Due date: **11:59 pm Thursday, December 12th**.

Please use <u>any programming language</u> (e.g., R, Python, MATLAB) to answer the tasks in this homework.

DOROTHEA data set is a drug discovery dataset. Chemical compounds represented by structural molecular features must be classified as active (binding to thrombin) or inactive. This is one of 5 datasets of the NIPS 2003 feature selection challenge. We mapped Active compounds to the target value +1 (positive examples) and Inactive compounds to the target value –1 (negative examples). Please see information about data below:

https://archive.ics.uci.edu/dataset/169/dorothea

1.1. Approach

To build our model, we need to perform several tasks, but we can summarize all these tasks within a four-phase process. Figure 1 shows an example of overall process for this project. This is an optional process, and you can do steps in a different order.

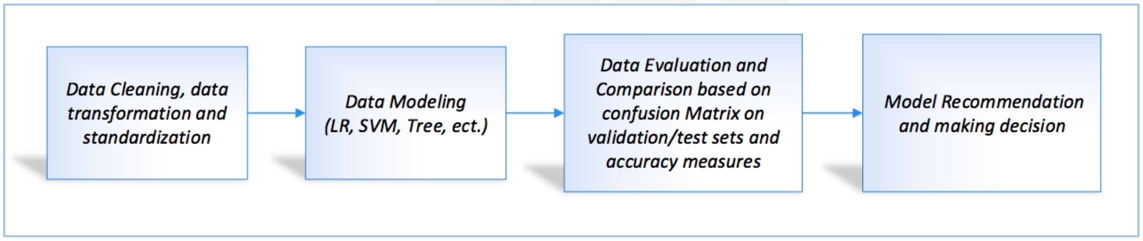
Data Cleaning, data transformation and standardization
Data Modeling (LR, SVM, Tree, ect.)
Data Evaluation and Comparison based on confusion Matrix on validation/test sets and accuracy measures
Model Recommendation and making decision

Figure 1: Our approach

1.2. Your Tasks

I have listed five tasks that allow you to practice your learning from the class on this dataset. To do the analysis, please use any programming language that you feel you are comfortable with.

1.2.1. Task 1

Check if the dataset has any missing values. If so, we will choose the simplest solution: we will get rid of the entire record!

1.2.2. Task 2

Do you need to standardize data? If yes. Please do it in a way the mean of each feature (column) is zero and the standard deviation is one.

1.2.3. Task 3

Run Logistic Regression (LR) and Support Vector Machine (SVM) using 10-fold cross validation and compute the performance metric (accuracy, sensitivity, specificity, G-mean, and ROC curve). Please refer to this [link] for the definition of accuracy, sensitivity, specificity. The G-mean is calculated as sqrt(sensitivity*specificity).

When implementing SVM, apply both linear and non-linear kernel functions (e.g., RBF and polynomial). You should perform hyperparameter tuning the identify the best hyperparameters for SVM.

1.2.4. Task 4

Run any feature selection method to decide the significant variables (risk factors) for prediction.

If you use the selected variables as the input of LR and SVM, would the performance results improve?

1.2.5. Task 5

Given the sensitivity, the specificity, accuracy, and the G-mean of the proposed models, discuss what classification method (in part 3 and 4) works best for this dataset.

1.3. Project Output

You are required to submit **one programming source file** and **one PDF file** compiled from your LaTeX or Word file.

1. Your **programming source file (e.g., R, Python, etc.)** must include the codes that you have implemented for each task.

2. Your **PDF file** must be a brief report explaining your approach, analysis, and computations results for each task. This file should also contain your answers to the questions asked in each part. Please note that your report must be written in a professional style (using charts, figures, and tables) with no grammatical or typo mistakes. If you are using **LaTeX file**, please submit both LaTeX file and its related **PDF** file.