UNIVERSITY OF GUELPH
School of Computer Science

CIS*6180/DATA*6300	**Analysis of Big Data**	**Winter 2023**
Total Marks: 14	**Assignment# 4**	**Sohail Habib**

Assignment date: March 20, 2023
Assignment due date: April 03, 2023, 3:00 pm

Overview

The aim of this assignment is to introduce you to MLlib, and you should complete it using Jupyter notebook.

Please use CourseLink for all communications. Ask a private question if necessary.

What to submit?

Please provide screenshots and/or a Jupyter notebook in both PDF and ipynb formats, and then zip all the files. The zipped file should be named as "first-name_last-name.zip".

Note: The time should be visible in all the screenshots

Classification:

You will be using the craft beer data set for performing a binary classification using Spark's MLib library. The classification task is to predict the beer style using the Alcohol by volume and International Bittering Units.

Import data into Postgres SQL (0.25 points):

You are provided with craft beer data set in three csv files beers_abv.csv, beers_oun.csv, breweries.csv). Explore the provided data and import the csv files to create three tables named BeersAbv, BeerOun and Breweries, with the same schema as the csv files.

Query the Postgres SQL to get the data (0.25 points):

Using the appropriate SQL queries only import appropriate the data required to solve this task into a spark DataFrame.

Data processing (1 point):

Explore the data for data quality issues ,clean and transform the data using appropriate methods. After the data cleaning split the data into train and test sets.

Binary classification (2 points):

Train and evaluate a Random Forest and a Linear Support vector machine classifier using five-fold cross validation.

Compare classification models (1 point):

Compare the both trained classification models using accuracy, precision, recall, f1-score, and area under the receiver operating characteristic curve (AuROC)

Regression

You will use regression analysis to explore relationship between water salinity & water temperature? The goal is to predict water temperature using salinity.

Import the data into MongoDB (0.25 points):

You are provided with CalCOFI dataset provided as a csv file (bottle.csv). Import this data into a MongoDB collection named Bottle

Query the MongoDB collection to get the data (0.25 points):

Using the appropriate MongoDB queries only import appropriate the data required to solve this task into a spark DataFrame.

Data processing (1 point):

Explore the data for data quality issues ,clean and transform the data using appropriate methods. After the data cleaning split the data into train and test sets.

Regression analysis (2 points):

Train and evaluate a Random Forest Regressor and Gradient-boosted Tree Regressor using five-fold cross validation.

Compare regression models (1 point):

Compare the both regression models using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-squared, and Explained Variance.

Clustering

You will use clustering analysis to analyse Facebook live sellers in Thailand dataset.

Data import and processing (1 point):

You are provided with Facebook live sellers in Thailand dataset as a csv file (Live.txt). You will use the ‘status_type’ column as the ground truth for checking the quality of clustering. Do not use the ‘status_id', 'status_published' columns for clustering analysis.

Import the data into a Spark DataFrame and Explore the data for data quality issues ,clean and transform the data using appropriate methods. Convert any categorical variables to integers.

Clustering Analysis (2 points):

Using the elbow method find the optimal number of clusters for k-means and Gaussian Mixture Models (GMM) clustering algorithms. Use the elbow method with both WSSSE (Within Set Sum of Squared Errors) and Silhouette Scores as evaluation metrics.

Compare clustering models (2 points):

Calculate the accuracy of clustering for both algorithms using both evaluation metrics, based on the optimal number of clusters determined, with the 'status_type' column serving as the ground truth.