DOUGLAS COLLEGE – Summer 2023

CSIS 3290 – 001 – Lab 3

- **Create a folder and name it according to naming convention stated below**
- **All the files you are required to submit for the assignment should be placed inside this folder.**
- **If cheating is determined (i.e., you shared your work with another student in the class), your work will a ZERO mark and you will face further consequences.**
- **Make sure to include all the necessary files to make sure that the code can run properly without producing any error**

BAMBANG SARIF

In this lab, we practice to Grid Search for different SVM classification algorithm. You need study the textbook, demo code and do your own research to make sure that you can perform all the tasks describe below.

1. Create a python notebook named as **Lab3_ABcXXXXX** with A signifies the first letter of your **first name**, Bc signifies the first two letters of your **last name** and XXXXX denotes the last five digits of your **student ID**.
2. Create a markdown cell at the top of the Jupyter notebook to state the lab, **your name and student ID**
3. **For each of the following section**, you need to create a **markdown heading cell** followed by a few code cells to complete the tasks. Please also put some comments in each code cell.
   a. **Load the python library** and **read the data**. Use the *salary.csv* file.
   b. **Drop the unneeded data.** Check if there is any null and drop them.
   c. **Create a new column.** Based on column salary, create a new column named as **salary_less50K** such that it will value "yes" if the value in the salary column is less than or equal to 50K, and "no" otherwise. You need to be careful in working in this column, there should only be two values: *<=50K* and *>50K*; there are other values with a "." (dot). Make sure to remove the dot from that column's values.
   d. **Modify column name and categorical content.** Change the column names to replace hypen '-' into underscore. Also, for all categorical columns, make sure to replace hypen '-' into underscore from all its values.
   e. **Reduce the number of unique values.** For column native_country, reduce the number of unique values such that if the values IS NOT *United_States* change it into *Other_Country*
   f. **Drop columns.** Column *capital_loss* and *capital_gain* are mostly zeros. Drop these columns.
   g. **Create dummy values for the dataframe**, do not forget to use the drop_first parameter.
   h. Create a grid search with cross validation (GridSearchCV) of SVM classifier with the following options/settings:
      - SelectFromModel with LogisticRegression as the estimator
      - Feature scaling (you can use any feature scaling method)
      - Classifiers options: kernel (linear, rbf), gamma (1, 0.1, 0.01), C (1, 10, 50)
   i. Provide the confusion matrix, classification report and analyze the precision/recall for the best model.

In order to reduce the computation power requirement, you should:

- **Start with a smaller dataset. Use a sample of 0.01 fraction that will give you around 400 data points.**
- **Use only 3 cross validations instead of the default 5.**
- **Increase the number of processors to execute the grid search. For this purpose, you need to set he n_jobs hyperparameter such that n_jobs = 4 signifies that you are going to use 4 processors core of your desktop/laptop; while n_jobs = –1 signifies that you will use all processors**

Once you can get results, make a copy of your notebook file and try to run the grid search with a bigger dataset. Note: For a 10th generation i7 with 32GB memory and 16 processor cores, the grid search cv with 0.05 fraction (5%) dataset sample still runs in a reasonable time.

Copyright © 2023 Bambang A.B. Sarif

Note on submission:

- Create a folder named as Lab3_ABcXXXXX following the naming convention.
- Put your Jupyter notebook and the original and cleaned dataset in this folder.
- Zip the file and submit it through the blackboard

LAB/ASSIGNMENT PRE-SUBMISSION CHECKLIST

- Did you follow the naming convention for your files and folder?!
- Does your submission work on another computer?!
- Double check **before** submitting

Copyright © 2023 Bambang A.B. Sarif and others. NOT FOR REDISTRIBUTION.
STUDENTS FOUND REDISTRIBUTING COURSE MATERIAL IS IN VIOLATION OF ACAMEDIC INTEGRITY POLICIES AND WILL FACE DISCIPLINARY ACTION BY THE COLLEGE ADMINISTRATION

BAMBANG SARIF