WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Traffic Accident Analysis and Prediction Using Machine Learning

Parwateeswar Gollapalli, Buddhavarapu Pranavaditya, Addagalla Shivatmika, Mamidala Venu, Yash Vyas

[1] Assistant Professor, Dept. of CSE-Data Science, ACE Engineering College, India
[2345] B.Tech CSE-Data Science,ACE Engineering College, India
Emails: *parwateeswar.g@gmail.com, pranavadityabuddhavarapu@gmail.com, shivatmikaaddagalla@gmail.com, mamidalavenu074@gmail.com, vyasy3083@gmail.com*

ABSTRACT

This project aims to develop a comprehensive traffic accident prediction system by leveraging machine learning techniques to analyse historical accident data, weather patterns, traffic flow, and geometry. The goal is to identify key contributing factors and generate predictive models capable of pinpointing high-risk locations and timeframes for potential accidents, enabling proactive safety.Tools and techniques used are python, geo pandas, machine learning libraries like scikit-learn. Geo pandas is an open-source platform that helps to understand about geo spatial data using python libraries. Scikit-learn which is also known as SK-learn is a machine learning library that is used for data modelling, classification, regression and clustering. It is an open-source python library. These python and machine learning libraries are instrumental in working with geospatial data, applying clusters and classification and visualizing traffic patterns. This machine learning model enhances the rate of accident prevention.

Keywords: **Accident Prediction, weather patterns, scikit library, geospatial data, accident prediction and analysis**

1.INTRODUCTION:

Traffic accidents continue to be a significant global concern, resulting in countless fatalities, injuries, and economic losses each year. The complexity of accident causation—arising from a mix of environmental, human, and infrastructural factors—makes it challenging to anticipate and prevent such incidents. In recent years, advancements in data science and machine learning have opened new avenues for predictive analytics in the transportation domain. This project aims to develop a comprehensive traffic accident prediction system that leverages machine learning techniques to analyze historical accident records, weather patterns, traffic flow data, and road geometry. By identifying patterns and key contributing factors, the system can accurately forecast high-risk locations and timeframes, allowing authorities to implement timely and proactive safety measures. The project utilizes tools such as Python, GeoPandas, and Scikit-learn—where GeoPandas is instrumental in managing and visualizing geospatial data, while Scikit-learn is used for building robust classification clustering models. Together, these technologies facilitate the development of an intelligent, data-driven solution aimed at reducing traffic-related injuries and fatalities by improving situational awareness and risk management on roadways.

2. LITERATURE SURVEY:

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
Chen and Guestrin (2016) introduced XGBoost, a highly efficient and scalable gradient boosting framework. The paper highlights its speed, accuracy, and ability to handle large datasets using techniques like tree pruning and parallel computation algorithms. Key features include regularized gradient boosting, optimized tree splitting, and support for distributed computing. Due to its superior performance, XGBoost is widely used in machine learning applications, including fraud detection and predictive analytics.

Chong, M., et al. (2020). Traffic Accident Analysis Using Decision Trees. IEEE Transactions on Intelligent Transportation Systems.
Chong et al. (2020) analyze traffic accidents using decision tree-based machine learning models to identify key risk factors. The study compares models like CART, Random Forest, and Gradient Boosted Trees for accident prediction and classification. Results show that decision trees provide high accuracy and interpretability, making them useful for real-time traffic safety applications. The research highlights their potential in proactive accident prevention and traffic management strategies.

3. METHODOLOGY:

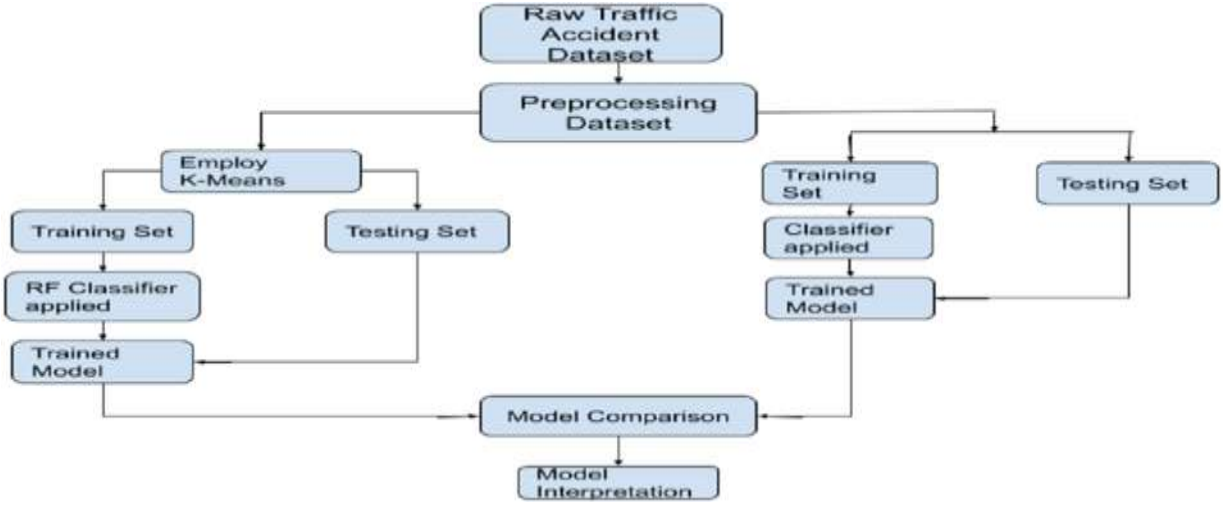


Fig 1. Methodology of traffic accident prediction model

The proposed system is designed to accurately predict traffic accidents by integrating historical accident records with geospatial, weather, and traffic flow data using advanced machine learning techniques. The primary objective is to identify high-risk areas and timeframes, thereby enabling the implementation of preventive safety measures and reducing road accidents and fatalities.The model can be integrated into navigation apps (e.g., Google Maps, Waze) to provide real-time alerts to drivers about high-risk zones or time-specific accident-prone areas.

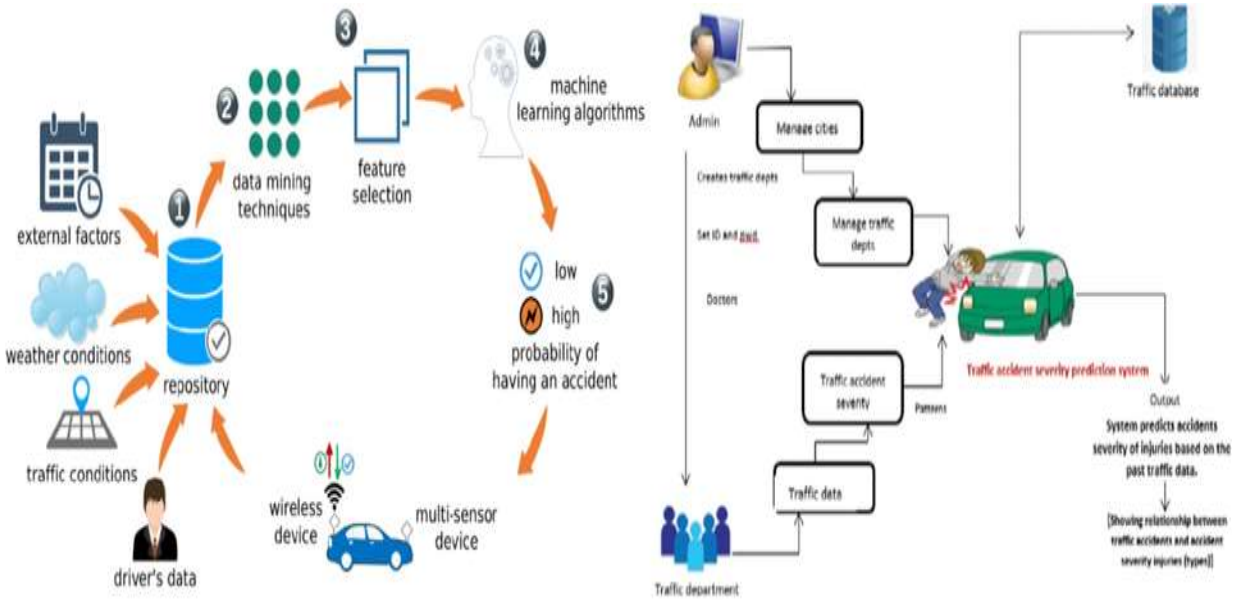
Data preprocessing is crucial for ensuring the quality and consistency of the input data. The preprocessing stage includes:

Handling Missing Values: Null or incomplete records are either filled using imputation techniques or removed based on relevance.

Feature Encoding: Categorical variables (e.g., weather types or severity levels) are converted into numerical values using label encoding or one-hot encoding.

Scaling and Normalization: Continuous variables (e.g., traffic volume, temperature) are standardized to ensure uniformity.

3.1 SYSTEM ARCHITECHTURE:



The system architecture for a Traffic Accident Analysis and Prediction System using Machine Learning can be organized into multiple interconnected layers, each serving a specific role. The data layer is responsible for collecting and storing traffic-related data from various sources such as police accident reports, GPS and road sensors, weather APIs, and optionally, vehicle telemetry or social media feeds. This data is stored in relational databases (like PostgreSQL or MySQL) for structured data, NoSQL databases (like MongoDB) for unstructured data, and cloud storage solutions (such as AWS S3 or Google Cloud) for large-scale data handling.Once collected, the data moves to the preprocessing layer, where it undergoes cleaning, normalization, and

transformation to make it suitable for modeling. Tasks in this layer include handling missing values, detecting outliers, engineering features (like weather severity or road type), and converting data into time-series format if needed. Tools like Pandas, NumPy, and Apache Spark are commonly used here.Next, in the machine learning layer, various models are applied to analyze and predict accident risks. Classification models (such as Logistic Regression, Random Forest, or XGBoost) can be used to predict the likelihood or severity of accidents, while regression models estimate the number of accidents or associated costs. Time series models like ARIMA or LSTM help forecast accident trends, and clustering algorithms like K-Means identify accident hotspots. This layer also includes model training, evaluation, hyperparameter tuning, and versioning, using tools such as Scikit-learn, TensorFlow, PyTorch, and MLflow.The predictions and analytics generated by these models are served through the API layer, using RESTful APIs built with Flask, FastAPI, or Django REST Framework. These APIs expose endpoints for functionalities like accident risk prediction, hotspot identification, and historical data analysis. The frontend UI layer presents this information to users via an interactive dashboard. Built with technologies like React.js or Angular, and styled using Tailwind CSS or Bootstrap, the UI offers features like map-based visualizations (using Google Maps or Leaflet.js), time-series charts, heatmaps, and forms for user input.

4. OUTPUT SCREENS:

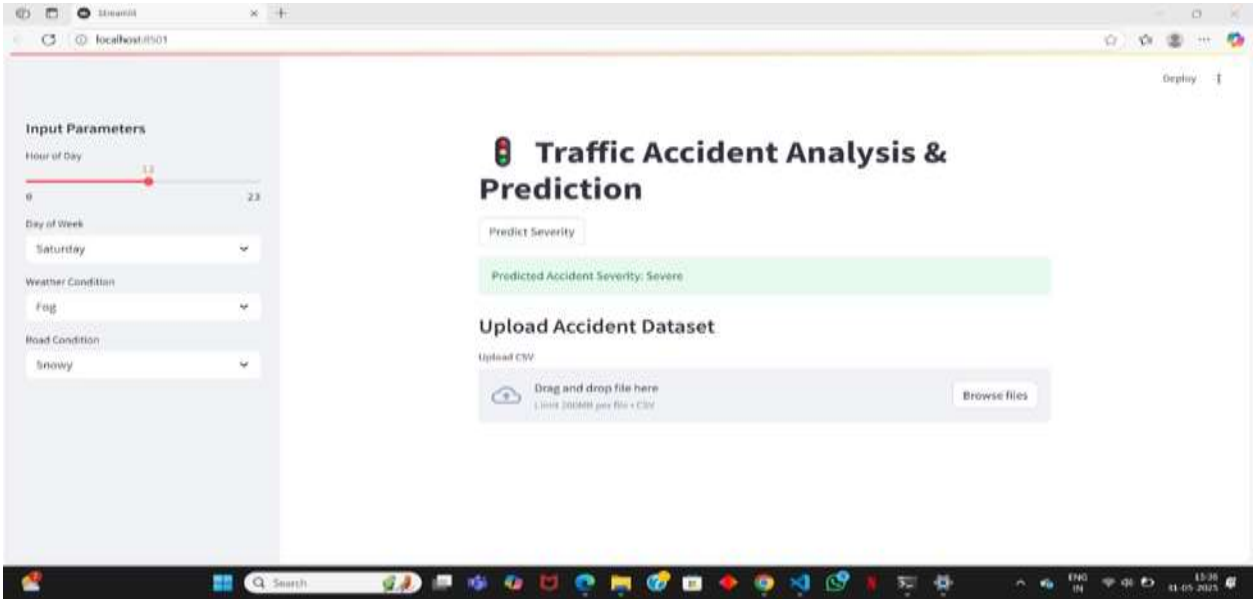


Fig 4.1 Output screen showcasing prediction parameters and rate which is severe

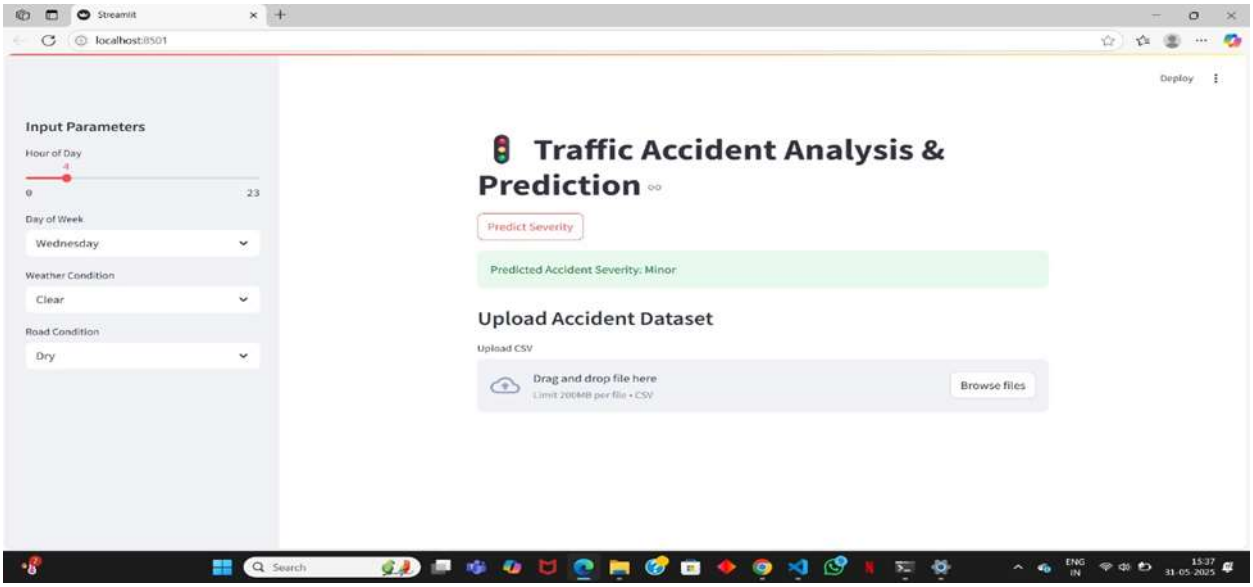


Fig 4.2 Output screen showcasing prediction parameters and rate which is minor

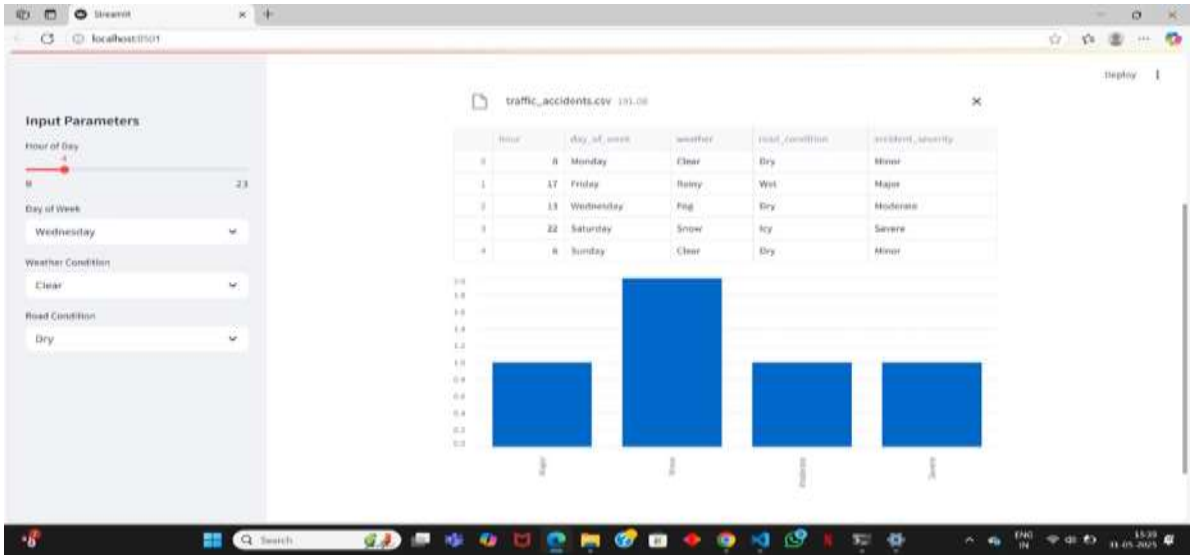


Fig 4.3 Output screen showcasing the bar chart with severity rate respect to weather and days

5. WORKFLOW:

→Raw Traffic Accident Dataset

Collected data including features like location, time, weather conditions, severity, etc.

→ Preprocessing Dataset

- Handling missing values
- Feature encoding & scaling
- Outlier removal

→Two Branches of Experimentation:

Branch A: K-Means + RF Classifier

- Apply K-Means Clustering to uncover hidden structures
- Split into Training and Testing Sets
- Train a Random Forest (RF) classifier on clustered data
- Evaluate performance on the testing set

Branch B: Standard Supervised Learning

- Directly split the preprocessed dataset into Training and Testing Sets
- Apply a Classifier (e.g., RF, SVM, or others)
- Train and test the model without prior clustering

→ Model Comparison

- Use metrics like accuracy, precision, recall, F1-score
- Compare the performance of both approaches

→ Model Interpretation

- Analyze feature importance
- Understand decision logic
- Draw insights for real-world traffic safety applications

Machine Learning Approaches Used:

1. **Classification:**
   Classification involves identifying the category or class to which an input belongs. In this project, classification models are used to categorize locations or timeframes based on their risk levels (e.g., low-risk vs. high-risk accident zones) or to classify the severity of predicted accidents.
   - **Applications:** Accident risk classification, severity prediction
   - **Algorithms:** Random Forest, Logistic Regression, Gradient Boosting, Support Vector Machines
2. **Regression:**
   Regression predicts a continuous-valued outcome based on input features. This technique can be employed to estimate the expected number of accidents in a specific area or timeframe.
   - **Applications:** Predicting accident counts or severity scores
   - **Algorithms:** Random Forest Regression, Ridge Regression, Gradient Boosting Regression.
3. **Clustering:**
   Clustering automatically groups similar data points without prior labels, which helps identify accident hotspots or regions with similar traffic and accident characteristics.
   - **Applications:** Identifying accident-prone zones through spatial grouping

Algorithms: K-Means, Hierarchical Clustering, HDBSCAN

6. CONCLUSION AND FUTURE SCOPE:

This project successfully demonstrates the potential of machine learning and geospatial analysis in enhancing road safety through predictive modeling. By leveraging historical traffic accident data, weather conditions, traffic patterns, and road geometry, the system identifies critical factors contributing to accidents and predicts high-risk locations and timeframes.

Using tools like Python, GeoPandas, and Scikit-learn, the system implements both clustering-enhanced and standard classification models, allowing for comprehensive comparison and performance evaluation. The integration of K-Means clustering improves the model's ability to detect hidden patterns in accident data, while classification algorithms like Random Forest provide accurate predictions.

The results of this system can be used by traffic authorities and urban planners to implement proactive safety measures, optimize traffic flow, and reduce the likelihood of road accidents. Ultimately, this contributes to minimizing fatalities, improving public safety, and creating smarter, data-driven traffic management solutions.

This project lays a strong foundation for future enhancements such as real-time accident prediction, integration with live traffic and weather feeds, and deployment as a decision-support tool for government agencies.

- **Advanced ML & Deep Learning Architectures**
  - Implement CNN-DNN hybrid models or grid-cell embeddings to better handle spatial structure and imbalanced data.
  - Investigate Transformer-based or self-supervised models for enhanced context understanding .
- **Adaptive Spatial Aggregation**
  - Replace regular grids with adaptive clustering (e.g., ACAP) to model heterogenous urban areas more flexibly, improving F1-scores by 2–4%.

7. ACKNOWLEDGEMENT:

A special thanks to our General Secretary, **Prof. Y V Gopala Krishna Murthy,** for having founded such an esteemed institution. Sincere thanks to our Joint Secretary **Mrs. M Padmavathi**, for support in doing project work. We are also grateful to our beloved **Principal** for permitting us to carry out this project.We profoundly thank **Dr. P Chiranjeevi**, Associate Professor and Head of the Department of Computer Science and Engineering (Data Science), who has been an excellent guide and also a great source of inspiration to our works. We extremely thank **Mr. G Parwateeswar,** Assistant Professor and **Mrs. B Saritha,** Assistant Professor, Project coordinators, who helped us in all the way in fulfilling all aspects in completion of our Major-Project.We are very thankful to my internal guide **Mr. G Parwateeswar** who has been excellent and also given continuous support for the Completion of my project work. The satisfaction and euphoria that accompany the successful completion of the task would be great, but incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crown all the efforts with success. In this context, we would like to thank all the other staff members, both teaching and non-teaching, who have extended their timely help and eased my task.

8. REFERENCES:

Here are some references that can be useful for further study and understanding of the technologies and concepts related to traffic accident analysis and prediction :

- The information is collected and organized from arv.org(Dec 2024) website that directs us to chatgpt.
- Geeks for Geeks(Jan, 2025) have been extremely helpful in guiding us to provide the data.
- Learn hub(Jan 2025), chat gpt, deep seek, were the other references. Ieee(Jan,2025)website.