Spark: Wrap Up Exercise

Exercise

Use the adult dataset (github.com/forons/BigDataExamples/files/adult.data)

1. Count the number of people that have the same age
2. Average age by marital-status
3. Maximum capital-gain by country
4. Perform classification for predicting the 'class' column
   a. Create a vector of features of all the numerical columns ('fnlwgt', 'age', 'education-num', 'capital-gain', 'capital-loss', 'hours-per-week') through the VectorAssembler
   b. Create an "numerical conversion" of the 'class' column with the StringIndexer
   c. Split the dataset in train and test set
   d. Apply RandomForestClassifier algorithm for predicting the numerical conversion of the ‘class’ column
   e. Convert the prediction again into string with the IndexToString
   f. Evaluate the model with the MulticlassClassificationEvaluator

Contacts

For any problem, send a mail to

daniele.foroni@unitn.it