A COURSE-BASED PROJECT

On

Road Accident Analysis

Submitted in partial fulfillment of Data Mining & Analytics Lab

GRIET Lab On Board (G-LOB)

By

P. Akshay Reddy	22241A3245
T. Sathwik Reddy	22241A3257
B. Shashank	22241A3252

Department of Data science

GOKARAJU RANGARAJU INSTITUTE OF ENGINEERING AND TECHNOLOGY

(Autonomous)

Bachupally, Kukatpally, Hyderabad, Telangana, India,500090

2024-2025

GOKARAJU RANGARAJU INSTITUTE OF ENGINEERING AND TECHNOLOGY
(Autonomous)

Hyderabad-500090

CERTIFICATE

This is to certify that the GLOB entitled "**Road Accident Analysis**" is submitted by **P.Akshay Reddy(22241A3245),T.Sathwik Reddy(22241A3257) and B.Shashank (22241A3252)** in partial fulfillment of the award of degree in BACHELOR OF TECHNOLOGY in Computer Science and Business System during Academic year 2024-2025.

Internal Guide
Ms. K. Kalpana
Assistant Professor

Head of the Department
Dr. S. Govinda Rao
Professor

ABSTRACT

This project conducts a comprehensive analysis of road accident data by integrating three large datasets—Accidents, Casualties, and Vehicles—through a shared identifier, `accident_index`. The data underwent extensive preprocessing, including handling missing values, converting time and date formats, and mapping coded variables to human-readable categories. Exploratory data analysis was performed using a variety of visualization tools. Sankey diagrams highlighted the flow between accident severity and vehicle types, while treemaps and sunburst charts revealed patterns in road types and severity levels. Geospatial scatter plots and Folium maps provided insight into the geographic distribution of accidents, and contour plots linked weather and road surface conditions to casualty counts. Time-based patterns were explored using monthly trend line plots and heatmaps of weekday and month-wise accidents. Additional analyses such as bubble charts, radar plots, and parallel coordinates helped uncover relationships among speed limits, casualty counts, and vehicle characteristics. The findings from this analysis can support traffic authorities, urban planners, and policymakers in identifying high-risk zones, understanding contributing factors, and developing evidence-based strategies to reduce accident rates and improve road safety.

Table of Contents

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction to the Project Work:

Road traffic accidents are a major concern worldwide, contributing to a significant number of deaths, injuries, and property damage each year. As urban populations grow and the number of vehicles increases, the risk of accidents rises, making road safety a critical area of study for government bodies, urban planners, and traffic authorities. The goal of this project is to conduct a detailed and data-driven analysis of road traffic accidents using three extensive datasets: **Accidents**, **Casualties**, and **Vehicles**. These datasets, when merged via the common identifier *accident index*, provide a holistic view of each traffic incident, including information about the location, severity, environmental conditions, vehicles involved, and individuals affected. To derive meaningful insights, the data undergoes preprocessing steps such as handling missing values, converting dates and times into usable formats, and mapping numerical codes to understandable categories. Once cleaned, the dataset is explored through a variety of visualization techniques to identify trends and relationships. For example, **Sankey diagrams** illustrate the flow between accident severity and vehicle types, while **sunburst charts** and **tree maps** break down severity levels across different road types. **Geospatial analysis** is employed to map accident locations and severity across regions, and **contour plots** explore how weather and road surface conditions relate to casualty numbers. Additionally, **bubble charts**, **parallel coordinates**, and **radar plots** examine how factors such as speed limits, engine capacity, and vehicle type influence accident outcomes. This analytical approach enables the identification of high-risk factors and accident hotspots, offering valuable guidance for the development of targeted interventions. By understanding which combinations of conditions most commonly lead to serious accidents, authorities can prioritize safety improvements, enforce relevant policies, and ultimately reduce the frequency and severity of road traffic incidents.

1.2 Significance of the Project:

The significance of this project lies in its potential to contribute to the reduction of road traffic accidents through data-driven insights. By analyzing extensive and real-world data on accidents, vehicles, and casualties, the project uncovers critical patterns and correlations that are often overlooked in traditional analysis. Through comprehensive visualizations and statistical interpretations, it identifies high-risk scenarios such as specific road types, vehicle categories, or environmental conditions associated with severe accidents. These insights can be instrumental for traffic management authorities, urban planners, and policymakers to make informed decisions about road design, traffic regulations, and public awareness campaigns. Furthermore, the integration of geospatial analysis helps pinpoint accident-prone locations, enabling targeted interventions. Overall, the project supports the broader objective of enhancing road safety, minimizing casualties, and optimizing emergency response strategies, thereby promoting safer and more efficient transportation systems.

LITERATURE SURVEY

2.1 Existing Approaches:

Road accident analysis has traditionally been conducted using standard data reporting systems maintained by traffic departments and government agencies. These systems often compile data annually, presenting summaries of total accidents, fatalities, and injuries based on limited variables such as location, time, and basic vehicle or casualty characteristics. Most of the analyses are descriptive in nature, relying on bar charts, pie charts, and tabular representations to communicate findings. While these methods provide a broad overview, they do not fully capture the complexity or multi-dimensional relationships among the factors contributing to accidents. Additionally, these datasets are often processed in isolation—for instance, accident data is studied separately from casualty and vehicle information—resulting in a lack of holistic understanding. Furthermore, these approaches generally lack automation and rely heavily on manual processes for data cleaning, integration, and analysis, which are time-consuming and prone to human error. The absence of interactive visualizations and advanced analytical techniques, such as machine learning, geospatial analysis, or network flow models, limits the potential for in-depth insight generation. Without the ability to drill down into the data or cross-reference variables dynamically, stakeholders are restricted to static interpretations that may overlook critical safety risks. As a result, traditional approaches struggle to support data-driven decision-making, proactive accident prevention, or efficient policy formulation aimed at improving road safety.

2.2 Drawbacks of Existing Approaches:

The existing approach to road accident analysis has several key limitations. First, it relies heavily on aggregated and historical data, often lacking real-time updates, which makes it difficult to address emerging issues or take timely action. Additionally, traditional methods typically focus on basic statistics and simple visualizations, such as bar charts and pie charts, which fail to uncover deeper insights or complex relationships between factors like weather conditions, road type, and vehicle characteristics. Another drawback is the lack of data integration. Accident, casualty, and vehicle information are often analyzed separately, preventing a comprehensive understanding of how these factors interact. Furthermore, the absence of advanced analytical techniques, like predictive modeling or machine learning, limits the ability to identify patterns and make data-driven predictions. Finally, these approaches are often resource-intensive and prone to human error due to manual data processing, making them reactive rather than proactive in improving road safety.

PROPOSED METHOD

3.1 Problem Statement

Road traffic accidents are a significant global issue, causing loss of life, injuries, and economic damage. Despite the availability of extensive datasets on traffic accidents, current analysis methods are limited in their ability to provide actionable insights that can drive effective policy and intervention. Traditional approaches often rely on aggregated, outdated data, and basic visualizations that fail to capture the complexity of accident causation, including the interactions between road conditions, vehicle types, weather, and driver behavior. Additionally, the lack of real-time data integration, advanced analytics, and interactive tools results in a reactive approach to road safety, where preventive measures are often implemented only after accidents occur.This project aims to address these limitations by developing a comprehensive, data-driven approach that combines accident, casualty, and vehicle datasets for deeper insights. By leveraging advanced visualization techniques, predictive analytics, and real-time data, the project seeks to improve the understanding of accident patterns and severity, identify high-risk zones, and ultimately contribute to more effective road safety measures and interventions.

3.2 Objectives of the Project:

The main objectives of this study, based on the exploratory analysis conducted in the Google Collab, are as follows:

1. **Comprehensive Data Integration**: To merge accident, casualty, and vehicle datasets into a unified dataset that allows for a holistic understanding of road traffic accidents.

2. **Data Cleaning and Preprocessing**: To clean and preprocess the data, including handling missing values, converting time and date columns, and normalizing numeric data for accurate analysis.

3. **Exploratory Data Analysis (EDA)**: To perform an in-depth analysis of accident patterns, identifying key factors such as weather conditions, road types, vehicle types, and accident severity that contribute to road traffic incidents.

4.**Advanced Visualization Techniques**: To create interactive and insightful visualizations, including Sankey diagrams, heatmaps, treemaps, and geospatial maps, that highlight the relationships between accident severity, road conditions, vehicle types, and other contributing factors.

5.**Predictive Modeling**: To develop predictive models that can forecast accident severity and identify high-risk zones or times based on historical data.

6.**Real-time Data Integration**: To explore the potential for integrating real-time data (if available) into the analysis for timely decision-making and more effective traffic management.

7. **Improved Decision-Making for Road Safety**: To provide actionable insights that can help traffic authorities and policymakers in developing targeted road safety measures, adjusting traffic laws, and improving infrastructure based on data-driven findings.

3.3 Explanation of Architecture Diagram :

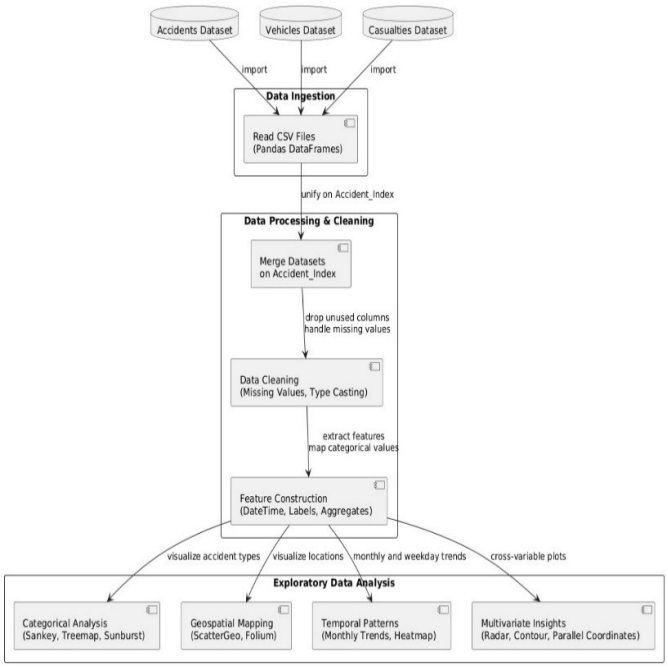

Fig1: Architecture Diagram

System Architecture (Summary)

1. **Raw Datasets:**
   Three files—Accidents dataset, Vehicles dataset, and Casualties dataset—are used as the primary data sources.
2. **Data Loading:**
   Data is loaded into the environment using Pandas for initial inspection and manipulation.
3. **Merging Datasets:**
   The datasets are merged on the common column `Accident_Index` to combine accident, vehicle, and casualty information.
4. **Data Processing & Cleaning:**
   Irrelevant columns are removed, missing values are handled, and the data is cleaned (e.g., replacing -1 with NaN, filling missing values).
5. **Exploratory Data Analysis (EDA):**
   Statistical summaries and trends are explored, such as accident severity, vehicle types, weather conditions, and road surface conditions.
6. **Data Visualization:**
   Visual tools like Plotly, Matplotlib, and Seaborn are used to generate Sankey diagrams, heatmaps, bubble charts, and geographical maps.
7. **Insights & Interpretation:**
   Key findings are extracted from the visualizations to understand patterns in accidents, such as high-risk areas, vehicle involvement, and accident severity.

CHAPTER 4

RESULTS AND DISCUSSION

The analysis reveals that most accidents occur in central and southern regions, with minor injuries being the most common, though serious and fatal cases are also prevalent in high-traffic urban areas, suggesting a need for targeted safety measures. The tree map shows that intersections and "other road types" account for the highest number of incidents, mainly minor, while straight and curved roads, though less frequent, are associated with more severe outcomes—likely due to higher speeds. Speed limit analysis across weekdays highlights Friday as having the highest accident count, while Sunday sees the lowest, with most accidents occurring in 30 km/h zones, indicating that even low-speed areas face significant risks. The radar chart emphasizes that accident severity and speed limit are the most influential factors, with higher engine capacities also correlating with greater casualty severity, underlining the importance of regulating speed and vehicle power, particularly in vulnerable zones.

Experimental Results:

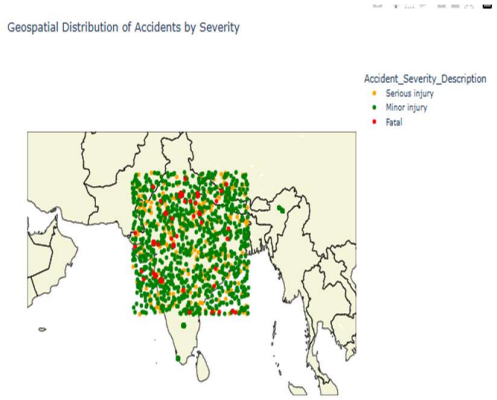

Fig 2:Geospatial Distribution of Accidents

Fig 3: Tree map of Accidents by Road type and severity

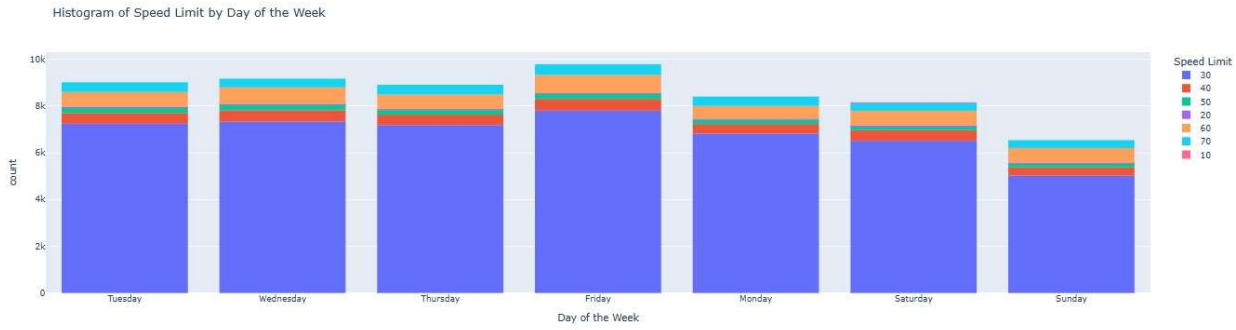

Fig 4: Histogram of Speed limit by Day of the Week

CHAPTER 5

CONCLUSION AND FUTURE ENHANCEMENTS

5.1 Conclusion

The traffic accident analysis project provides valuable insights into patterns and trends associated with road accidents. By merging datasets on accident details, vehicle types, and casualties, and performing thorough data cleaning and feature engineering, we were able to generate meaningful visualizations. Tools like Sankey diagrams, treemaps, and geospatial mapping effectively highlighted key patterns such as accident severity based on vehicle type, accident locations, and temporal trends. Furthermore, the project provides actionable insights that could be used for better road safety measures, targeted interventions, and policymaking. Ultimately, this analysis contributes to a deeper understanding of factors affecting road safety and offers a foundation for further investigations or predictive models in the domain.

5.2 Future Enhancements

While the current traffic accident analysis provides a strong foundation, several enhancements can further improve its utility and impact. Future work could include:

1. **Integrate Machine Learning Models:**
   Predict accident severity, casualty numbers, or accident hotspots using machine learning models.
2. **Incorporate External Datasets:**
   Add datasets like weather conditions, traffic volume, or road maintenance history for deeper analysis.
3. **Interactive Dashboard:**
   Convert the analysis into an interactive dashboard using tools like Streamlit or Power BI for real-time insights.
4. **Add Filtering Options:**
   Enable users to filter the data by factors such as time, location, or vehicle type for customized exploration.
5. **Advanced Statistical Analysis:**
   Use hypothesis testing or regression models to identify key factors influencing accident severity and frequency.
6. **Automate Data Updates:**
   Set up automated systems to update the dataset with new traffic accident data, ensuring up-to-date insights.
7. **Focus on Specific Accident Types:**
   Conduct deeper analysis on specific accident types or categories, such as pedestrian accidents or collisions involving heavy vehicles.

REFERENCES

[1]. S. Gupta, R. Mishra and K. Mehta, "Data Analytics in Healthcare: A Review of Techniques and Applications," Journal of Data Science and Analytics, vol. 8, no. 2, pp. 58-65, 2020.

[2]. A. Sharma, R. Soni and N. Verma, "Machine Learning Approaches for Predictive Analytics in Data Mining: A Survey," International Journal of Data Science and Analysis, vol. 5, no. 3, pp. 215-223, 2018.

[3]. M. Ali, L. Patel and P. Arora, "A Comparative Study of Regression Techniques for Big Data Analytics," International Journal of Big Data Analysis, vol. 12, no. 4, pp. 99-107, 2021.

[4]. T. Bansal, S. Saxena and A. Gupta, "Data Analytics for Business Decision Making: Techniques and Trends," Journal of Business Analytics, vol. 4, no. 1, pp. 45-52, 2022.

[5]. J. Lee, H. Kim and S. Park, "A Review of Data Mining and Machine Learning Algorithms for Data Analysis in Business Applications," International Journal of Business Intelligence and Data Mining, vol. 10, no. 2, pp. 133-142, 2019.

[6]. P. Sharma, A. Agarwal and S. Rawat, "Data-driven Decision Making for Smart Cities: A Comprehensive Review of Methods and Applications," Journal of Urban Data Analysis, vol. 7, no. 6, pp. 211-220, 2020.

[7]. N. Yadav, K. Singh and R. Srivastava, "Big Data Analytics for Industrial Applications: A Review," Journal of Industrial Data Science, vol. 15, no. 2, pp. 35-40, 2021.

[8]. A. Roy, M. Banerjee and S. Choudhury, "Sentiment Analysis Using Machine Learning Algorithms: A Survey and Comparative Study," International Journal of Data Mining and Machine Learning, vol. 3, no. 9, pp. 78-85, 2018.

[9]. S. Kumar, P. Singh and R. Tiwari, "Data Analysis for Predictive Analytics: Challenges and Approaches," Journal of Predictive Data Science, vol. 4, no. 8, pp. 110-118, 2020.

[10]. K. Verma, N. Agarwal and P. K. Mishra, "The Role of Data Analytics in Supply Chain Management: A Review," International Journal of Supply Chain Analytics, vol. 11, no. 1, pp. 24-32, 2022.