SEAS 8515 - Data Engineering for AI
Project 3

Due Date: May 19, 2024 (11:00pm EST)

Overview

This project involves developing end-to-end machine learning models using the Apache Spark library. The project is divided into two parts:

- **Project 3a (70 points):** Developing a machine learning model using your own dataset.
- **Project 3b (30 points):** Building a Song Genre classification model using the beatsdataset.csv dataset.

Project 3a (70 points)

For this part, you will develop an end-to-end machine learning model using your dataset. The steps are as follows:

1. **Identify the Machine Learning Problem:**
   - Define the problem type (supervised, unsupervised, semi-supervised).
   - Clearly frame the problem statement.
2. **Data Preparation and Cleaning:**
   - Use Spark APIs to clean and prepare your data.
   - Utilize libraries such as `VectorAssembler`, `StringIndexer`, and `OneHotEncoder` for feature engineering.
3. **Model Development:**
   - Combine transformers and estimators using the Spark ML Pipeline.
   - Develop the model using multiple machine learning algorithms.
   - Save and reload the model for inference.
4. **Model Evaluation:**
   - Evaluate models using built-in Spark model evaluators.

Project 3b (30 points)

In this part, you will build a machine learning model using the beatsdataset.csv data. This dataset is designed for a Song Genre classification model and includes audio features such as Zero Crossing Rate, Spectral Centroid, MFCCs, Chroma Vectors, and BPM.

1. **Data Analysis:**
   - Analyze the data and derive initial insights using Spark.
   - Address any defect in the dataset (missing data, class imbalance problem, outliers etc)
   - Find the correlation between different features.
2. **Model Development:**
   - Develop classification models using Spark MLlib.
   - Evaluate the models using appropriate metrics.
3. **Feature Importance and Insights:**
   - Determine feature importance and derive insights from the analysis.

Submission Requirements

The submission for both Project 3a and Project 3b should include:

- **Documentation:** Include well-commented source code and comprehensive documentation to explain your code and workflow.
- **Presentation:** Prepare a clear and concise presentation summarizing key aspects and outcomes of the project. The presentation should not exceed 10 slides.

Evaluation Criteria

Projects will be evaluated based on:

1. The complexity of the analysis performed.
2. Correctness of the machine learning approach.
3. Utility of Spark APIs for all aspects of the project.
4. Clarity of documentation.
5. Effectiveness of the presentation in conveying the project's results.

Grading Rubric and Criteria

Criteria	Excellent (90-100)	Good (80-89)	Satisfactory (70-79)	Needs Improvement (0-69)
Design and Planning	Comprehensive design with clear understanding of structured data processing and machine learning principles.	Solid design with good understanding of concepts.	Basic design with some gaps in understanding structured data.	Inadequate design, lacking comprehension of structured data processing.
Effective Use of Spark MLlib and Data Processing Skills	Extensive use of Spark MLlib with machine learning principles, and excellent data processing and analysis using Apache Spark and SQL.	Good use of Spark MLlib and understanding of machine learning principles, with competent use of Spark APIs.	Basic use of Spark MLlib and adequate data processing skills with limited application of advanced features.	Poor utilization of Spark MLlib, ineffective data processing, and little to no application of machine learning principles.
Code Quality and Implementation	High-quality code, efficient, well-organized, and fully integrated.	Good coding with minor issues in efficiency or organization.	Functional code but lacking in efficiency or organization.	Poorly written or non-functional code, major integration issues.
Scalability and Reliability	Excellent scalability, handling large data volumes efficiently.	Good scalability for moderate data volumes.	Sufficient for small to moderate data volumes, some scalability issues.	Poor scalability, struggles with large data volumes.
Documentation and Presentation	Comprehensive and clear documentation and presentation.	Good documentation and effective presentation.	Basic documentation and presentation, lacking some clarity.	Poor documentation and unclear presentation.