CMPT459 Fall 2020
Data Mining
Martin Ester
TAs: Madana Krishnan Vadakandara Krishnan and Rhea Rodriguez

Milestone 3 and report of the Course Project

Deadline: December 15

Total marks: 100

Introduction

In the second milestone, you have built 2 or 3 baseline classification models. In the final milestone you will be tuning the hyperparameters, performing cross validation and comparing the models based on their performances. You will also be submitting a complete project report which includes all the steps that you have performed as a part of the project.

More often than not, the baseline model is not the best performing model. Different algorithms have different hyperparameters that can be adjusted to get a better performing model. Depending upon the problem statement, a particular metric (ex: Recall) might be more important than the other (ex: Accuracy). For this project, the goal is to predict the outcome (hospitalized, nonhospitalized, recovered, deceased) of a case. It is crucial to predict the 'deceased' label correctly, with few 'False Negatives', i.e. the 'recall' for class 'deceased' should be high.

Tasks

3.1 Hyperparameter tuning (**30 marks**)

Split the dataset into training dataset and testing dataset with train to test ratio 75:25, similar to what you did in milestone 2. Using only the training dataset, perform hyperparameter tuning for all of your baseline classification models. To do so, perform cross validation on the training dataset. **Find and report the parameters of the model that give the best result for 'recall' on class 'deceased'**. Report the results for each model in this form:

Hyperparameters	Accuracy	Overall Recall	Recall on 'deceased'
hyp1=val1, hyp2=val2	---	---	---
hyp1=val3, hyp2=val4	---	---	---
.	.	.	.
.	.	.	.

Discuss the technique used (ex: GridSearchCV, RandomSearchCV, BayesianOptimization etc) for hyperparameter tuning. If the model has many hyperparameters, you need not tune more than 3 hyperparameters (this will help reduce the runtime).

NOTE: The overall recall of the model and recall for 'deceased' are different. The documentation link below provides a nice example of how to use multiple metrics and 'refit'.
https://scikit-learn.org/stable/auto_examples/model_selection/plot_multi_metric_evaluation.html - running-gridsearchcv-using-multiple-evaluation-metrics

3.2 Comparative Study (**10 marks**)

From task 3.1, you now have 2 or 3 tuned models. Evaluate the performance of each of these tuned models on the test dataset (25% of the original data). Use accuracy, precision, recall (from confusion matrix and/or classification report) as performance metrics. Compare the performance of the different models and discuss their advantages and disadvantages. Finally, discuss which is the best model for the goal specified in the Introduction of this document.

3.3 Report **(60 marks)**

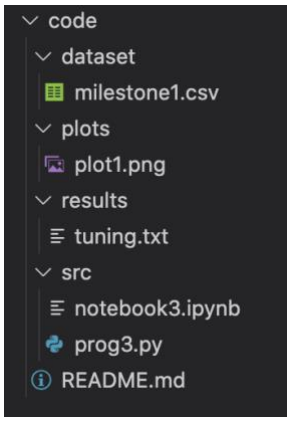
As a part of the final milestone, you need to submit a project report. The project report should cover all tasks covered in the 3 milestones. Use the following outline for the report.

- **Project Title**
- **Problem Statement** – Define the goal of this project
- **Dataset description and EDA** – Insights derived from the dataset using visualizations or statistical techniques
- **Data preparation** – Discuss steps taken in data cleaning, outlier detection, merging datasets
- **Classification models** – Which classification techniques were used in this project? Justify why you chose a particular model
- **Initial evaluation and overfitting** – Brief description about results obtained for baseline models and checking for overfitting
- **Hyperparameter tuning** – Fine tune models and discuss approaches taken
- **Results** – Report and discuss results obtained during model tuning (accuracy, overall recall, recall for 'deceased')
- **Conclusion** – Compare performance of the models on the test dataset using standard metrics and report the best fit
- **Lessons learnt and future work** – What lessons did you learn during this project? How can it be improved further?
- **Contributions** – Briefly describe the contributions and tasks completed by each group member, from the start to the end of the project

In the report, please include plots, diagrams and tables that will help understand the work done. Ex: plots from EDA, training and evaluation. The page limit for the report is 8 pages. Reports should **NOT** exceed the page limit. The font size should be 12, and the page margins at least 2 cm.

Submission (Code + Report)

A). Code
Submit a 'code.zip' file with the following contents. It should contain the code that you have written for the hyperparameter tuning and comparative study task. You can include any plots that you might generate. Please note that in order for the TAs to run the code for milestone 3, you need to provide the *dataset* that you used. Include the *dataset* within a folder named 'dataset' as shown in the figure. Include the results from hyperparameter tuning task within the 'results' folder in the form of .txt, .csv or images. You do NOT need to save and submit the models. The structure can look like this.



B). Report
Submit a ‘report.pdf’ file, as described above.