used (give also References to the used algorithms).
   - Results:
     - Screenshot your F1 score of your model on the **Internal validation set**.
     - Plot ROC Curve and count AUROC on **Internal validation set**.
     - Top 20 features with highest importance.
     - Note: Since you don't have the answers of class labels for the testing set, the F1 score and ROC in the report should be based on your results on the internal validation set.

# 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. An output file on the testing set with the name "**testing_result.csv**"
   a. Just output the result of your **model**.
   b. Data format example:

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

   c. Note: You need to **sort (ascending) the data according to patient_id**.

# Grading

1. Report (70%)

    - Please refer to the “Evaluation” part for the sections you need to include in the report.
    - Note: Make all descriptions as clear and complete as possible. Your score will be deducted if the report is not organized well (including the clarity and completeness).
2. Performance Scoring (30%)
    - Baseline score (F1-Score) on the testing set (20%)
        - You need to pass a baseline score on the testing set.
        - The baseline score will be announced on E3 later.
    - Ranking (10%)
        - The points will be determined by your ranking of F1-Score (on testing set) compared among all classmates.
        - Note: Before ranking, you need to pass the baseline score mentioned above. Otherwise, you’ll get 0 points for this ranking part.
    - **Note: If you have the wrong format on the output file, you may not get any score on this 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.**

# Important Date

- Deadline: 12/16 (Fri) 23:59:59 (Firmed Deadline; No Postponement)

# Penalty

- Format error
    - The report is not in pdf. (-5%)
    - Any turn-in files have format errors. (-5%)
- Late submission
    - If your work is submitted within one day after the deadline, a penalty of 20 percentage marks will be applied.
    - If your work is submitted within two days after the deadline, a penalty of 50 percentage marks will be applied.
    - If your work is submitted over two days after the deadline, you will get a score of 0 on this homework.