Big Data Analysis

A Project Submitted

By

AMAN GUPTA

Under the Guidance of

PROF. PRABAKARAN M V

DHANALAKSHMI SRINIVASAN

COLLEGE OF ENGINEERING AND TECHNOLOGY

ECR, MAMALLAPURAM. KANCHIPURAM DISTRICT.

Approved by AICTE, NEW DELHI | Affiliated to Anna University, Chennai.

Phase 4 Documentation: Development Part 2

Project: Big Data Analysis

Phase 4 Submission by: Aman Gupta
Year: 3rd Year, Computer Science and Engineering (CSE)

Table of Contents

1. Introduction

This document represents Phase 4 of the "Big Data Analysis" project, submitted by Aman Gupta, a 3rd-year student of Computer Science and Engineering (CSE). Phase 4, titled "Development Part 2," is a continuation of the project, focusing on advanced analysis techniques and visualization of results.

2. Advanced Analysis

Objective: In this phase, we continue building the big data analysis solution by applying advanced analysis techniques to the dataset.

Random Forest Regression

What to Do:
- The dataset is further analyzed using Random Forest Regression to make predictions.
- Model evaluation is performed using the Mean Squared Error metric.
- The results are visualized, comparing actual and predicted rainfall values in a scatter plot

Principal Component Analysis (PCA)

What to Do:
- Dimensionality reduction is applied using Principal Component Analysis (PCA) to explore patterns in the data.
- Reduced-dimensional data is visualized in a scatter plot, showing the first two principal components.

Time Series Analysis

What to Do:
- Time series analysis is conducted, assuming a time series dataset.
- An ARIMA model is fitted to the data, and Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots are visualized.

Advanced Visualizations

What to Do:
- More advanced visualizations are applied, such as a pairplot to explore relationships between features in the dataset.
- Pairplots are used to visualize relationships between specific features, with color-coding based on a target category.

3. Conclusion

Phase 4, "Development Part 2," continues to build on the big data analysis project by applying advanced analysis techniques and visualization methods to the dataset. The use of Random Forest Regression, PCA, time series analysis, and advanced visualizations provides deeper insights into the data. These techniques contribute to the project's overall goal of uncovering valuable insights from big data.