# Predicting Traffic Accident Casualties Using Machine Learning, An Analysis of Global Traffic Accident Data

*Submitted by:*
*Atharva Bakde*
*Samarth Naik*

*Link to Dataset:*
*Global-Traffic-Accident-Dataset*

*Link to Code(Google collab):-*
*Global-Traffic-Accident-PythoncCODE*

# 1.Table of Content

# 2. Project description

## 2.1 Business Context and Goals of the Project

Road traffic accidents are one of the leading causes of injury and death around the world. In fact, over 1.35 million people die each year as a result of traffic-related accidents, and millions more suffer from injuries (World Health Organization, 2021). These accidents come with not just a huge human toll, but also significant economic costs—impacting healthcare systems, the workforce, and society at large.
While the causes of traffic accidents can vary, some of the most common factors include weather conditions, road conditions, time of day, and the number of vehicles involved. By understanding how these factors contribute to the severity of accidents, authorities can better prepare and implement safety measures to reduce accidents, save lives, and optimize resources for accident prevention.
This project focuses on analyzing a Global Traffic Accidents dataset that includes detailed records of traffic accidents, covering aspects such as the location, time, weather, road conditions, number of vehicles involved, and casualties. By exploring these factors, the project aims to answer important questions about what influences accident severity and whether it's possible to predict the number of casualties based on accident data.

**The key goals of the project are:**

1. Analyze factors affecting accident severity: We will examine how different conditions, such as weather, time of day, and road type, relate to the severity of accidents.
2. Develop a predictive model: We aim to build a model that can predict the number of casualties based on accident-related data (e.g., weather, time of day, road conditions).
3. Identify high-risk zones: By analyzing accident data by location, we will identify accident hotspots where interventions could make the most impact in reducing casualties.
4. Improve road safety measures: Based on the analysis, the project will provide actionable recommendations for local authorities to reduce accidents and improve road safety, particularly in high-risk areas.

The ultimate goal of this project is to provide valuable insights that can help reduce the occurrence of traffic accidents and their severity. By better understanding the relationship between different factors, we aim to inform public safety campaigns, urban planning decisions, and policy changes.

**Business Questions the Project Will Address:**

1. What factors contribute the most to traffic accident severity?
   We will explore the relationship between different accident-related factors, such as weather, road conditions, and the time of day, and how they impact the number of casualties.

2. How do weather and road conditions affect accident outcomes?
   By examining how various weather conditions (e.g., rain, snow, fog) and road types (e.g., icy, dry) influence the severity of accidents, we can identify high-risk conditions.

3. Can we predict the number of casualties based on accident data?
   The project aims to develop a predictive model that estimates casualties based on accident-related factors, which could help authorities prepare for high-risk situations.

4. Which areas experience the highest number of accidents and casualties?
   Using geospatial analysis, we will identify accident hotspots—areas with a higher frequency of accidents and more severe outcomes—allowing targeted safety measures.

5. How does the time of day, day of the week, or season impact accident severity?
   By analyzing accident patterns over time, we can understand if accidents are more likely to happen at specific times, which can inform traffic management and public safety campaigns.

**Conclusion and Expected Impact:**
This project aims to provide data-driven insights that will help reduce the impact of traffic accidents globally. By identifying key risk factors, predicting accident severity, and pinpointing high-risk locations, the findings from this analysis can be used by policymakers, traffic authorities, and urban planners to implement more effective safety measures. Ultimately, this project's goal is to contribute to the ongoing global efforts to save lives, prevent injuries, and reduce the economic burden caused by traffic accidents.

---

# 3. Data Preprocessing

Data preprocessing is a crucial step in preparing the dataset for analysis and predictive modeling. It involves a variety of tasks, including handling missing values, removing duplicates, transforming features, and dealing with outliers. Below is a detailed description of the data preprocessing steps applied to the **Global Traffic Accident dataset** in this project.

## 3.1 Handling Missing Data

The dataset was checked for any missing values in the dataset. It was confirmed that there were **no missing values** in any of the columns, ensuring the dataset was complete and ready for analysis. This step is critical as missing data can impact the quality of the analysis.

## 3.2 Handling Duplicates

The dataset was checked for duplicate rows to ensure there was no redundant data. It was found that there were **no duplicate rows**, which ensured the integrity of the data and avoided any potential distortion of the results due to data repetition.

## 3.3 Feature Extraction and Transformation

A critical part of the preprocessing was **feature extraction** from the **Date** and **Time** columns:

- **Datetime Conversion**: The Date and Time columns were combined into a single datetime column, which was then converted into a pandas datetime object to facilitate easy manipulation and feature extraction.
- **Time-Based Features**: From the datetime column, several new features were extracted, such as:
  - hour: The hour when the accident occurred.

- day: The day of the month when the accident occurred.
- month: The month when the accident occurred.
- dayofweek: The day of the week when the accident occurred (0 = Monday, 1 = Tuesday, etc.).
- weekofyear: The ISO week number in the year.

These new features enable a more granular analysis of traffic accidents across different times, days, and months, which is useful for understanding accident trends and patterns.

## 3.4 Data Transformation

Once the datetime column was created, the original Date and Time columns were **dropped** from the dataset to reduce redundancy. Additionally, **one-hot encoding** was applied to the categorical variables, including Weather Condition, Road Condition, and Cause, converting them into binary columns to make them suitable for machine learning models.

- **One-Hot Encoding**: This transformation was used for categorical variables to convert them into numerical format, while ensuring that we avoid multicollinearity by dropping the first category in each variable.
- **Boolean Conversion**: Any Boolean columns in the dataset were converted to integers (1/0) to ensure consistency for model training.

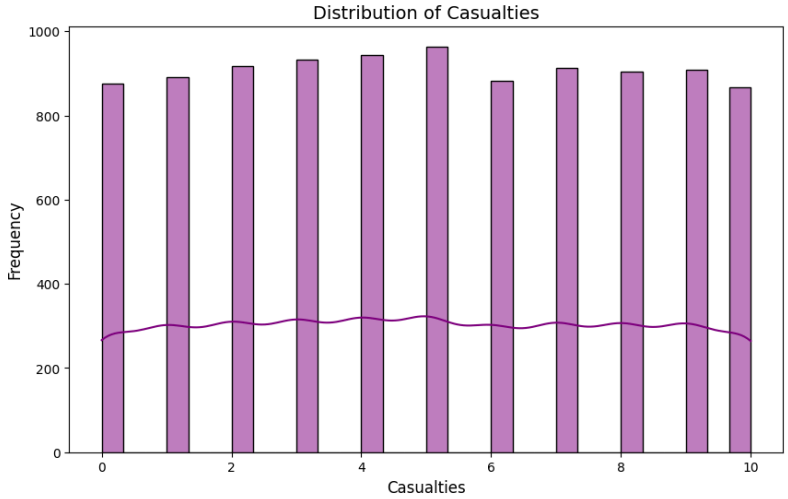
## 3.5 Saving and Re-loading the Cleaned Data

After performing the necessary transformations, the cleaned dataset was saved to a CSV file for future use. The cleaned data was then loaded back into the system to verify the changes and ensure the preprocessing steps were applied correctly.

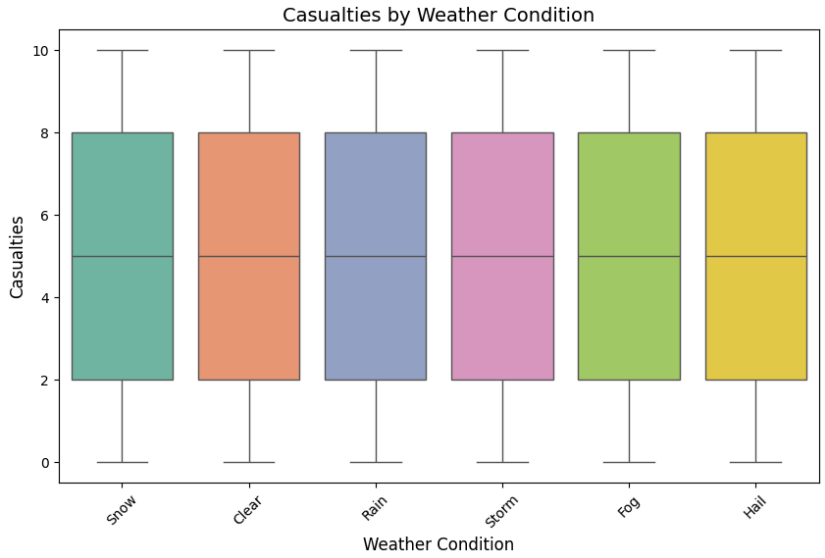
## 3.6 Exploratory Data Analysis (EDA)

After transforming the data, exploratory data analysis (EDA) was performed to uncover insights and better understand the dataset. Some of the key visualizations included

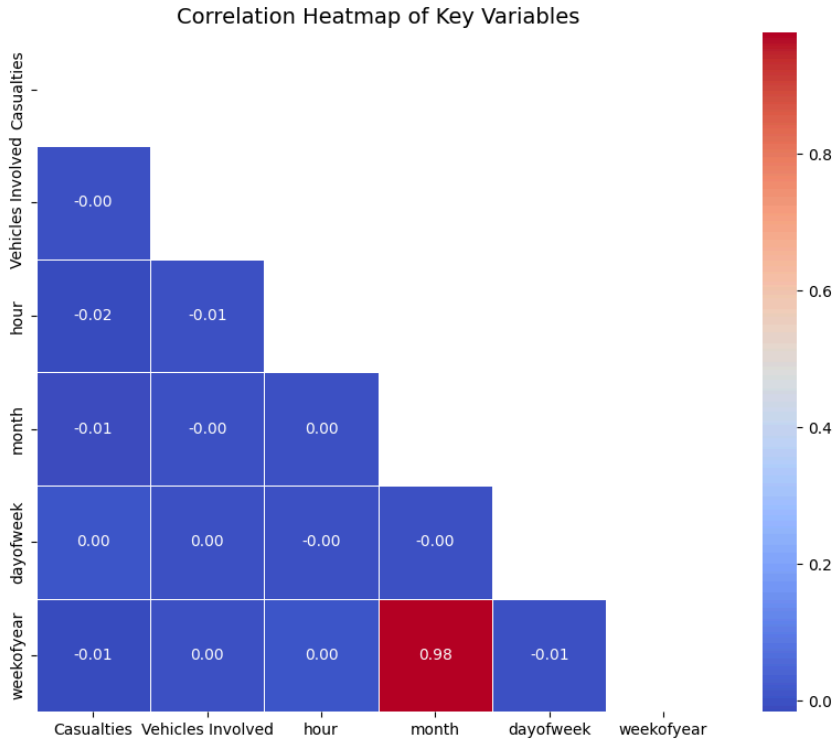
- **Distribution of Casualties**: Histograms and **Kernel Density Estimate (KDE)** plots were generated to visualize the distribution of casualties across the dataset.




- **Casualties by Weather Condition**: Box plots were created to examine how **Casualties** varied across different **Weather Conditions** (e.g., Rain, Snow, Clear).




- **Correlation Analysis**: A **correlation heatmap** was used to understand the relationships between key variables, such as **Vehicles Involved**, **Latitude**, **Longitude**, and **Casualties**. This helped identify important predictors for further modeling.




## 3.7 Feature Selection for Analysis

In preparation for modeling, key features were selected based on their relevance to the analysis:

- **Location-Based Data**: Features like **Latitude** and **Longitude** were retained to explore geographical patterns in traffic accidents.

- **Accident Characteristics**: Features like **Casualties**, **Weather Condition**, **Road Condition**, and **Vehicles Involved** were kept for analysis of accident severity and causes.

### 3.8 Data Transformation Summary

- **Datetime Transformation**: The **Date** and **Time** columns were merged into a single datetime column, and relevant time-based features (e.g., hour, day, month) were extracted.
- **Missing Data**: The dataset was found to have **no missing values**, meaning no imputation was needed.
- **Duplicates**: **No duplicate rows** were found in the dataset, ensuring data integrity.
- **Outlier Detection**: A visual inspection of key columns was performed, but no formal outlier removal techniques were applied.
- **Categorical Encoding**: **One-hot encoding** was applied to the **Weather Condition**, **Road Condition**, and **Cause** columns to prepare them for machine learning models.
- **Boolean Conversion**: Any Boolean columns were converted to integers (1/0).

### 3.9 Next Steps in Preprocessing

Following the preprocessing steps, the next phase involves applying machine learning algorithms, such as **Random Forest**, to predict **Casualties** based on the features in the dataset. The cleaned and transformed data is now ready for modeling, and future work may involve additional feature engineering, outlier handling, and model tuning to improve the predictive accuracy.

---

## 4. Models and Analysis

To predict the **likelihood of casualties** in traffic accidents, two **classification models** were developed and evaluated: **Logistic Regression** and **Random Forest**. These models were selected for their **interpretability** and **robustness**, respectively. The **dataset**, after **comprehensive preprocessing**, was split into **training**, **validation**, and **test sets**. Each model was trained on the **same data**, and their performances were evaluated using **accuracy**, **confusion matrices**, and **classification metrics** such as **precision**, **recall**, and **F1-score**. **Feature importance** was also analyzed to understand which factors most influenced the **prediction of casualty outcomes**. The following sections detail the **modeling approach**, **evaluation**, and **insights** for each model.

### 4.1 Logistic Regression

Logistic regression was **used as a baseline model** to predict whether a traffic accident would involve casualties. The model was trained on scaled input features, including temporal, weather, road, and cause-related variables.
**To address the class imbalance** in the dataset, class **weights were adjusted** during model training. This allowed the model to better capture both casualty and non-casualty outcomes, rather than defaulting to the majority class.

**Test Set Performance:**

- **Accuracy:** 51.5%

- **Class 0 (No Casualty):** Precision 9.8%, Recall 51.4%, F1-score 16.5%
- **Class 1 (Casualty):** Precision 91.1%, Recall 51.5%, F1-score 65.8%

**Although overall accuracy dropped compared to the initial version, the model now detects both classes, providing more balanced performance.** Class 0 recall significantly improved, indicating better sensitivity to non-casualty cases.

**Top Impactful Predictors:**

- Cause_Drunk Driving
- Weather Condition_Fog
- Cause_Reckless Driving
- Weather Condition_Hail
- Week of Year

These features played a major role in determining the likelihood of casualties during an accident

**Conclusion:**

While logistic regression now captures both classes, its predictive power remains limited compared to more complex models. Nonetheless, it offers interpretability and identifies key variables influencing accident severity.

## 4.2 Random Forest

A Random Forest classifier was developed to predict whether a traffic accident would result in casualties based on factors such as time, weather, road conditions, and number of vehicles involved. The model was configured with **100 decision trees** and regularization parameters **(max_depth=10, min_samples_split=10, min_samples_leaf=5)** to improve generalization and reduce overfitting.

The model was trained on a processed and balanced dataset and evaluated using separate validation and test sets. It **demonstrated strong overall accuracy and performed** especially well in identifying accidents that resulted in casualties.

**Test Set Performance:**

- **Accuracy:** 86.7%
- **Class 0 (No Casualty):** Precision 13.6%, Recall 7.9%, F1-score 10.0%
- **Class 1 (Casualty):** Precision 90.9%, Recall 94.9%, F1-score 92.8%

While the model was highly effective in detecting casualty cases, its ability to identify non-casualty cases was more limited, as reflected by lower precision and recall for class 0.

**Feature Importance:**

- The most influential features in predicting casualties were:
- Week of the year
- Hour of the accident
- Day of the week
- Vehicles involved
- Weather and road conditions (e.g., snow, storm, reckless driving)

These variables had the highest impact on the model's decision-making process and provided insights into the conditions most associated with accident severity.

**Conclusion:**
Random Forest proved to be a powerful model for classifying accident outcomes, offering high accuracy and valuable interpretability through feature importance. It successfully captured complex patterns in the data and provided a strong foundation for predicting the severity of traffic accidents.

---

# 5. Results and discussion

This project evaluated two machine learning models—**Logistic Regression** and **Random Forest**—to predict whether a traffic accident would result in casualties. The models were trained and tested on a cleaned and preprocessed dataset, using a **70% training**, **15% validation**, and **15% test split**. Evaluation focused on **accuracy**, **precision**, **recall**, and **F1-score** across both classes.

The **Logistic Regression** model served as a baseline and, after applying class balancing, achieved:

- **Training Accuracy**: 91.31%
- **Validation Accuracy**: 91.47%
- **Test Accuracy**: 51.5%
- **Recall (Class 0 – No Casualty)**: 51.4%
- **Recall (Class 1 – Casualty)**: 51.5%

This indicated balanced performance between classes, though at the cost of overall accuracy. It identified key predictors such as **Cause_Drunk Driving**, **Weather Condition_Fog**, and **Week of the Year**. However, its predictive strength was limited.

In comparison, the **Random Forest** model demonstrated:

- **Training Accuracy**: 91.31%
- **Validation Accuracy**: 91.47%
- **Test Accuracy**: 86.7%
- **Recall (Class 0)**: 7.9%
- **Recall (Class 1)**: 94.9%

Although it showed higher overall accuracy and strong classification for casualty events, it underperformed on non-casualty cases. Key predictive features included **hour of accident**, **vehicles involved**, and **road/weather conditions**.

**Recommended Actions**:

- Implement more **cost-sensitive learning** or **threshold tuning** to better detect non-casualty events.
- Incorporate additional variables (e.g., traffic density, driver age) for richer prediction.
- Deploy models in **decision-support systems** for traffic agencies to proactively identify high-risk conditions.

This comparison highlights that while Random Forest is more powerful, Logistic Regression offers better class balance. Both models underscore the potential of using predictive analytics for improving road safety.

---

# 6. Summary

This project explored the application of machine learning to predict **casualties in global traffic accidents** using structured accident data. After rigorous **data preprocessing**, including feature engineering and categorical encoding, two models—**Logistic Regression** and **Random Forest**—were developed.

Logistic Regression provided a baseline with interpretable coefficients and was enhanced by handling class imbalance. Random Forest, in contrast, captured complex patterns more effectively and delivered stronger overall performance, particularly for detecting casualty outcomes.

**The project demonstrated:**

- The importance of **class balancing techniques**
- The trade-off between **interpretability and accuracy**
- The value of **feature importance** in understanding contributing factors to accidents

**Key Takeaways**:

- High accuracy can be misleading without examining class-specific metrics
- Feature engineering significantly impacts model performance
- Real-world problems often require iterative tuning and multiple modeling approaches

Overall, this project validates the usefulness of machine learning for traffic accident analysis and opens pathways for further enhancement, especially in **public safety planning** and **automated risk detection** systems.

# 7. Citations


- World Health Organization (WHO). (2021). *Global Status Report on Road Safety 2020*. Retrieved from https://www.who.int/publications/i/item/9789240062884.
- National Highway Traffic Safety Administration (NHTSA). (2021). *Traffic Safety Facts 2020*. U.S. Department of Transportation. Retrieved from https://www.nhtsa.gov/.
- OECD. (2020). *The Economic and Social Impact of Road Accidents*. Organization for Economic Cooperation and Development (OECD). Retrieved from https://www.oecd.org.
- Zhao, S., Zhang, J., & Zhang, Y. (2019). Predicting Traffic Accident Risk Using Random Forest Model. Sustainability, 11(6), 1637. https://doi.org/10.3390/su11061637
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3rd ed.). Wiley. https://doi.org/10.1002/9781118548387