# Road Accidents Fatality Data Report

Novermber 2024

## TABLE OF CONTENT



### 1. Business Understanding

**Objective:**

The primary goal of this project is to develop a machine learning model that can predict the likelihood of fatality occurring in accidents based on specific factors, such as location, time, and other contextual data. The aim is to provide insights into high-risk areas and times, helping to allocate resources efficiently and implement preventive measures. Ultimately, the project seeks to support traffic safety authorities in reducing accidents through data-driven decision-making.

Key Questions:

- What are the main factors contributing to accidents?
- Can we accurately predict accidents and identify high-risk areas or times?
- How can the predictive model be deployed for real-time applications?

2. Data Understanding

Data Collection:

The dataset consists of crash data, including geographical coordinates (latitude and longitude), temporal details, and various categorical and numerical features related to crash characteristics.

Exploratory Data Analysis (EDA):

- Geospatial Analysis: Latitude and longitude data was analyzed to identify high-density crash zones, with a heatmap providing visual insights into crash hotspots.
- Temporal Patterns: Analysis of crash occurrences over time helped to identify high-risk hours and days.
- Feature Distributions: Histograms and pair plots were used to understand the distribution of key features (e.g., crash reports, presence of fatality words, specific vehicle mentions like motorcycles or matatus).
- Correlations: Relationships between features were explored to identify any dependencies or patterns.

Insights from EDA:

- High-density clusters indicate specific geographic regions with frequent accidents.
- Certain times and days have higher accident occurrences, suggesting temporal trends.
- Some keywords (e.g., “fatality”) were noted to be indicators of more severe crashes.

## 3. Data Preparation

**Preprocessing Steps**:

- Feature Engineering: New features were created to capture contextual data, such as indicators for mentions of keywords (e.g., “pedestrian,” “matatu,” “fatality”), and transformations were applied to latitude and longitude data for geospatial analysis.
- Handling Missing Values: Any missing values were addressed through imputation or exclusion, as required.
- Scaling: Numerical features were scaled to ensure model stability and improve performance, especially for distance-based models.
- Encoding: Categorical features were one-hot encoded to make them suitable for machine learning models.
- Splitting: The dataset was split into training and testing sets to validate the models effectively.

**Challenges:**

- Imbalanced Data: Certain categories, such as keywords indicating fatalities, were less represented, leading to a class imbalance.

- High Dimensionality: The large number of categorical variables led to high dimensionality after encoding.

**4. Modeling**

**Model Selection**:

Various models were selected to evaluate different approaches to prediction:

- Baseline: A Dummy Classifier was used as a benchmark.
- Machine Learning Models:
- Random Forest, Logistic Regression, Gradient Boosting, and SVM were trained with cross-validation and hyperparameter tuning.
- Ensemble Method: A Voting Classifier was used to combine predictions from multiple models, enhancing robustness.
- Deep Learning Model: A neural network was implemented to capture complex patterns in the data, showing high predictive accuracy.

**Evaluation Metrics**:

- Accuracy, Precision, Recall, F1-score, and ROC-AUC were used to assess model performance.
- Calibration Curve: Assessed the reliability of probability estimates, ensuring that the predicted probabilities were consistent with observed outcomes.

Model Performance:

- The Random Forest model performed well, with strong accuracy and stability.
- Ensemble Voting Classifier provided robust generalization by combining multiple models.
- The Neural Network achieved high accuracy, indicating potential for future enhancement with more data.

**5. Evaluation**

**Key Insights:**

- High-Risk Areas: Geospatial analysis revealed specific regions with frequent accidents, enabling targeted safety interventions.
- Influential Features: Certain features, such as time of day, location, and specific keywords (e.g., “fatality,” “motorcycle”), were significant predictors of accident severity.

- Model Interpretability: Random Forest and Logistic Regression offered insights into feature importance, aiding interpretability.

**Conclusion:**

The models provided effective predictions, with the Random Forest and ensemble models showing high performance. The deep learning model achieved promising results, suggesting it may be further optimized with additional data.

**Limitations:**

- The dataset's imbalanced nature, particularly around keywords like "fatality" or "pedestrian," could impact predictive performance.
- Some important contextual data may be missing, limiting the models' capacity to capture all accident-related nuances.

## 6. Deployment

**Deployment Strategy:**

- Model Saving: Models and preprocessing objects were serialized using joblib for easy loading during deployment.
- API Development: A REST API was built using Flask to serve model predictions, allowing for integration with web or mobile applications.
- Containerization: A Dockerfile was included to create a containerized version of the API, ensuring consistent deployment across environments.
- Cloud Deployment: Guidelines were provided for deploying the Docker container on a cloud platform (e.g., AWS, Google Cloud), enabling real-time access to predictions.

**Monitoring and Maintenance:**

- Monitoring: Suggested monitoring of API usage and model predictions to ensure performance and identify data drift.
- Retraining: Recommendations for periodic retraining based on new data to maintain model relevance and accuracy.

## 7. Recommendations for Future Work

**Data Collection:**

- Collect more data, particularly with underrepresented keywords or categories, to improve model accuracy and robustness.

**Feature Engineering:**

•Incorporate additional geospatial data, such as proximity to intersections or road types, for enhanced context.

**Advanced Modeling**:

•Explore other ensemble techniques, like Stacking or Boosting, and advanced deep learning architectures to further boost performance.

**Real-Time Data Integration:**

•Integrate real-time data sources, such as weather or traffic conditions, to improve the timeliness and relevance of predictions.

**Stakeholder Communication:**

•Develop a dashboard to visualize high-risk areas, peak times, and model predictions in a user-friendly format for non-technical stakeholders.