UNIVERSITY OF LOUISIANA Lafayette

College of Engineering
Department of
Chemical Engineering

P.O. Box 44130
Lafayette, LA 70504-4130
Office: (337) 482-6562

Université des Acadiens

CHEE 426(G) – Machine Learning for Chemical Engineers
SPRING Semester 2024
Faculty-in-Charge: Dr. Dhan Lord B. Fortela

Project 2: Classification of Drugs by Training on Molecular Data
Topics Covered: Chapter 3 & 4 - Logistic Regression and Classifier Evaluaiton Metrics

Overview:
The development of a new drug can cost well over a $1 billion, so any way to predict if a molecule will fail during clinical trials is highly valuable. This **Project 2** covers the **development of a Classifier to predict the Approval (or Non-approval) by FDA (Food and Drug Administration) of Drugs by training the model on molecular data**. This is also a project where you will compete with each other via Kaggle to develop the best Classifier for this project. Hence, this project has its Kaggle competition page were you will submit your notebook and your predictions to a Test Set that only I know the correct labels (true Y). Therefore, your Kaggle performance from this project will be counted as a seatwork - this is separate from your project grade even though we use the same problem for Project and seatwork Competition via Kaggle.

What is expected from you given we have a template Notebook?
Having this template notebook for the project, your task then is to improve the model. Here are possible ways to improve the model according to the discussion of our textbook author:

1. Use only a subset of the X variables.
2. Transform some of the X variables.
3. Use models other than Logistic Regression.

Referring to (3), you may also use techniques not covered by the textbook (note that the textbook is focused on few project examples - few examples but deep analysis and good demo) and there are many algorithms out there other than Logistic Regression to develop a Classifier.

Where Do You Go from Here?
Use this template notebook as a Base Case then include your improved model development by extending the notebook (add starting at the end of this template notebook). Requirement: Use Evaluation Metrics for Classification (Chapter 4) to numerically show the improved performance of your own model compared to the Base Case model.

Data:
The dataset will be downloaded from a GitHub repository via a Pandas function. The dataset was prepared by the MoleculeNet group (Zhenqin Wu, et al. Moleculenet: a benchmark for molecular machine learning. Chemical science, 9(2):513–530, 2018.). The GitHub repo of the data is https://github.com/whitead/dmol-book/raw/main/data/clintox.csv.gz. It is a collection of molecules that succeeded or failed in clinical trials.

Required Outputs:

(1) Jupyter Notebook in its native format (.ipynb); *Add a your Name and ULID at the top via markdown text.*
(2) Jupyter Notebook in (1) saved as PDF after you decide your notebook is final (.pdf)
(3) (To be recorded as your Competition grade) Submit an entry (Notebook and Predictions) to the Kaggle competition for the project.

Constraints:
This is an individual project, so make sure that your submissions are products of your own work. You may discuss the project with your classmates and others, but make sure that eventually you include in your submission files only your own work. You may use the provided solutions in the textbook resources and by the instructor during class sessions as templates for your work.

(1) Your project must clearly show significant change over the Base Case model shown in the Template notebook.
(2) Use Evaluation Metrics for Classification (Chapter 4) to numerically show the improved performance of your own model compared to the Base Case model.

(End of Project Statement)