**Historic Road Accidents in France – A Study**

Carlos Natale

Stephen Waller

Ehsan Jafari

**Report 1: Exploration, Data Visualization, and Data Pre-Processing**

**Context**

For this project, we are tasked with conducting a detailed analysis of the history of road accidents in France using data from the French government through the Ministry of the Interior and Overseas Territories.

The objective is to analyze the available data, clean and format it technically, and to understand any correlation between data items that could provide insight and statistical significance of the data objects. This can lead to improved predictions of the likelihood of road accidents and their severity. We will also utilize advanced data analytics techniques, such as machine learning, to build predictive models and derive actionable insights.

In the long term, this analysis could benefit both the French population and the French government. Reducing accidents or their severity would have far-reaching benefits due to the socio-economic impact of road accidents—such as road closures, injuries, or even death. Accident prevention will help reduce healthcare costs, alleviate the financial burden on victims and their families, and minimize the economic impact of traffic disruptions. Furthermore, enhanced road safety will lead to increased productivity and economic efficiency.

From a scientific standpoint, this project contributes to road safety research by providing a data-driven approach to understanding accident patterns and their causes. It supports the development of predictive models and safety interventions based on empirical evidence, enhancing our ability to prevent accidents and save lives.

**Objectives**

The main objective of this project/report is to analyze, process, clean, and combine the data as required to create a singular dataset that can be used for complex statistical analysis and modeling and to create predictive models to forecast accident hotspots and severity.

All group members have differing levels of expertise in data science, but we are all working towards improving our knowledge as part of the training with DataScientest.

**Understanding and manipulation of data**

**Framework**

The data used for this project is sourced from the Ministry of the Interior and Overseas territories and is freely published by the French government on the following webpage. [*Bases de données annuelles des accidents corporels de la circulation routière - Années de 2005 à 2022 -* ***data.gouv.fr***](https://www.data.gouv.fr/en/datasets/bases-de-donnees-annuelles-des-accidents-corporels-de-la-circulation-routiere-annees-de-2005-a-2022/)

The data on this site contains details of road accidents from 2005 to 2022 – for the purposes of this project, as a team, we have agreed to focus on the data sets from 2019-2022, as 4 years’ worth of recent data would be able to give a valuable enough insight, and reducing the risk of results being skewed by using more historic data where the severity of an accident may be affected by differing standards in vehicle safety, etc.

The datasets themselves consisted of 4 .csv files per year - “usagers”, “vehicules”, “lieux” and “carcteristiques” - users, vehicles, locations, and characteristics related to an accident.

Considering the years 2019 to 2022, the volume of records for each dataset is:

* characteristics: **218,404** rows
* locations: **218404** rows
* vehicles: **373,584** rows
* users: **494,182** rows

## **Relevance**

As a team we have agreed on the “grav” (Severity) field as the target variable for this project as a whole, this field describes the severity of injuries as a result of this accident. Ranging from “uninjured” to “Fatal” accidents, this target was chosen as the severity of an accident can be directly linked to a number of factors in the datasets, such as location and speed.

There are limitations to the data, as this data is only present where an accident has been logged by a law enforcement unit (Police etc.), so any accidents that have occurred without a report being written, or attendance of law enforcement, would not be present on the dataset(s), so a complete picture may not be possible.

**Relevant Variables:** The fields were selected according to the items described in “Data Cleaning and Processing”.

|  |  |
| --- | --- |
| **Category** | **Columns** |
| **Identifiers** | AccID, vehicleID, num\_veh |
| **Temporal Variables** | birth\_year, day, month, year, time |
| **Spatial Variables** | lat, long |
| **Accident Characteristics Variables** | collision\_type, initial\_impact\_point, fixed\_obstacle, mobile\_obstacle, accident\_situation |
| **Environmental Variables** | lum, atm\_condition |
| **Demographic Variables** | user\_category, gender, age |
| **Vehicle-Specific Variables** | vehicle\_category, motor |
| **Human Factors Variables** | maximum\_speed, manv, seat, reason\_travel, safety\_equipment1 |
| **Target Variable** | **gravity** |
| **Road Characteristics Variables** | route\_category, traffic\_regime, total\_number\_lanes, upstream\_terminal\_number, distance\_upstream\_terminal, plan, surface\_condition, infra, traffic\_direction, reserved\_lane\_code, longitudinal\_profile |

**Target Variable:**

The primary target variable is the severity of injuries sustained, categorized into:

* Indemne (Uninjured)
* Tué (Fatal)
* Blessé hospitalisé (Hospitalized injury)
* Blessé léger (Minor injury)

**Dataset Features:**

The dataset provides a comprehensive view of each accident, including:

* Detailed accident descriptions (time, location, conditions).
* Demographic information of involved parties.
* Vehicle details.

## 

## **Pre-processing and feature engineering**

**Data Cleaning and Processing**

Extensive data cleaning and processing were necessary to prepare the dataset for analysis. The treatment process involved key steps:

* **Merging Datasets**

The preprocessing began by merging several CSV files across four distinct datasets—accidents, locations, users, and vehicles—collected from 2019 to 2022. These files were loaded and consolidated into a single dataset, ensuring all relevant data related to each accident was included for comprehensive analysis.

* **Standardizing Column Names**

To facilitate readability and collaboration, especially within an English-speaking team, all column names were translated from French to English. This ensured consistency and clarity across the different datasets, enabling easier understanding and processing of the data.

* **Handling Missing and Outlier Values & Irrelevant Fields**

Replacement of Specific Values: In cases where the dataset contained `1` to indicate "Not specified" values (such as in the `reason\_travel` column), these were replaced with more meaningful values like `'0'`, converting "Not specified" to "Unknown." Additionally, all occurrences of `1` across the dataset were replaced with `NaN` to standardize missing data representation.

Column Removal: The dataset contained some columns deemed irrelevant for the analysis, such as `id\_usager`. These columns were removed. Furthermore, columns with more than 30% missing values were dropped to ensure the integrity of the data and avoid bias caused by incomplete data.

* **Duplicate and Outlier Removal**

**Duplicates**: Any duplicate records in the dataset were identified and removed to ensure that each accident was represented only once.

**Outliers**: Extreme or implausible values were also treated. For example:

**Speed**: Values lower than 5 km/h or greater than 125 km/h were removed from the `maximum\_speed` column as they were considered unrealistic.

**Age**: Based on statistical testing, values below 0 and above 97 were removed as outliers in the `age` column.

**Geographical Codes:** Outlier values in certain geographical columns, such as `dep\_code`, were also removed based on predefined rules to ensure data consistency.

* **Latitude and Longitude Conversion**

To maintain consistency in geographic coordinates, all commas (`','`) used as decimal points in the latitude and longitude columns were converted to periods (`'.'`). This ensured proper handling of geographic data and prevented errors during subsequent geospatial analysis.

* **Missing Value Imputation**

Missing values in critical fields were addressed through imputation. Rather than simply dropping rows with missing values, imputation was performed based on the distribution of existing values in each column. This allowed the dataset to maintain its integrity without introducing significant bias or reducing the overall size of the dataset.

* **Deletion of Redundant Location Fields**

While the dataset contained multiple fields related to accident location, only the latitude and longitude columns were retained. Other location fields were removed to avoid redundancy and to focus on the most critical geographical information for the analysis.

This preprocessing phase ensured that the data was clean, consistent, and ready for use in subsequent analytical and modeling tasks aimed at predicting accident severity and identifying key factors contributing to road accidents in France.

|  |  |
| --- | --- |
| **Raw Data** | **Processed Data** |
|  |  |

We used Ridge regression to shrink coefficients and reduce overfitting to provide insight into how each feature contributes to the target variable. Below are the regression coefficients for each feature in the model.

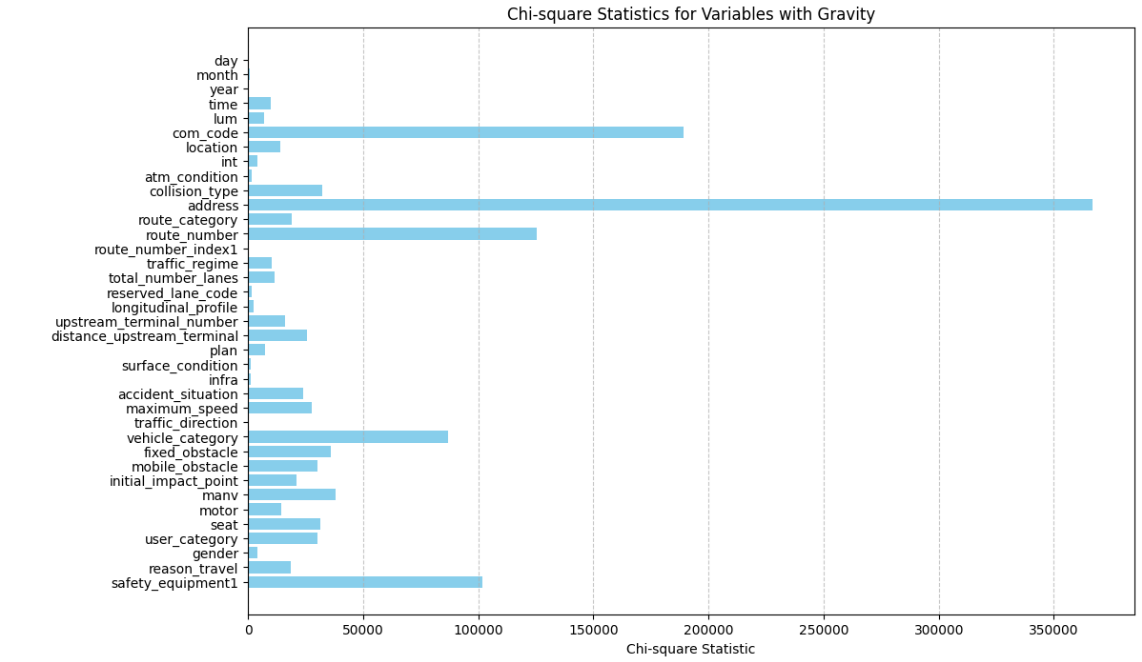
* **Positive Coefficients**: Features with positive coefficients increase the predicted target value (e.g., accident severity) as their values increase.
  + **user\_category (0.366)**: The type of user (driver, passenger, pedestrian) has the strongest positive influence on the target.
  + **motor (0.208)** and **gender (0.196)**: These factors also show significant positive contributions, indicating their importance in predicting accident severity.
  + **maximum\_speed (0.061)**: Higher speed limits slightly increase the predicted severity of accidents.
  + **seat (0.046)**, **fixed\_obstacle (0.044)**, and **plan (0.042)**: These variables contribute positively but less substantially, possibly linked to road conditions or safety measures.
* **Negative Coefficients**: Features with negative coefficients decrease the predicted target value.
  + **mobile\_obstacle (-0.094)**: This has the most significant negative impact, suggesting that mobile obstacles reduce the severity of accidents.
  + **collision\_type (-0.055)**: Certain collision types are associated with lower severity.
  + **reserved\_lane\_code (-0.044)**: Reserved lanes may reduce accident severity due to better traffic management.
  + **age (-0.012)** and **year (-0.016)**: Older individuals and more recent years correlate with decreased accident severity.
* **Near-Zero Coefficients**: Features with coefficients near zero have minimal impact on the target variable.
  + **time (-9.73e-10)** and **upstream\_terminal\_number (-6.92e-05)**: These features are insignificant in predicting accident severity.

In the next step, we will analyze the variables using the correlation **matrix methods, the mutual information (MI) scores, and the feature importance scores of a Random Forest model** to confirm or identify the most significant predictors and understand the relationships between the variables.

After all the pre-Processing, the files have been merged and the total number of records in the Dataset is **447,670** rows.

## **Visualizations and Statistics**

Relationships between our target variable of “grav” (Severity) have been identified via various methods of analysis, for example chi-square analysis (see graph below) has been done on the overall dataset against the “grav” field showing that there are key variables that are significant to be used to model against in the future – and example of this would be the “route number” which gives details on specific roads / road types – which have a correlation with the severity of an accident. As part of modeling, this would be a good data item to use, to see how strong the correlation between road types and severity – a hypothesis of this would be more rural roads would have a higher severity of accident.



**Chart 1 – Number of accidents by Accident Severity**

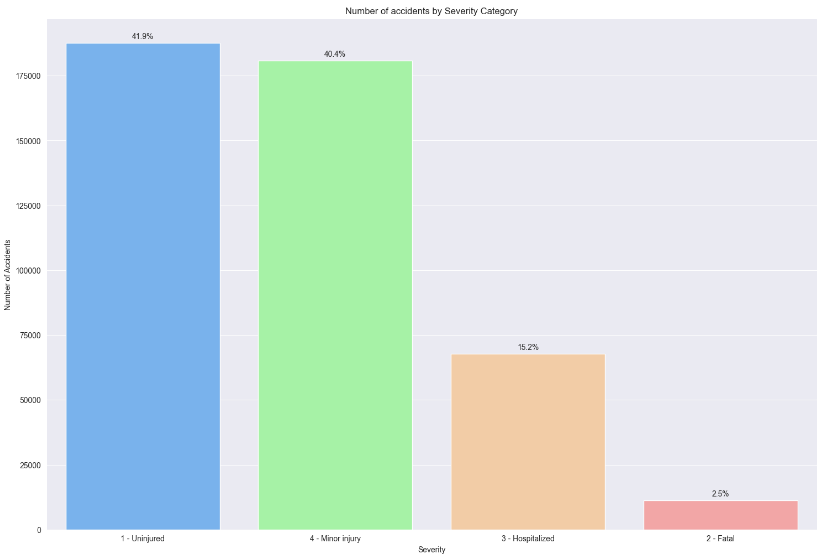


Chart 1 illustrates that the majority of accidents led to either no injury or only minor injuries, accounting for over 80% of the cases. A smaller proportion of accidents resulted in hospitalization, indicating more severe injuries. Fatal accidents are rare, making up a very small percentage of the total. This distribution suggests that while accidents are frequent, the majority do not result in severe harm. Only a small fraction of accidents lead to hospitalization or death.

**Chart 2 – Accident Severity distribution by Year**

|  |  |
| --- | --- |
|  |  |

Chart 2 illustrates a consistent pattern where most accidents result in either no injury or minor injuries, in line with general traffic accident trends. The percentage of fatal accidents is very low but slightly increases over time, which may require closer examination to identify potential causes. The stability in the percentages of hospitalized individuals suggests that while the severity of injuries remains relatively unchanged, the absolute number may vary with the total number of accidents each year.

The combined visualization effectively communicates both the absolute counts and relative percentages of accident severity over the years, providing insights into the nature and trends of traffic accidents during this period.

**Table 1 – Accident Severity distribution by Lightning condition**

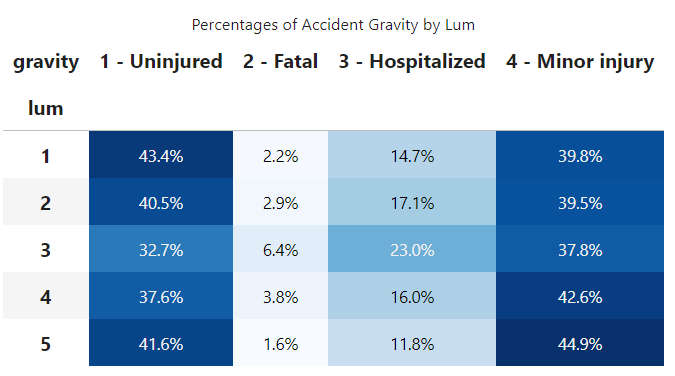


Table 1 indicates a relationship between lighting conditions and the seriousness of accidents. Lum 3 seems to be linked to more severe outcomes (higher fatality and hospitalization rates), which may be linked to poor or challenging lighting conditions (night without public lighting). Lum 1, likely representing optimal lighting (Daylight), shows a higher proportion of uninjured and minor injuries, which aligns with safer driving conditions. Lum 5, while displaying the highest percentage of minor injuries, also has the lowest fatality rate, which could suggest good but not perfect lighting conditions (night with public lighting).

**Chart 3 – Accidents Severity Distribution by Gender**

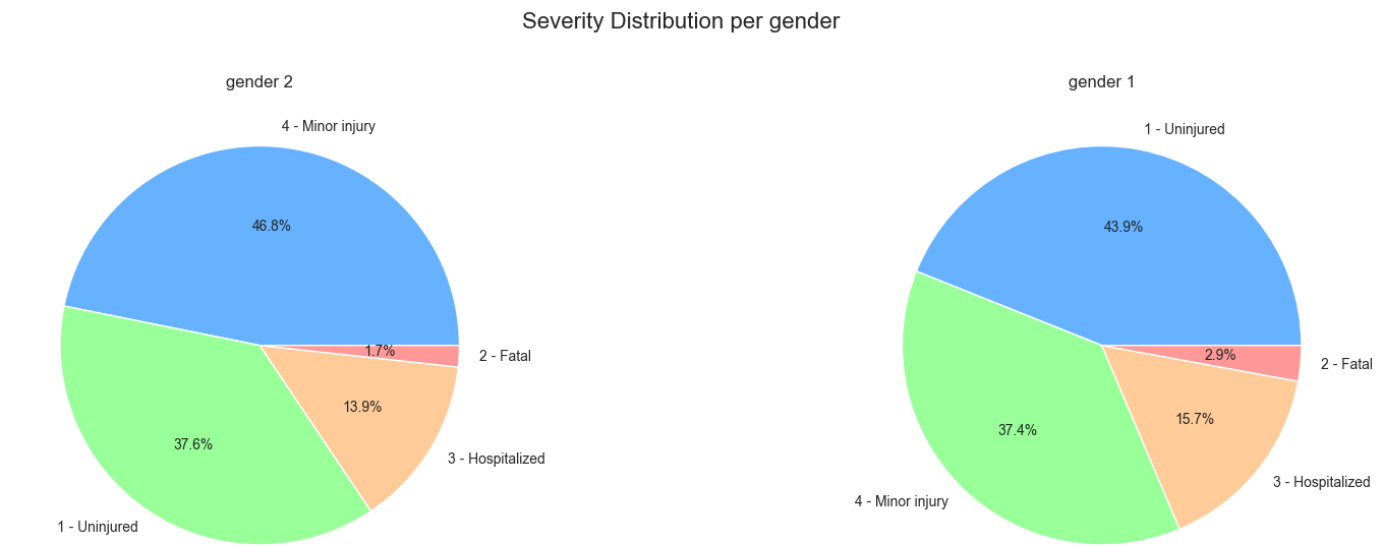


Chart 3 indicates that Gender 2 (Female) has a higher percentage of minor injuries and a lower percentage of fatalities compared to Gender 1 (Male). Gender 1 has a higher percentage of accidents resulting in hospitalization and fatalities, indicating that accidents for this group tend to be more severe. Gender 1 also has a higher distribution of uninjured accidents, while minor injuries are more common in Gender 2. These differences may reflect factors such as driving behavior, risk exposure, or vehicle types associated with each gender.

**Chart 4 – Age Distribution by Accidents Severity**

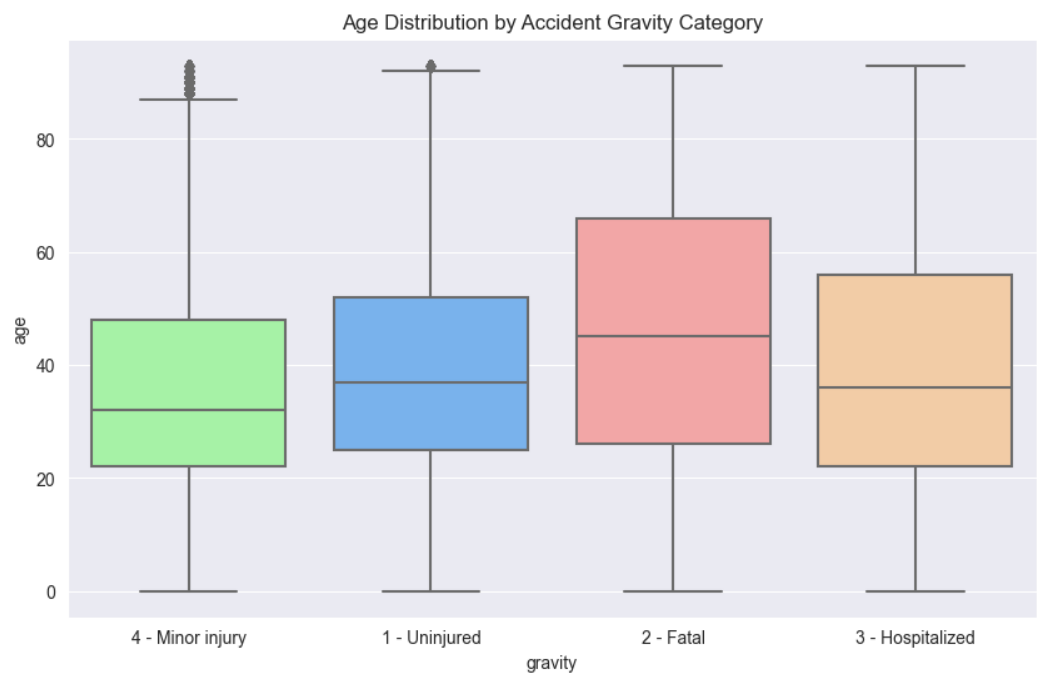


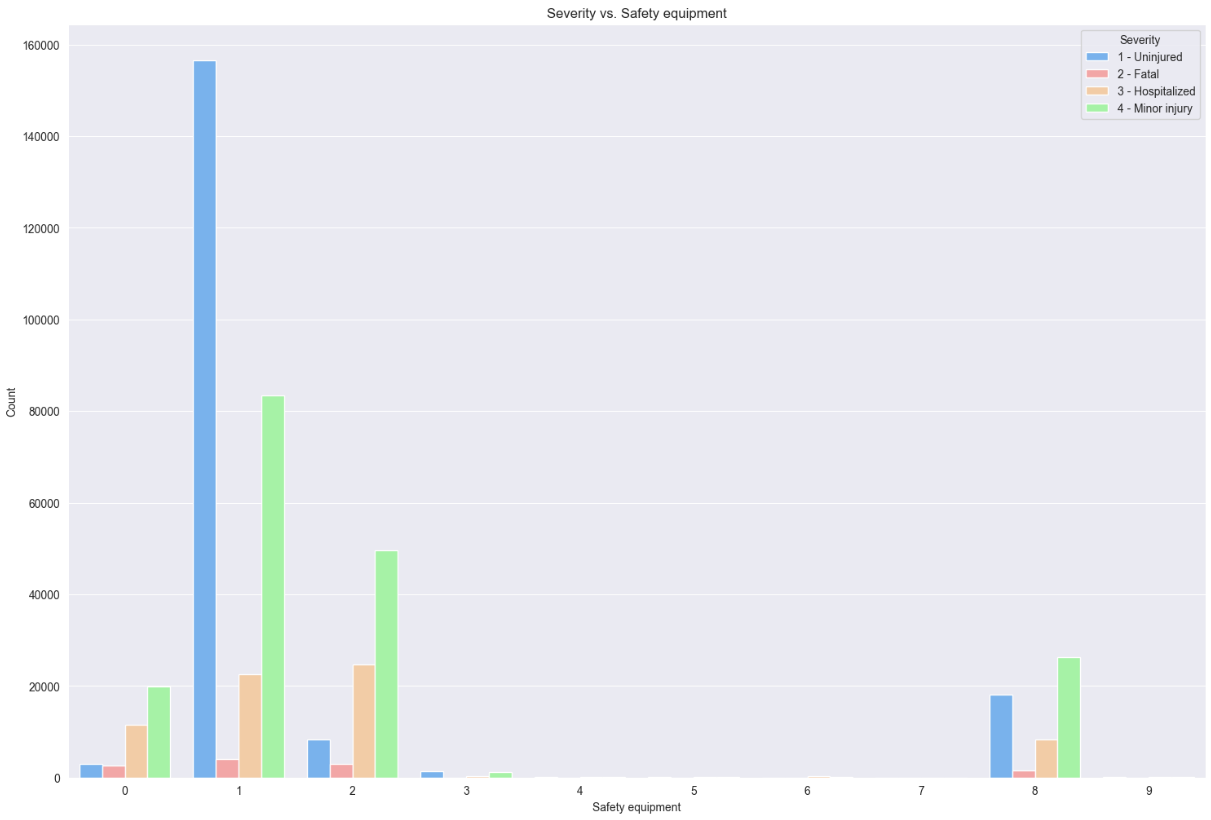
Chart 4 shows that fatal accidents are more common among older individuals, with the highest median age compared to other categories. Uninjured and Minor Injury categories have similar age distributions, with younger median ages. Hospitalization affects a wide age range, with a median age slightly lower than fatal accidents but higher than minor injuries and uninjured cases. The presence of outliers in the Minor Injury and Uninjured categories suggests that accidents affect individuals of all ages, though the severity of outcomes tends to increase with age. This graph suggests the existence of a relationship between age and accident severity, showing that older age groups are more susceptible to severe outcomes such as fatalities.

**Chart 5 – Accident Severity by Age vs. Maximum Speed**



Chart 5 shows that higher speeds consistently result in fatal accidents across all age groups, especially among individuals aged 20-60. Among younger individuals, different speeds indicate a higher risk of severe accidents (fatal or requiring hospitalization) when speed is a factor. As individuals get older, the maximum speed decreases, and the severity of accidents tends to level off, indicating that speed has less influence on accident outcomes. This chart highlights the strong connection between higher speeds and the severity of accidents, especially fatal ones, across most age groups. It emphasizes the crucial role of speed management in reducing severe outcomes, particularly for younger and middle-aged drivers.

**Chart 6 - Accidents Severity Distribution by Safety Equipment**

****

In Chart 6, Category 1: This is the most effective safety equipment, with the highest rate of uninjured individuals, making it the best choice for preventing injuries in accidents. Categories 0, 2, 5, 6, and 9 show varying levels of effectiveness. The chart highlights the importance of safety equipment in reducing the severity of accidents, though it also shows that certain types of equipment (or lack thereof) are associated with a higher incidence of minor injuries and hospitalizations.

**Chart 7 - Accidents Severity Distribution by Seat in accident moment**

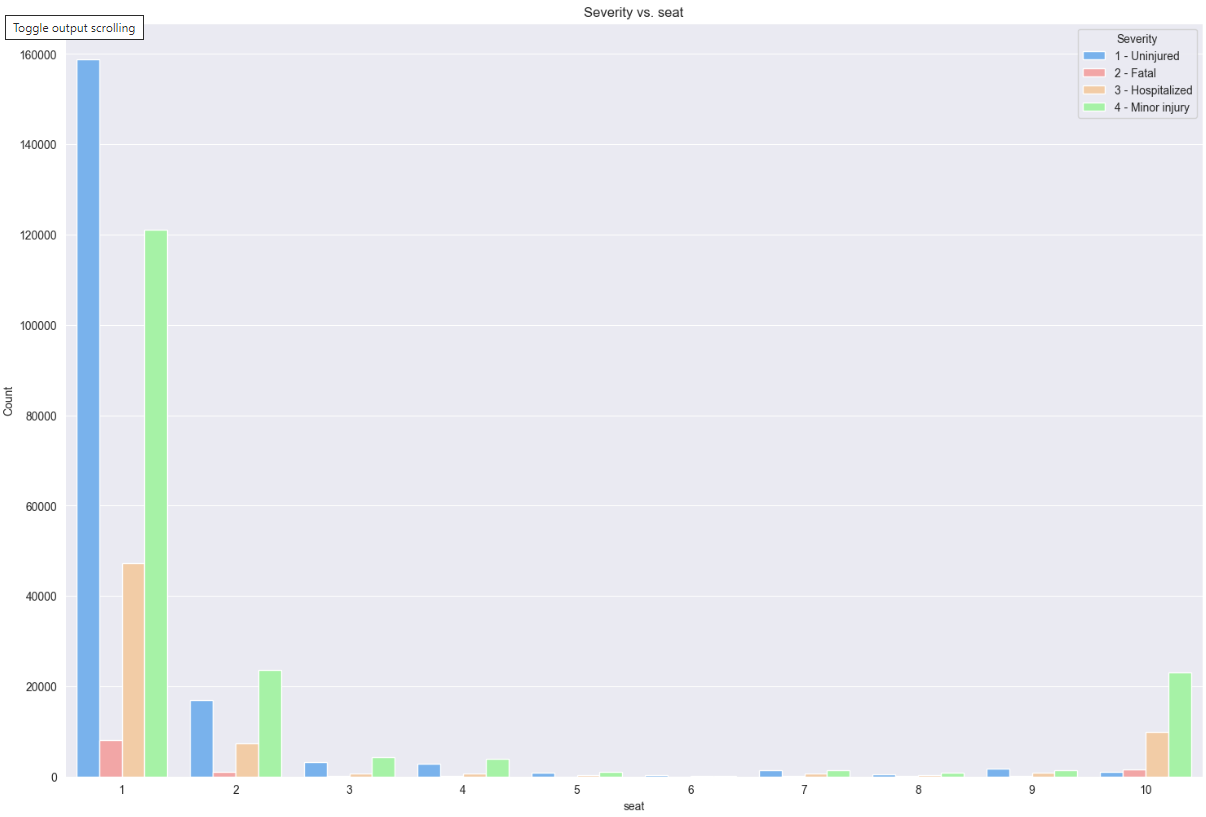
****

Chart 7 indicates that Seats 0 and 1, likely the driver and front passenger seats, have the highest number of accidents. Most of these accidents resulted in no injuries or minor injuries, but there were also a significant number of hospitalizations and some fatalities. Seat 2, possibly a rear seat, shows a more balanced distribution of accident severity outcomes. Seats 3 to 10 had fewer recorded accidents, suggesting that these seats might be less occupied or less prone to severe outcomes. Seat 10 stands out slightly for minor injuries. This chart demonstrates the distribution of accident outcomes across different seat positions, indicating that the front seats are associated with a higher exposure to both minor and severe accidents. This information can be useful in understanding the risks associated with different seat positions and can help inform vehicle safety designs.

**Chart 8 - Accidents Severity Distribution by User Category involved in accident moment**

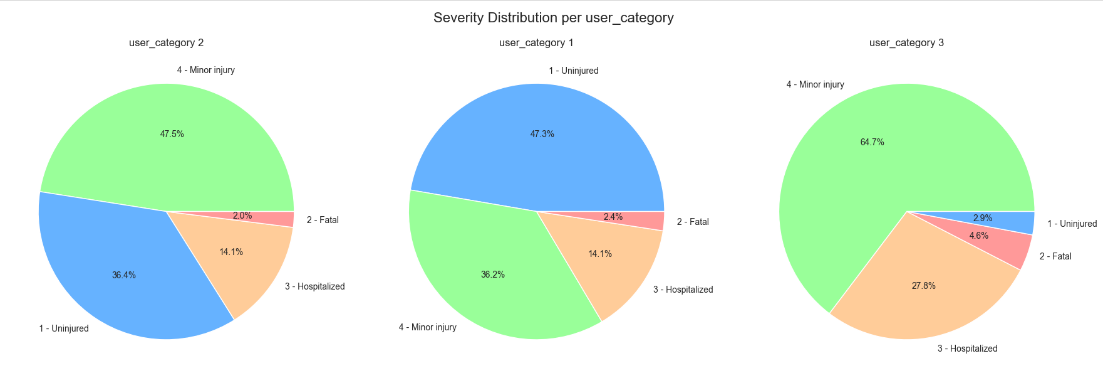


Chart 8, User Category 1 (Driver) appears to be safer, with a higher percentage of uninjured outcomes and lower fatality rates. User Category 2 (Passenger) shows a slightly higher risk compared to User Category 1, with more minor injuries and a similar hospitalization rate. User Category 3 (Pedestrian) seems to be the most vulnerable, with a higher rate of minor injuries, hospitalizations, and fatalities. This suggests that this group, possibly pedestrians or cyclists, faces greater risks in accidents. This visualization highlights the varying risk profiles across different user categories, indicating that safety measures may need to be tailored to better protect more vulnerable groups, such as those in User Category 3.

## 

## **Next stage of modeling**

Now that the final variables have been decided upon, and there is a “clean” (but still raw) dataset that encompasses the variable that are best suited to affect the “grav” (severity) target variable, the next stage of or project is to begin the modeling task to see how effective these variables are in predicting our target.

As part of this modeling process there will be further feature engineering covered - to create more variables - for example a creation of an overall date/time field that covers the accident, as well as averages of these - this feature would then be used as part of historic/date-time modeling, likely using the ARIMA method - this could provide insight as part of the conclusion of this project to see if the severity of accidents are changing over time, as well as the average number in a given time-frame. This would allow predictions into the future, with (hopefully) a reasonable degree of accuracy around possible numbers of accidents, and their predicted severity.

Other field amendments would be amending the categorical data we have in our dataset to prepare for predictive modeling, methods being considered is Label Encoding (setting categories to a unique integer) or Frequency Encoding where the frequency of the data is important for the next stage of modeling. In addition, Normalization/Standardization and Data Balancing techniques would be applied.

Initial use of a standard test/train split would be done, then this outcome would be applied to various modeling techniques available via Python libraries, such as Gradient Boosting, Decision Trees, and Random Forests. This will enable a further prediction of the target “severity” based upon changing variables that feed in, with a high degree of accuracy.

The outcomes of these will be tuned as appropriate and then re-run to give the most accurate predictions possible, using most of the data we have.

## 

## **Conclusion**

This study provides a comprehensive analysis of road accidents in France using recent data from 2019 to 2022. The primary objective was to explore, clean, and preprocess various datasets related to accidents, vehicles, users, and locations to understand the factors influencing accident severity and eventually develop predictive machine learning models that could forecast accident severity.

Extensive data cleaning and preprocessing were crucial steps in this analysis. This involved merging multiple datasets, standardizing column names, handling missing and outlier values, and removing irrelevant fields. Key preprocessing steps included replacing inconsistent values, eliminating duplicates, and converting geographic coordinates to a consistent format.

Our analysis identified significant correlations between accident severity and various factors, such as road types, lighting conditions, demographics, and safety equipment use. Visualization techniques helped uncover patterns, showing, for instance, that fatal accidents are more likely on certain road types and under poor lighting conditions, and that older individuals are more susceptible to severe accident outcomes.

The next phase of the project will involve advanced modeling using machine learning techniques, such as Gradient Boosting, Decision Trees, and Random Forests. Further feature engineering will be conducted, including encoding categorical data to enhance model performance. These models will allow us to predict accident severity with a high degree of accuracy, providing valuable insights for policymakers to implement safety measures aimed at reducing accidents and their associated socio-economic impacts.

Through this data-driven approach, the project aims to contribute significantly to road safety research, offering predictive insights that could help save lives and improve overall traffic safety in France.

**Report 2: Data Processing and ML Models Evaluation**

**Data Preprocessing**

# **Random Forest Implementation and Results Analysis**

**Overview of the Study**

This report outlines the application of a Random Forest algorithm to classify accident severity into two primary categories: Severely Injured (classes 2 and 3) and Slightly Injured (classes 1 and 4). The study focuses on optimizing model performance through normalization techniques and hyperparameter tuning to handle a high-dimensional dataset with numerous features.

**Data Preprocessing and Normalization**

Preprocessing was a crucial step to ensure the model could effectively interpret the input data. The following steps were taken:

* Handling Missing Values: Missing data points were either imputed based on existing distributions or removed based on a missingness threshold.
* Normalization: All features were scaled to fall within a similar range, ensuring that no feature disproportionately affected the model’s decision-making process.
* Handling Imbalanced Classes: The dataset was imbalanced, with fewer slightly injured cases. Adjusting class weights enhanced the Random Forest classifier’s ability to handle class imbalance.

**Model Selection and Hyperparameter Tuning**

A Random Forest classifier was selected due to its ability to handle high-dimensional data and capture complex interactions between features. Key hyperparameters were tuned using a grid search method:

* Number of Trees (n\_estimators): Set to 100 to maintain model performance while avoiding overfitting.
* Maximum Depth of Trees (max\_depth): Set to 10 to ensure the trees do not become overly complex and overfit to the training data.
* Minimum Samples Split (min\_samples\_split): Set to 2, meaning nodes must have at least two samples before split further.
* Class Weights: Adjusted to handle class imbalance, giving higher weight to the severely injured class to improve recall for this critical category.

**Model Training and Evaluation**

**Model Training**

The Random Forest classifier was trained on the normalized dataset, utilizing a 70/30 train-test split. The training process involved growing multiple decision trees, where each tree was trained on a random subset of the data and a random subset of the features.

**Model Evaluation Metrics**

The best performance of the Random Forest model is summarized below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Random Forest** | | | | |
| **Class(gravity)** | **precision** | **recall** | **f1-score** | **support** |
| **1** | 0.73 | 0.84 | 0.78 | 37371 |
| **2** | 0.33 | 0.16 | 0.22 | 2335 |
| **3** | 0.49 | 0.50 | 0.49 | 13737 |
| **4** | 0.70 | 0.61 | 0.65 | 36091 |
| **Non-Fatal** | 0.98 | 0.96 | 0.97 | 87199 |
| **Fatal** | 0.22 | 0.39 | 0.28 | 2335 |
| **Slightly Injured** | 0.91 | 0.90 | 0.91 | 73462 |
| **Severely Injured** | 0.56 | 0.58 | 0.57 | 16072 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Class (gravity)** | **accuracy** | **precision** | | **recall** | | **f1-score** | |
| **macro** | **weighted** | **macro** | **weighted** | **macro** | **weighted** |
| **4 severities** | 0.68 | 0.56 | 0.67 | 0.53 | 0.68 | 0.53 | 0.67 |
| **Non-Fatal vs Fatal** | 0.95 | 0.60 | 0.96 | 0.68 | 0.96 | 0.63 | 0.93 |
| **Slightly Injured Severely Injured** | 0.85 | 0.74 | 0.85 | 0.74 | 0.85 | 0.74 | 0.85 |

The model performed best in identifying non-fatal cases (classes 1 and 4) with an impressive F1-score of 0.97. However, performance drops significantly for the severely injured cases (classes 2 and 3), with an F1-score of 0.57. While recall for the severely injured category improved, precision remains a challenge.

**Analysis of Results**

The Random Forest model showed strong performance in classifying non-fatal and slightly injured cases. However, certain patterns and limitations were observed:

* Severely Injured Classification: The model's ability to detect severely injured cases improved with a recall of 0.58. However, the precision remains low (0.56), suggesting that many non-severely injured cases are being misclassified as severe.
* Slightly Injured Classification: The model demonstrated excellent performance in identifying slightly injured cases, with an F1 Score of 0.91.
* Fatal Case Detection: Identifying fatal accidents remains challenging, with a precision of 0.22 and an F1-score of 0.28.

**Threshold Tuning for Severely Injured Cases**

Further tuning the classification threshold for the severely injured cases improved recall, though at the expense of precision.

Adjusted Threshold Results for Severely Injured:

* Precision: 0.56
* Recall: 0.58
* F1-Score: 0.57

This adjustment improved recall for the severely injured group, ensuring the model captured more true positives, but precision remains lower than desired.

**Strengths and Weaknesses of the Model**

Strengths:

* High Accuracy for Slightly Injured Cases: The model achieved strong performance for non-fatal and slightly injured cases, with a high F1-score of 0.91.
* Feature Interpretability: Random Forest's feature importance ranking provided transparency in the decision-making process, offering insights into the most influential features.
* Resilience to Overfitting: The model’s structure and hyperparameter tuning reduced the risk of overfitting, even with high-dimensional data.

Weaknesses:

* Severely Injured Case Classification: Despite adjustments, the model struggled to identify severely injured cases, particularly with precision, which remains relatively low.
* Class Imbalance: The dataset's class imbalance, despite weight adjustments, continues to challenge the model's performance for the underrepresented classes.

**Conclusion**

The Random Forest model, combined with normalization and hyperparameter tuning, strongly predicted accident severity. While the accuracy and precision for non-fatal and slightly injured cases are high, the model faces challenges in identifying severely injured and fatal cases. Future work should address these challenges through advanced sampling, ensemble methods, and continued feature engineering to enhance performance further.

# **XBoost Implementation and Results Analysis**

**Overview of the Study**

This report outlines the application of the XGBoost algorithm to classify accident severity into two primary categories: Severely Injured (classes 2 and 3) and Slightly Injured (classes 1 and 4). The XGBoost model was chosen due to its strong performance in classification tasks, particularly when handling imbalanced datasets. The study focuses on optimizing model performance through normalization techniques, hyperparameter tuning, and evaluation of classification results.

**Data Preprocessing and Normalization**

To ensure the model can effectively interpret the input data, comprehensive data preprocessing was conducted:

* Missing Values: To ensure data quality, any missing values were imputed or removed based on a predefined threshold.
* Normalization: Normalization was applied to standardize the range of input features, ensuring that no feature disproportionately influenced the model’s predictions.
* Handling Class Imbalance: The dataset had a significant class imbalance, especially with fewer severely injured cases. To mitigate this imbalance, adjustments were made to the model using techniques such as weight balancing.

**Model Selection and Hyperparameter Tuning**

XGBoost was chosen because of its ability to efficiently handle classification problems with imbalanced data. Key hyperparameters were tuned to enhance model performance:

* Learning Rate: Set to control the step size of updates during training, ensuring convergence to the optimal solution.
* Maximum Depth: The maximum depth of the trees was set to limit the complexity of the model and prevent overfitting.
* Number of Boosting Rounds: The number of boosting rounds was optimized to ensure the model received sufficient training without overfitting to the data.
* Class Weights: The model was trained with custom class weights, giving more importance to the severely injured class.

**Model Training and Evaluation**

Model Training

The XGBoost model was trained using normalized data, with the training set split into 70% training and 30% testing. The model utilized gradient boosting to iteratively improve performance, focusing on misclassified instances.

Model Evaluation Metrics

The performance of the XGBoost model is summarized in the table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **XGBoost** | | | | |
| **Class(gravity)** | **precision** | **recall** | **f1-score** | **support** |
| **1** | 0.74 | 0.81 | 0.77 | 37371 |
| **2** | 0.30 | 0.17 | 0.22 | 2335 |
| **3** | 0.48 | 0.46 | 0.47 | 13737 |
| **4** | 0.67 | 0.62 | 0.64 | 36091 |
| **Non-Fatal** | 0.98 | 0.98 | 0.98 | 87199 |
| **Fatal** | 0.25 | 0.27 | 0.26 | 2335 |
| **Slightly Injured** | 0.95 | 0.79 | 0.86 | 73462 |
| **Severely Injured** | 0.46 | 0.82 | 0.59 | 16072 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Class (gravity)** | **accuracy** | **precision** | | **recall** | | **f1-score** | |
| **macro** | **weighted** | **macro** | **weighted** | **macro** | **weighted** |
| **4 severities** | 0.67 | 0.54 | 0.66 | 0.46 | 0.62 | 0.53 | 0.66 |
| **Non-Fatal vs Fatal** | 0.96 | 0.62 | 0.96 | 0.63 | 0.96 | 0.62 | 0.96 |
| **Slightly Injured Severely Injured** | 0.80 | 0.71 | 0.87 | 0.81 | 0.80 | 0.73 | 0.82 |

The XGBoost model demonstrated strong performance for non-fatal (slightly injured) cases, with an F1-score of **0.98**. However, the model faced challenges in identifying fatal and severely injured cases, as shown by the lower precision and recall for these categories.

**Analysis of Results**

* **Severely Injured Classification**: The XGBoost model achieved a recall of **0.82** for the severely injured category, which improved over previous models. However, the precision remained lower at **0.46**, indicating that many non-severely injured cases were misclassified as severe. This resulted in an F1-score of **0.59**.
* **Slightly Injured Classification**: The model performed very well in identifying slightly injured cases, with an F1-score of **0.86** and a recall of **0.79**. These metrics indicate that the model can reliably predict slightly injured instances, effectively balancing precision and recall.
* **Fatal Case Detection**: The model's performance in identifying fatal accidents (classes 2 and 3) remains limited, with an F1-score of **0.26** and precision at **0.25**. This indicates significant room for improvement in detecting more severe cases, where misclassifications are frequent.

**Threshold Tuning for Severely Injured Cases**

To further enhance recall for the severely injured category, threshold tuning was applied. The adjustments improved recall but came at the cost of precision:

* **Adjusted Threshold Results for Severely Injured**:
  + Precision: **0.46**
  + Recall: **0.82**
  + F1-Score: **0.59**

The threshold adjustment ensured more severely injured cases were captured, with a substantial boost in recall. However, the lower precision highlights that the model still struggles with false positives in this category.

**Strengths and Weaknesses of the Model**

Strengths:

* Strong Performance for Slightly Injured Cases: The model achieved high accuracy and recall for the slightly injured class, which was expected due to the more significant number of instances.
* Efficiency of XGBoost: XGBoost's gradient boosting technique helped the model efficiently focus on the most challenging cases.
* Feature Importance Insights: XGBoost provides feature importance scores, allowing a better understanding of which features were most influential in the classification decisions.

Weaknesses:

* Performance on Severely Injured Cases: Despite tuning, the model struggled with precision for severely injured cases, primarily due to the class imbalance.
* Class Imbalance: Even with class weights, the XGBoost model faced challenges in classifying the minority class, with a lower F1-score for the fatal category.

**Conclusion**

The XGBoost model, in conjunction with normalization and hyperparameter tuning, provided strong performance for classifying accident severity. While the model achieved high accuracy and recall for slightly injured cases, it faced challenges with the severely injured class. Further improvements through advanced sampling techniques and continued model optimization are recommended to enhance recall and precision for severely injured cases.

# **AdaBooster Classification Implementation and Results Analysis**

**Overview of the Study**

This report presents the results of using the AdaBoost algorithm to classify accident severity into two categories: Severely Injured (classes 2 and 3) and Slightly Injured (classes 1 and 4). AdaBoost was chosen due to its ability to focus on misclassified instances and its flexibility in ensemble methods. The study incorporates data normalization, hyperparameter tuning, and performance evaluation to improve classification accuracy.

**Data Preprocessing and Normalization**

Preprocessing was critical to ensure data quality and model performance:

* **Handling Missing Data**: Missing values were addressed using imputation methods or were removed when appropriate based on thresholds.
* **Normalization**: Features were scaled using normalization techniques to ensure no feature disproportionately influenced the classifier.
* **Addressing Class Imbalance**: The dataset was imbalanced, with fewer instances in the severely injured category. The class weights were adjusted to give more importance to the underrepresented class.

**Model Selection and Hyperparameter Tuning**

AdaBoost was selected due to its ability to combine weak learners and iteratively correct misclassified samples. The following hyperparameters were optimized to improve model performance:

* **Number of Estimators**: The number of weak learners was set to ensure sufficient boosting iterations while controlling for overfitting.
* **Learning Rate**: This parameter was adjusted to control the contribution of each weak learner and to ensure that the ensemble model converges efficiently.
* **Base Estimator**: The default base estimator used was a decision tree with shallow depth, optimized for identifying simple patterns in the data.

**Model Training and Evaluation**

**Model Training**

The AdaBoost model was trained on the normalized data using a 70/30 train-test split. The model focused on improving performance for the severely injured class, which was underrepresented in the dataset. AdaBoost iteratively adjusted the weights of misclassified instances, allowing the model to pay more attention to harder-to-classify examples.

**Model Evaluation Metrics**

The AdaBoost model’s performance metrics are summarized below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **AdaBoostClassifier** | | | | |
| **Class(gravity)** | **precision** | **recall** | **f1-score** | **support** |
| **1** | 0.73 | 0.79 | 0.76 | 37371 |
| **2** | 0.13 | 0.62 | 0.22 | 2335 |
| **3** | 0.38 | 0.42 | 0.40 | 13737 |
| **4** | 0.71 | 0.46 | 0.55 | 36091 |
| **Non-Fatal** | 0.98 | 0.98 | 0.98 | 87199 |
| **Fatal** | 0.16 | 0.16 | 0.16 | 2335 |
| **Slightly Injured** | 0.95 | 0.77 | 0.85 | 73462 |
| **Severely Injured** | 0.43 | 0.80 | 0.56 | 16072 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Class (gravity)** | **accuracy** | **precision** | | **recall** | | **f1-score** | |
| **macro** | **weighted** | **macro** | **weighted** | **macro** | **weighted** |
| **4 severities** | 0.59 | 0.49 | 0.65 | 0.57 | 0.59 | 0.48 | 0.61 |
| **Non-Fatal vs Fatal** | 0.96 | 0.57 | 0.96 | 0.57 | 0.96 | 0.57 | 0.96 |
| **Slightly Injured Severely Injured** | 0.78 | 0.69 | 0.85 | 0.78 | 0.78 | 0.70 | 0.80 |

The AdaBoost model performed well in classifying slightly injured and non-fatal cases, with an F1-score of **0.98** for non-fatal cases. However, the classification of severely injured cases remains challenging due to the lower precision and recall for this group.

**Analysis of Results**

* **Severely Injured Classification**: The AdaBoost model achieved a recall of **0.80** for severely injured cases, which indicates the model was able to capture most of these cases. However, the precision was relatively low at **0.43**, leading to a lower F1-score of **0.56**, as many non-severely injured cases were misclassified as severe.
* **Slightly Injured Classification**: The model demonstrated strong performance in identifying slightly injured cases, achieving an F1-score of **0.85** with high precision (**0.95**) and recall (**0.77**). This shows that the model consistently classified slightly injured cases accurately.
* **Fatal Case Detection**: The model's performance detecting fatal accidents (classes 2 and 3) was suboptimal. With an F1-score of **0.16** for fatal cases, further tuning and balancing are required to improve this classification.

**Threshold Tuning for Severely Injured Cases**

To improve recall for the severely injured class, the classification threshold was adjusted. The adjustments helped increase recall at the cost of precision:

* **Adjusted Threshold Results for Severely Injured**:
  + Precision: **0.43**
  + Recall: **0.80**
  + F1-Score: **0.56**

This adjustment helped capture more severely injured cases, but precision remains a limiting factor as the model still struggles with false positives in this category.

**Strengths and Weaknesses of the Model**

**Strengths:**

* **High Accuracy for Slightly Injured Cases**: AdaBoost performed well in the non-fatal class, with an F1 Score of **0.97**.
* **Focused Learning on Hard Cases**: The model effectively adjusted its learning to focus on misclassified instances, improving its ability to capture difficult cases.
* **Interpretability**: AdaBoost allows straightforward interpretation of how weak learners are combined to form the final classification decision.

**Weaknesses:**

* **Challenges with Severely Injured Cases**: The model’s precision for severely injured cases remains relatively low, despite efforts to adjust class weights and thresholds.
* **Class Imbalance**: The significant imbalance in the dataset affected the model’s ability to detect the minority class (severely injured and fatal accidents).

**Conclusion**

The AdaBoost model performed strongly in predicting slightly injured and non-fatal cases. However, severely injured and fatal cases remain challenging, with room for improvement in precision. Future work should focus on balancing the dataset more effectively and applying advanced ensemble techniques to improve classification accuracy for the minority class.