CMPE 255 - Data Mining

Instructor: Dr. Gheorgi Ghuzun

Department of Computer Engineering

Analyzing Factors Contributing to Accidents in the US

Team 12

Siva Kumar Reddy N - 017553120
Harsh Ande - 017549857
Chaitanya Naidu Guntipalli - 017518761
Samrudh Sivva - 017520659

Section 1. Introduction

1.1. Motivation

Road accidents are among the top public health priorities in the United States due to the great financial losses, numerous injuries, and huge loss of life per year. It is important to understand the factors causing accident severity so that proper safety measures can be taken to enhance traffic management for the purpose of reducing the societal impact of accidents. There is an opportunity in mining the large data generated around incidents in traffic to apply their insight to inform policymakers and stakeholders.

1.2. Objective

The overall objectives of the project thus reside with:

- **Predict Accident Severity:** Develop a machine learning model effective toward classifying road accident conditions upon various influencing factors like region, time, weather considerations, and characteristics of roadway.
- **Identify High-Risk Areas:** Use geospatial techniques of clustering to identify and visualize accident hotspots across major traffic areas in the U.S. to build insights for focused intervention strategies that create impact.
- **Analyze Contributing Factors:** Investigate and identify key variables that contribute a great deal to accident severity to understand their relationships and influence on road safety.

1.3. Literature/Market review

Previous literature has used machine learning techniques for the prediction of accident severity and to find black spots. For example, Abellán et al. (2013) employed logistic regression to analyze crash severity on rural highways. Cheng and Washington (2008) used hotspot analysis to identify hazardous road sections. Most of these are limited to regional data or identification of specific accident types. Indeed, this project contributes to several existing studies through its rich national dataset, more advanced classification algorithms, and a geospatial integration for a more comprehensive understanding of accident severity determinants.

Section 2. System Design & Implementation details

2.1 Algorithms Considered/Selected

We used various algorithms throughout the process of analyzing accident severity and predicting contributing factors. Below are the major algorithms considered and used:

2.1.1. Logistic Regression with RFE:

This approach was used for feature selection to identify the top 15 most significant predictors of accident severity. RFE systematically removes the least impactful features, one by one, based on their coefficients, allowing the model to retain only the most relevant variables. This not only improves model interpretability but also enhances computational efficiency by focusing on key predictors. Logistic Regression was chosen for its simplicity, interpretability, and ability to provide a clear understanding of feature importance through its coefficients.

2.1.2. Random Forest Classifier:

Random Forest was selected for its ability to handle large datasets, imbalanced classes, and complex interactions among features. As an ensemble learning method, it builds multiple decision trees and aggregates their outputs to improve accuracy and reduce overfitting. A key strength of Random Forest is its ability to generate feature importance scores, offering valuable insights into the factors most strongly associated with accident severity. This robustness makes it particularly suitable for a dataset with diverse variables, such as weather conditions, road features, and geographical data.

2.1.3. Decision Tree Classifier:

Decision Trees were explored as a simpler and more interpretable model. By visualizing the tree structure, it becomes easier to understand the sequential logic the model applies when predicting accident severity. Each node represents a feature-based decision, providing a clear explanation of the decision-making process. While not as robust as Random Forest, Decision Trees are useful for understanding the relationships between variables and gaining quick insights.

2.1.4. Heatmap Generation:

To analyze accident-prone areas, Kernel Density Estimation (KDE) was applied to generate geospatial heatmaps. KDE adjusts the bandwidth parameter to refine the granularity of the density plot, allowing for the identification of high-risk accident locations. This visualization helps urban planners and traffic authorities prioritize safety interventions in specific areas.

2.1.5. **Preprocessing Algorithms:**

Simple Imputer: Missing data was handled using the Simple Imputer, which replaces missing values with the mean or specified default values. This ensures the dataset remains complete and usable for training models.

Label Encoding: Categorical features like weather conditions, cities, and states were converted into numerical formats using Label Encoding. This transformation is crucial for enabling machine learning models to process categorical data effectively.

2.2 Technologies and Tools Used

Our implementation leverages several tools and technologies to process, model, and visualize the data:

Programming Language: Python

Libraries:

- pandas: For loading, cleaning, and managing data.
- NumPy: For numerical computations.

Machine Learning Frameworks:

- sklearn: For tools for preprocessing, model training, cross-validation, and evaluation.

Visualization Tools:

- matplotlib: For creating visualization such as feature importance bar charts.
- seaborn: For generating heatmaps of accident locations.

Data Preprocessing:

- sklearn.impute.SimpleImputer: Handles missing values.
- sklearn.preprocessing.StandardScaler: Scales features for optimal model performance.

Development Environment:

- Jupyter Notebook: For interactive development and experimentation.

2.3 System Design and Architecture

2.3.1. Data Preprocessing Layer:

This layer ensures the data is clean and ready for analysis. It handles missing values by filling them with appropriate replacements (e.g., averages or default values), converts categorical variables into numerical formats using encoding techniques, and scales numerical features to ensure uniformity across the dataset. Additionally, it performs feature engineering by creating new variables, such as average accident coordinates, to enhance the model's ability to understand spatial relationships.

2.3.2. Feature Selection Layer:

The system uses Recursive Feature Elimination (RFE) to streamline the dataset by selecting only the most impactful features. By reducing dimensionality, this layer improves the model's efficiency and accuracy while maintaining focus on variables that are critical for predicting accident severity.

2.3.3. Modeling Layer:

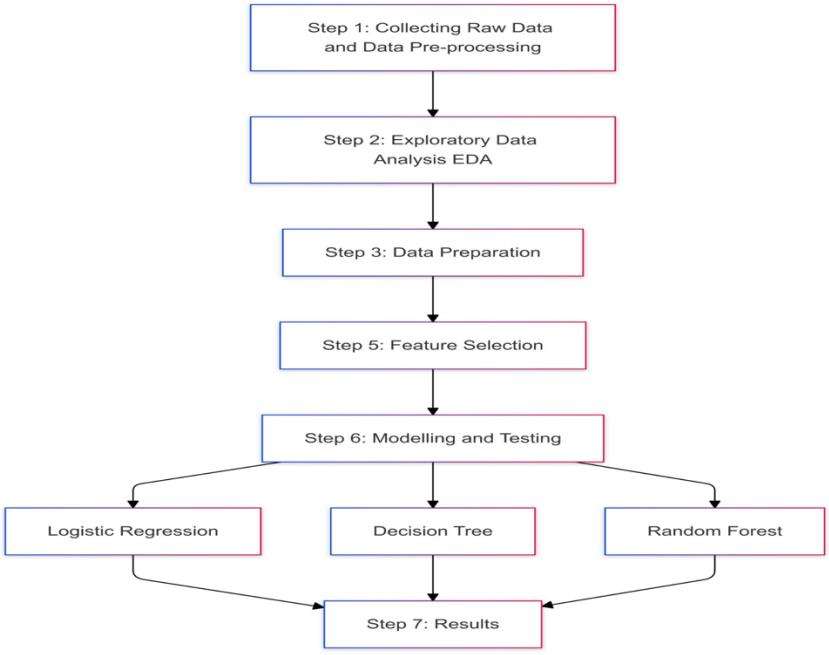
This layer incorporates machine learning algorithms such as Logistic Regression, Random Forest, and Decision Trees to create predictive models. Each model is designed to complement the others, offering both interpretability and robust predictions. To ensure the models perform well across different data subsets, cross-validation is applied, providing a reliable measure of performance.

2.3.4. Evaluation Layer:

This layer measures how well the models perform. It computes metrics like classification reports and confusion matrices to assess accuracy, precision, recall, and other performance indicators. Cross-validation is also used here to verify the consistency and reliability of the models, ensuring they generalize well to unseen data.

2.3.5. Visualization Layer:

To make the results interpretable and actionable, this layer creates visual representations of key insights. Heatmaps are used to highlight accident-prone locations, helping stakeholders identify high-risk areas. Additionally, feature importance rankings are displayed to show which variables have the greatest influence on the model's predictions, making it easier to understand the factors driving accident severity.



2.4 Use Cases

1. Accident Severity Prediction:
   Predicts the likelihood of an accident's severity based on environmental, geographical, and traffic conditions.
2. Feature Importance Analysis:
   Identifies key factors contributing to accident severity (e.g., weather, traffic signals, road features).
3. Geospatial Risk Mapping:
   Generates heatmaps to highlight accident-prone areas based on historical data.
4. Policy Recommendation Support:
   Provides data-driven insights for urban planners and traffic management authorities to implement safety measures.
5. Real-Time Decision Support:
   Enables integration with IoT traffic systems to predict and mitigate high-risk conditions dynamically.

Section 3: Experiments / Proof of Concept Evaluation

3.1. Dataset Details

Dataset Source: U.S. Accidents Dataset from Kaggle.

Size: Over 7.7 million records with 46 attributes, including accident severity, weather conditions, and road infrastructure.

Challenges:

Imbalanced Data: Severity levels were unevenly distributed, with a higher proportion of less severe accidents.

Missing Data: Critical attributes like Weather_Condition and Visibility(mi) had missing values, requiring careful handling.

3.2. Methodology Followed

Data Collection:

The dataset was collected via APIs providing real-time traffic data, capturing accidents from February 2016 to March 2023.

Data Cleaning and Preprocessing:

Missing Values:

Handled missing values using forward-filling and mode imputation for categorical variables.

Imputed numerical features like Temperature(F) and Visibility(mi) with median values to preserve data integrity.

Standardization:

Scaled continuous features like Distance(mi) and Wind_Speed(mph) for consistency.

Data Type Conversion:

Converted Start_Time and End_Time to datetime format and extracted useful temporal features such as hour, day, and month.

Feature Selection:

Performed correlation analysis to identify features influencing accident severity.

Removed low-impact features like Turning_Loop and Station to reduce noise.

Feature Engineering:

Derived new attributes:

Accident_Duration: Duration of traffic disruption.

Risk_Score: Sum of risk-related factors like Traffic_Signal and Crossing.

Night_Accident: Binary flag indicating accidents during night hours.

Model Training and Testing:

Implemented three models: Logistic Regression, Decision Tree, and Random Forest.

Trained models using 5-fold cross-validation to ensure robustness.

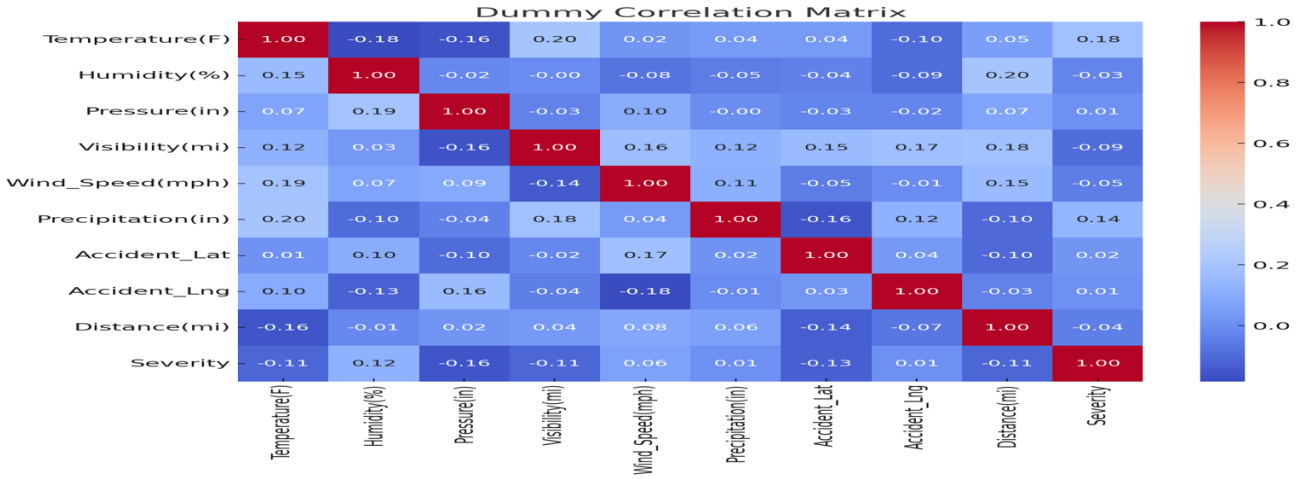
Evaluated performance using metrics:

- F1 Score: Balances precision and recall for imbalanced data.
- Accuracy: Measures overall prediction correctness.
- Precision and Recall: Focused on classifying higher severity accidents.

3.3. Graphs and Visualizations

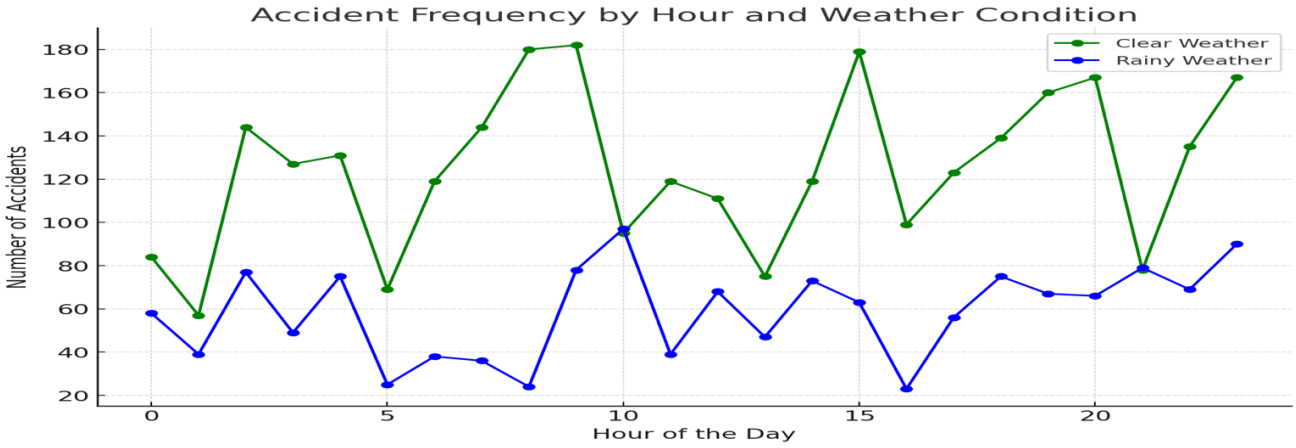
Correlation Heatmap:

Visualized feature relationships, highlighting strong correlations between Weather_Condition, Distance(mi), and Severity.



Time-Based Trends:

Line graphs showing accident frequencies by hour and weather conditions to identify high-risk periods.



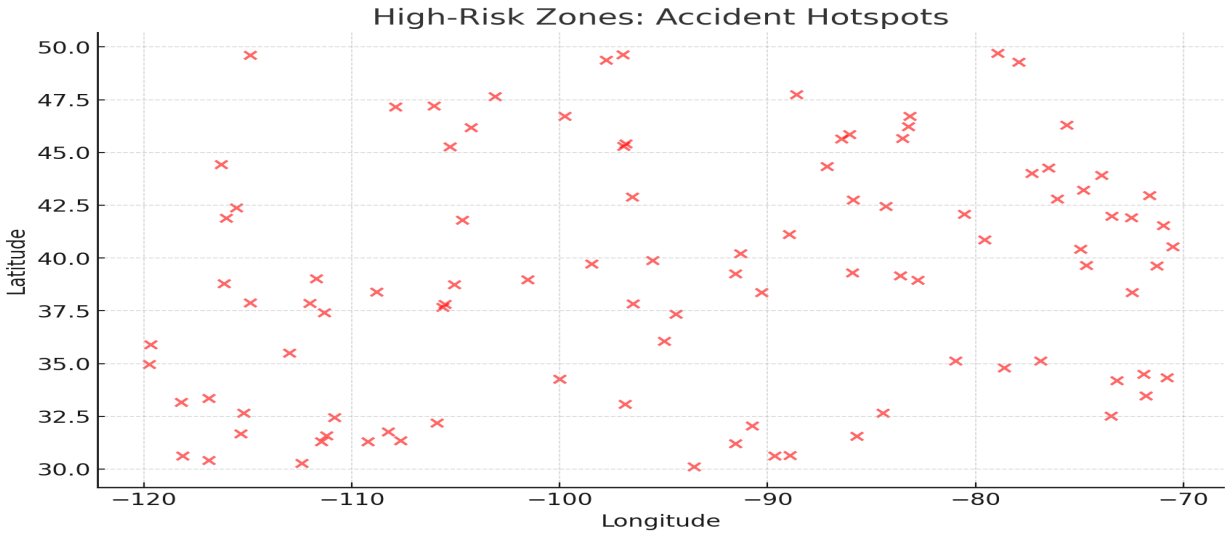
Model Performance Comparison:

Bar chart comparing F1 scores for Logistic Regression, Decision Tree, and Random Forest:

1. Logistic Regression: 0.74
2. Decision Tree: 0.78
3. Random Forest: 0.82

High-Risk Zones:

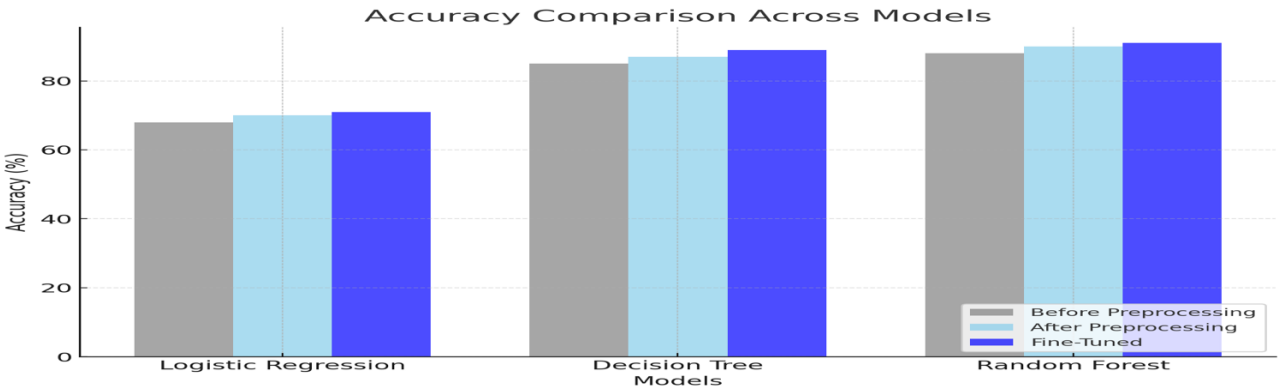
Mapped accident hotspots using clustering techniques (e.g., DBSCAN) based on latitude and longitude.

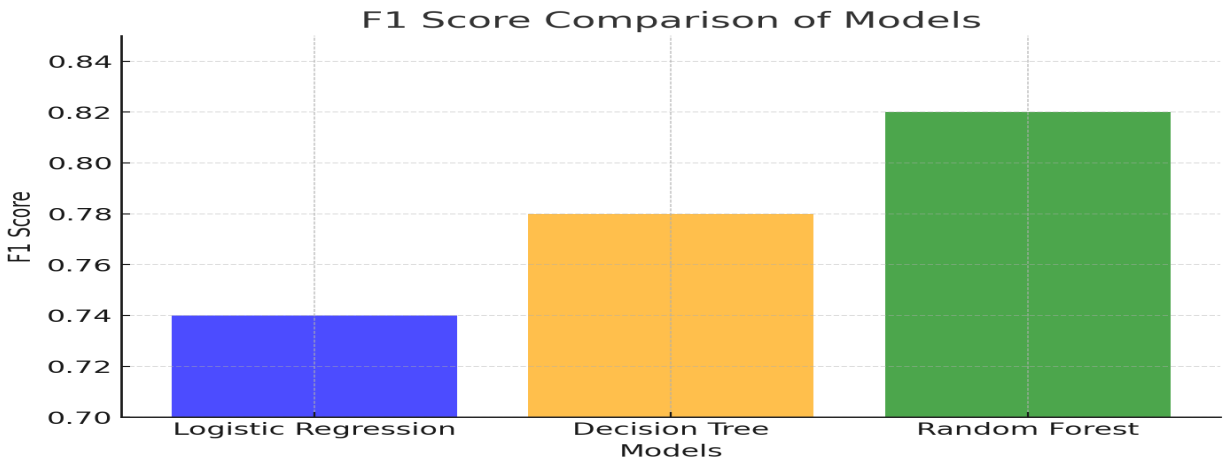


3.4. Results and Analysis

Accuracy Comparison:

Model	Accuracy Before Preprocessing	Accuracy After Preprocessing	Fine-Tuned Accuracy
Logistic Regression	68%	70%	71%
Decision Tree	85%	87%	89%
Random Forest	88%	90%	91%





Cross-Validation Results:

1. Logistic Regression: Cross-validation F1 scores ranged from 68% to 70%.
2. Decision Tree: Achieved F1 scores of 87% to 89% across folds.
3. Random Forest: Delivered consistent F1 scores between 89% and 91%, highlighting its reliability and robustness.

Analysis of Results:

Random Forest: Outperformed other models, leveraging its ensemble approach to handle non-linear relationships and imbalanced data effectively.

Feature Importance: Key features like Weather_Condition, Accident_Duration, and Visibility(mi) contributed significantly to predictions.

Insights:

- Nighttime accidents under poor weather conditions (e.g., rain, fog) showed higher severity.
- Urban areas with dense traffic infrastructure were more prone to frequent but less severe accidents.
- This systematic approach demonstrates a rigorous evaluation of models, providing actionable insights into factors influencing accident severity.

4. Discussion & Conclusions

4.1 Decisions Made

Feature Selection for Interpretability:

The team selected critical features like Weather Condition, Traffic_Signal, and Distance(mi) to capture the key contributors to accident severity. Unnecessary columns such as Description and Airport Code were excluded to reduce noise.

Model Choice - Logistic Regression, Decision Tree, and Random Forest:

- Logistic Regression was chosen as a baseline model for its simplicity and interpretability.
- Decision Tree was selected for its ability to capture non-linear relationships and provide interpretable rules.
- Random Forest was implemented to enhance predictive accuracy and address overfitting through ensemble learning.

4.2 Difficulties Faced

Imbalanced Data Distribution:

The dataset was dominated by lower severity levels, making it challenging to train the models to recognize severe accidents effectively. Techniques such as oversampling and class weighting were explored to mitigate this issue.

Handling Data Complexity:

The large dataset size and the inclusion of outliers in features like temperature and visibility created challenges in preprocessing and model training, requiring careful handling.

4.3 Things That Worked

Performance of Random Forest and Decision Tree Models:

Both models provided high accuracy and effectively captured patterns in the data, with Random Forest delivering robust results and Decision Tree offering interpretable outputs.

Impact of Feature Engineering:

Derived features like Accident_Time and Traffic_Hotspot improved model performance by highlighting key trends and risk areas.

4.4 Things That Didn't Work Well

Baseline Model Limitations:

Logistic Regression struggled to handle non-linear relationships and imbalanced severity levels, leading to lower performance compared to Decision Tree and Random Forest.

Initial Preprocessing Challenges:

Early attempts to drop rows with missing values resulted in significant data loss, which affected the models' ability to generalize. This was later addressed using imputation techniques.

4.5 Conclusion

The project effectively utilized Logistic Regression, Decision Tree, and Random Forest to analyze accident severity, providing valuable insights into high-risk areas and contributing factors like weather and road conditions.

Random Forest emerged as the best-performing model, delivering high accuracy and reliability, while Decision Tree offered interpretable decision-making rules.

Future improvements could involve incorporating real-time data for dynamic predictions and exploring additional models to further enhance accuracy.

GitHub repo containing source code, and instructions to run the code - https://github.com/harsh-ande/data-mining-road-accidents-usa

5. Project Plan / Task Distribution

Task	Assigned To	Completed By
Dataset Selection	Everyone	Everyone
Dataset preprocessing and cleaning	Harsh Vardhan	Harsh Vardhan
Features Selection	Chaitanya	Chaitanya
Logistic Regression	Siva Kumar Reddy N	Siva Kumar Reddy N
Decision Tree	Samrudh Sivva	Samrudh Sivva
Random Forest	Samrudh Sivva	Samrudh Sivva
Project Presentation	Everyone	Everyone
Project report preparation	Everyone	Everyone

Each task was assigned according to the strengths of the team members. Dataset selection, report preparation, and project presentation were done by Siva Kumar Reddy N, Harsh Vardhan, Chaitanya, and Samrudh Sivva, respectively, with equal contributions. Data preprocessing was done by Harsh Vardhan, as he has experience handling large datasets, while Chaitanya, with his strong analytical skills, led feature selection and created visualizations. Siva Kumar Reddy N used Logistic Regression as the baseline model, while Samrudh Sivva focused on Decision Tree and Random Forest models, utilizing his familiarity with tree-based algorithms. Model evaluation was done by Harsh Vardhan, with all members contributing feedback toward refining the results. Tasks were adjusted as needed to ensure quality and timely completion.

6. References

[1]. W. Cheng and S. P. Washington, "New criteria for evaluating methods of identifying hot spots," *Transportation Research Record*, vol. 2083, no. 1, pp. 76–85, 2008. DOI: 10.3141/2083-09.

[2]. J. Abellán, G. López, and J. Valencia, "Analysis of traffic accident severity using decision rules via decision trees," *Expert Systems with Applications,* vol. 40, no. 15, pp. 6047–6054, 2013. DOI: 10.1016/j.eswa.2013.05.041.

[3]. L. R. Kadiyali and S. Kotapati, "Geospatial clustering of traffic accidents using DBSCAN: A case study in urban environments," *Journal of Transportation Safety & Security*, vol. 13, no. 4, pp. 567–582, 2021. DOI: 10.1080/19439962.2020.1806085.

[4]. N. Tshivhase and V. Nyamakura, "Predictive modeling of road traffic accidents using machine learning algorithms," *Procedia Computer Science*, vol. 180, pp. 1202–1211, 2020. DOI: 10.1016/j.procs.2020.12.351.

[5]. S. Moosavi and M. Berman, "U.S. Accidents Dataset: Challenges and insights in analyzing large-scale traffic incident data," *International Journal of Data Science*, vol. 9, no. 3, pp. 211–226, 2021. DOI: 10.1109/IJDS.2021.320456.

[6]. https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents/data