**Regulatory Affairs of Road Accident Data 2020 India**

Submitted by: ABHIJIT SINHA

Email: sinhaabhijit12@yahoo.com

UNID: UMID08062541597

# Abstract

This project analyzes the causes and outcomes of road accidents in million-plus cities of India (2020). The dataset was obtained from data.gov.in and contains details such as accident causes, outcomes (injuries, deaths), and counts. The analysis leverages Python, SQL, Excel, and Machine Learning models, integrated into a Streamlit app for interactive exploration.

# Dataset Description

The dataset is sourced from data.gov.in and contains 9,550 records across 5 columns:  
1. Million Plus Cities – City names  
2. Cause Category – Primary accident cause category  
3. Cause Subcategory – Detailed accident cause  
4. Outcome of Incident – Accident outcome (e.g., Persons Killed, Injured)  
5. Count – Number of cases

A close-up of a list

AI-generated content may be incorrect.

# Technology Stack

* Languages: Python, SQL
* Libraries: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, openpyxl
* Database: MySQL (via SQLAlchemy)
* Visualization: Streamlit Dashboard, Excel exports
* Other: GitHub for version control

# Project Workflow and Implementation Steps

# 1. Data Preprocessing

The dataset was cleaned and standardized before analysis. Steps included  
 - Removing missing values  
 - Standardizing column names  
 - Handling inconsistent text entries  
 - Converting numerical columns to proper data types

# 2. Exploratory Data Analysis (EDA)

* Top 10 accident-prone cities
* Distribution of causes & outcomes
* Accident Causes vs Outcomes

**3. SQL Queries & Insights**

* Total by outcome
* Top 10 cities (Persons Killed)
* Top 20 causes (Persons Killed)

# 4. Machine Learning Models

* Train models to predict **Counts of Persons Killed**
* Compare **Linear Regression vs Random Forest**
* Evaluate using **MAE and R²**

# 5. Streamlit Application

An interactive Streamlit app was developed with

* dataset preview,
* SQL query execution,
* charts,
* ML evaluation.

It allows users to explore the dataset dynamically and view real-time results.

# Analysis & Results

1. **The Preprocessing Flow Diagram.**

A group of colorful text boxes

AI-generated content may be incorrect.

1. **EDA Charts**

A graph of a number of cities

AI-generated content may be incorrect.

A graph with orange and white bars

AI-generated content may be incorrect.

A graph of injury and injury

AI-generated content may be incorrect.

**3. SQL Query Result Visualizations**

A graph with orange bars

AI-generated content may be incorrect.

A graph of a bar graph

AI-generated content may be incorrect.

A graph of a number of different colored bars

AI-generated content may be incorrect.

**4.ML Model Training Visualizations**

A graph with red and black lines

AI-generated content may be incorrect.

**Caption: -** ML Model Performance Comparison (Cross-Validation R² Scores).

**Interpretation: -** The **Random Forest model consistently outperforms Linear Regression** with higher predictive accuracy (R²). This suggests that accident count prediction benefits from capturing complex non-linear relationships between cities, cause categories, and subcategories.

# 5. Stream lit app screenshots

# a) Overview Page

**A screenshot of a computer

AI-generated content may be incorrect.**

**b) SQL Queries page**

**A screenshot of a computer

AI-generated content may be incorrect.**

**c) SQL Charts and visualizations page**

**A graph of a bar graph

AI-generated content may be incorrect.**

**d) SQL Machine Learning page**

**A screenshot of a computer

AI-generated content may be incorrect.**

# Insights & Future Work

Key findings:  
- Certain cities reported higher accident counts, requiring targeted policy action.  
- Accident causes were dominated by road features and impacting vehicles.  
- Fatalities and injuries varied significantly by cause category.  
- Traffic Control & Road Features were leading causes of accidents.

Add time-series forecasting for accident trends Include geospatial analysis with maps

# Conclusion

The project successfully demonstrated the integration of Python, SQL, ML, and visualization tools to analyze road accident data.  
• Metropolitan cities reported the highest accident numbers.  
• Road features and traffic violations were leading causes.  
• Random Forest outperformed Linear Regression with lower MAE and higher R².

The interactive Streamlit app makes it easy to explore insights and predictions.

# References

**Dataset:** C:\Users\Hp\OneDrive\Desktop\python\Road\_Accident\_Analysis\Regulatory Affairs of Road Accident Data 2020 India.csv  
**GitHub Project:** https://github.com/abhi-1009/Regulatory\_Affairs\_Of\_Road\_Accidents