**George Brown College**

**Centre for Arts, Design and Information Technology**

**Machine Learning Process – Working with pandas – Assignment 05**

Please follow the steps below and have the resulting file on the day of the final practice exam. This file will be corrected and marked along with the final exam.

1. Download the csv file from the following URL and store it in a DataFrame and call it myDF.

http://tiny.cc/assignment5a

1. By using info() method, calculate the DF size in two approximate and precise states.
2. Calculate the memory usage of each Series in two approximate and precise states.
3. Change the ‘Continent’ column to a category type and check the memory usage after the modification.
4. Please show how are category values stored in the DF, using codes() method.
5. Do you see any other candidate for typecasting to the category type? Why?
6. Reload the csv file from the following URL and store it again in a DataFrame and call it myDF.

http://tiny.cc/assignment5a

Then define Continent values as dummy variables, deleting the first column.

1. Download the csv file from the following URL and store it in a DataFrame and call it mySchool.

http://tiny.cc/assignment5b

1. Split the mySchool. DataFrame into 70% random rows for the training purpose and 30% that are not in the training set for the testing purpose. Save them as training.csv and testing.csv, respectively.
2. Consider 'Math', 'Physics' columns as feature columns and save them into X.
3. Consider the Result column as the label of the dataset and save it as Y.
4. Use the logistic regression to find the relationship between X and Y.
5. Save the actual result for the test result and name it Y\_actual\_test
6. Predict the Result for the testing data and name it as Y\_pred\_test.
7. Print Y\_actual\_test and Y\_pred\_test and compare them.
8. Bonus:
   1. Calculate the accuracy of the mechanism.
   2. Import decision tree and repeat the experiment, using the decision tree.
   3. Calculate the accuracy of the mechanism and compare it with the logistic regression. Which one works better?
9. Pickle the train and test data into train.pkl and test.pkl, respectively.
10. Convert the ‘Time’ column to datetime type. Check the DF data type and show the first five rows.
11. Fetch the hour and the year of the Time column and save them in new columns, named ‘Hour’ and ‘Year’. Now, let’s see the first five rows.
12. Create a timestamp, called ts and assign '1/1/2015' to it. Now, filter those rows that their time is greater than the timestamp.
13. Using the value\_counts() method, show the number of occurrence in each year, sorted based on the year. Then plot the result.
14. Show the duplicates rows that their ‘Math’ and ‘Physics’ contents are duplicated, keeping the first values.
15. Show the duplicates rows that their ‘Math’ and ‘Physics’ contents are duplicated, refusing to keep duplicated values and show all the duplicated ones.
16. What is the maximum number of rows that we can see by default? Change it to infinity. Then reset it to the default value.
17. What is the maximum number of columns that we can see by default? Change it to 500. Then reset it to the default value.
18. Try to change the floating numbers using comma separated format. Does it change the way your data looks? Does it change the underlying data? Reset all the options to their default values.
19. Create a DF that includes the random numbers between 0 to 100 in 3 Rows and 4 Columns. The column names should be Red, Green, Blue, and Yellow. The indexes should be 10, 20, and 30.
20. Change all the DF content of question 27 to the floating point.
21. Show the maximum value in each row.
22. Show the column that contains the maximum value in each row.

Thank you and good luck,

Reza

I know you’re tired but come, this is the way. ~Rumi