DESCRIPTION OF COURSEWORK

Course Code	SWE404
Course Name	Big Data Analytics
Lecturer	Lu Yang
Academic Session	2020/04
Assessment Title	Assignment 3

A. Introduction/ Situation/ Background Information

This assignment evaluates the students' understanding of Lecture 7-8, which are mainly about MLlib and machine learning algorithms.

B. Course Learning Outcomes (CLO) covered

At the end of this assessment, students are able to:

CLO 2 Demonstrate appropriate knowledge of Hadoop, Spark, Storm and Map Reduce framework and apply them to build a VM-based environment.

CLO 3 Integrate knowledge and understanding of the basic principles, techniques and methodologies of organizing and searching Big Data, and apply them to create value with business insight.

C. University Policy on Academic Misconduct

1. Academic misconduct is a serious offense in Xiamen University Malaysia. It can be defined as any of the following:
   i. **Plagiarism** is submitting or presenting someone else's work, words, ideas, data or information as your own intentionally or unintentionally. This includes incorporating published and unpublished material, whether in manuscript, printed or electronic form into your work without acknowledging the source (the person and the work).

ii. **Collusion** is two or more people collaborating on a piece of work (in part or whole) which is intended to be wholly individual and passed it off as own individual work.

iii. **Cheating** is an act of dishonesty or fraud in order to gain an unfair advantage in an assessment. This includes using or attempting to use, or assisting another to use materials that are prohibited or inappropriate, commissioning work from a third party, falsifying data, or breaching any examination rules.

2. All the assessment submitted must be the outcome of the student. Any form of academic misconduct is a serious offense which will be penalised by being given a zero mark for the entire assessment in question or part of the assessment in question. If there is more than one guilty party as in the case of collusion, both you and your collusion partner(s) will be subjected to the same penalty.

D. Instruction to Students

This assignment is an **individual** assignment. Each student should submit Jupyter Notebook file on Moodle, named as "Assignment3_YourStudentID.ipynb".

The deadline is **18:00, 15th June**. Overdue penalty will be given to the assignment that is submitted after the deadline.

* Your codes will be sent to a **Plagiarism** detection system for duplication checking. Please write your codes independently. (Modify your code if you copy some fragment from the Internet because your classmates may copy the same fragment.)

E. Evaluation Breakdown

No.	Component Title	Percentage (%)
1.	Assignment 3	100
	TOTAL	**100**

F. Task(s)

1. Predict whether income exceeds $50K/yr based on census data. There are two files:
    1) adult.data: the training data with 32,561 samples.
    2) adult.test: the test data with 16,281 samples.

    There are 14 features and the information is as follows:

    age: continuous.

    workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

    fnlwgt: continuous.

    education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.

    education-num: continuous.

    marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

    occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

    relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

    race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

    sex: Female, Male.

    capital-gain: continuous.

    capital-loss: continuous.

    hours-per-week: continuous.

    native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

You are required to do the following tasks by PySpark with MLlib:

a. Use RandomForestClassifier to build a classification model on the training data. Tune the hyperparameters numTrees, subsamplingRate, and featureSubsetStrategy. What are the best hyperparameters for this dataset? (20 marks)
b. By checking featureImportances, which features are the most important? Try to give an analysis on your results. (20 marks)
c. Compare RandomForestClassifier with GBTClassifier. (You can use sklearn: https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html)
    i. Compare them in terms of accuracy, F1 score and AUC. (30 marks)
    ii. Draw the ROC curves of testing results. (30 marks)