DATA CENTER SCALE COMPUTING - LAB 4

Objective - This lab is designed to help you learn and implement pyspark MlLib and spark ml on GCP. The outcome of this assignment will be

- Start a dataproc cluster, enable gateway and use Jupyter Notebook to run spark
- Implement Machine Learning algorithm using RDD
- Implement Machine Learning algorithm using spark DataFrame
- Understand the difference between both the implementations

Dataset

The datasets are available in the repository. You must use “studentData.txt” for Task1 and “studentData.csv” for Task2

Schema - Data Set Description

Attribute	**Description**
ID	The identification number
Major	Computer Science, Electric Engineering, Mathematics, Information Science, Liberal Arts, and others
Gender	M (for Male), F (for Female)
C01-C10	The score of a course C01, C02, …, C10.
Academic	The academic performance (0 to 100)
Campus	Campus evolvement attribute record how many student-organizations joined, how many events participated/organized, etc.
Internship	Internship record how many internship the student have taken. To quantify it, we use the # of months for this record. Full time internship contributes 100% month and parr-time contributes 50% of time. For example, a student have a part-time internship for 2 months, and a full-time internship for 3 months, the total month will be 2/2 + 3 = 4

AtRisk_{academic,campus,internship}:	Student may need help in {}, 0 indicates no help needed, 1 indicates help needed
AtRisk	We may identify students at risk and reach out to help them (0 as no risk, 3 as high risk)
Graduate_program	The likelihood of the student will continue in a graduate program
Government	The likelihood of the student will take a government position
Industry	The likelihood of the student will take an industry position
Placement	The possible placement includes graduate program, industry positions,government. It will be measured by 0 to 3 where 0 is no placement, and 3 is highest.
Annual	The annual salary of the student.

Task - 1 - RDD Computation [40 points]

- You are required to use any classification method (Linear Regression, SVM) on the dataset.
- You must use **RDD computation only** and not spark Dataframe
- Use AtRisk_Academic as the target variable
- Use columns major, gender, c01,c02 , c03, c04, c05 , c06, c07, c08 , c09, c10,academic, campus, internship
- Change the categorical columns to numeric columns.
- Remove rest of the unwanted columns before modeling
- Report Rsquared error and accuracy
- Each function or code snippet should be self explanatory with necessary comments
- You may use spark DataFrame at a penalty of 20 points

Task - 2 - Spark DataFrame [40 points]

- You are required to use Logistic Regression or Random forest to perform a multiclass classification
- Perform EDA
- Compute correlation and plot the correlation Matrix
- Use required columns(choose yourselves) to get the best Metrics (Accuracy/F1 Score)
- Report metrics
- Perform Hyperparameter tuning
- Report best Hyperparameters
- Each function or code snippet should be self explanatory with necessary comments
- You must use spark ML pipeline and mention each stage

Reasoning Questions [20 points]

1. Explain the difference between the RDD Computation and using Spark DataFrame for Machine Learning in your own words(from your experience)
2. Explain ML pipeline. Your answer must include the stages in the ML pipeline, its usage and the advantages.