THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Data Mining
DATS 6103
Date: *June-20-2022*

Exam 2

- **Show ALL Work, Neatly and in Order.**
- **No** credit for Answers Without Work.
- Open Books, Open Notes.

1 Introduction

This exam consists in using a Data Mining algorithm to predict the level of comfortability in life. This dataset is gathered from people in certain countries and ask them about their education, their work class, gender, race, hours of work they put in the week and other features which are important in the quality of life.

One of the goals of the government is to make sure their citizens are living under very good conditions. This dataset is the sample of people's response and having a good predictive model is very helpful for those people in power to make appropriate decisions.

The format of this exam is a competition style between you and the rest of your classmates. You are given a training set and there is a held out set which we will test your models on. You need to submit your results, so that we can test it on our private held-out set.

As you will find out, this dataset is very imbalanced, so the accuracy is not the best metric to use for the ranking. Instead, we will use the F1-score in the ranking.

2 Dataset and Submission file

- Download the exam dataset from BB Test section Exam 2.
- Please check the sample submission file called ***Test_submission_nedid.csv***. You need to fill the comfortability column with 0 and 1 predicated values from your model.

3 Rules of Competition

Please read these rules **carefully** and if you have any questions please send an email to me directly.

- You can **only** use Sklearn for training. This means only sklearn models can be used in this competition. **No Keras** models or other deep learning models (CNN,RNN).
- You can use Pandas, numpy, matplotlib, seaborn, scipy, matplotlib in this competition.

- You can only use the data you are given. Using additional data from any other sources is not allowed.
- You can do any kind of pre-processing with the training data, which you should split into at least training and testing. You may use whichever library you want for this purpose.
- You are not allowed to share your results with others. If we find out you will get **zero** grade for the Exam.
- You are not allowed to copy code or ideas from any students in the class. If we find out you will get **zero** grade for the Exam.
- You are allowed to search in the internet and find out ideas. You can use any external GitHub but you need to **cite** it. If we found any violation of this rule you get a reduce grade.

4 Clarifications on preprocessing

- You are allowed to do any pre processing on training and test set.
- **Note**: Makes sure you preserve the order of the submission file, since the grader code read the csv file and then compare it with the ground true values. If the shuffle the order of the data then your results will not have any meaning.

5 Deliverables

1. A single submission file `Test_submission_netid.csv` which netid is your GWU netid.
2. A training code and all of its subroutine if you have one.
3. You need to submit item 1 and item 2 to Blackboard.