**UK Road Safety Analysis 2023**

Understanding Accidents & Predicting Severity

**Table of contents**

[Introduction 3](#_Toc182929526)

[Research Problem 3](#_Toc182929527)

[Research Objectives 3](#_Toc182929528)

[Methodology and Framework 4](#_Toc182929531)

[Data Collection 4](#_Toc182929532)

[Data Cleaning 4](#_Toc182929533)

[Data Analysis 5-12](#_Toc182929534)

[Python Analysis 5-12](#_Toc182929535)

[Tableau Analysis 5-12](#_Toc182929536)

[Statistical Analysis…………………………………………………………………………...13  
 Descreptive Analysis 13](#_Toc182929537)  
 Hypothetical testing………………………………………………………………… 13-18

[Modeling](#_Toc182929538) 18-20

Challenges faced…………………………………………………………………………..20

[Conclusion 20-21](#_Toc182929539)

[Model selection, Tuning, Deployment 21-22](#_Toc182929540)

References……………………………………………………………………………………22

Acknowledgement……………………………………………………………………………22

**Group 8:**

Abdul Adnan

Sahil Sahil

Shubham Shubham

Vaibhav Vaibhav

Mohit Rexwal

Shalu Choudhary

**Introduction:**Road accidents are a major public safety concern, causing injuries, fatalities, and economic losses. This project by **Group 8** analyses and predicts road accident severity in the UK for 2023 using **machine learning and data visualization**.

We utilize **Python for data preprocessing, Tableau for insights, and PyCaret for model selection**, addressing challenges like class imbalance with **SMOTE**. Statistical methods help identify key accident factors, such as weather, road conditions, and speed limits.

Our goal is to **build accurate predictive models** to classify accidents as Minor, Serious, or Fatal, providing **data-driven recommendations** to improve road safety policies.  
  
**Research Problem**

Road accidents are a major public concern, causing significant injuries, fatalities and economic losses. Understanding the factors that contribute to these accidents is crucial for developing key strategies that improve road safety.   
  
The goal of this business problem is to uncover key metrics and trends to design focused safety interventions, including road enhancements, driver education initiatives, and enforcement strategies. By identifying the factors contributing to road accidents, we aim to support authorities and policymakers in implementing data-driven solutions to reduce accident severity and improve overall road safety.

## **Research Objective**

By leveraging data and AI, we aim to uncover patterns and build a system that can predict accident severity and support decision-making for safer roads. Our project focuses on predicting road accident severity whether an accident is likely to be Minor, Serious, or Fatal, using machine learning models.

* The key objective is to build a highly accurate model that can predict accident severity, and to understand which factors contribute most to severe accidents.
* To create a data visualization dashboard highlighting important insights.
* Train and test various machine learning models to predict accident severity.
* Provide data-driven suggestions for accident prevention and mitigation.

# **Methodology and Framework**

## **Data Collection:**

We obtained our dataset from the **UK government website** and the **Road Transport Department of the UK**. The dataset contains 104,259 rows and 36 columns. The key attributes include:

* Accident Severity: Minor, Serious, Fatal
* Number of vehicles involved
* Weather & Road Conditions
* Speed limits
* Road Types
* Location Data (Longitude, Latitude)

## **Data Cleaning:**

To prepare the data for modeling, we performed the following steps:

* **Handling Missing Data:** Removed incomplete records to enhance model reliability.
* **Class Imbalance Handling:** Used **SMOTE (Synthetic Minority Over-sampling Technique)** to address imbalanced classes in 'accident severity'. It was used to balance the classes, resampling all to **63,392 each**. This ensures the model learns patterns from **Serious** and **Fatal** cases, improving its prediction accuracy for rare but critical events

# **Data Analysis**

## **EDA Using Python:**

### Plot 1

A graph of accident with multiple colored squares

AI-generated content may be incorrect.

### Plot 2

A graph of blue rectangular shapes

AI-generated content may be incorrect.

### Plot 3

A graph of a accident

AI-generated content may be incorrect.

### Plot 4

A graph of accident

AI-generated content may be incorrect.

### Plot 5

A red dot on a white background

AI-generated content may be incorrect.

### Plot 6

A graph of a number of people with different colored bars

AI-generated content may be incorrect.

### Plot 7

A graph of a speed limit

AI-generated content may be incorrect.

### Plot 8

A graph of a number of vehicles

AI-generated content may be incorrect.

### Plot 9

A graph of a graph

AI-generated content may be incorrect.

### Plot 10

A screenshot of a weather forecast

AI-generated content may be incorrect.

### Plot 11

A graph of different colored bars

AI-generated content may be incorrect.

### Plot 12

A graph with text on it

AI-generated content may be incorrect.

### Plot 13

A graph of different colored rectangles

AI-generated content may be incorrect.

### Plot 14

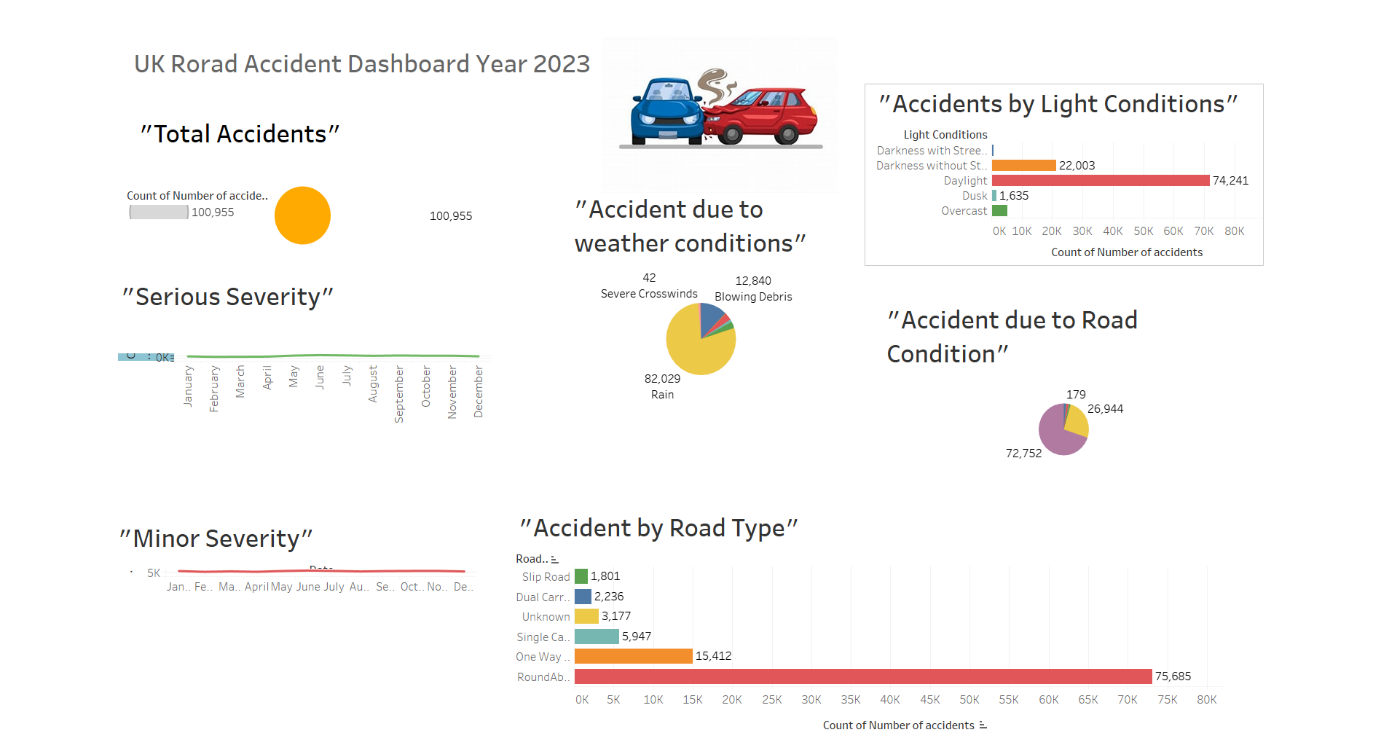
A screenshot of a calendar

AI-generated content may be incorrect.

**Our exploratory data analysis revealed the following insights**:

* The majority of accidents are significantly **Minor** than any other category.
* **Saturday** has the highest number of accidents, suggesting weekends might be riskier due to increased traffic, leisure activities, or impaired driving.
* Most **Minor** and **Serious** accidents occur on **wet/damp roads**, with fewer on dry or icy surfaces.
* Accidents peak between **3 PM and 6 PM**, with fewer incidents during early morning hours.
* A geographic plot shows accident locations, densely clustered in **central and southern regions**.
* Most accidents happen during **rain**, with fewer incidents in severe weather like hail, fog, or snow.
* **Higher speed limits** are linked to more severe accidents, with **fatal accidents** occurring at higher average speeds.
* Accidents with more vehicles tend to have more casualties, though most incidents involve fewer than **6 vehicles**.
* The number of accidents fluctuates throughout the year, peaking around **June** and dipping in **February**.
* Most **Minor** accidents happen during **daylight or rain**, with some occurring in **darkness without streetlights** or overcast conditions.
* **Rounabout** has maximum number of accidents than **one way street.**
* Construction areas has more minor accidents than serious accidents.

**Tableau Analysis:**



**Statistical Analysis:**

**Descriptive Analysis:**A screenshot of a graph

AI-generated content may be incorrect.

**Hypothesis testing:**

A blue background with white text

AI-generated content may be incorrect.

The **Chi-Square test** was performed to examine the relationship between **weather conditions and accident severity** as well as **light conditions and accident severity**. The test yielded the following results:

1. **Weather conditions vs. Accident severity**
   * **p-value:** 1.35e-74 (an extremely small value)
   * **Interpretation:** This indicates a **highly significant relationship** between weather conditions and accident severity. In simple terms, the chances of this relationship occurring purely by random chance are astronomically small. Therefore, we can confidently say that weather conditions—such as rain, snow, or fog—have a meaningful impact on the severity of road accidents.
2. **Light conditions vs. Accident severity**
   * **p-value:** 1.18e-132 (even smaller than the previous test)
   * **Interpretation:** This shows an **even stronger statistical significance** between light conditions and accident severity. Essentially, the relationship is so strong that we can firmly conclude that light conditions—whether it’s daylight, darkness with or without streetlights—are closely tied to how severe an accident is. Poor lighting, for example, seems to heighten the severity of accidents.

In both cases, the extremely small p-values indicate that weather and light conditions are not independent of accident severity they play a crucial role in determining how serious an accident might be.

A screenshot of a test

AI-generated content may be incorrect.

**Weather Conditions vs. Accident Severity:**

* **Hypothesis Decision:** Since the p-value is **much smaller than 0.05**, we **reject the null hypothesis (H₀)**.
* **Conclusion:** This result shows **weather conditions have a significant impact on accident severity**. The relationship is not due to random chance, meaning factors like rain, fog, snow, or clear skies clearly influence how severe an accident is.

**Light Conditions vs. Accident Severity:**

* **Hypothesis Decision:** With a p-value well below **0.05**, we confidently **reject the null hypothesis (H₀)**.

**Conclusion:** There is a strong, statistically significant relationship between light conditions and accident severity. Whether it’s daylight, darkness with streetlights, or complete darkness without streetlights these factors clearly play a crucial role in determining how severe accidents tend to be.

* In both tests, rejecting the null hypothesis means that weather and light conditions are not independent of accident severity — they are, in fact, key factors influencing how serious accidents are. This reinforces the idea that road safety measures must account for both environmental and visibility factors to mitigate accident severity.

**Future Selection**

**Random Forest**

A screenshot of a computer

AI-generated content may be incorrect.

To identify the most impactful variables for predicting **accident severity**, we applied multiple **feature selection techniques** — including the **LASSO Regression**, and **Random Forest**. These methods allowed us to isolate the features that have the strongest influence on accident outcomes.

The **top 5 features** identified were:

* **Speed Limit** (Coefficient: 0.04): Higher speed limits often correlate with more severe accidents, emphasizing the critical role of speed regulations.
* **Junction Control** (Coefficient: 0.027): The presence or absence of junction controls (like traffic lights or stop signs) influences accident severity, particularly at intersections.
* **Number of Vehicles** (Coefficient: 0.020): Accidents involving multiple vehicles tend to have more severe outcomes.
* **First Road Class** (Coefficient: 0.0151): The classification of the primary road (motorway, A-road, etc.) affects accident risk and severity.
* **Second Road Class** (Coefficient: 0.0135): The type of intersecting or secondary road also plays a role in determining accident impact.
* **First Road Number** (Coefficient: 0.0131): Specific Road identifiers may indicate high-risk zones or accident-prone routes.

These features were selected based on their contribution to the model’s predictive accuracy, ensuring our model focuses on the most influential factors when determining accident severity.

**LASSO**

A graph with blue lines

AI-generated content may be incorrect.

The **Embedded Method** was used for feature selection, identifying the most impactful variables for predicting **accident severity**. This method integrates feature selection directly into the model training process, prioritizing features that contribute the most to model accuracy.

The results highlighted the following as the most important features:

* **Speed Limit:** Plays a crucial role in accident severity, with higher speed limits often correlating with more severe outcomes.
* **Number of Vehicles:** Directly influences accident dynamics — more vehicles involved typically result in a higher risk of serious or fatal accidents.
* **Number of Casualties:** Strongly linked to accident severity, as incidents with more casualties are generally classified as more severe.

These features were selected based on their **strong contribution to model performance**, meaning they have a significant impact on predicting accident severity.

**Modeling**

**PyCaret Implementation:**

A screenshot of a computer

AI-generated content may be incorrect.

**A screenshot of a computer screen

AI-generated content may be incorrect.**

**Results**

In our project, we utilized **PyCaret**, an open-source low-code machine learning library, to streamline the process of model selection and evaluation for predicting accident severity. PyCaret automatically compared multiple classification models using essential performance metrics such as Accuracy, AUC (Area Under the Curve), Recall, Precision, F1 Score, Kappa, and MCC.

Our target variable, 'accident\_severity', was imbalanced, with a majority of cases categorized as 'Minor' and fewer labeled as 'Serious' and 'Fatal'. To address this, we applied SMOTE (Synthetic Minority Over-sampling Technique) to balance the classes, ensuring that the model did not favor the majority class.

From PyCaret's model comparison, we observed that the Extra Trees Classifier (et) achieved the highest accuracy (81.50%) and strong recall (81.50%), making it the most reliable model for identifying accident severity across all classes. The Random Forest Classifier (rf) followed closely, also performing well in terms of both accuracy and AUC.

The selection of the best model was not solely based on accuracy; we carefully considered recall and F1 score since predicting severe and fatal accidents correctly is more critical than merely optimizing overall accuracy. A model with high recall reduces the risk of underestimating the severity of an accident — an important aspect of public safety.

Ultimately, PyCaret's automated approach helped us efficiently test and rank models, guiding us to choose Extra Trees as the final model due to its balanced performance across all key metrics, especially for the minority classes. This model will be instrumental in predicting accident severity with higher confidence, allowing us to focus on preventing the most critical accidents.

**Challenges Faced**

Throughout the project, we encountered and addressed several challenges:

* **Class Imbalance:** The 'Fatal' category was significantly underrepresented, requiring the use of **SMOTE**.
* **Noise in Data:** Outliers and irrelevant features had to be carefully handled to prevent skewed model performance.
* **Model Overfitting:** Simpler models tended to underfit, while complex ones risked overfitting, necessitating fine-tuned regularization.

**Conclusion:**

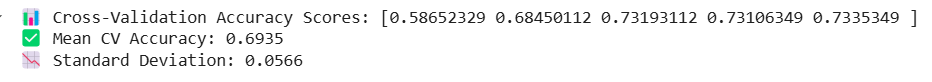
* In this project, we have successfully analyzed a dataset related to road accidents, with the goal of predicting the severity of accidents. Using PyCaret, we preprocessed the data by removing incomplete records, encoding categorical variables, and discretizing continuous variables into deciles. These preprocessing steps improved the performance of our classification model by streamlining the feature set and reducing noise.
* Our model demonstrated satisfactory performance in predicting accident severity, with metrics such as accuracy, precision, recall, and F1-score providing a comprehensive evaluation of its effectiveness. The confusion matrix and AUC-ROC curve further confirmed that the model is capable of distinguishing between different levels of accident severity. While the model performed well, the results highlighted the importance of data quality, especially in dealing with missing values and imbalanced classes.
* This project provides valuable insights into how data preprocessing and machine learning techniques can be applied to real-world road safety problems. The model can potentially aid authorities in identifying high-risk areas or improving response strategies in the event of accidents.

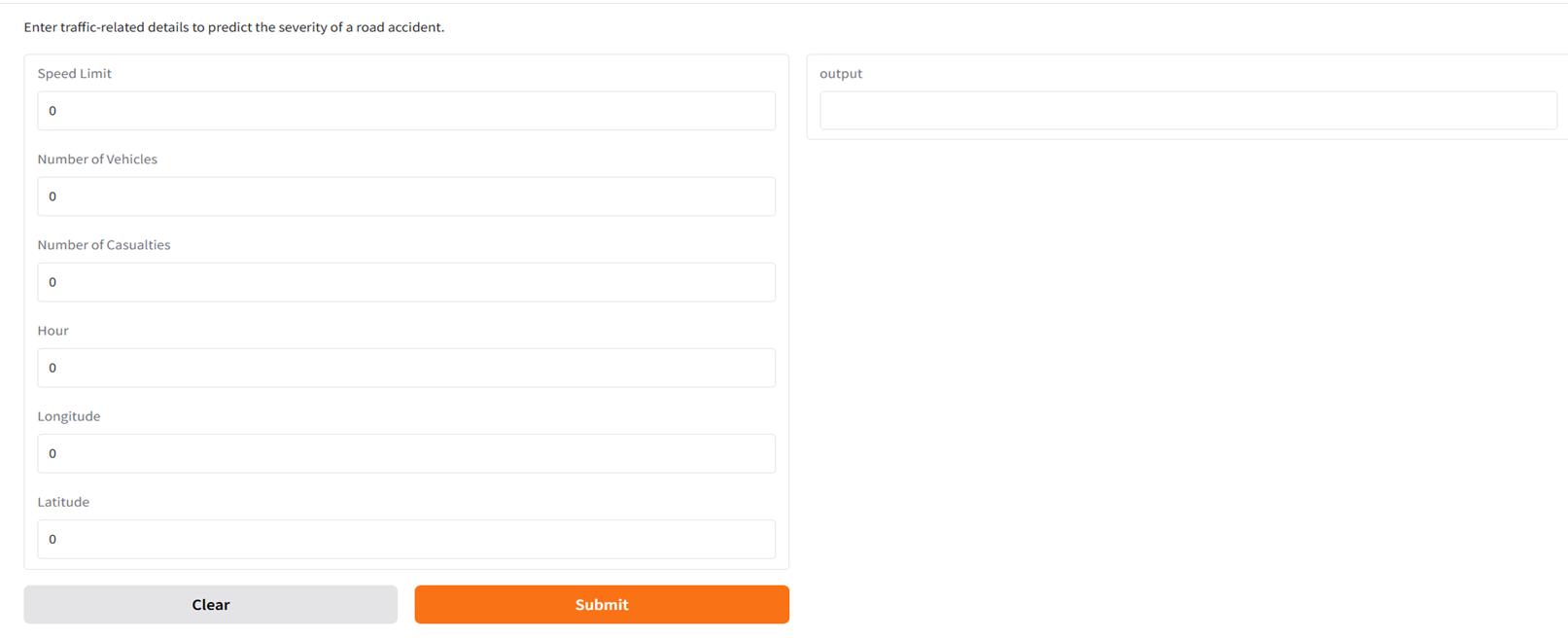
**Model Selection, Tuning and Deployment**

After analyzing the PyCaret results, we decided to further evaluate the top six machine learning models. Among these, Randim forest Classifier wa selected as the final modelbased on its balanced performance across accuracy, recall, precision and F1-score metrics.

We then performed hyperparameter tuning on the Random Forest model t optimize parameters such as the number od estimators, Maximun depth, and minimum samples split ensuring better generalization and reducing overfitting. To validate the tuned model, we applied Cross validation. Finally we deployed the trained and tuned Random Forest model using Gradio, an open source Python library. This deployment allows users to input accident related features and receive a prediction on the expected accident severity.

|  |  |
| --- | --- |
| A screenshot of a number of numbers | A screenshot of a graph |





“ Github link of our Project “

<https://github.com/adnan8055/UKroadAccident>

**Refrences:**

* PyCaret Development Team. (2021). *PyCaret: A low-code machine learning library in Python*. <https://pycaret.org/>
* Towards Data Science. (2021). "Feature selection techniques in machine learning with Python." *Towards Data Science*. <https://towardsdatascience.com/>
* Uk government road accidents surveys

**Acknowledgement**

We would like to extend our heartfelt thanks to our professor, [Abiodun Sodiq Shofoluwe], for their ongoing support and constructive feedback throughout the development of this project. Their expert guidance has been instrumental in shaping our understanding of the techniques involved, particularly in the areas of feature selection, machine learning algorithms, and model evaluation. The feedback provided at each stage of the project has been invaluable in refining our approach and ensuring that we remain on the right track.