CSCI 6612 - Visual Analytics
Summer 2019 - Assignment 2

Due 11:55 PM, Friday, June 07, 2019.

Instructions:

- Submit your assignment in Brightspace (https://dal.brightspace.com).
- Read the Dalhousie Policy on Plagiarism.
- The assignment must be done **individually**.
- Properly cite any external source that you used.
- You are not allowed to use or reuse any piece of code from other students.
- For this assignment, all implementation should be in Python (v3+ or newer). You may use any Python library that you consider necessary. You may use *Jupyter Notebook*.
- **If you are not** using Jupyter Notebook, submit also a PDF (max 3 pages) with the graphs, results, and descriptions required.
- Keep your code in a single file.
- Your code submission must contain all code necessary to find the results you reported. **We will probably run your code!**

Problem:

Now that you have fixed the dataset (Assignment 1), let's try to use it! The dataset contains classes (expected *salary*) for each person. The data you already have has a classification for all rows and you can use it as training to classify unseen profiles on a new dataset (**dataset1_test.csv**). You **should use your preprocessed** dataset to achieve the best **average class accuracy** possible!

You don't know the true class for the test set, so you'll need to split your training data into a validation set so you can estimate how well your model is performing! You are allowed to change your preprocessing and use any Machine Learning (ML) model as long as it can be trained and evaluated in less than 10 minutes **without** a GPU, so don't dive into a deep Neural Network!

[100 Marks] Requirements:

Code will be marked based on functionality, structure, reusability, best practices, and documentation. Some items below require you to describe, plot or report some results. You can include them as part of your Jupyter Notebook if you are using it, otherwise, create and submit a PDF including them.

1. [15 Marks] Load your preprocessed dataset from assignment 1 and convert categorical columns into multiple binary or numerical columns as you prefer. Normalize the numerical columns. Explain your final decision for each column. You can be brief.
2. [15 Marks] Split the data into 2 parts: train and validation.

3. [25 Marks] Using the train part of the dataset, train at least **3 different machine learning classification** algorithms and evaluate their accuracy. Report the *accuracy* and the *average class accuracy* on training and validation sets. Plot **one** bar graph to compare these results for each algorithm. Discuss the results.
4. [20 Marks] Choose **one** of the ML algorithms and plot one line chart to show how the variation of one parameter of the selected ML algorithm affects the *average class accuracy* on the validation set (accuracy vs parameter).
5. [20 Marks] From the algorithms you tested, select the one you believe is the best for this task, justify and try to achieve the best performance possible over your validation set. Report your best result (average class accuracy). Explain any decisions you consider relevant during the training of your ML model. You may try to do changes to your pre-processing (including the imputation of the first assignment or columns transformations performed here) and/or tune the hyperparameters of your model. Describe the steps you have taken to find the configuration that yields the best result.
6. [5 Marks] Using your best model, predict the classes for the profiles provided in the test file **dataset1_test.csv** and generate an output file (submit as ***B00XXXXXX*_prediction.csv**) containing only the predicted class for each row. One entry class per line, e.g.,

```
B0099999_prediction.csv:
      <=50K
      >50K
      >50K
      <=50K
      …
```

Note: The file dataset1_test.csv DOES NOT include the last column (the class: **salary**). Make sure to keep the row order in your output!!