Paul Polsinelli

IFT 410 - 22485

3/7/2021

# Lab 6: Model Training & Cross Validation in Spark ML

Q3. Using your databricks account, run the classification code on the dataset file provided. List the obtained results. *(Make sure to comment out the last two lines of code that save & load the model.)*

Test Error = 0.36363636363636365

Learned classification tree model:

DecisionTreeModel classifier of depth 3 with 7 nodes

If (feature 376 <= -0.198863)

Predict: 0.0

Else (feature 376 > -0.198863)

If (feature 63 <= 0.6485415)

If (feature 8 <= -2.7191799999999997)

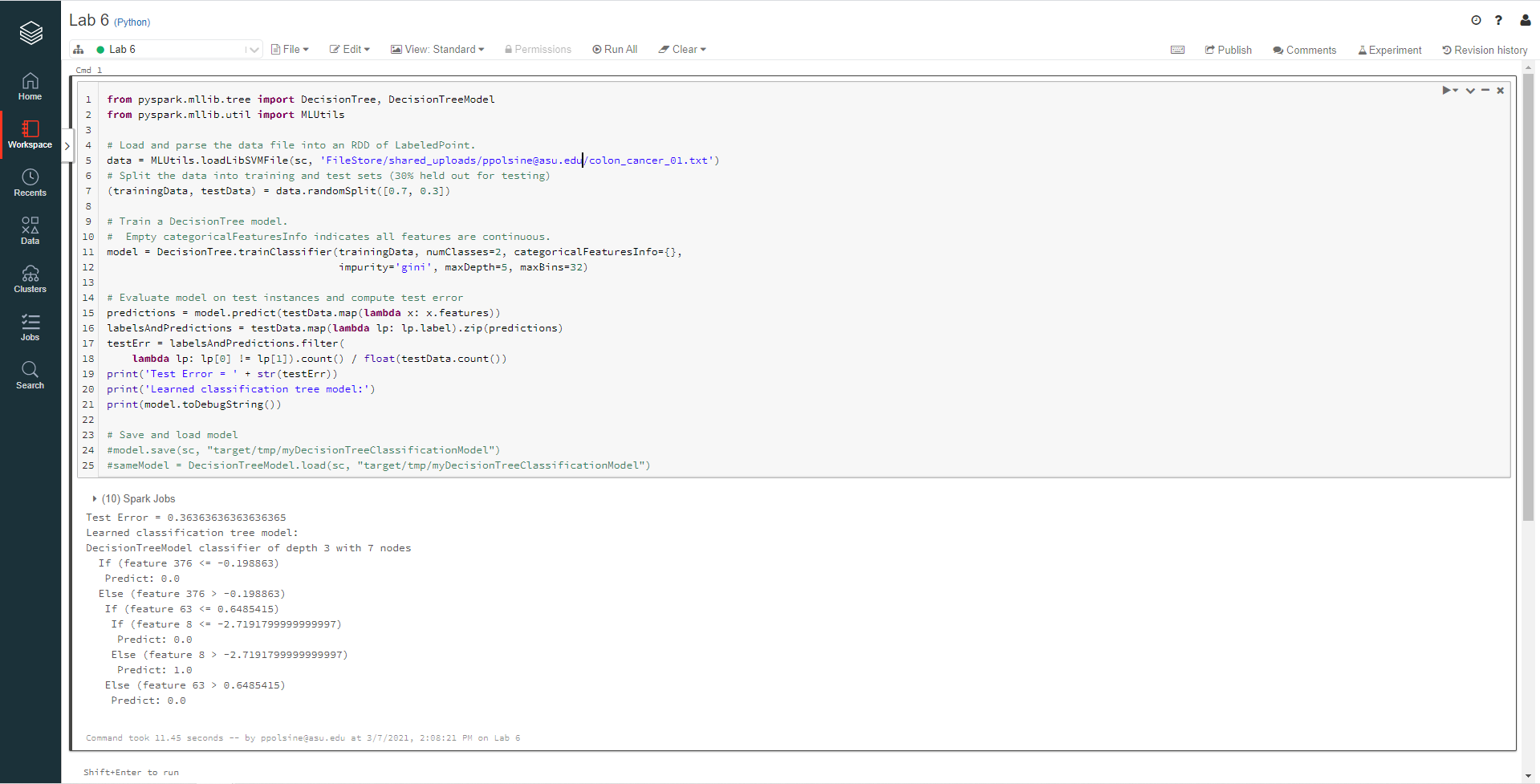
Predict: 0.0

Else (feature 8 > -2.7191799999999997)

Predict: 1.0

Else (feature 63 > 0.6485415)

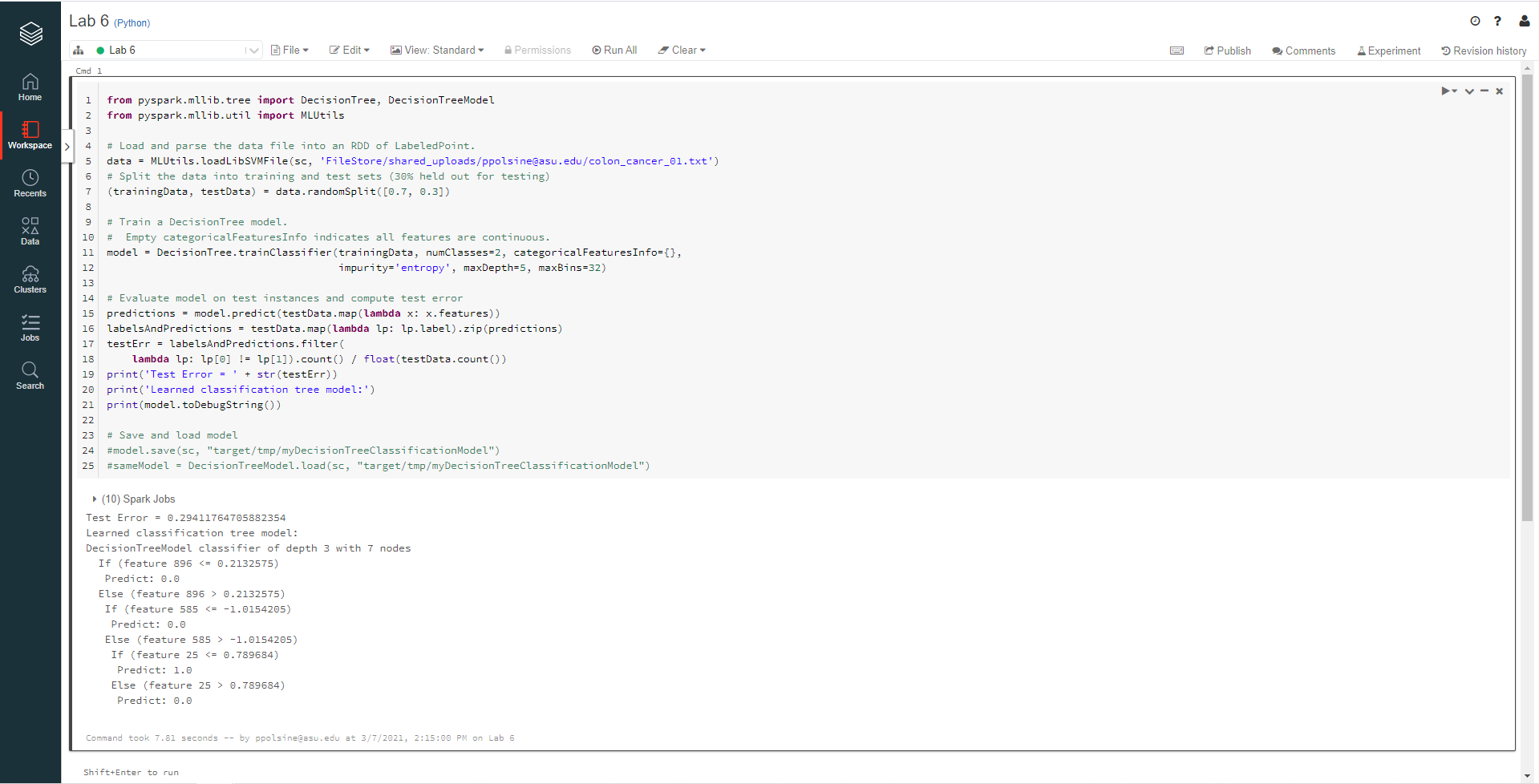
Predict: 0.0



Q4. Change any of the tunable attributes in the code. For example,

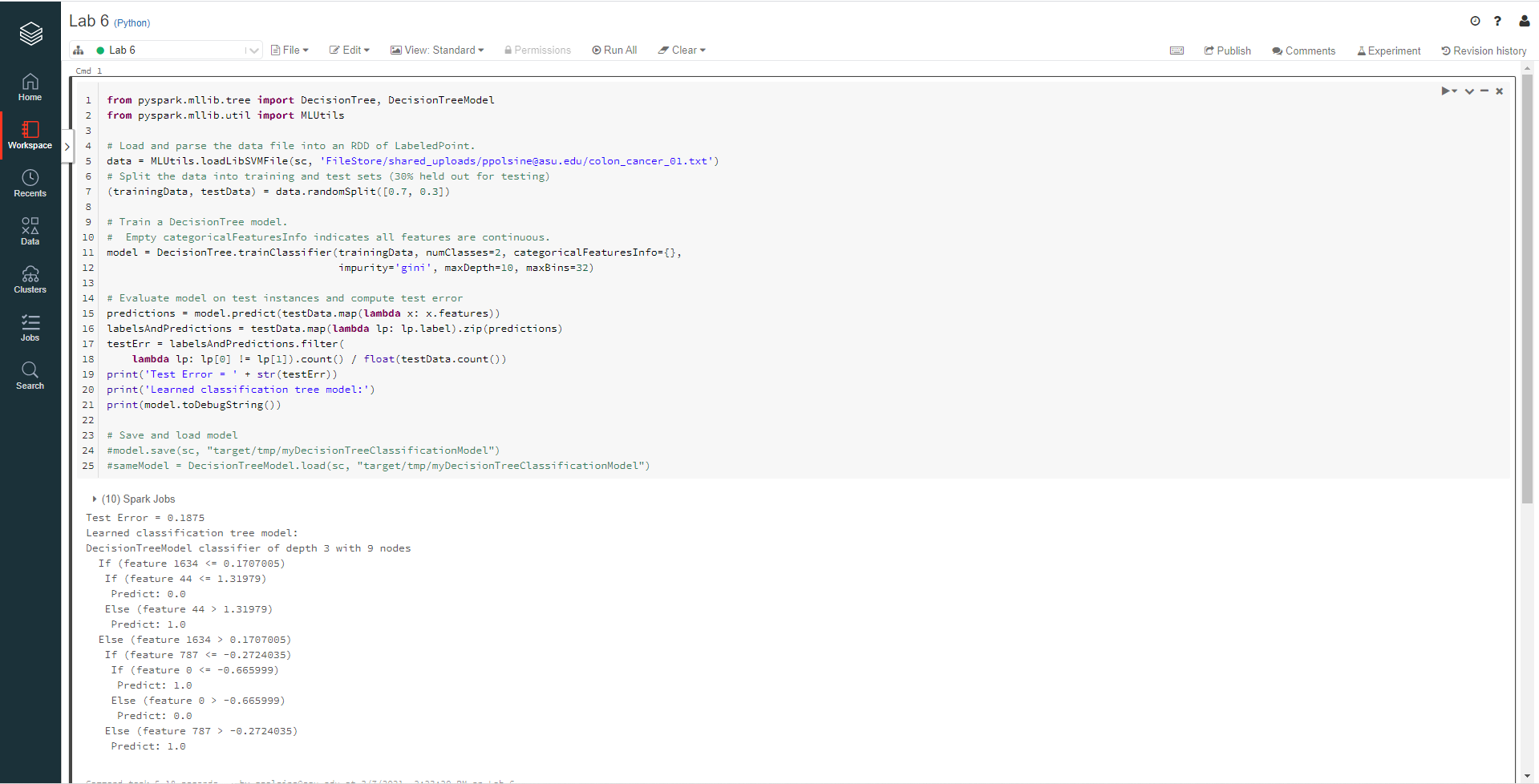
* Change the *impurity*function from *'gini'* to *'entropy'*.

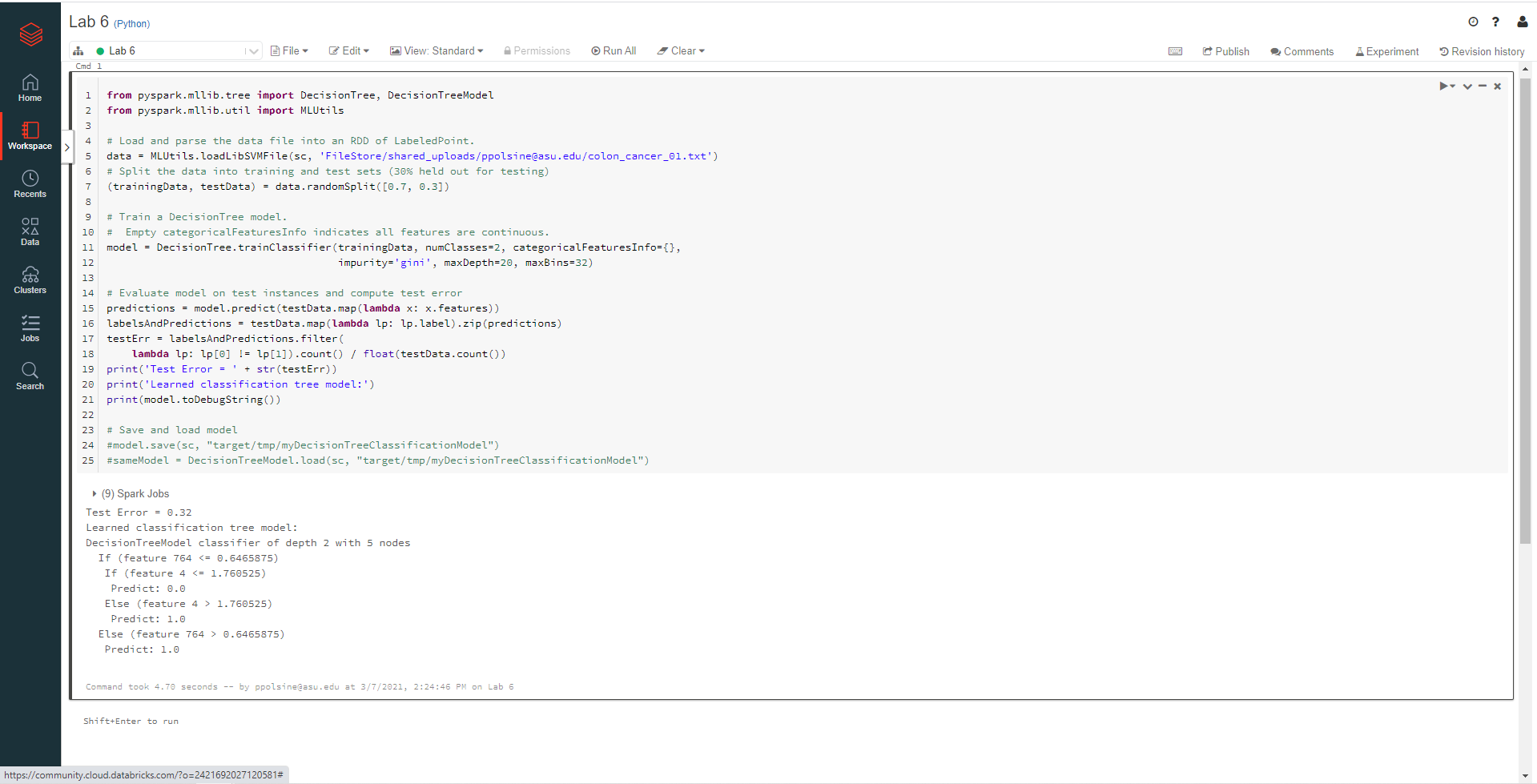
The test error decreased to 0.29411764705882354



* Change the *maxDept*to *10*then *20*.

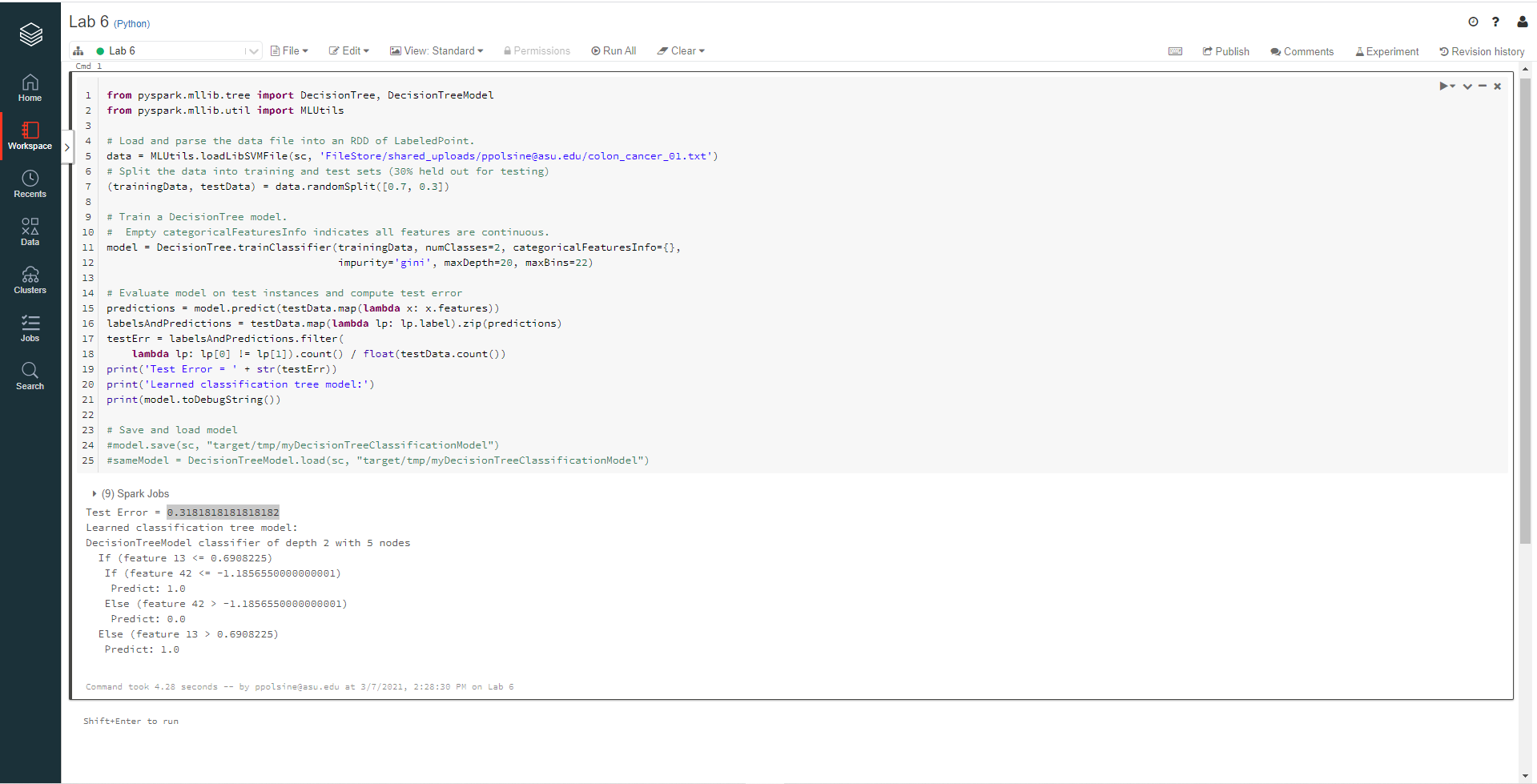
From 5 to 10 the error rate decreased to 0.1875, from 10 to 20 it increased to 0.32

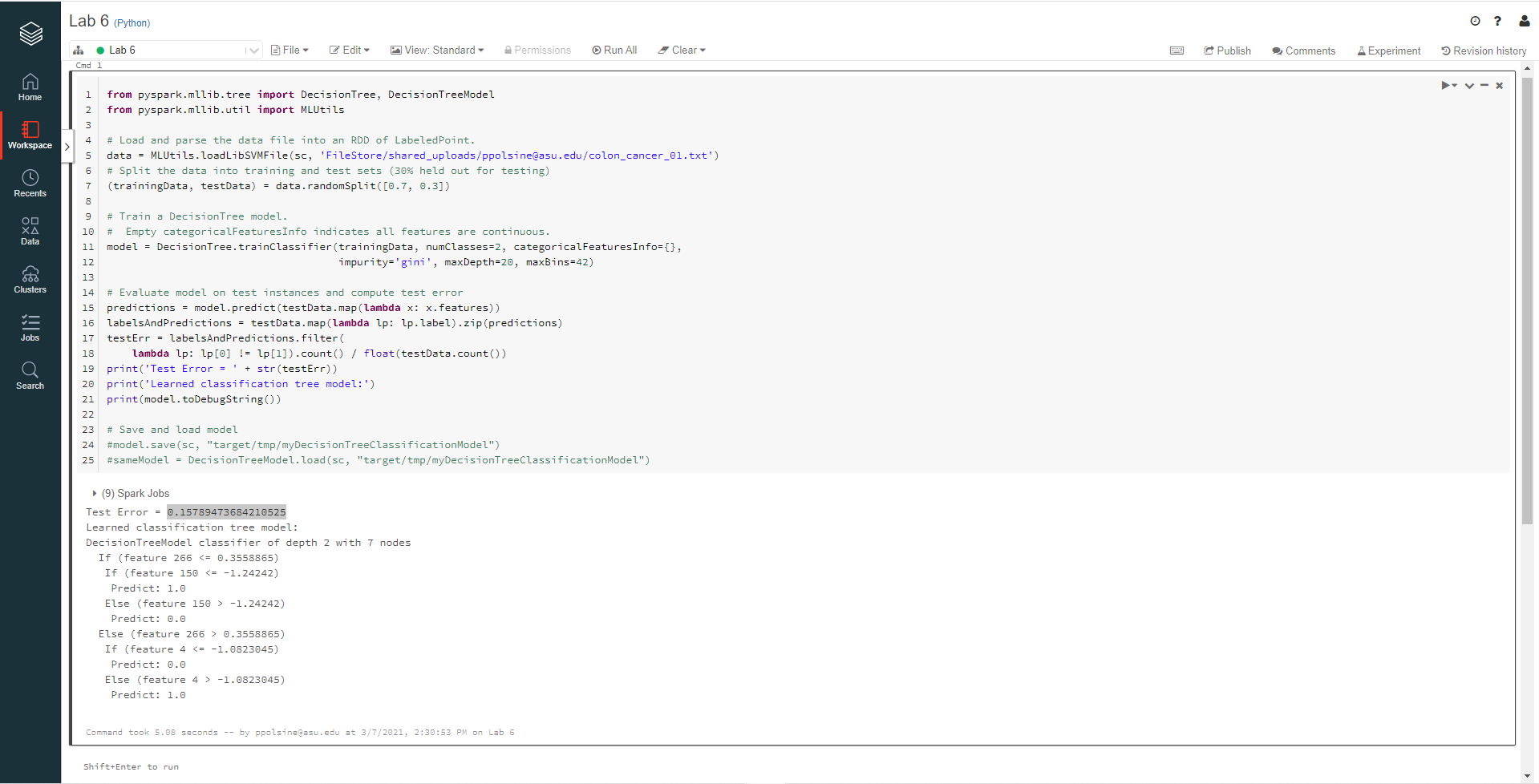




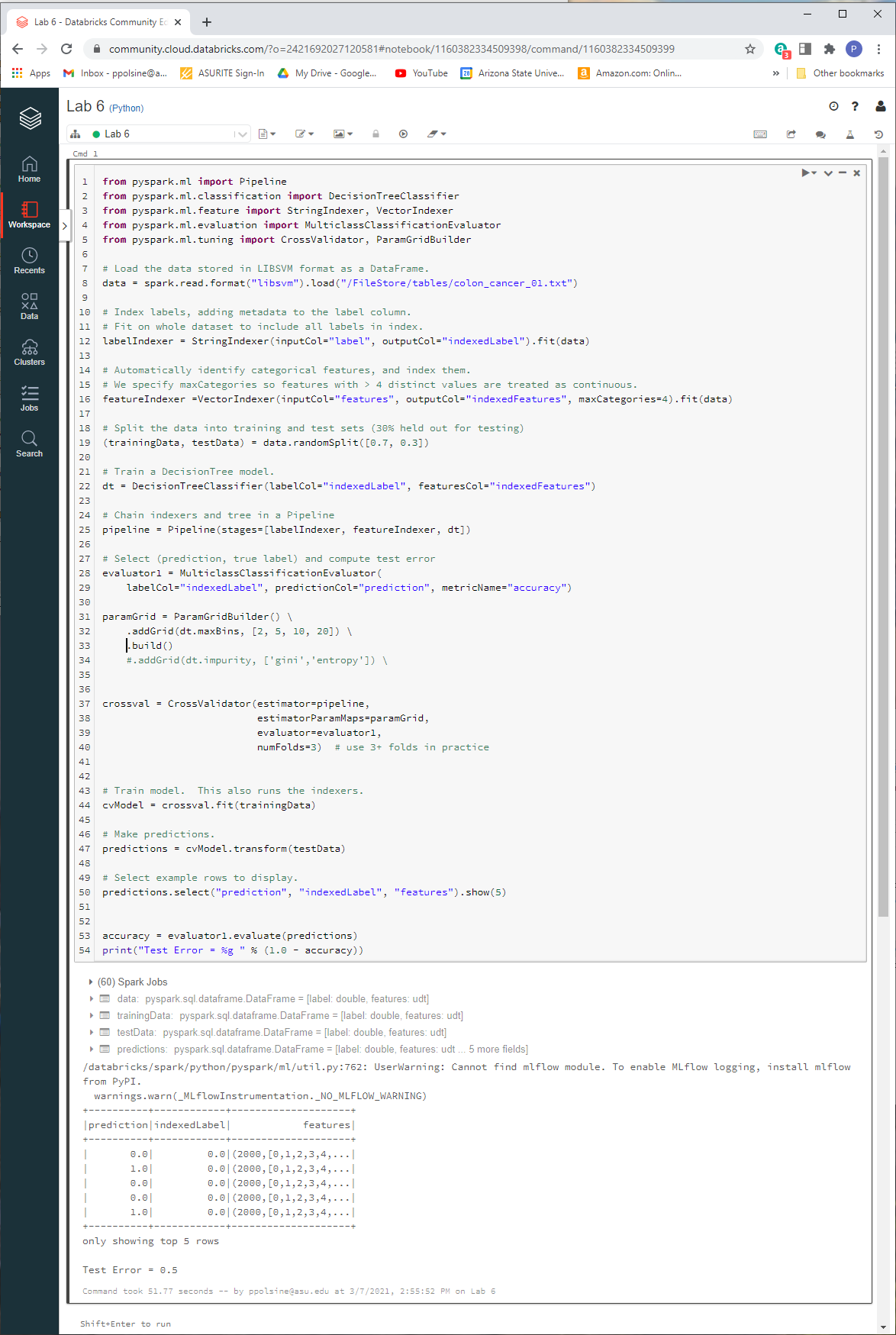
* Change the *maxBins*to *22*then *42*.

From 32 to 22 the error rate decreased to 0.3181818181818182, from 32 to 42 it decreased to 0.15789473684210525

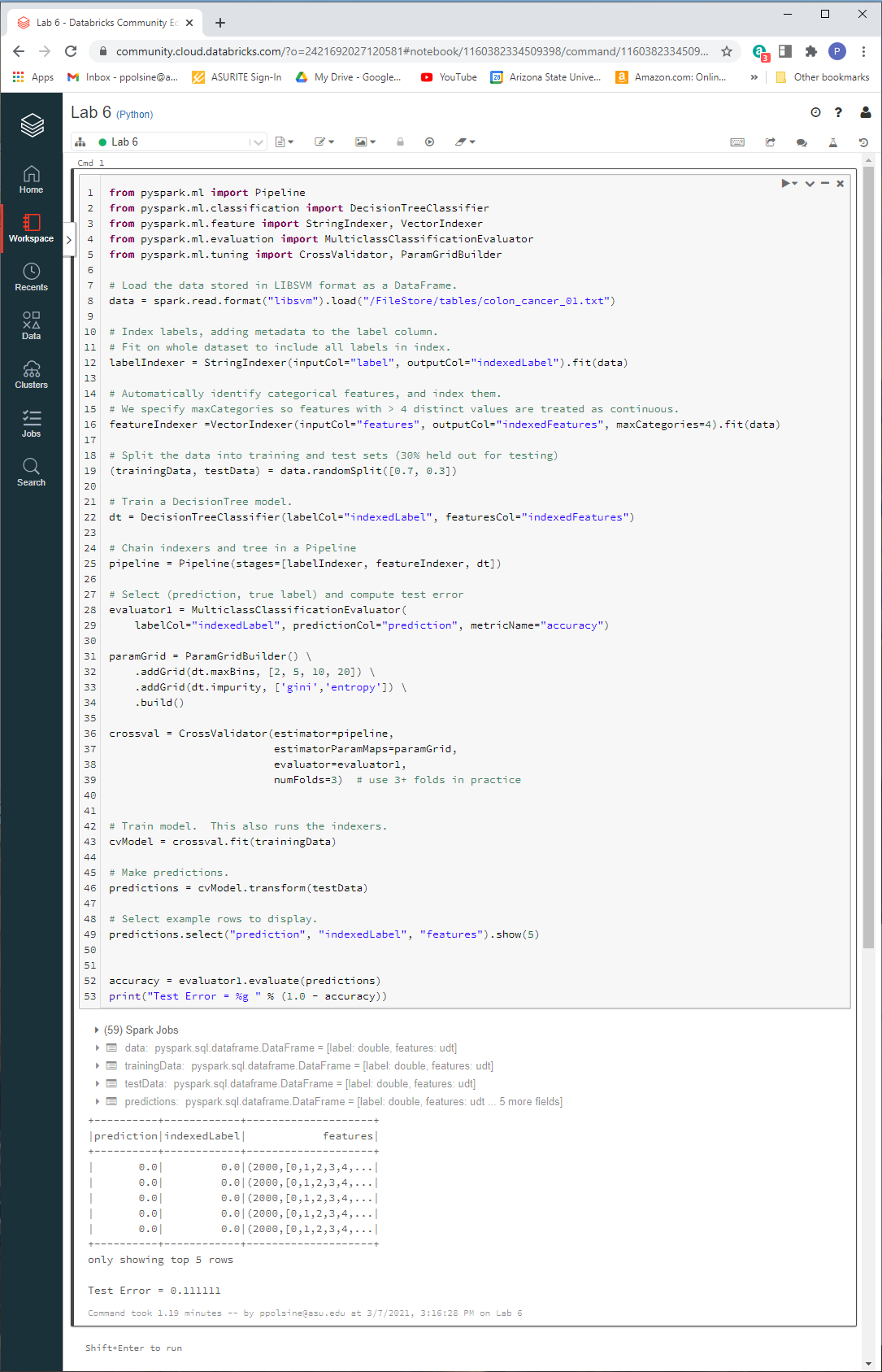




* Changing all the above parameters to each of their lowest error outpus; ‘gini’ to ‘entropy’, *maxDept* to 10, and *maxBins*to *42*, the error rate only slightly decreased from 0.36363636363636365 to 0.32 suggesting a more complex interaction amongst the various parameters then the expected maximized error reducing combined effect.

Part II: Decision Tree Training & Selection with K-Fold Cross Validation 

What is the obtained test error? 0.5

Can you insert another grid to experiment with the impurity measures 'gini' vs 'entropy'? 

What are the obtained results? Test Error = 0.111111