CCZG506 - API-driven Cloud Native Solutions

Assignment I

Group Details

Group No: <Fill Group No>

Group Member Names with Contribution

Sl. No	BITS ID	Name	Contribution of team member (Qualitative)	Percentage Contribution (Out of 100) Quantitative
1				
2				
3				
4				
5				
6				

General Instructions

- It is a Group Assignment.
- You can form groups of min 2 students and max upto 6 students, and execute the assignment.
- Weightage - 15 marks
- Duration -> 08-Sep-24 (3 pm) till 14-Oct-24 (11.55pm).
- Submissions after the due date will result in deduction of marks.
- URL for Group Formation - https://docs.google.com/spreadsheets/d/1sigoSALFjIVhJkBdY-WQm0fpbFnQklSzjU2LKdoUHEM/edit?usp=sharing

Project Details

Objective: Create a Cloud-based Data Science / Machine Learning application, that includes:

1. Design and Development of a Data Pipeline: This includes data ingestion, preprocessing, exploratory data analysis, and monitoring activities.
2. Design and Development of a Machine Learning Pipeline: This involves model preparation, training, assessment, and monitoring activities.
3. API Access: Access the application's details using APIs.

Sub-Objective 1: Design and Development of a Data Pipeline

Weight: 8 marks

Activities

1.1 Business Understanding: Identify a business problem in the area of data science.

1.2 Data Ingestion: For the identified problem, find an appropriate dataset from a public repository (Ex: Kaggle). Ensure that the dataset has sufficient records to carry out a meaningful data science experiment.

1.3 Data Pre-processing: Perform activities such as displaying summary statistics, checking for missing values, imputing missing data for numeric columns, displaying data types, and normalizing data.

1.4 Exploratory Data Analysis (EDA): Conduct EDA to include calculating correlation coefficients, identifying correlations between numeric and/or categorical features, binning, encoding, assessing feature importance, and visualizing data (using charts and graphs for univariate and bivariate analyses).

1.5 DataOps: Implement workflows to automate activities from steps 1.3 and 1.4 within a data pipeline. Schedule these workflows to run every 2 minutes, logging all activity details and displaying them on a Cloud dashboard.

Sub-Objective 2: Design and Development of a Machine Learning Pipeline

Weight: 5 marks

Activities

2.1 Model Preparation: Identify suitable machine learning algorithms for solving the business problem based on the dataset. Select any two algorithms.

2.2 Model Training: Split the dataset into training (70%) and testing (30%) sets and train the models.

2.3 Model Evaluation: Evaluate the models using at least one metric (e.g., accuracy for classification models).

2.4 MLOps: Monitor the model and log relevant metrics (at least four, such as accuracy, precision, recall, F1 score, etc.).

Sub-Objective 3: API Access

Weight: 2 marks

Activities

3.1 Retrieve Key Application Details: Use Built-in APIs to access important application information (e.g., flow, deployment etc.)

3.2 Display Application Details: Present at least four application details retrieved via APIs.

Submission Guidelines

a. Prepare a word document containing the project details (screenshots of the application are required, along with explanation). Highlight clearly the contribution of each group member in executing the assignment. Upload the file as <groupid.docx> or <groupid.pdf> into the portal. Note that only one student needs to upload the file representing the group. No need of Individual submissions.

b. Prepare a video demonstrating the entire project. Upload the Video into a google drive that will be shared by Dr. Shreyas.

Note

1. No extension of deadline will be given under any circumstances. Manage your time and complete the assignment within the deadline.
2. Since Assignment-I is part of EC-1 component, there is NO MAKEUP for this component.
3. Any questions or clarifications regarding the assignment can be emailed to "shreyas.rao@pilani.bits-pilani.ac.in".

All the Best!