Big Data Analytics

A Project Submitted

By

Batch 2

Under the Guidance of

Kunal

BMS INSTITUTE OF TECHNOLOGY AND MANAGEMENT

YELAHANKA BANGALORE

Phase 3 Documentation

Project: Big Data Analytics – Rainfall Analysis
Phase 3 Submission by: Batch 2
Year: 4th Year Artificial Intelligence And Machine Learning
Institute: BMS Institute of Technology and Management, Yelahanka, Bangalore

---

Objective

The goal of this development phase is to build an end-to-end data analysis and prediction pipeline using Python. This phase integrates data ingestion from IBM Db2, data transformation, exploratory data analysis (EDA), and implementation of machine learning models — including regression, classification, clustering, dimensionality reduction, and time series forecasting — to enhance predictive accuracy and analytical insights.

1.Data Ingestion

Tools & Libraries Used:

- ibm_db for database connection to IBM Db2

Key Tasks:

- Established a secure connection to IBM Db2
- Parsed and inserted rainfall dataset records into the INDIA Rainfall table
- Used prepared statements and parameter binding for efficiency and security

Highlights:

- Handled credentials securely
- Verified data insertion success/failure

2.Data Transformation

Tools & Libraries Used:

- pandas for manipulation

Key Tasks:

- Retrieved data using SQL queries
- Applied transformations like column scaling and computed fields

Highlights:

- Enhanced dataset cleanliness
- Generated derived columns for better analysis

3.Exploratory Data Analysis (EDA)

Tools & Libraries Used:

- pandas, matplotlib, seaborn

Key Tasks:

- Generated descriptive statistics
- Created visualizations.

- Histogram
- Scatter Plot
- Correlation Heatmap
- Boxplot grouped by category

Highlights:

- Identified data distributions, patterns, and outliers
- Revealed key correlations and anomalies

4.Machine Learning Implementation

Tools & Libraries Used:

- scikit-learn, seaborn, statsmodels

---

a) Linear Regression

- Predicted continuous values (e.g., rainfall)
- Evaluated using **Mean Squared Error (MSE)**

---

b) Random Forest Classifier & Regressor

- Used for both classification and regression tasks
- Evaluated with:
  - **Accuracy**
  - **Confusion Matrix**
  - **Classification Report**
  - **MSE** for regression

---

c) K-Means Clustering

- Applied unsupervised clustering with n_clusters=3
- Visualized cluster distribution using Seaborn scatter plot

5.Dimensionality Reduction with PCA

Tools:

- sklearn.decomposition.PCA

Tasks:

- Reduced features into two principal components
- Visualized 2D scatterplot of PCA-transformed data

Benefits:

- Simplified complex datasets
- Enabled better visual understanding of clustering

6.Time Series Analysis with ARIMA

Tools:

- statsmodels.tsa.ARIMA

Tasks:

- Fitted ARIMA(1,1,1) on univariate rainfall data
- Analyzed residuals using ACF and PACF plots

Insights:

- Detected trends and seasonality
- Validated model fit with statistical residual plots

7.Advanced Visualization

Tools:

- seaborn.pairplot

Tasks:

- Plotted relationships between multiple features
- Used hue to categorize by target class (customizable)

Challenges & Solutions

Challenge	Solution
Node.js vs browser JavaScript environment	Used docs to understand global and module contexts
Asynchronous fetch in frontend/backend	Implemented async/await and Promises
CORS issues with React-Node interaction	Set appropriate CORS headers on the backend
JSX syntax and component design	Refined using planned structure and practice
Form validation in React	Used controlled components with useState
State management complexity	Applied useReducer and Context API
Non-linear model input preparation	Validated features and formats
Residual interpretation for ARIMA	Used ACF and PACF for accurate residual analysis
Choosing PCA components	Started with 2; extensible based on variance explained

---

Evaluation Metrics

- **MSE** – For regression models
- **Accuracy, Confusion Matrix, Classification Report** – For classification
- **Cluster visualizations** – For K-Means
- **ACF/PACF residual plots** – For ARIMA
- **Visual comparisons** – Predicted vs. Actual

Outputs

- Regression predictions and actual comparisons
- Cluster-labeled scatterplots (K-Means)
- PCA-reduced 2D visualizations
- Time series residual analysis (ACF/PACF)

- Seaborn pairplots showing multi-feature relationships

Conclusion

Phase 3 marks the transition from data exploration to intelligent modeling. The integration of regression, classification, clustering, PCA, and ARIMA allows the system to uncover rainfall patterns and forecast future trends. Robust metrics and visualization techniques support interpretability and performance validation.

This phase establishes a scalable, analytical core ready for deployment, UI integration, and real-time data processing in subsequent stages.