CSc 740 Spring 2022
Assignment 2
Due Date: 11:59 pm, April 3, 2022

Read and Follow Assignment Instructions Carefully

1. This is an individual assignment. All work submitted must be your own.

2. First read the entire assignment description to get the big picture; create your own notes about the control flow, expected functionality of the various methods and why you are being asked to implement specific items. To repeat: **First, read the entire assignment completely and think about how the different parts are connected.**

3. You are allowed to use the Scikit-Learn library and standard machine learning and python libraries. You are **Not allowed** to **use any AutoML solutions or packages**.

4. Deliverables: the code and answers should be written in a Jupyter notebook named **'<lastname>_<firstname>_assignment2.ipynb'**. The notebook should also include short write-ups using markdown (2-5 sentences) summarizing results. Additionally, please submit predictions for the sample I will share and name it **"<lastname>_<firstname>_pred2.csv"**

5. Make sure you copy each question with the question number as a Markdown Cell in your Jupyter notebook and have the code response right below it. Points will be deducted if it is difficult to locate the question and response.

6. Make sure you comment your code. Points will be deducted if code logic is not apparent.

7. The written sections will be graded on correctness and preciseness while code will be graded on structure, implementation and correctness.

About the assignment: This assignment is intended to build the following skills:

1. Practical Machine Learning
2. Data preprocessing for numeric, categorical, and text data in correspondence to algorithms used
3. Model stacking
4. Model evaluation techniques and Results Summary

Practical ML: This assignment focuses on practical machine learning and multi-modal modeling (using text, categorical, and numeric data and handling missing data - note "multi-modal" usually refers to images, text, and audio, but I prefer to extend use here). Please find **"8k_diabetes.csv"** on Blackboard and you will be using the features in this data (all but "readmitted") to predict the target ("readmitted"). The grading for this assignment is outlined below.

Part A: Model Code and Exploration (100 pts)

1. Perform Exploratory Data Analysis (EDA) and discuss the data and what you observe prior to beginning modeling and how impact how to proceed [10 pts]
2. Pre-processed categorical data for use in the model and justified pre-processing method. Note this may be different for each algorithm you try. [10 pts]
3. Pre-processed numerical data appropriately including handling missing data and justified methods used. Note this may be different for each algorithm you try. [10 pts]
4. Implement a model to make predictions using text data using tf-idf [20 pts]
5. Use model stacking to incorporate tf-idf predictions for all 3 text fields (so 3 models unless you elect to concatenate the text fields into 1 - need to justify if so) in downstream algorithm the uses non-text features [20 pts]
6. Perform experimentation for multiple modeling algorithms and justify why you selected the experiments you chose [20 pts]
7. Final model selection and discussion of your model choice and the model weaknesses (generally, where model doesn't perform well, etc.) [10 pts]

Part B: Model Performance (100 pts)

This section your grade will be entirely based on the performance of your model. Please submit a "<lastname>_<firstname>_pred2.csv" with the predictions of readmission for the sample I share 5 days before the submission deadline. We will use AUC as our performance metric. Please see the grading table below:

AUC	**Points**
>=.6	25
>=.65	50
>=.66	60
>=.67	70
>=.68	80
>=.69	90
>=.695	100