**GEOSPATIAL NETWORK ANALYSIS AND SEVERITY RANKING OF TRAFFIC ACCIDENTS**

**Group Members:**

**Akshay S**

**Anagha K K**

**Sreelakshmi M**

1. **INTRODUCTION**

**1.1 Background**

Traffic accidents represent a critical public safety issue. The primary goal of this study is to predict the severity of traffic accidents by analyzing various factors related to road type, weather conditions, traffic controls, and other road features. Insights from this analysis can guide preventive measures and resource allocation to reduce the occurrence of severe accidents.

**1.2 Problem Statement**

The aim of this project is to predict the severity of road accidents in Kerala using various features like road type, traffic control, weather conditions, and collision type. This predictive tool will help the traffic police to proactively manage high-risk areas and provide timely interventions.

**1.3** **Objectives**

* To perform exploratory data analysis (EDA) on accident data.
* To engineer features that are critical to predicting accident severity.
* To build and validate machine learning models for severity prediction.
* To develop a user-friendly web application that enables police officers to input accident conditions and get severity predictions.

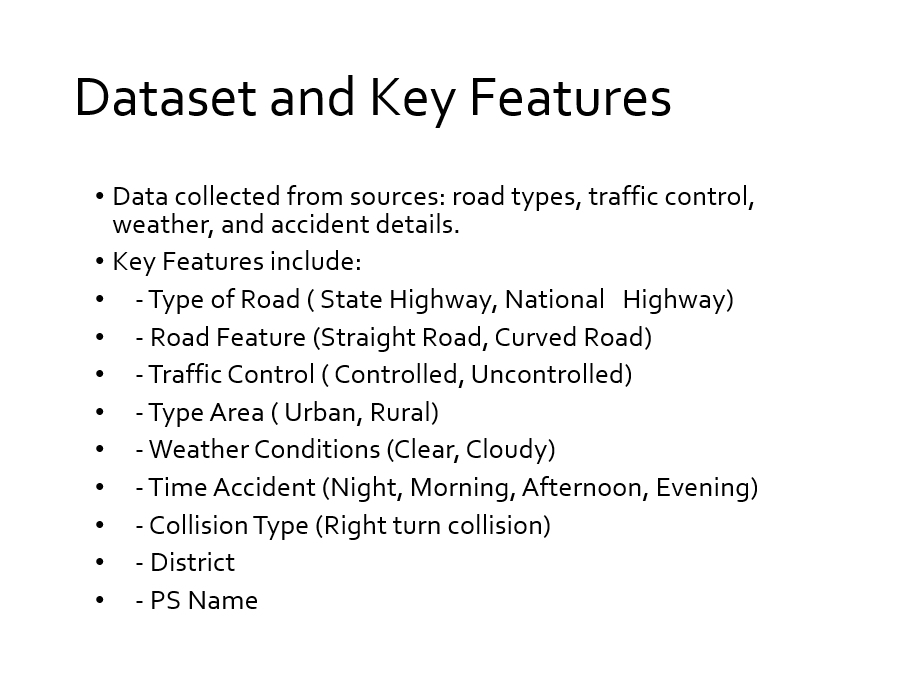
**2 Data Description**

**2.1 Dataset Overview**

* **Source:** Dummy Traffic data for Kerala.
* **Number of Entries:** 20,066 entries (After combining data for January Month for 5 years).
* **Number of Columns:** 18 columns.
* **Target variable:** Severity with four levels
* 0: Non-Injury
* 1: Minor Injury
* 2: Grievous Injury
* 3: Fatal

**2.2 Feature Description**

* **Categorical Variable:** Road Type, Road Features, Traffic Control, Type Area, Weather, Time Accident , Collision, District, PS Name
* **Numerical Variable:** Severity Score



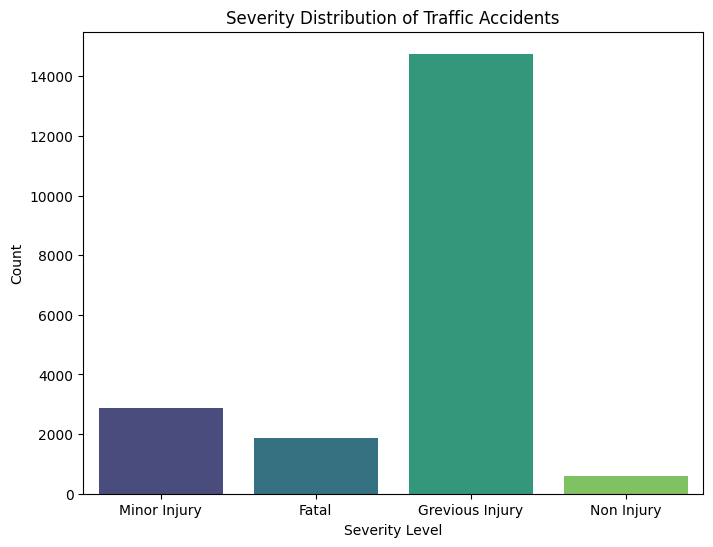
**3 Methodology**

**3.1 Data preprocessing**

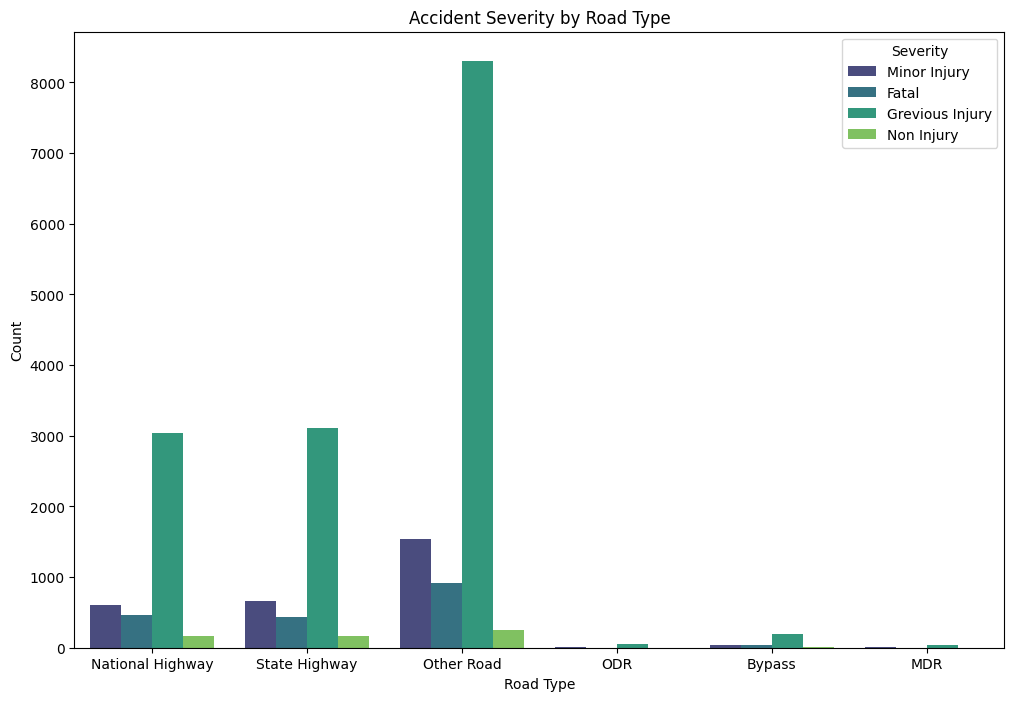
* **Missing Values**: Handled missing data through imputation methods and dropping columns with excessive missing values. (T-junction)
* **Duplicate Column Removal**: Duplicates, such as repeated categorical columns, were removed to prevent redundancy in the data.
* **Outlier Detection and Removal**: Outliers in features such as accident count and time of day were identified and managed, as extreme values could bias the model.

**3.2 Explanatory Data Analysis**

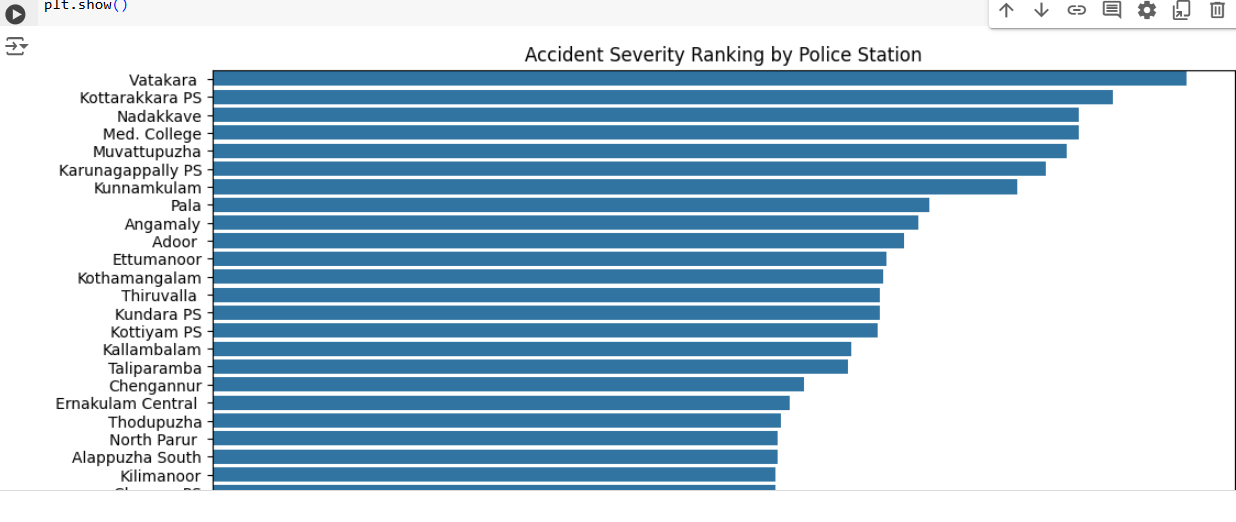
* **Severity Distribution (Bar Chart):**



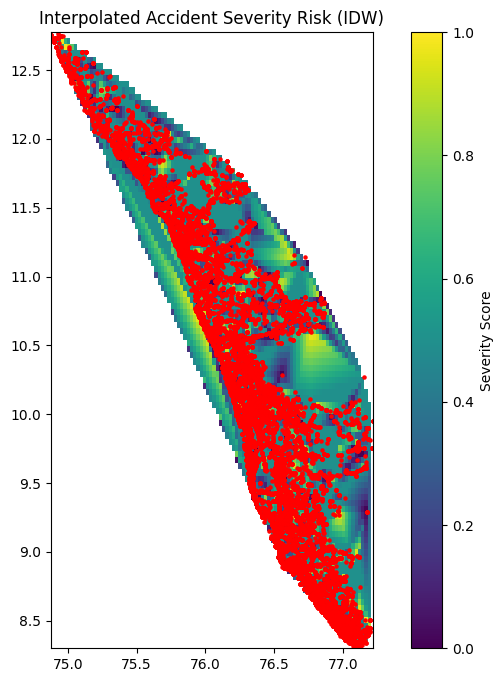
* **Accident Severity by Road Type and District (Grouped Bar Chart):**



* **Accident Severity Ranking by Police Station**



* **Interpolated Accident Risk**



**3.2 Feature Encoding**

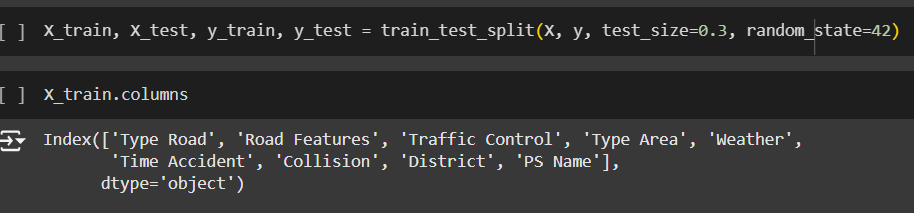
Since the dataset contains categorical variables, feature encoding was necessary to convert these into numerical values**:**

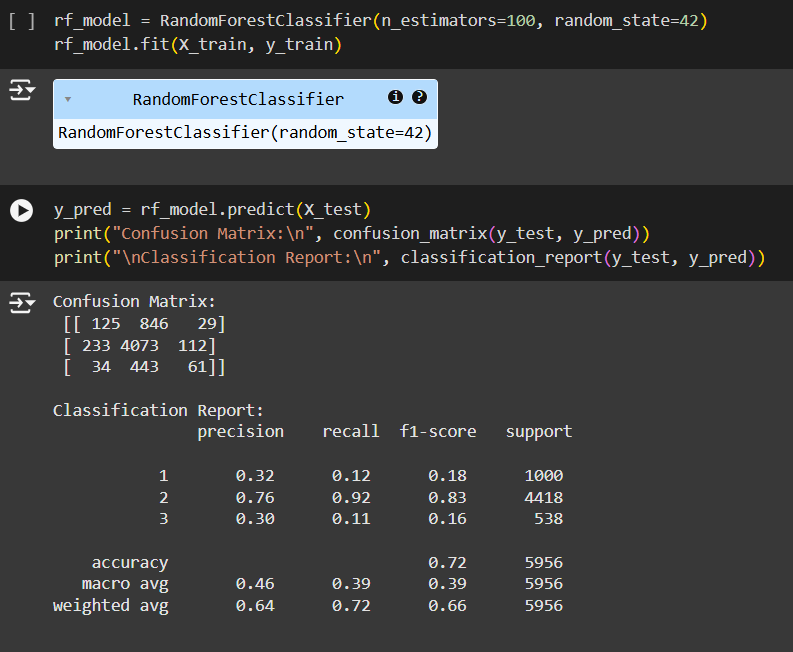
* One-Hot Encoding: Some categorical variables with limited categories, such as Weather and Traffic Control etc. were one-hot encoded to create a binary representation.
* Label Encoding: Encode categorical variables such as Type Road, Road features, Type Area, weather, Time Accident, Collision, District, PS Name into numerical formats suitable for machine learning algorithms.

**3.3 Model Selection**

Several models were considered, including logistic regression and decision trees. However, the Random Forest Classifier was chosen for its robust performance on structured data and its ability to handle a mix of categorical and numerical features. Key model details:

* **Model used:** Random Forest Classifier
* **Train-Test Split**: The data was split into training (70%) and testing (30%) sets to evaluate the model’s generalization ability.





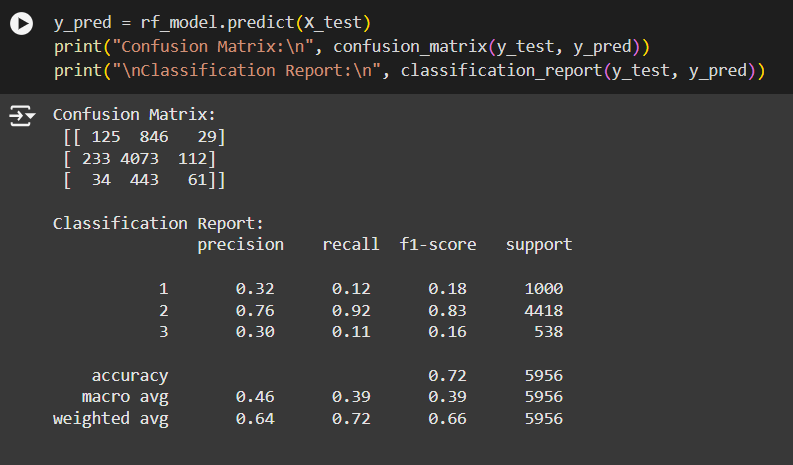
* **Hyperparameter Tuning:** Grid search was used to optimize parameters, such as the number of trees and maximum depth, to improve classification accuracy.

**4. Results**

**4.1 Model Performance**

The Random Forest model performance was evaluated based on accuracy, precision, recall, and F1 score:

* **Accuracy**: The model achieved an accuracy score of 72%, indicating its overall effectiveness in distinguishing high-risk and low-risk areas.
* **Confusion Matrix**: Analysis of the confusion matrix revealed the model’s ability to correctly classify high-risk zones while minimizing false positives and negatives. This is crucial in ensuring accurate severity predictions.
* **Classification Report**: Detailed metrics, including precision, recall, and F1 scores, were calculated for both high-risk and low-risk zones, with high precision in identifying high-risk areas.



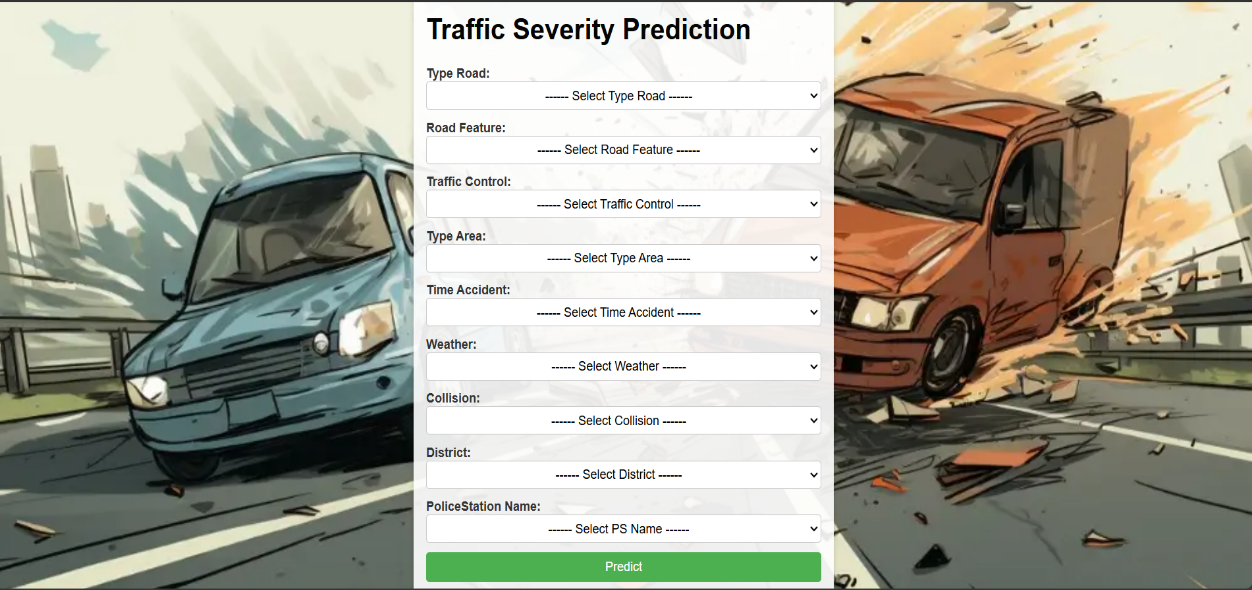
**5. Web Application Development**

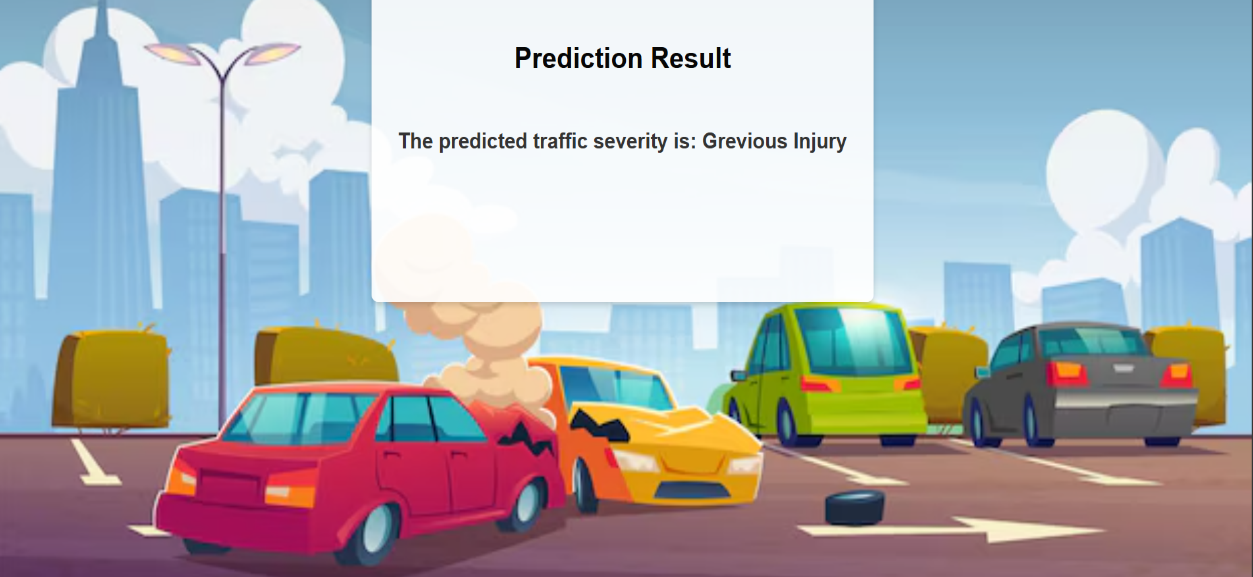
**5.1 Overview**

The project aims to develop a user-friendly web application that predicts traffic accident severity based on inputs like road type, weather, and traffic controls. Designed for road safety officials, it allows quick data entry to generate real-time severity predictions (e.g., Minor, Grievous, Fatal), supporting data-driven decisions and prioritizing high-risk areas. The application’s backend integrates the predictive model, enhancing safety measures and resource allocation for accident prevention.

**5.2 Web App Functionality**

* **Input Fields:** Fields for Type Road, Traffic Control, Time Accident, District, PS Name and other relevant features.
* **Output:** Displays the predicted severity.





**6. Conclusion**

**6.1 Key Insights**

* The model effectively predicts accident severity, offering valuable insights into high-risk conditions.
* It enables strategic, data-driven decisions for targeted accident prevention and safety improvements.
* By identifying factors influencing severity, the model helps prioritize high-risk areas and resource allocation.
* Future enhancements, like adding traffic density data, could further refine accuracy and impact.
* Overall, the model supports proactive safety measures, contributing to improved road safety outcomes.

**6.2 Challenges and Insights**

* **Class Imbalance:** High-risk zones comprised a smaller proportion of the dataset, necessitating strategies like balanced class weights to prevent bias in predictions.
* **Data Quality:** Missing and inconsistent data presented a challenge, requiring careful preprocessing to ensure the model’s accuracy and reliability.
* **Encoding Complexity**: Some categorical features required advanced encoding techniques, such as mean encoding, which added to the complexity of the preprocessing pipeline.

**6.3 Future Work**

* **Advanced Model Tuning and Testing:** Further model tuning and testing with algorithms like Gradient Boosting could improve predictive accuracy for complex scenarios.
* **Real-Time Data Integration**: Incorporating live traffic and weather data would enhance the model’s responsiveness, making it more adaptable to current conditions.
* **Expanded Web Application Features:** Adding functions like real-time accident reporting, automated alerts, and integration with GIS data for real-time mapping could increase the application’s value for end-users.

**CONTRIBUTION:**

**Akshay S:** Encoding, Feature Importance Analysis and Web-app Development

**Anagha K K:** Model Testing, Hyperparameter tuning, EDA and Basic Data Visualization

**Sreelakshmi M:** Background Research, Data Preprocessing, Feature Engineering and basic Modelling