|  |  |
| --- | --- |
| Logo, company name  Description automatically generated | **Comsats University islamabad**  **Department of Computer Science**  **Lab Term Examination, SPRING 2025** |

Class/Section: - BDS – 6 & 7 Marks: - 50

Subject: - Big Data Analytics

---------------------------------------------------------------------------------------------------------------------

**Project Title:Comparative Analysis of Big Data Models on a Real-World Dataset Using Cloud Deployment**

**Objective:**

To apply big data analytics techniques on a large dataset using PySpark, perform extensive data analysis, build and evaluate multiple machine learning models, and deploy the solution on **AWS** or **Azure** cloud.

**Project Tasks (Based on CLOs):**

**CLO 4: Apply various data analysis techniques on a given dataset**

You must demonstrate the ability to apply the following techniques:

1. **Data Acquisition**: Load and preprocess a large dataset from Kaggle, UCI, or open data portals.
2. **Exploratory Data Analysis (EDA)**:
   * Use PySpark for analysis.
   * Visualizations with libraries like Matplotlib, Seaborn.
3. **Data Cleaning & Transformation**:
   * Handle missing values, outliers, and anomalies.
   * Perform normalization, encoding, and feature engineering.
4. **Feature Selection/Dimensionality Reduction**:
   * Apply techniques such as PCA, correlation analysis, or Chi-square test.
5. **Data Partitioning**:
   * Split the dataset into training and test sets using appropriate strategies (e.g., time-based, stratified sampling).

**CLO 5: Design an experiment to compare performance of multiple models**

You are required to:

1. **Build Multiple Models**:
   * Use at least two machine learning algorithms (e.g., Logistic Regression, Decision Tree, Random Forest, XGBoost, KMeans, etc.).
2. **Compare Models**:
   * Evaluate & Compare the models using techniques as suitable.

**Cloud Deployment (Mandatory):**

* Your final notebook/script must be deployed on **AWS (SageMaker/EC2/EMR)** or **Azure (ML Studio/VM/DataBricks)**.
* The notebook must be live on cloud when you come for viva.

**Deliverables:**

1. **Jupyter Notebook / Python Script** with all code and comments.
2. **PDF Report** with:
   * Introduction, Dataset description
   * Methodology & techniques used (CLO 4)
   * Model comparison (CLO 5)
   * Cloud deployment summary
   * Conclusion
   * Cloud Deployment Screenshot