**Homework #4 (200 points, due date: by 9 a.m., Nov. 16; late homework - by 9 a.m., Nov. 17: 30% off)**

**Compress your codes and report in a zip, and upload to the e3 system** (<http://dcpc.nctu.edu.tw/>)**.**

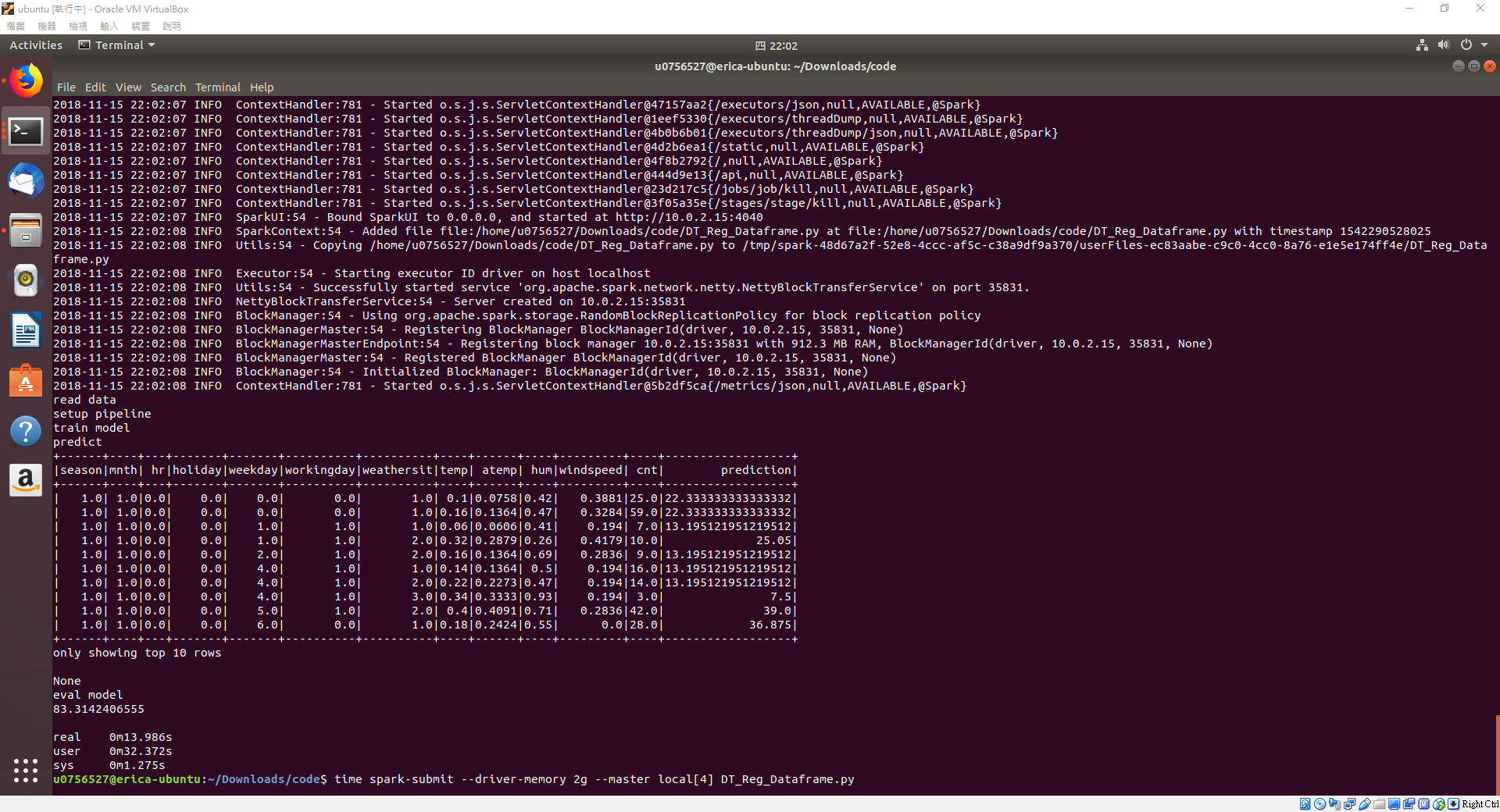
**HW #4 Spark MLlib & ML Pipelines:**

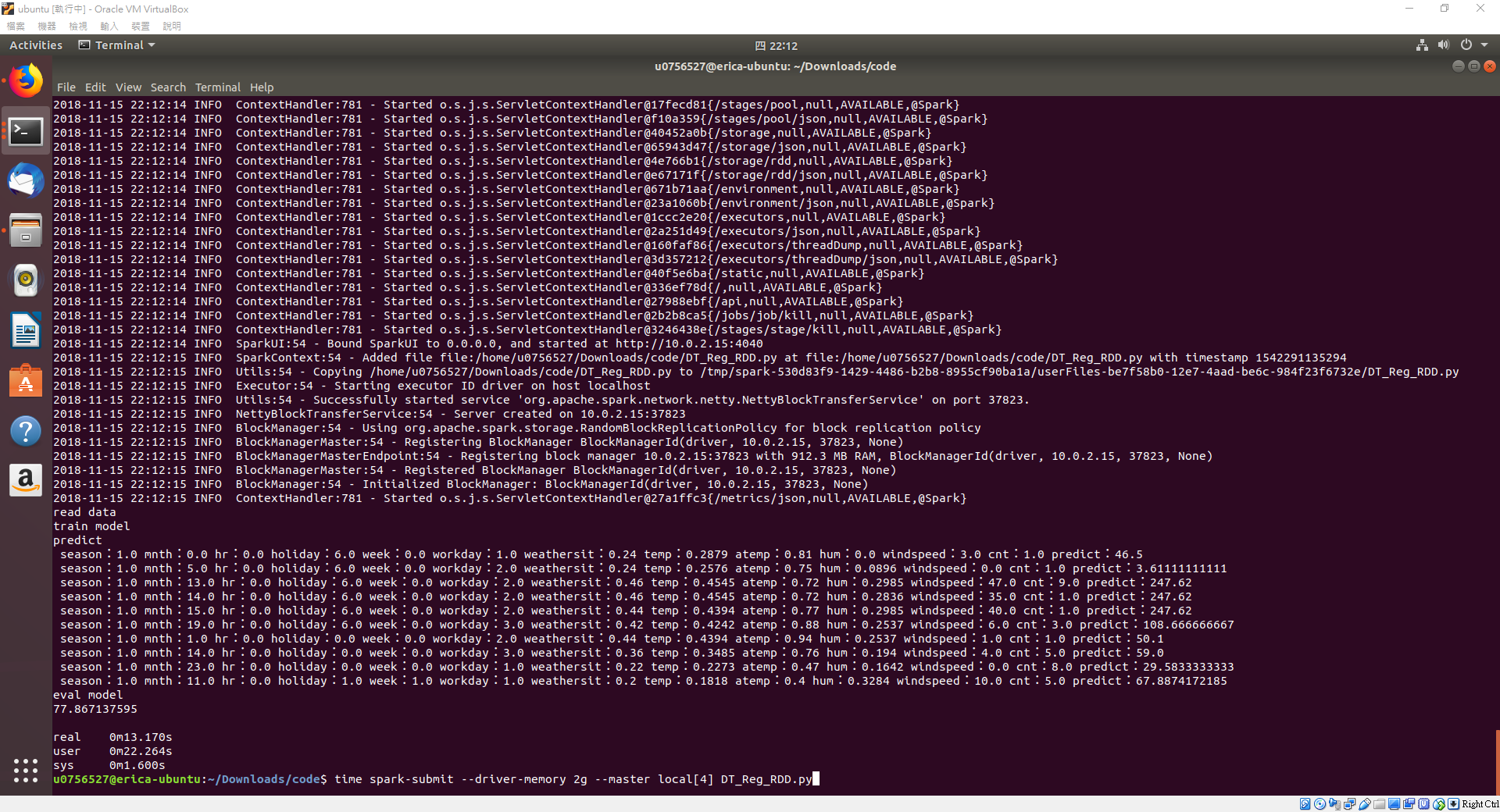
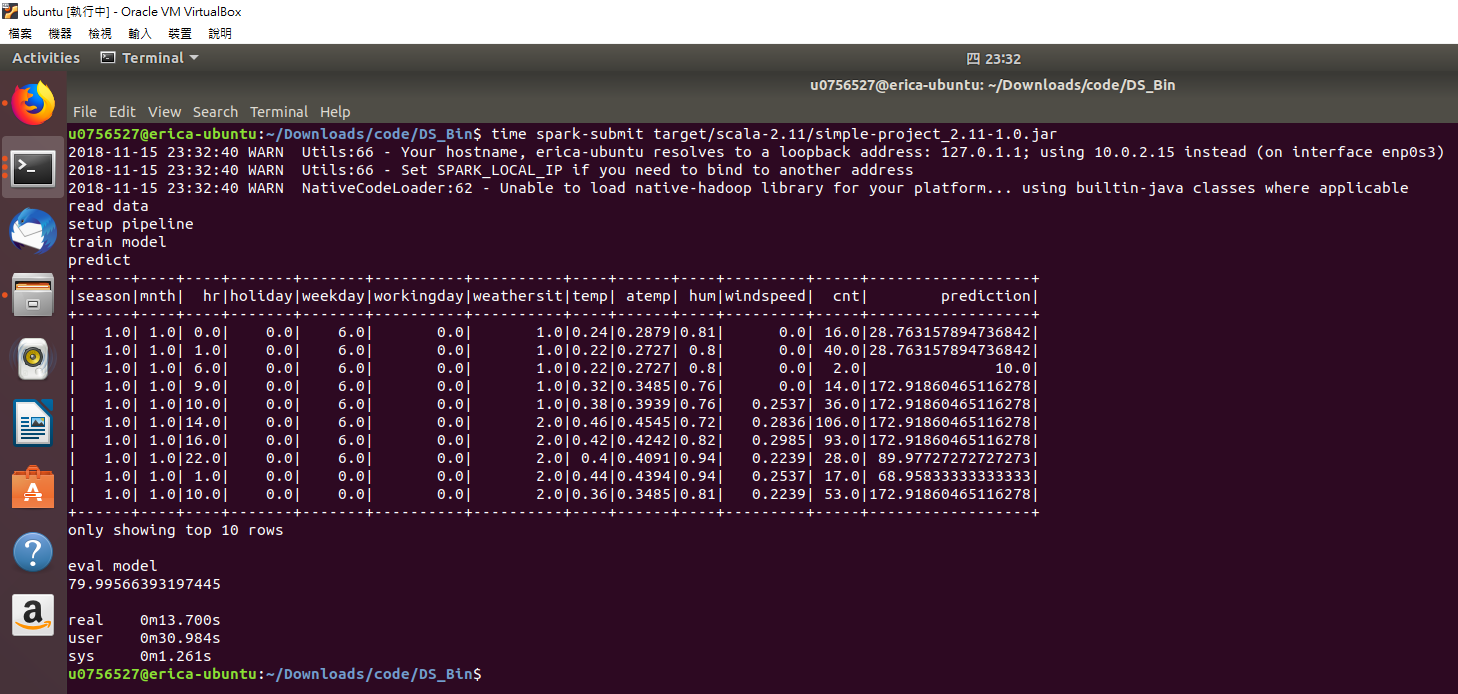
**Specifications are as follows:**

**Part I: Spark MLlib Decision Tree - Regression version (140 points):**

**Write three Decision Tree - Regression programs for** [**Bike Sharing DataSet**](https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset) **and using RDD (40 points), DataFrame (60 points), and DataSet (40 points) in Spark MLlib.**

**In this homework, we only use hour.csv to train and evaluate your model. 70% of hour.csv are used to train your model, and the rest is used to evaluate. Then, show 10 predictions, and use Spark Build-In Evaluator to exanimate your model in RMSE (root-mean-square error), and capture the results by taking screenshots. Like the following figures, your screenshots should contain your student ID in the command line prompt, predictions from your model, and an RMSE value. In addition, the RMSE value should be less than 85.**





**Hints:**

1. **Drop ”instant”, “dteday", “yr”, "casual", and "registered" columns.**
2. **For the arguments of DecisionTreeRegressor & DecisionTree.trainRegressor, you may use the following values:**

**maxDepth: 10**

**impurity: "variance"**

**maxBins: 100**

**Keywords (APIs)**

**RDD:**[**RegressionMetrics**](https://spark.apache.org/docs/2.3.0/api/python/pyspark.mllib.html#pyspark.mllib.evaluation.RegressionMetrics)**,** [**rootMeanSquaredError**](https://spark.apache.org/docs/2.3.0/api/python/pyspark.mllib.html#pyspark.mllib.evaluation.RegressionMetrics)[**DecisionTree.trainRegressor**](https://spark.apache.org/docs/2.3.0/api/python/pyspark.mllib.html#pyspark.mllib.tree.DecisionTree.trainRegressor)

**DataFrame & DataSet:**[**VectorAssembler**](https://spark.apache.org/docs/2.1.0/api/python/pyspark.ml.html#pyspark.ml.feature.VectorAssembler)**,** [**VectorIndexer**](https://spark.apache.org/docs/2.1.0/api/python/pyspark.ml.html#pyspark.ml.feature.VectorIndexer)**,** [**DecisionTreeRegressor**](https://spark.apache.org/docs/2.3.0/api/python/pyspark.ml.html#pyspark.ml.regression.DecisionTreeRegressor)

[**RegressionEvaluator**](http://spark.apache.org/docs/2.3.0/api/python/pyspark.ml.html#pyspark.ml.evaluation.RegressionEvaluator)

**Part II: (60 points)**

**Write a 1-page report that answers the following questions and describes your user experience:**

1. **Please answer the following three questions:**
   * **Use Bash *time* command to compare the execution time between your RDD version code and your DataFrame version code, and describe what you find from this comparison. (10 points)**
     1. **Usage: 0656051@u0656051-VM:~$ time spark-submit <<your code>>**
   * **Why do we use Decision Tree - the regression version in this homework, instead of using Decision Tree - the classification version? (10 points)**
   * **(bonus) What is the mail difference between DataSet and DataFrame? Use your code to explain it. (Hint: Think of the main difference between statically typed & dynamically typed programming languages) (10 points)**

**Ans. (a-1) RDD version code: 0m13.170s**

**DataFrame version code: 0m13.986s**

1. **What you have learned, what problems you encountered, and how the problems were resolved. (40 points)**