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)