AI Term Project #3 - Candidemia Mortality Prediction Guidelines

Due Date: June 22, 2024, 23:59:59

Introduction

Candidemia represents a critical health challenge, increasingly recognized as a significant cause of mortality globally, especially in immunocompromised individuals and patients in intensive care units. The rising incidence of candidemia is associated with several risk factors including immunosuppression, severe illnesses, aging populations, cancers, the usage of broad-spectrum antibiotics, and the presence of in-dwelling venous catheters. Despite advancements in treatment, mortality rates remain alarmingly high.

The project aims to develop the predictive model(s) using machine learning (ML) techniques to estimate the survival outcomes of patients suffering from candidemia within 14 days of diagnosis. The model(s) will consider multiple variables and patient data, utilizing advanced ML methods to overcome the challenges posed by the complex nature of medical datasets.

Objectives:

This AI Term Project requires participants to develop the binary classifier(s) using **Python** (scikit-learn) or **Weka** (Waikato Environment for Knowledge Analysis) to predict **14-day mortality** among patients with candidemia. Performance metrics to be used include the **F1 score**, Matthews Correlation Coefficient (**MCC**), and the Area Under the Curve (**AUROC**). This project is an **individual assignment** intended to enhance understanding of ML algorithms, including data preprocessing, predictive modeling, parameter tuning, and model evaluation.

Dataset Description:

Participants will utilize a dataset composed of real-world data from candidemia patients described by 77 features (F1-F77), including a binary outcome (Deadin_D14: death within 14 days of diagnosis). The project involves training and validating models with provided datasets (train_X.xlsx & train_y.xlsx) and **final testing on an independent test set (test_X.xlsx & test_y.xlsx) to be released on June 21, 2024**. Detailed information on patient features is available in **Data Description.docx**.

Submission Requirements:

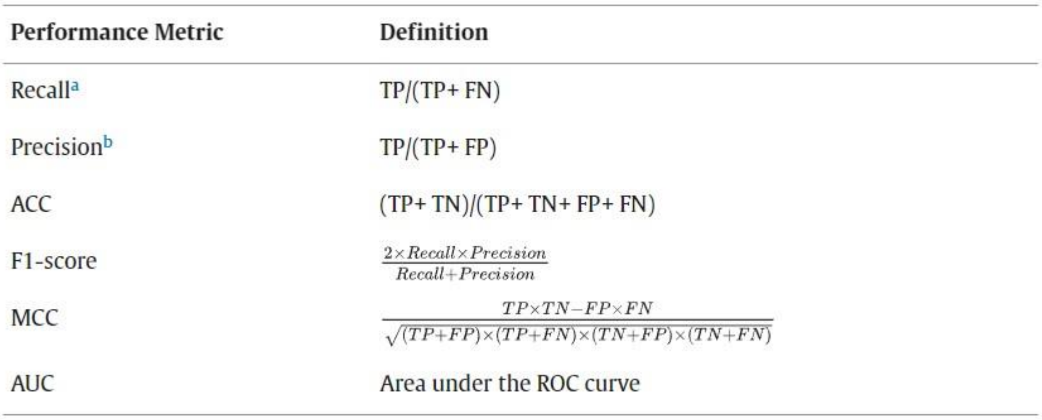
1. **Project Report**: Submit a comprehensive report detailing the experimental process, classification methods employed, data preprocessing steps such as imputation, oversampling, undersampling, cross-validation, and challenges encountered with corresponding solutions. The report should be saved as **studentID.pdf**. (for example, **0886123.pdf**)

2. **Results**: Submit the **test_results.xlsx** generated from the one best-performing prediction model. It should contain two columns: prediction (of death) and probability (of death). The upload format should follow the example file **Results_Example.xlsx**.

Note: Combine all the required files into one zip file, named **studentID.zip** (for example, **0886123**.zip).

Performance measure

To assess the performance of ML models in predicting candidemia, we have provided a brief summary of essential performance metrics. These measures and their definitions are as follows:



Performance Metric	Definition
Recall[a]	TP/(TP+ FN)
Precision[b]	TP/(TP+ FP)
ACC	(TP+ TN)/(TP+ TN+ FP+ FN)
F1-score	$\frac{2\times Recall\times Precision}{Recall+Precision}$
MCC	$\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}}$
AUC	Area under the ROC curve

TP, true positive; TN, true negative; FP, false positive; FN, false negative; MCC, Matthews correlation coefficient; ACC, accuracy; AUROC, area under the curve; ROC, receiver operating characteristic. [a] Recall is equivalent to sensitivity in its definition. [b] Precision is equivalent to positive predictive value in its definition.

Grading Policy

1. Performance Measurements (60%):

The PR score will be determined based on the highest metrics (**F1-score**, **MCC**, **AUROC**) achieved by the classifier using an independent test set, which will be provided on June 21, 2024

PR Scores:

- PR 99: 60/60 points
- PR 90: 54/60 points
- PR 85: 51/60 points
- PR 80: 48/60 points
- PR 75: 45/60 points
- Below PR 75: Scores decrease progressively (e.g., a PR of 70 might receive 18/60 points).

Note that a complete failure in prediction accuracy will result in a PR score of 0.

2. Report (40%):

Implementation Details: Include a description of preprocessing techniques, data handling, and the architectures of the classifier(s), detailing hyperparameters and other relevant specifics.

Discussion: Provide an analysis of implementation challenges, prediction results, and key insights, including a comparative assessment of the models.

Plagiarism Policy:

Post-submission, reports from ten randomly selected students will undergo a plagiarism check. Any detected plagiarism will result in a score of zero for the project.