Analysis and Model Creation for Finding Accident Hotspots

Group Members:

Fathima Shemeema

Uthara Dileep

Nayana. O

1. Introduction

1.1 Background

Traffic accidents continue to be a major global public safety concern, contributing to thousands of fatalities, injuries, and substantial economic costs each year. Understanding the patterns and causes of these accidents is essential for public safety planning and preventive measures. Identifying high-risk areas, or accident hotspots, enables targeted enforcement and infrastructure improvements, ultimately reducing accident occurrences. This project uses data science and machine learning techniques to analyse traffic accident data and classify high-risk and low-risk zones, thus assisting authorities in making informed decisions.

1.2 Problem Statement

Increased traffic accident rates and their adverse consequences demand effective solutions to identify and manage accident-prone areas. This project aims to classify traffic accident hotspots and predict potential high-risk zones by analysing historical accident data. By implementing predictive clustering models, the project identifies key accident factors and classifies specific areas based on their risk levels. This clustering provides a foundation for targeted interventions and resource allocations to improve road safety.

1.3 Objectives

- To analyse historical traffic accident data and classify geographic locations into high-risk and low-risk clusters.
- To determine significant features and conditions associated with accident occurrences, such as weather, time, road type, and traffic control measures.
- To develop a machine learning model capable of accurately classifying new areas as high-risk or low-risk.
- To create a user-friendly web application that visualizes accident hotspots and allows users to interact with the predictive model in real time.

2. Data Description

2.1 Data Overview

The dataset used for this project consists of detailed records of traffic accidents, provided by the traffic police department. It includes multiple attributes that describe the accident's location, date, time, weather conditions, and involved vehicles. The dataset offers a comprehensive view of each accident, making it suitable for clustering and risk classification based on different situational and environmental factors.

2.2 Feature Description

The key features used in this analysis include:

- **District**: The administrative area where each accident occurred, which helps group data by regional characteristics.
- **PS Name**: The police station jurisdiction covering the accident location, assisting in identifying police station-based clusters of incidents.
- **Date Accident**: The date of the accident, useful for analysing seasonal trends or special event impacts on accident frequency.
- **Time Accident**: Time of day when the accident occurred, relevant for exploring traffic patterns and peak accident timings.
- **Weather**: Weather conditions, as adverse weather like rain or fog often correlates with increased accident rates.
- **Accident Type**: Specifies the nature of the accident, such as a collision or pedestrian-related incident.
- **Road Features**: Describes physical road characteristics (e.g., intersections, curves) that may contribute to accident likelihood.
- **Traffic Control**: The presence or absence of traffic control devices (signals, signage), which may influence accident occurrence.
- **Latitude and Longitude**: Coordinates to map accident locations, allowing for spatial analysis and clustering visualization.

3. Methodology

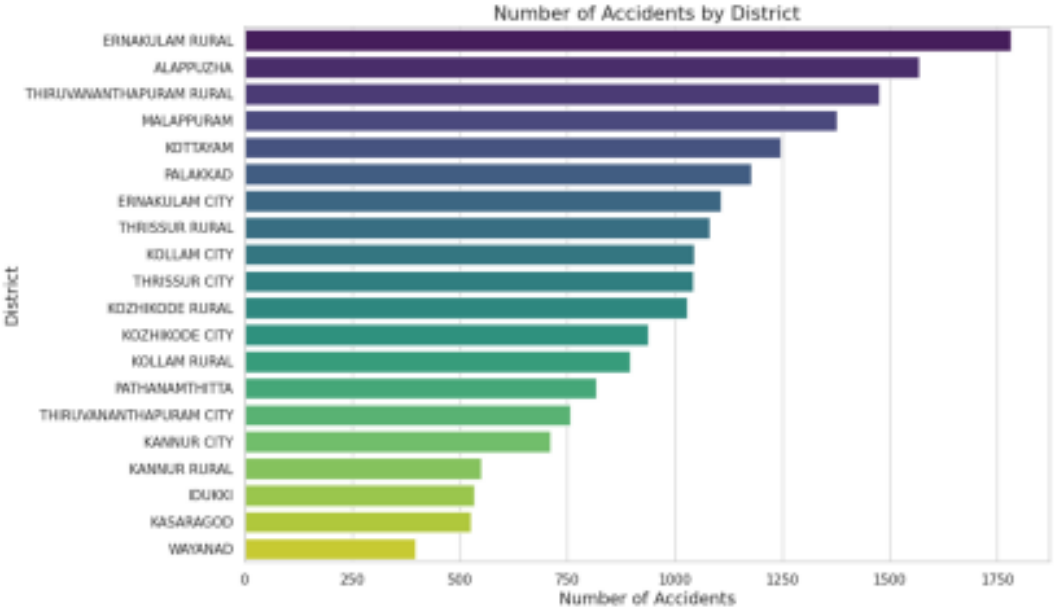
3.1 Data Preprocessing

Initial data preprocessing involved:

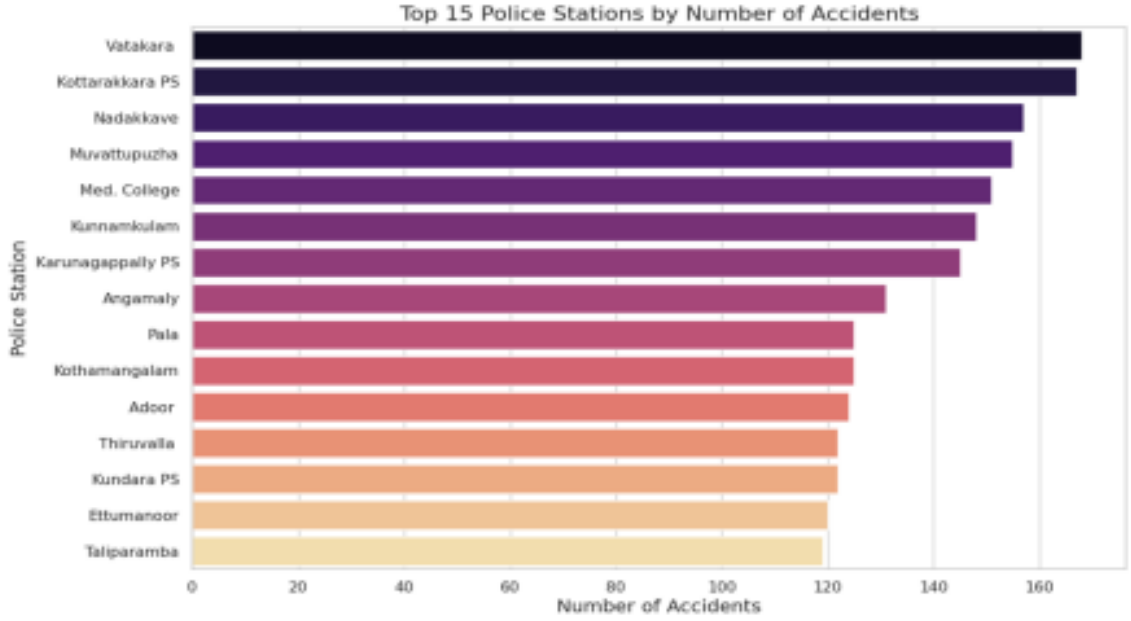
- **Handling Missing Values**: Missing values were addressed, either by removing entries with incomplete records to maintain data integrity.
- **Outlier Detection and Removal**: Outliers in features such as accident count and time of day were identified and managed, as extreme values could bias the model.
- **Duplicate Column Removal**: Duplicates, such as repeated categorical columns, were removed to prevent redundancy in the data.

3.2 Explanatory Data Analysis

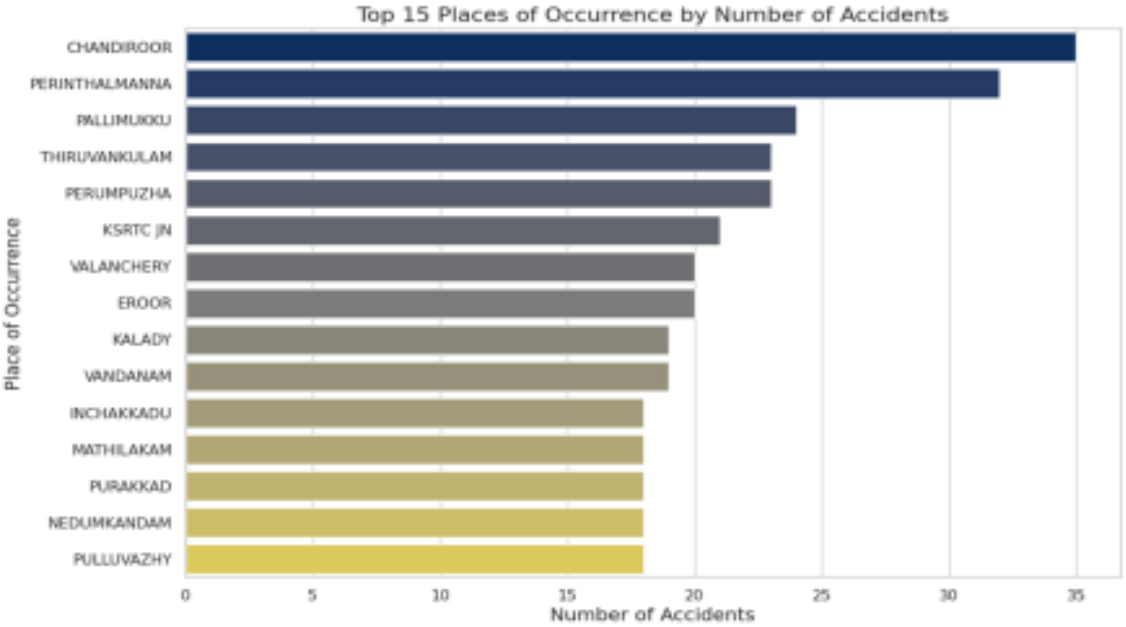
- **Accident count by District**



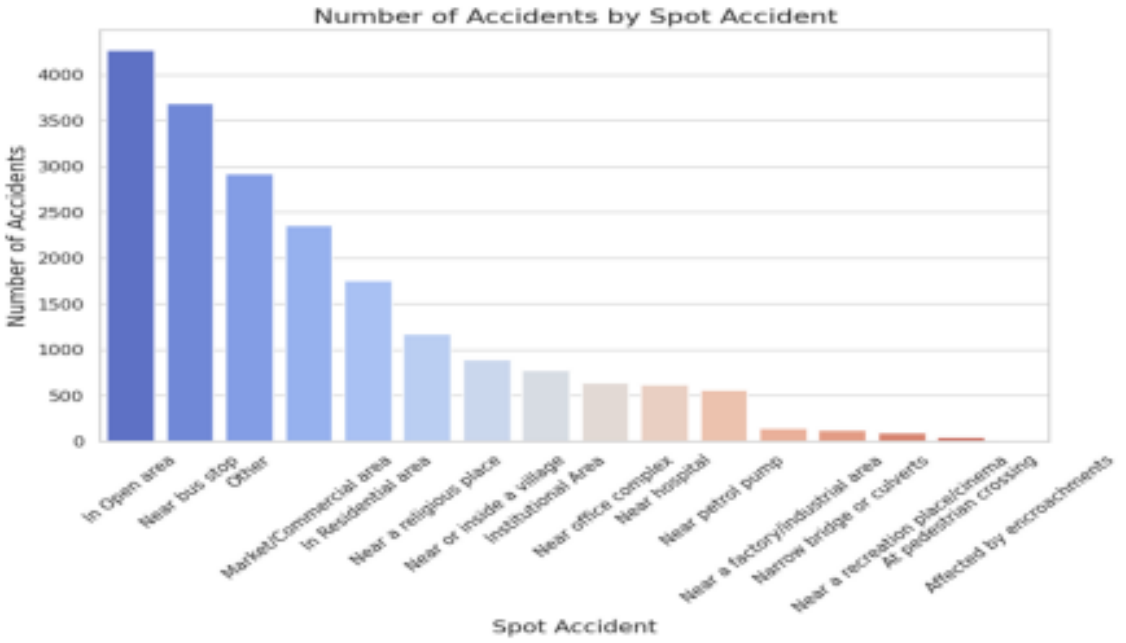
- **Accident Count by PS Name**



- **Accident count by Place of Occurrence**



- **Accident count by Accident Spot**



3.3 Feature Encoding

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

- **Mean Encoding**: For certain categorical features, such as District and PS Name, mean encoding based on accident count was applied. This allowed for a representation of how each category contributes to accident occurrences.

3.4 Feature Engineering

In order to identify accident hotspots, a new feature, **"Accident Count,"** was engineered to capture the frequency of accidents at specific locations. This feature was designed based on the guidelines from the Ministry of Road Transport and Highways (MoRTH), which classify a location as a hotspot if it records at least 5 accidents in a single year. The following steps outline the feature engineering process:

- **Grouping by Key Attributes**:
  - The accident data was grouped by critical location-based attributes: **District**, **Spot of Occurrence**, and **Police Station**. This grouping allowed for a granular analysis of accident trends across different areas.
  - Each unique combination of these attributes represents a specific location, enabling a localized focus on accident frequency for each area.
- **Accident Count Calculation**:
  - A new column, labelled **"Accident Count,"** was created to store the count of accidents for each unique location combination. This column was populated by calculating the total number of accidents recorded for each grouped location within the dataset.
- **Hotspot Classification**:
  - Based on the MoRTH guidelines, any location with an **Accident Count of 5 or more** was classified as a **"hotspot."**
  - This threshold was applied to each location group to assign a binary hotspot classification: locations with an accident count of 5 or above were labelled as **high-risk (hotspot)**, while those with fewer than 5 accidents were labelled as **low-risk (non-hotspot)**.
- **Additional Processing**:
  - The engineered **Accident Count** feature, along with the binary **hotspot classification**, provides key input for clustering and classification models, helping the model learn patterns associated with high-risk areas.
  - These derived features enable the model to distinguish between high-risk and low-risk zones effectively, leveraging accident frequency data as a significant predictor for accident-prone locations.

3.5 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:

- **Balanced Class Weights**: To address class imbalance, where high-risk zones may be fewer in number, balanced class weights were used to ensure equitable model training.
- **Train-Test Split**: The data was split into training (70%) and testing (30%) sets to evaluate the model's generalization ability.

4. Results

4.1 Model Performance

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

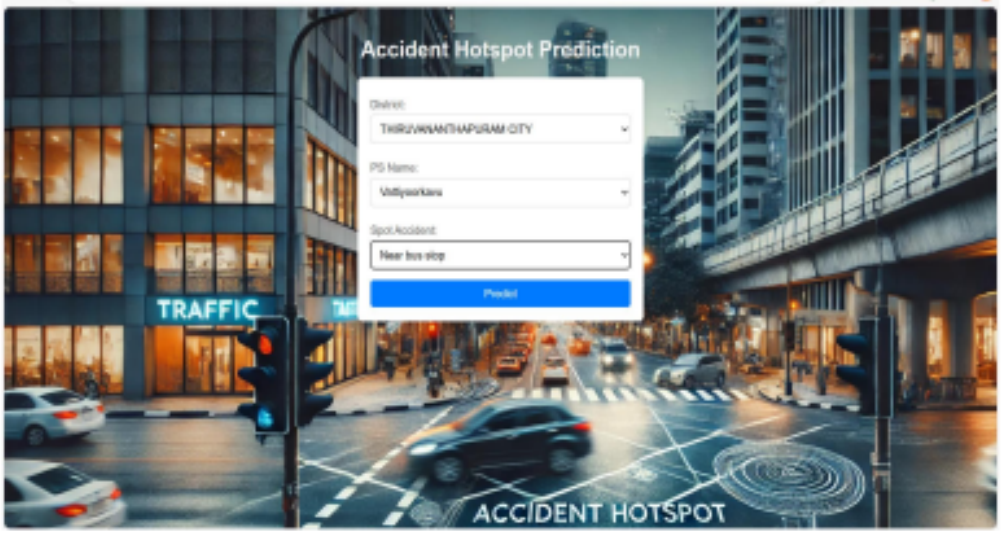
- **Accuracy**: The model achieved an accuracy score of 82%, indicating its overall effectiveness in distinguishing high-risk and low-risk areas.
- **Confusion Matrix**: Analysis of the confusion matrix revealed the model's ability to correctly classify high-risk zones while minimizing false positives and negatives. This is crucial in ensuring accurate hotspot 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

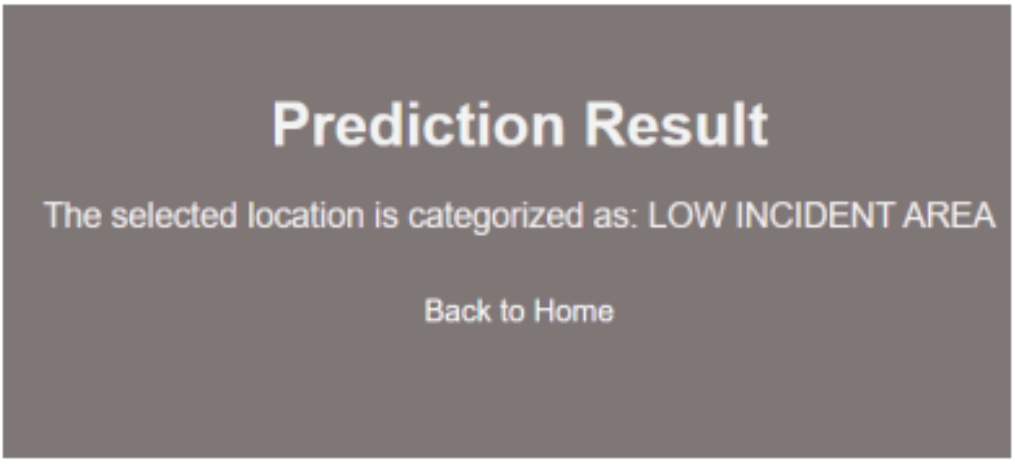
A web application was developed to allow users to interact with the data and visualize accident hotspots. Key features include:

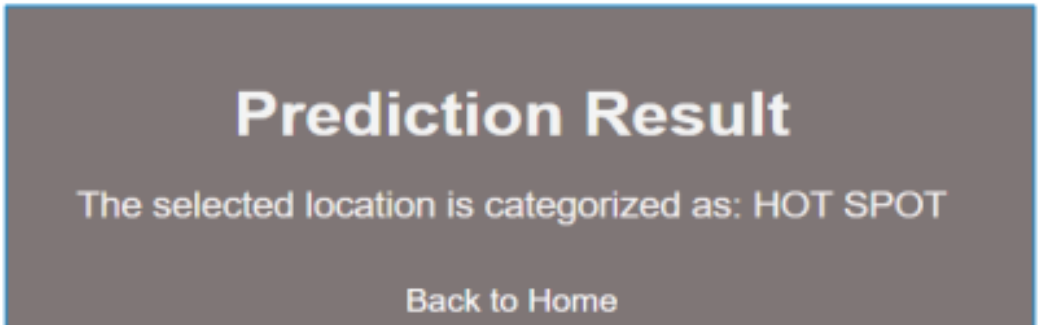
- **Interactive Mapping**: Using Folium, accident locations are visualized on an interactive map, with red markers representing high-risk areas and blue markers for low-risk areas.
- **Cluster Classification Tool**: Users can input specific details, such as district and accident location, to receive a real-time prediction on the location's risk level.
- **Information Pop-ups**: Each map marker includes pop-ups displaying details such as accident count, area classification, and other relevant features.

- **User Interface**: The app is designed to be user-friendly, with a clean layout and accessible information tabs for navigating the map and input fields.



By entering the set of input values, it predicts whether the particular district is a hotspot or a low incident area.





6. Conclusion

6.1 Key Insights

- **Accident Concentration**: The analysis highlights specific districts and police station jurisdictions with higher accident rates, where additional safety measures could be prioritized.
- **Feature Importance**: Weather, time of day, and road features emerged as significant predictors of accident likelihood, guiding future strategies for road safety.
- **Model Applicability**: The Random Forest model effectively differentiates between high-risk and low-risk areas, proving the feasibility of using predictive clustering for traffic safety interventions.

6.2 Challenges and Limitations

- **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

Fathima Shemeema :Encoding, Feature importance analysis and Web-app development, model testing

Uthara : Data preprocessing, Feature Engineering and basic modelling, model evaluation, model testing

Nayana: Background Research, EDA and basic visualization, model testing