BDA Mini Project Report

This report presents the Big Data Analytics (BDA) Mini Project conducted using **PySpark** and **Matplotlib**. The primary goal of this project is to analyze **pesticides dealers license data** to uncover insights related to the distribution of licensed pesticide dealers across different districts and taluks. The project demonstrates data processing using PySpark, along with visual representation of trends through various graphs.

Objectives:
- To perform large-scale data analysis on pesticides dealers license data using PySpark.
- To calculate and visualize the total number of dealers by **district** and **taluk**.
- To identify **top-performing districts and taluks** based on the total number of dealers.
- To generate meaningful visual insights through **graphs and charts**.

Methodology:
1. The dataset was loaded into a **PySpark DataFrame** for distributed processing.
2. The **total number of licensed pesticide dealers** was computed for each district and taluk.
3. **Aggregation and ranking** were performed to determine the **top districts and taluks** with the highest number of dealers.
4. The processed data was converted into **Pandas DataFrame** for visualization using **Matplotlib**.
5. **Five graphs** were created to represent different analytical views of the dataset.

Results and Visualizations:
The project generated five key visual outputs:
1. **Pie Chart:** Distribution of total licensed pesticide dealers among top 5 districts.
2. **Bar Chart:** Top 5 districts by total number of pesticide dealers.
3. **Line Chart:** Trend of pesticide dealer counts across all districts in descending order.
4. **Horizontal Bar Chart:** Top 5 taluks with the highest number of pesticide dealers.
5. **Bar Chart:** Comparison of dealer counts across all districts for overall distribution analysis.

Conclusion:
This BDA mini project successfully demonstrates how **PySpark** can be integrated with **Matplotlib** to perform big data analysis and visualization efficiently. The insights derived from the **pesticides dealers license data** enable a better understanding of **geographical concentration and distribution of licensed dealers across districts and taluks**. Such insights can help policymakers, agricultural departments, and suppliers identify **key regions of activity** and plan resource distribution more effectively.

Tools Used: PySpark, Pandas, Matplotlib
Environment: Jupyter Notebook
Dataset: Pesticides Dealers License Report (CSV)