CSC 4760/6760 DSCI 4760 Big Data Programming

Assignment 4

Due Date: 10:30 pm ET, 11/27/2022

Problem 1. (100 points)

On Spark ML - Please use the provided Decision Tree Machine Learning Algorithm to predict the Test Accuracy on the provided dataset.

About the dataset: Iris.csv : the dataset classifies flower species based on their sepal and petal length. There are three classification labels (setosa, versicolor, and virginica)

Report:

Implementation:

Implement a PySpark program to solve the problem. We have provided an *almost complete* python code. You must implement the code in a JuPyter Notebook.

"*almost complete*" phrasing is important here – some Hints:

1. **The python code provided does not include the required statements on the findspark package needed on your JuPyter Notebook cell.**
2. **The python code, although correct, may or may not be ready to use as is (Python is strict with indentation issues, for example). It is strictly part of your assignment to verify.**

Report and Submission Materials :

Please write a report illustrating your steps of execution. **Uploading screenshots to iCollege is NOT writing a report**.

In the report, you should include the answers to the following questions.

1) Your Explanation of the provided source code, in addition to the comments provided -with screenshots

a. If required - how you clean your data.
b. How you encode your data.
c. What classification label to choose in the Decision Tree classifier algorithm.
d. What the accuracy of the algorithm's prediction is.

Submission Materials:

a) Your report in PDF containing the screenshots of the outputs

b) Source code (JuPyter Notebook)