**Accident Prediction Model**

**based on Weather Conditions**

*A project report submitted to ICT Academy of Kerala*

*in partial fulfillment of the requirements*

*for the certification of*

**CERTIFIED SPECIALIST**

**IN**

**DATA SCIENCE & ANALYTICS**

submitted by

**Hiba S S**

**Athira S**



**ICT ACADEMY OF KERALA**

**THIRUVANANTHAPURAM, KERALA, INDIA**

**JUNE 2024**

**List of Figures**

| **Figure** | **Page No** |
| --- | --- |
| Fig 3.1 Description of the data | 10 |
| fig 3.2 Boxplot of outliers | 10 |
| Fig 4.1 histogram for weather influence | 11 |
| fig 4.2 correlation matrix of numerical features  fig 5.1.Lime Tabular Explainer | 12  15 |
| Fig 5.2 Histograms for various features | 16 |

**List of Abbreviations**

1. **LIME -** Local Interpretable Model-Agnostic Explanations
2. **XGBoost -** Extreme Gradient Boosting
3. **EDA -** Exploratory Data Analysis

**Table of Contents**

| **S no** | **Title** | **Page no** |
| --- | --- | --- |
| **1** | **Problem Definition** | **6** |
| **2** | **Introduction** | **8** |
| **3** | **Data Description** | **9** |
| **4** | **Data analysis** | **10** |
| **5** | **Exploratory Data Analysis (EDA)** | **12** |
| **6** | **Modelling** | **13** |
| **7** | **Result** | **15** |
| **8** | **Conclusion** | **16** |

**Abstract**

Road safety is a pressing global concern, with traffic accidents being one of the leading causes of fatalities and injuries worldwide. Each year, countless lives are lost, and substantial economic resources are drained due to road accidents. Addressing this issue requires proactive measures that go beyond traditional safety protocols, focusing instead on leveraging modern technologies to predict and prevent accidents before they occur.

This project aims to harness the power of data science and machine learning to predict the likelihood of road accidents. By integrating real-time data on weather conditions, visibility, road surface characteristics, and geographic information, we strive to create a robust predictive framework. The ultimate goal is to provide decision-makers with actionable insights that can be used to enhance road safety measures.

Moreover, this project emphasizes the importance of understanding the underlying factors contributing to accidents. Through techniques such as LIME (Local Interpretable Model-Agnostic Explanations), we analyze and explain which features, such as weather patterns or visibility, are the most significant contributors to accidents. This level of transparency helps build trust in the predictive model and enables stakeholders to address the root causes effectively.

**1. Problem Definition**

**1.1 Overview**

Road safety remains one of the most pressing concerns worldwide, with traffic accidents accounting for a substantial loss of human lives and imposing a severe economic burden on individuals, families, and nations.

The complexity of road safety lies in its dependence on various dynamic factors, such as weather conditions, road surface quality, visibility, and traffic density. Adverse weather conditions, including rain, fog, and snow, significantly increase the risk of accidents by reducing visibility and making road surfaces slippery. Poorly maintained roads and unexpected obstacles further exacerbate these risks, especially in areas with high traffic volumes or insufficient traffic management systems.

In this context, the ability to predict accidents based on real-time weather and road conditions can be a transformative step toward reducing accidents and saving lives. By leveraging advanced technologies such as machine learning and real-time data analysis, it becomes possible to identify high-risk scenarios and implement preventive measures.

The importance of such predictive capabilities extends beyond immediate accident prevention. They also provide insights for policymakers, urban planners, and traffic management authorities to address the root causes of accidents and design long-term safety strategies. Ultimately, adopting a data-driven approach to road safety can save lives, reduce economic losses, and promote safer and more efficient transportation systems globally.

**1.2 Problem Statement**

The primary challenge in ensuring road safety is the dynamic and unpredictable nature of factors that contribute to accidents. Real-time variables such as adverse weather conditions, fluctuating visibility levels, and diverse road features, including surface quality and traffic density, significantly influence the likelihood of accidents. These factors often change rapidly and can interact in complex ways, making it difficult to assess risk accurately and implement timely interventions.

Traditional approaches to road safety typically rely on historical data or reactive measures, such as deploying traffic enforcement after accidents occur. While these methods are valuable, they fail to provide the proactive insights needed to prevent accidents before they happen. This gap highlights the urgent need for innovative solutions that integrate real-time data analysis and predictive modeling to anticipate high-risk scenarios and enable preemptive actions.

This project seeks to bridge this critical gap by developing a comprehensive predictive model that uses machine learning to analyze real-time weather conditions, road visibility, and other key features. The model identifies patterns and correlations that contribute to accidents, providing actionable insights for stakeholders. By addressing the challenges of real-time prediction, this project aims to shift the focus from reactive responses to proactive measures, ultimately reducing accident rates and creating safer roads for all users.

**2. Introduction**

This project combines the power of data analysis, machine learning, and application development to tackle the pressing issue of road safety. With millions of accidents occurring globally each year, there is a critical need for innovative approaches to minimize risks and prevent incidents. By leveraging modern technologies, this project aims to provide a data-driven solution to identify and mitigate accident-prone scenarios effectively.

The core objective of this initiative is to develop a predictive model that uses key factors such as weather conditions, road features, and visibility to forecast the likelihood of road accidents. The model will analyze complex interactions among these variables to identify patterns and correlations that contribute to higher accident risks. By incorporating real-time data, the model aims to provide accurate and actionable insights, enabling timely interventions.

Through the integration of data science, advanced algorithms, and user-centric application design, this project aspires to create a comprehensive solution to road safety challenges. By predicting accidents and offering visual insights, it empowers stakeholders to take proactive measures, ultimately contributing to a safer and more efficient transportation system.

# 3. Data Description

# 

Fig 3.1 Description of the data

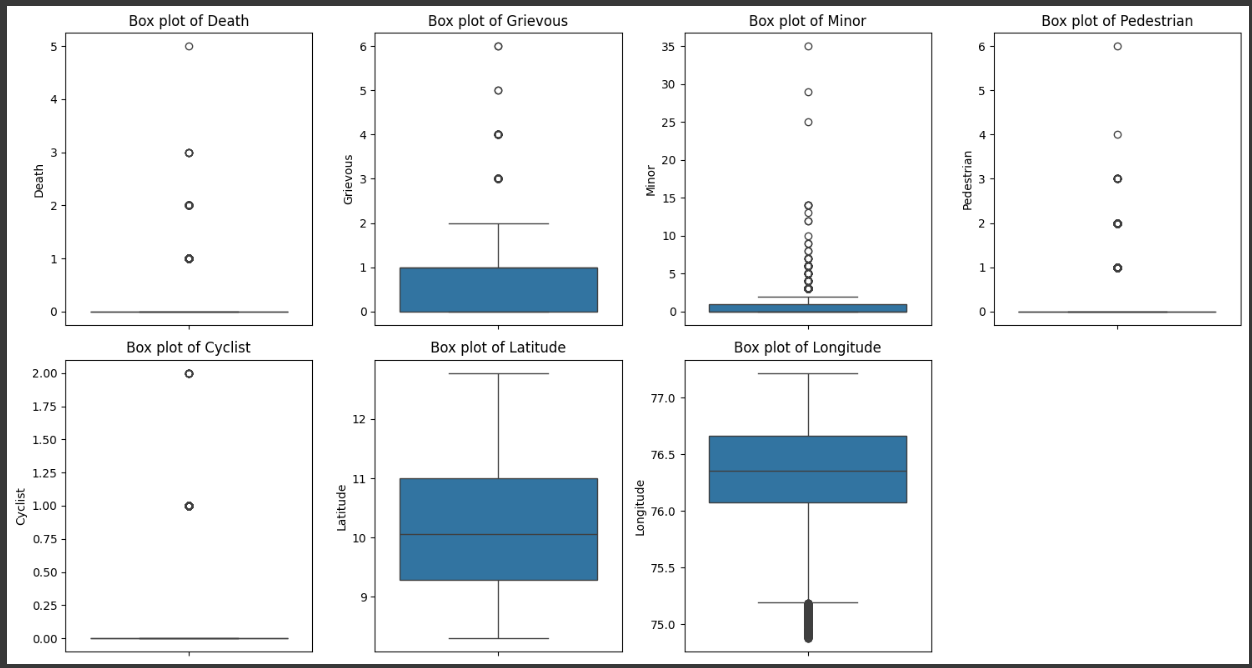


fig 3.2 Boxplot of outliers

# 4. Data Analysis

## 4.1 Exploratory Data Analysis

Exploratory Data Analysis (EDA) provided valuable insights into the key factors influencing road accidents, helping to uncover patterns and correlations essential for building a robust predictive model.

To ensure reliable model performance, extensive data cleaning and preprocessing were carried out. This included handling missing values, removing outliers, and normalizing feature distributions. Categorical variables, such as weather conditions and road types, were encoded for compatibility with machine learning algorithms. Moreover, feature engineering was employed to derive meaningful variables, such as combining visibility and weather conditions to create a composite risk factor.

These steps ensured that the dataset was not only clean and comprehensive but also well-suited for training predictive models. The insights gained during EDA provided a strong foundation for feature selection and model development, contributing significantly to the project’s overall success.

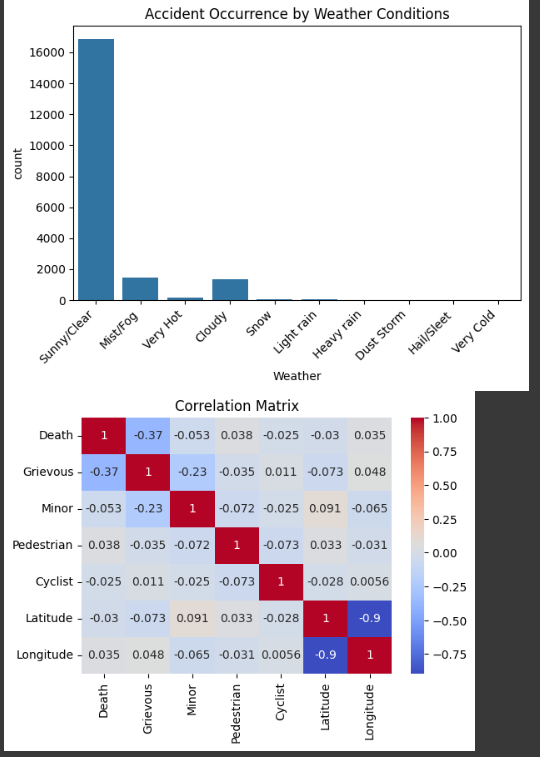


Fig 4.1 histogram for weather influence

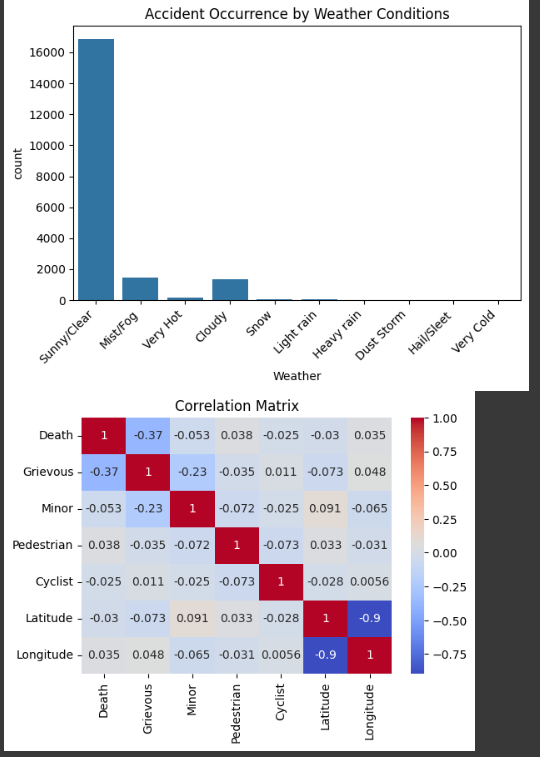


fig 4.2 correlation matrix of numerical features

## 4.2 Feature Selection

The selection of features plays a pivotal role in the success of any predictive modeling project, and this study carefully identified variables with the greatest influence on road accidents. The chosen features—weather conditions, visibility, road type, and accident location—were determined to be significant predictors based on comprehensive correlation analysis and domain understanding.

* **Weather Conditions**: This feature captures variations such as rain, fog, snow, and clear weather. It emerged as a critical factor due to its direct impact on road surface quality and driver visibility, both of which are closely linked to accident frequency.
* **Visibility**: Representing the clarity of the driver’s view, this feature was found to have a strong inverse correlation with accident likelihood.
* **Road Type**: Roads differ in design and purpose, such as highways, urban streets, and rural roads. This feature provided insights into how accident risks vary with traffic density, speed limits, and infrastructure.
* **Accident Location**: Geospatial data on where accidents occurred helped identify hotspots. Incorporating this feature enabled the model to associate certain areas with higher accident probabilities, likely due to design flaws, congestion, or environmental conditions.

The selection process relied heavily on statistical methods, such as correlation matrices and feature importance analysis, to ensure that these features were not only statistically significant but also intuitively aligned with real-world factors affecting road safety. This thoughtful feature selection laid a solid foundation for building an accurate and interpretable predictive model.

# 

# 5. Modeling

A Voting Classifier was utilized for this project to leverage the strengths of multiple machine learning algorithms and achieve improved predictive accuracy. The Voting Classifier combines the predictions of three models: Random Forest, Logistic Regression, and XGBoost. This ensemble approach ensures that the final predictions benefit from the complementary strengths of these algorithms, resulting in a more robust and reliable model.

* **Random Forest**: Known for its ability to handle non-linear relationships and high-dimensional data, Random Forest contributed its robustness and capacity to manage complex interactions among features such as weather conditions and road types.
* **Logistic Regression**: As a simple yet powerful linear model, Logistic Regression provided a probabilistic perspective and ensured that the ensemble could capture linear relationships effectively. Its inclusion added interpretability to the overall model.
* **XGBoost (Extreme Gradient Boosting)**: This advanced boosting algorithm excels at handling large datasets and minimizing prediction errors. Its capacity for fine-grained optimization and regularization made it an essential component for capturing subtle patterns in the data.

The Voting Classifier aggregates the predictions of these three models, either by majority voting (classification) or averaging probabilities. This approach reduces the likelihood of individual model biases impacting the final prediction.

LIME (Local Interpretable Model-Agnostic Explanations) was used to interpret the results of the Voting Classifier, providing insights into the contribution of each feature to the model's predictions. By analyzing individual predictions, LIME helped identify the most critical factors influencing accident risks, such as poor visibility, adverse weather, and accident-prone locations.

This ensemble methodology ensured that the model delivered both high accuracy and interpretability, making it a powerful tool for predicting and preventing road accidents.

A screenshot of a computer

Description automatically generated

fig 5.1.Lime Tabular Explainer

**6. Result**

**A map of a road

Description automatically generatedA screenshot of a computer

Description automatically generated**

Fig 6.1. Application interface

**7. Conclusion**

In this project, a comprehensive approach was adopted to address the critical issue of road safety through the integration of data analysis, machine learning, and application development. By leveraging a Voting Classifier that combined Random Forest, Logistic Regression, and XGBoost, the model effectively captured both linear and non-linear relationships, providing accurate predictions of road accidents based on dynamic factors such as weather conditions, visibility, and road type. The use of LIME for model interpretability added transparency, ensuring that stakeholders could understand and trust the predictions while gaining actionable insights into accident causation.

the project successfully demonstrated the potential of predictive analytics in enhancing road safety, there remains scope for future improvements. These include integrating additional real-time data sources, refining the model with more diverse datasets, and expanding the application to support broader geographic areas. Overall, this project represents a significant step toward reducing road accidents and promoting safer travel environments.