Enhancing road safety with ai driven traffic accident analysis and prediction:-

**NAME:** SUBASHINI S

**REG.NO:** 422623106003

**INSTITUTION:** University College Of Engineering Panruti

**DEPARTMENT:** Electronics And Communication Engineering

**DATE OF SUBMISSION:** 08.05.2025

**Problem Statement:**

The complex, nonlinear interactions among the various elements impacting accident occurrence and severity are often not captured by traditional approaches for traffic accident analysis and forecasting. The majority of present road safety solutions are reactive rather than proactive, even with the increasing availability of large-scale traffic, environmental, and behavioural data. With over 1.19 million deaths and large financial losses each year, traffic accidents pose a serious threat to worldwide public health and society.

There is a critical need for advanced, AI-driven solutions that can:

* Using machine learning models trained on a variety of data sources, including temporal, environmental, infrastructural, and behavioural aspects, forecast the probability, frequency, and severity of traffic accidents.
* To facilitate focused, data-driven actions and resource allocation, identify high-risk areas, times, and contributing variables.
* Assist stakeholders and policymakers in converting model projections into practical safety regulations and policies by offering them understandable insights.The challenge lies in developing robust, interpretable AI models that not only achieve high predictive accuracy but also uncover actionable patterns and causal relationships within complex traffic accident data, ultimately enabling proactive strategies to reduce accident rates and enhance road safety for all users.

**Objectives of the Project:**

1. **Predict Accident Risk and Severity**

Using a variety of data sources, including past accident records, weather, traffic flow, and road infrastructure, create machine learning models (such as Random Forest and ensemble methods) to predict the probability, frequency, and severity of traffic accidents. These models seek to identify high-risk areas and times with high accuracy (e.g., >85% precision).

1. **Determine the Contributing Elements:**Examine environmental, temporal, and spatial factors (such as weather, road layout, and driver conduct) to identify the underlying causes of collisions and measure how these affect risk levels. For instance, give priority to elements like poor visibility or abrupt turns in regions that are prone to accidents.
2. **Facilitate Proactive Interventions:**

Develop real-time prediction systems to notify drivers and authorities of new hazards, enabling the prompt rerouting of traffic or the deployment of emergency services. Connect with navigation app and traffic management agency platforms.

1. **Optimise Resource Allocation:**

Give legislators practical advice on how to best deploy funds and prioritise infrastructure improvements (such as putting up signs in high-risk areas). Reduced accident rates (e.g., 20% decrease in deaths after 2 years) are a good way to gauge progress.

1. **Improve Explainability and Transparency:**

Use interpretable AI methods to help stakeholders understand model choices, fostering confidence and enabling evidence-based policy reforms.

**Flowchart of the Project Workflow:**

Data Collection

↓

Data Preprocessing & Cleaning

↓

Exploratory Data Analysis (EDA)

↓

Model Selection & Training

↓

Model Evaluation & Validation

↓

Deployment & Integration

↓

Real-Time Prediction & Alerts

↓

Decision Support & Preventive Actions

↓

Continuous Learning & System Improvement

**Data Description:**

# Dataset name and origin

# Global Traffic Accidents Dataset (Kaggle):

The Global Traffic Accidents Dataset (Kaggle) offers a variety of accident scenarios for model training and analysis, including 10,000 traffic accident records gathered from different locations across the globe.

* **Traffic Accident Detection Video Dataset for AI-Driven Computer Vision Systems:**

Developed by Victor Adewopo, this dataset includes approximately 5,700 video clips from traffic surveillance and dash cameras, categorized into multiple accident types and normal traffic conditions. It is designed to support real-time accident detection in smart city environments and is publicly available for research.

* **US Accidents Dataset (Kaggle**):

A large-scale dataset covering 49 US states with around 1.5 million records of traffic accidents collected from February 2016 to March 2023. It integrates data from state transportation departments and other official sources.

* **ML4RoadSafety Dataset:**

A repository containing about 9 million accident records from 8 US states, collected from publicly available sources. It supports graph neural network modeling for accident analysis.

* **Car Crash Dataset (CCD):**

Comprises real dashcam videos of traffic accidents, annotated for various environmental and accident conditions. It is used mainly for computer vision and autonomous vehicle safety research.

# Type of data:

* Tabular Data
* Video Data
* Textual Data
* Environmental and Contextual Data
* Sensor and IoT Data (in some datasets)

# Number of records and features.

* **Global Traffic Accidents Dataset:**

Contains 10,000 records of traffic accident reports from various locations worldwide, providing insights into accident patterns and causes.

* **Global Road Accidents Dataset:**

A large dataset with 132,000 records (rows) and 30 diverse features. It includes detailed attributes such as accident severity, weather conditions, vehicle involvement, driver characteristics, geographic regions, and more. This dataset is ideal for exploratory data analysis, feature engineering, and machine learning to predict accident severity and other outcomes.

* **Road Traffic Accidents Dataset:**

Contains 12,316 instances with 32 features, including accident details and environmental factors, after data encoding and cleaning.

These datasets collectively offer a rich set of features capturing temporal, environmental, vehicular, and human factors critical for accurate accident analysis and prediction.

# Static or dynamic dataset.

* **Static Data:**

Road infrastructure details (road width, terrain, and surface type), geographic and policy-level meta-features, and environmental context are examples of static data. These are fixed attributes that don't change often. These variables offer a consistent foundation for comprehending the causes and hazards of accidents.

* **Dynamic Data:**

Traffic flow, lane occupancy, average vehicle speed, visibility, weather, and incident occurrences recorded with temporal granularity (e.g., hourly or daily) are examples of time-series data that make up dynamic data. For the purposes of real-time prediction and post-event analysis, this data records the changing traffic circumstances and incident dynamics.

# Target variable

he target variable in AI-driven traffic accident analysis and prediction projects is primarily focused on accident severity or fatality outcome. This variable represents the level of injury or damage resulting from a traffic accident and is used as the main prediction goal for machine learning models. Common forms of the target variable include

**Accident Severity Levels:** Categorized into classes such as minor, moderate, severe, or fatal injuries.

**Binary Fatality Outcome:** Fatal vs. non-fatal accidents, predicting whether an accident results in death or not.

**Casualty Class:** Differentiating types of casualties (drivers, passengers, pedestrians) and their injury severity.

**Data Preprocessing:**

# Handle missing values :

**Decision Tree-Based Imputation with Correlation Sampling:** This method builds decision trees (e.g., using C4.5 algorithm) on complete records to predict missing values in incomplete records by exploiting attribute correlations within subsets of data. It has been shown to outperform traditional imputation methods on large traffic accident datasets, especially when many attributes are categorical.

**Fuzzy C-Means Clustering with Genetic Algorithm:** Missing traffic data values are imputed by clustering similar days at the same station using fuzzy c-means, then minimizing imputation errors with a genetic algorithm. This approach is effective for temporal traffic data imputation.

**Spatiotemporal Imputation:** Combines temporal imputation followed by spatial imputation to fill missing traffic data by leveraging both time-series and spatial correlations, improving accuracy in traffic datasets.

# Remove or justify duplicate records:

* Identify duplicates based on key features
* Use SQL queries or data processing tools
* Data processing libraries

# Detect and treat outliers.

**Detect outliers:**

* Exploratory Data Analysis (EDA)
* Statistical Methods:
* Domain-Specific Thresholds
* Machine Learning-Based Detection

**treat outliers:**

* Data Cleaning
* Winsorization or Trimming
* Weight Adjustment

# Encode categorical variables

* One-Hot Encoding
* Dummy Encoding
* Effect Encoding
* Binary Encoding
* Other Techniques

# Convert data types and ensure consistency.

* Data Collection from Diverse Sources
* Standardizing Numeric and Categorical Data
* Uniform Date-Time Formatting
* Handling Missing and Inconsistent Data
* Automated Data Preprocessing Pipelines

**Exploratory Data Analysis (EDA):**

* **Visualizing Accident Distributions**: Using histograms, bar charts, and heatmaps to explore the frequency of accidents by location, time, weather conditions, and severity[2](https://www.tandfonline.com/doi/full/10.1080/23311916.2020.1834659)[3](https://rpubs.com/hossein_glm/traffic_accident_eda)[4](https://github.com/shubamsumbria/us-accidents-analysis).
* **Identifying Contributing Factors:** Analyzing correlations between variables such as road type, traffic volume, weather, and accident severity to identify important predictors[1](https://www.sciencedirect.com/science/article/pii/S2590198223000611)[8](https://www.mdpi.com/2071-1050/15/7/5939).
* **Handling Missing and Outlier Data:** Detecting anomalies or missing values that could bias analysis, and deciding on treatment methods to ensure data quality[3](https://rpubs.com/hossein_glm/traffic_accident_eda).
* **Temporal and Spatial Analysis:** Examining accident trends over time and mapping hotspots using geospatial techniques to reveal high-risk areas and temporal patterns like peak accident hours or seasonal effects[2](https://www.tandfonline.com/doi/full/10.1080/23311916.2020.1834659)[4](https://github.com/shubamsumbria/us-accidents-analysis)[5](https://onlinelibrary.wiley.com/doi/10.1155/2023/6643412).
* **Feature Exploration for Model Building:** Assessing variable distributions and relationships to select meaningful features for machine learning models predicting accident occurrence or severity

**Feature Engineering:**

**1.**Feature Identification and Selection

* Selecting key factors that influence accident occurrence and severity is crucial. Studies have identified features such as weather conditions, road type, traffic volume, time of day, vehicle type, and driver behavior as significant predictors[1](https://www.mdpi.com/2624-8921/7/2/38)[2](https://www.sciencedirect.com/science/article/pii/S2590198223000611)[4](https://www.civilejournal.org/index.php/cej/article/view/5812).

**2. Feature Creation and Transformation**

* Creating new features by combining or transforming existing variables can capture complex relationships. For example, temporal features like “rush hour” or “weekend vs. weekday,” spatial features such as proximity to intersections or high-risk zones, and weather condition indicators enhance predictive power[3](https://journals.sagepub.com/doi/10.1177/03611981231217497?icid=int.sj-full-text.similar-articles.5)[5](https://ijsrset.com/index.php/home/article/view/IJSRSET2411446).

**3. Encoding and Normalization**

* Categorical variables (e.g., road type, weather) are encoded into numerical formats using techniques like one-hot encoding or label encoding to make them compatible with machine learning algorithms[5](https://ijsrset.com/index.php/home/article/view/IJSRSET2411446).

**4. Handling Imbalanced Data**

* Accident datasets often have imbalanced classes (e.g., fewer severe accidents). Feature engineering combined with techniques like SMOTE (Synthetic Minority Over-sampling Technique) helps balance the data for better model training.

**5. Use of Advanced Models for Feature Extraction**

* Deep learning models such as CNNs, LSTMs, and hybrid architectures automatically extract complex features from multi-resolution traffic data, trajectory data, and sensor inputs, enhancing real-time crash risk prediction.

**Model Building:**

# Common Machine Learning Approaches

* Random Forest (RF): Widely used due to its robustness and ability to handle nonlinear interactions. RF models incorporate features such as location, time, weather, road conditions, and other relevant factors to predict accident risk or severity
* Support Vector Machines (SVM) and Artificial Neural Networks (ANN): Often used alongside RF, with performance depending on data quality and preprocessing
* Ensemble and Hybrid Models: Combining multiple ML algorithms or integrating clustering methods (e.g., SVM-Fuzzy C-Means) to improve prediction accuracy and handle class imbalance.
* Deep Learning Models: Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), and Transformer-based models excel in time series forecasting of accident occurrences (e.g., LSTM achieving RMSE ~0.18 monthly).

# Feature Engineering and Data Preparation

* Incorporate diverse features: temporal (time of day, season), spatial (location, proximity to intersections), environmental (weather, road conditions), and vehicle/driver attributes.
* Address class imbalance and missing data through oversampling (SMOTE) and imputation.
* Encode categorical variables and normalize numerical features to improve model training.

**Visualization of Results & Model Insights:**

* Feature Importance Visualization  
  Using models like LightGBM combined with SHAP (SHapley Additive exPlanations) values, researchers visualize the impact of features on accident severity. For example, spatial features such as **Longitude** and **Latitude**, along with temporal features like **Hour**, often emerge as the most influential factors affecting accident outcomes. SHAP plots display how variations in these features increase or decrease predicted severity, providing interpretable insights into model behavior.
* Spatial Visualization and Hotspot Mapping  
  Geographic Information Systems (GIS) and heatmaps are used to identify accident hotspots in urban areas. Mapping accident locations reveals high-risk zones, enabling targeted interventions. Visual tools also show spatial distribution patterns, helping authorities prioritize resources.
* Severity and Risk Prediction Outputs  
  Visual dashboards present predicted accident severity or risk scores for specific locations and times. These outputs often include classification results (e.g., low, medium, high risk) with confidence levels, aiding real-time decision-making and alert systems[1](https://www.mdpi.com/2071-1050/15/7/5939)[2](https://www.sciencedirect.com/science/article/pii/S2590198223000611)[6](https://arxiv.org/pdf/2406.13968.pdf).
* Model Performance Metrics Visualization  
  ROC curves, confusion matrices, and precision-recall plots are used to assess and communicate model accuracy, sensitivity, and specificity. These visualizations help compare different models (e.g., Random Forest, Decision Trees, Logistic Regression) and
  + select the best-performing approach[1](https://www.mdpi.com/2071-1050/15/7/5939)[2](https://www.sciencedirect.com/science/article/pii/S2590198223000611).
* Multi-Feature Relationship Exploration  
  Pair plots, correlation matrices, and scatter plots enable exploration of relationships between multiple variables (e.g., weather conditions, road type, traffic volume) and their joint effect on accident occurrence or severity

**Tools and Technologies Used :**

Mention all tools used in this phase of the project.

* Programming Language: Python .
* IDE/Notebook: Google Colab, Jupyter Notebook, VS Code, etc.
* Libraries: pandas, numpy, seaborn, matplotlib, scikit-learn, XGBoost, etc.
* Visualization Tools: Plotly, Tableau, Power BI.

**Team Members and Contributions:**

1. **SUJITHRA M** :Project Lead : Oversees the entire project, coordinates tasks, and ensures deadlines are met.

2.**SUBASHINI S**: Handles data collection, cleaning, and exploratory data analysis.

3**. MARIYA MOUNIYA A**: Focuses on building and training the AI models for accident prediction.

4.**REKA R**: Manages data storage, preprocessing, and feature engineering.

5.**NITHYASRI N**: Works on deploying the AI model into a web application and ensures it runs smoothly.