Department of Information Technology
Program of Data Science and Artificial Intelligence (DSAI)

Course ID: 2010042360 **Course Desc.: BigData** **Fall 2023/2024**

Assignment: 2

Distributed Machine Learning

Due Date: (Mon) 08-01-2024 11:59 PM

Max. Score: 5 points

Submission Guidelines:

1- Your submission structure should adhere to the following:
   a. Create one ***PDF*** file that shows all the requirements below; named ***{FirstName_1*** }_{***FirstName_2***}.pdf (e.g. Mohammad_Ahmad.pdf).
2- Do not use scikit-learn for this assignment. Ensure that your code is compatible with a distributed computing environment, as we are focusing on leveraging the capabilities of PySpark for large-scale data processing.
3- For each of the following, take screenshots of your solution on the virtual machine as evidence that you have completed the required task. These screenshots will serve as documentation for your work and may be requested for evaluation purposes. Make sure to capture key steps and outputs to demonstrate your understanding and execution of the assigned tasks. *(See the style of my PPT slides)*
4- Submit the group's PDF file through MS Teams Chat to me. Ensure that your submission includes all required details and evidence for the requirements listed below. Random answers are not accepted and could result in zero points for a given question. If, for example, you need to optimize the K value in the K-mean clustering, you have to show evidence (*e.g. Silhouette plot against different values of K*) that this value would be the best choice based on your dataset.

Defense Requirments

No defense or discussion is required for this assignment, just submit your PDF report.

Problem 1

Consider the attached "HAR_3000.csv" which represents a sample of a dataset related to Human Activity Recognition (HAR) generated from wearable devices. This collection of data captures various physical activities performed by individuals. These datasets are commonly used in machine learning and data science for developing models that can recognize and classify human activities (e.g. Walking, laying, etc) based on sensor data.

(*you may need to apply "*StringIndexer*" on the Activity column*)

Example:

. . .

```
ML_data = data.select(data.features, data.Activity)
from pyspark.ml.feature import StringIndexer
# Assuming "Activity" is the name of your label column
indexer = StringIndexer(inputCol="Activity", outputCol="Activity_index")
indexed_data = indexer.fit(ML_data).transform(ML_data)
```

1. Use PySpark to read the data into a DataFrame and print the number of unique classes present in the 'Activity' column.

2. What are the dimensions of this dataset?

3. Prepare the dataset for the logistic regression classification algorithm.

4. Split the dataset into 80% for training and 20 for testing (use `seed=3`).

5. On the training dataset, apply logistic regression through cross-validation with a 10-fold.

6. Evaluate the best model on the **unseen** testing dataset generated above.

7. Use multiclass classification evaluation and print out the accuracy, precision, recall, and F1-score.

8. On the same training dataset and testing dataset apply the RandomForest algorithm:

8.1 Use the grid search method to set the number of trees where the searching space is [10, 15, 20, 25] and the max depth where the searching space is [3, 5, 7, 9] that maximize the accuracy.

8.2 Which combination shows the maximum accuracy?

8.3 Extract the top 50 important features according to the RandomForest model you developed.

8.4 Subset the original dataset and keep only the top 50 features selected above + the Activity column.

9. Prepare the new subset dataset (50 features + Activity label), apply logistic regression again, and evaluate the model on the unseen testing dataset.

10. Compare the accuracy you got in (9) and the one you got in (7) above. What do you think?

Problem 2

Consider the attached "ben_inf_allFeatures_balanced_FS01.csv" which represents a sample of a dataset related to a Cybersecurity project. This collection of data captures various network activities and potential threats. It may include data such as log files, network traffic patterns, system vulnerabilities, and other relevant information that security analysts use to monitor, analyze, and respond to cybersecurity events. The dataset provides valuable insights into the cybersecurity landscape, aiding in the development of machine learning models and algorithms for threat detection, anomaly detection, and overall cybersecurity defense strategies.

1. Read the dataset, create a data frame, and prepare the dataset for clustering using the PySpark k-mean algorithm. Please use **seed= 99**.
2. What is the value of K that maximizes the overall Silhouette Score? You may need to try values between 2 and 10. Remember, you need to show the evidence. A good one could be the figure in Ch05-Slide number 31.

---------- Good Luck ----------

Dr. Majdi Maabreh