Traffic Accident Analysis in India using PySpark

Subject: Big Data Analytics Mini Project

Name: Varshitha Gosala
Rollno: 2211CS010699
Section: Section_1_91
Semester: 7 (CSE)

Source: "Kaggle – India Road Accident Dataset Predictive Analysis"
Repository: "Traffic Accidents Analysis"

Executive Summary
Road traffic accidents remain a major public safety concern in India, contributing to thousands of fatalities and injuries annually. This project leverages **Big Data analytics using PySpark** to perform large-scale analysis of accident data, uncovering key trends and factors influencing accident severity. Through distributed processing and advanced visualization, the study identifies patterns related to time, weather, vehicle type, and geography.
The analysis provides actionable insights for policymakers and traffic authorities to improve road safety and reduce casualties through data-driven interventions.

---

1. Introduction & Business Context

1.1 Problem Statement

India records one of the highest numbers of road accidents globally. Traditional accident reporting systems and small-scale analyses fail to capture large-scale trends across diverse states and years.
This project aims to address this challenge by applying **Big Data technologies (PySpark)** to analyze extensive accident datasets, detect patterns, and generate insights that can inform prevention and safety strategies.

1.2 Business Impact

- Reduce accident rates through data-driven policy decisions
- Identify high-risk locations and time windows for targeted intervention
- Support better infrastructure planning and road safety awareness programs
- Enable continuous monitoring through scalable data pipelines

---

2. Dataset & Objectives

2.1 Dataset Description

The dataset used in this study captures road traffic accident details across multiple Indian states, including factors such as location, time, weather, vehicle type, and casualties.

2.2 Key Characteristics

- **Total Records:** ~400,000 accident entries
- **Features:** 10+ attributes (State, Month, Weather, Severity, Vehicle Type, Time, Casualties, etc.)
- **Time Period:** 2020–2023
- **Data Source:** Kaggle – Traffic Accident Data India

2.3 Feature Breakdown

Feature	Description
State	Indian state where the accident occurred
Month	Month of the year (used for temporal trends)
Weather_Condition	Weather at the time of the accident
Vehicle_Type	Type of vehicle involved
Severity	Level of accident severity (Minor, Serious, Fatal)
Casualties	Number of casualties in the accident
Time_of_Day	Time category (Morning, Afternoon, Evening, Night)

2.4 Primary Objectives

1. Perform **state-wise and temporal analysis** of accidents
2. Analyze **relationships between weather, time, and severity**
3. Identify **high-risk states and vehicle categories**
4. Build **visualizations** to communicate insights effectively
5. Demonstrate **scalable data processing using PySpark**

---

3. Methodology & Technologies

3.1 Technology Stack

- **PySpark:** For distributed data processing and transformations
- **Python:** For scripting and analysis
- **Pandas:** For final aggregation and visualization
- **Matplotlib / Seaborn:** For data visualization
- **Jupyter Notebook:** Development and analysis environment

3.2 Analytical Approach

Phase 1 – Data Preprocessing

- Load and clean large-scale accident dataset using PySpark
- Handle missing or inconsistent values
- Convert categorical fields into standard formats

Phase 2 – Exploratory Data Analysis (EDA)

- Perform aggregation and grouping using Spark SQL
- Derive insights by severity, weather, time, and region

Phase 3 – Visualization

- Generate plots such as heatmaps, bar charts, pie charts, and line plots
- Correlate multiple factors (e.g., weather vs severity, vehicle type vs casualties)

Phase 4 – Insights & Recommendations

- Summarize key findings and propose actionable safety measures

---

4. Data Preprocessing & Feature Engineering

4.1 Data Cleaning

- Removed null or duplicate entries

- Standardized state and weather labels
- Converted “Severity” into categorical numeric form (1–3 scale)

4.2 Data Partitioning

Data was partitioned by **state** and **year** using Spark’s distributed architecture, enabling parallel analysis across large records efficiently.

4.3 Feature Engineering

- Derived **Time_of_Day** from timestamps
- Calculated **Casualties per Accident** metric
- Created **Severity_Index** for weighted severity scoring

---

5. Exploratory Data Analysis & Key Insights

5.1 Visualization 1: State-wise Accident Heatmap

Chart Type: Heatmap

Purpose: Show accident intensity across Indian states

Key Findings:

- High accident densities in **Uttar Pradesh, Maharashtra, and Tamil Nadu**
- Northeast and smaller states show fewer but often more severe accidents

---

5.2 Visualization 2: Accident Count by Month

Chart Type: Bar Chart

Purpose: Analyze monthly trends

Key Findings:

- Accident rates spike during **monsoon months (June–September)**
- Correlation with weather-related visibility and road conditions

---

5.3 Visualization 3: Accident Severity Distribution

Chart Type: Pie Chart

Purpose: Show proportions of minor, serious, and fatal accidents

Key Findings:

- **Minor accidents:** ~65%
- **Serious accidents:** ~25%
- **Fatal accidents:** ~10%
- Emphasizes need for better emergency response for severe cases

---

5.4 Visualization 4: Weather vs Severity

Chart Type: Grouped Bar Chart

Purpose: Show how weather influences severity

Key Findings:

- **Rainy and foggy** conditions correlate with higher severity
- **Clear weather** still accounts for majority due to higher traffic volume

---

5.5 Visualization 5: Vehicle Type vs Casualty Count

Chart Type: Boxplot
Purpose: Analyze casualties by vehicle type
Key Findings:

- **Two-wheelers** contribute to the highest casualty rates
- **Heavy vehicles** show fewer accidents but higher fatality rates per event

---

5.6 Visualization 6: Time of Day vs Accident Frequency
Chart Type: Line Plot
Purpose: Identify accident frequency across day periods
Key Findings:

- Peaks during **evening and late-night hours**
- Likely causes: fatigue, poor lighting, and high-speed travel

---

6. Predictive Insights & Interpretation
Though not a machine learning prediction task, the analysis yields **predictive insights** for accident prevention. By understanding temporal, spatial, and weather-based correlations, authorities can anticipate and prevent high-risk scenarios.
Example Insights:

- Deploy **traffic enforcement** during high-risk time windows
- Install **warning systems and reflectors** in fog-prone zones
- Implement **awareness programs** for two-wheeler safety

---

7. Conclusion & Strategic Recommendations
Summary of Key Findings

1. **High-risk States:** Uttar Pradesh, Tamil Nadu, Maharashtra
2. **Peak Accident Periods:** June–September (Monsoon)
3. **Vehicle Type:** Two-wheelers most vulnerable
4. **Weather Impact:** Rainy/foggy weather increases severity
5. **Time Factor:** Evenings and nights most accident-prone

Business Value Proposition
Implementing data-driven road safety programs can:

- Reduce road fatalities by 20–30%
- Improve urban planning for safer transport corridors
- Optimize resource allocation for emergency response
- Enable predictive monitoring of accident-prone zones

Strategic Recommendations

1. **Infrastructure Improvements**
    - Enhance lighting, signage, and road quality in high-accident regions.
    - Build better drainage systems for monsoon months.
2. **Predictive Safety Systems**

    - Develop a **real-time monitoring dashboard** using PySpark + Power BI.
    - Use predictive models to issue alerts for high-risk conditions.
3. **Public Awareness Campaigns**
    - Target two-wheeler and heavy-vehicle drivers with safety education.
4. **Policy and Enforcement**
    - Increase patrolling during high-risk hours.
    - Impose stricter penalties for speeding and drunk driving.

Future Enhancements

- Integrate **live traffic data and weather APIs** for real-time analysis.
- Explore **accident prediction models** using MLlib.
- Deploy **interactive dashboards** for traffic authorities.
- Include **geospatial mapping (GIS)** for location-based insights.

☑ **In summary:**

This PySpark-based Big Data project demonstrates the power of distributed analysis for large-scale public safety applications.

The insights generated form a solid foundation for data-driven policy-making in Indian road safety management.