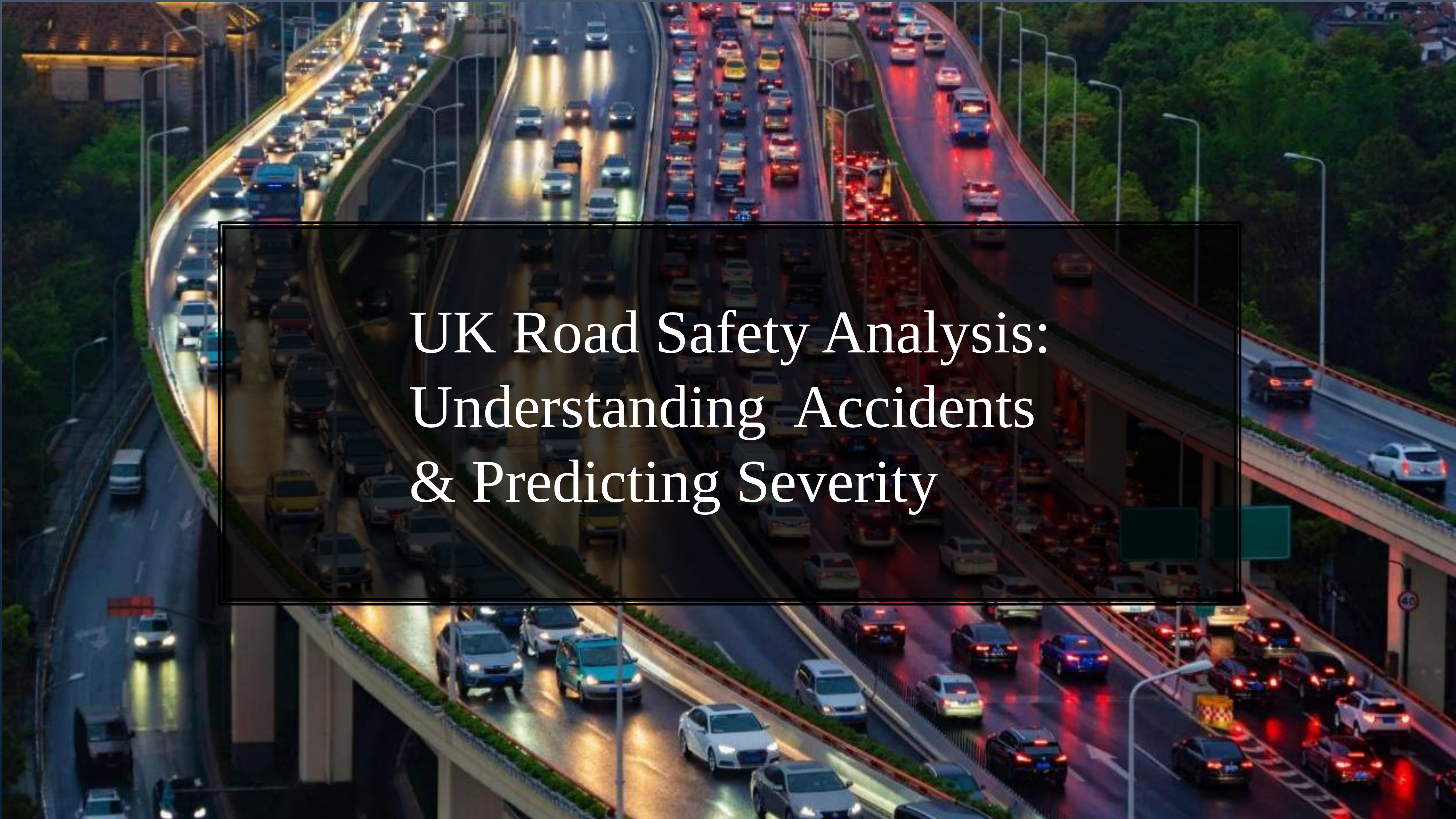


UK Road Safety Analysis: Understanding Accidents & Predicting Severity

UK road accident Analysis, Statistical Insights and Feature selection

Exploring accidents trends and key influencing factors

Presented by: Group 8

Group Members: Abdul Adnan
Sahil Sahil
Shubham Shubham
Vaibhav Vaibhav
Shalu Choudhary
Mohit Rexwal

Introduction

• Road accidents are a significant global concern, causing injury, death, and economic loss.

- 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.

UK road accident 2023 Dataset Overview

Sourced : UK government Website & Road Transport dept of UK
Total Records: 104259 Rows & 36 Columns

Key Attributes:

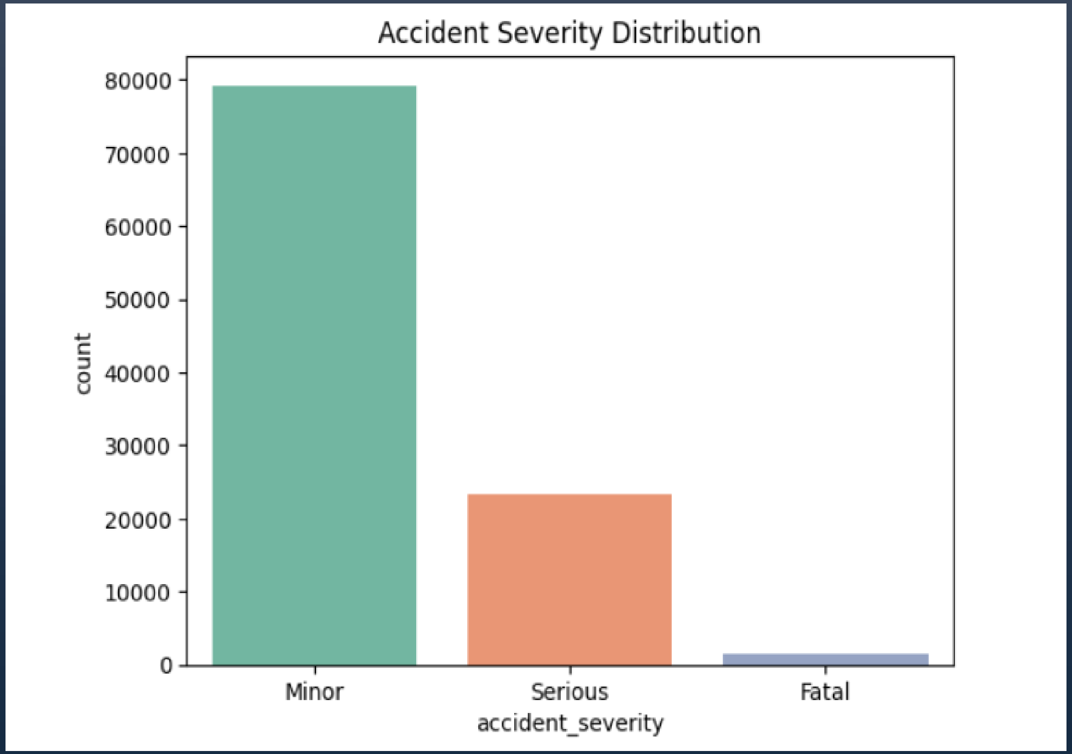
- Accident Severity : Minor, Serious, Fatal
- Number of vehicles involved
- Weather & Road Conditions
- Speed limits
- Road Types
- Location Data (Longitude, Latitude)

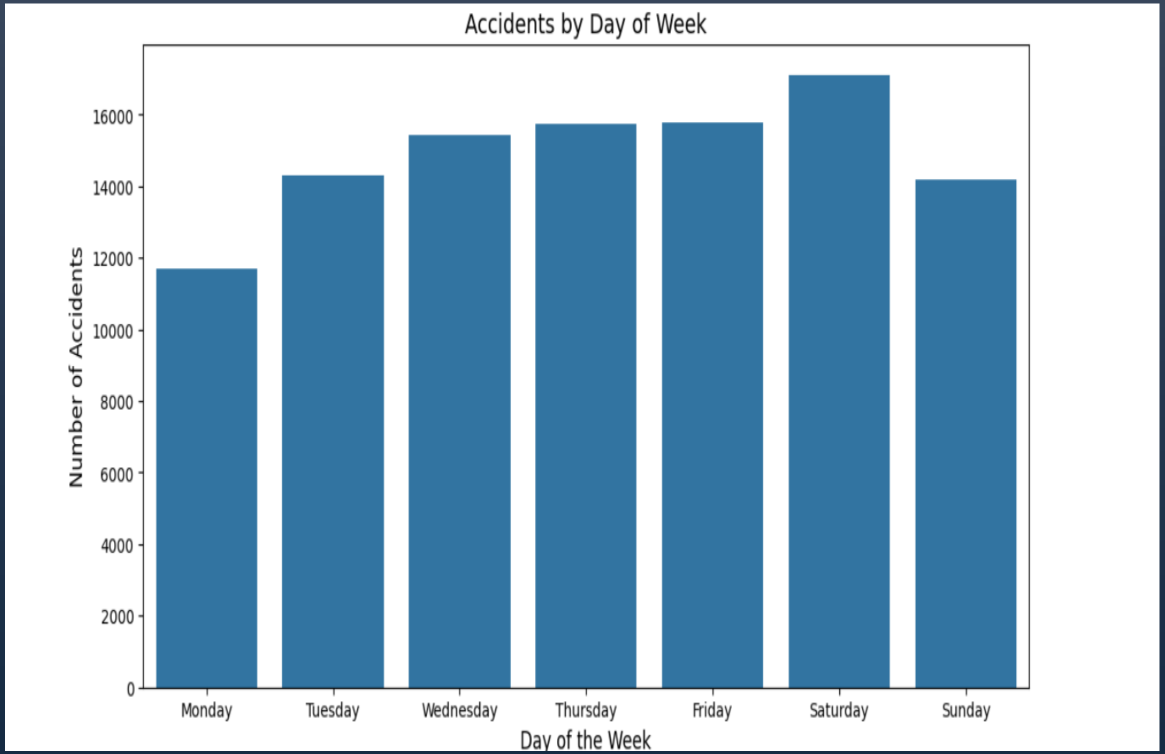
- Objective: Identifying significant accident patterns and key influencing factors

light_conditions	weather_conditions	road_surface_conditions	special_conditions_at_site	carriageway_hazards
Darkness without Streetlights	Hail	Snow	Construction	Broken Road Signs
Darkness without Streetlights	Rain	Wet/Damp	Construction	Broken Road Signs
Darkness without Streetlights	Rain	Wet/Damp	Construction	Broken Road Signs
Darkness without Streetlights	Flooding	Wet/Damp	Construction	Broken Road Signs
Darkness without Streetlights	Rain	Wet/Damp	Construction	Broken Road Signs
...	...	...	...	...
Darkness with Streetlights	Blowing Debris	Snow	Road Closure	Dislodged Vehicle
Daylight	Flooding	Wet/Damp	Construction	Broken Road Signs
Darkness without Streetlights	Blowing Debris	Snow	Road Closure	Dislodged Vehicle
Darkness without Streetlights	Hail	Wet/Damp	Road Closure	Dislodged Vehicle

Exploratory Data Analysis.

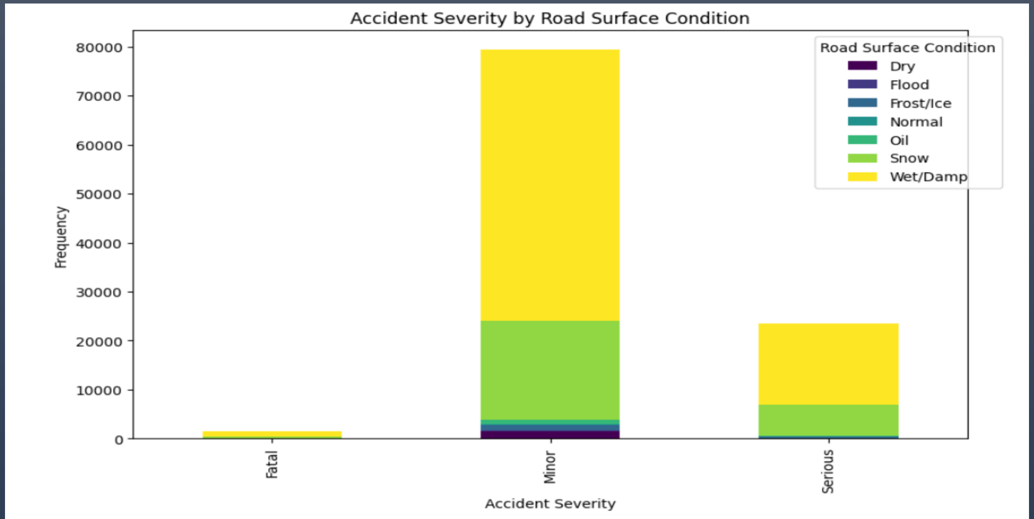
Showing Accident Severity based on different factors:



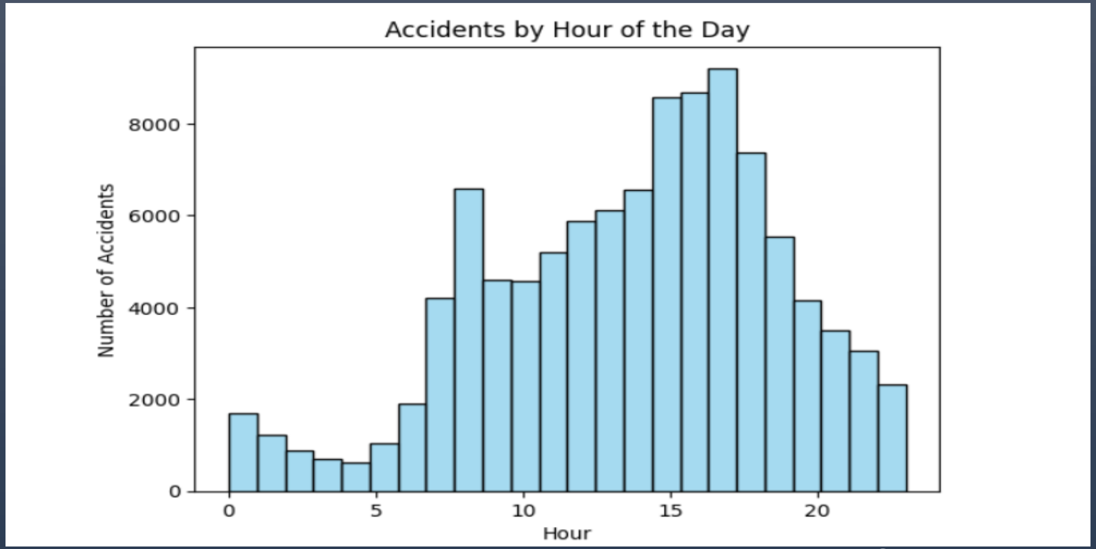


The majority of accidents are significantly **Minor** than any other category.

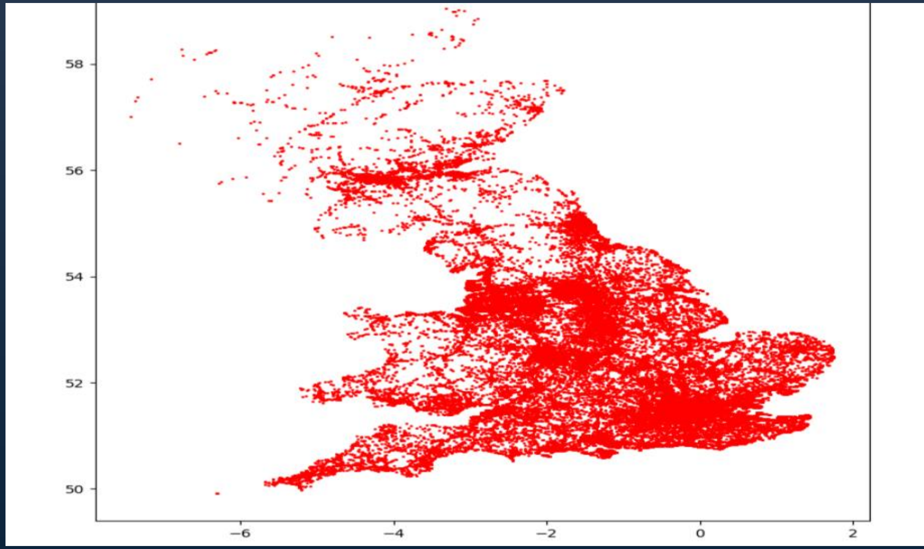
Saturday has the highest number of accidents, suggesting weekends might be riskier due to increased traffic, leisure activities, or impaired driving.



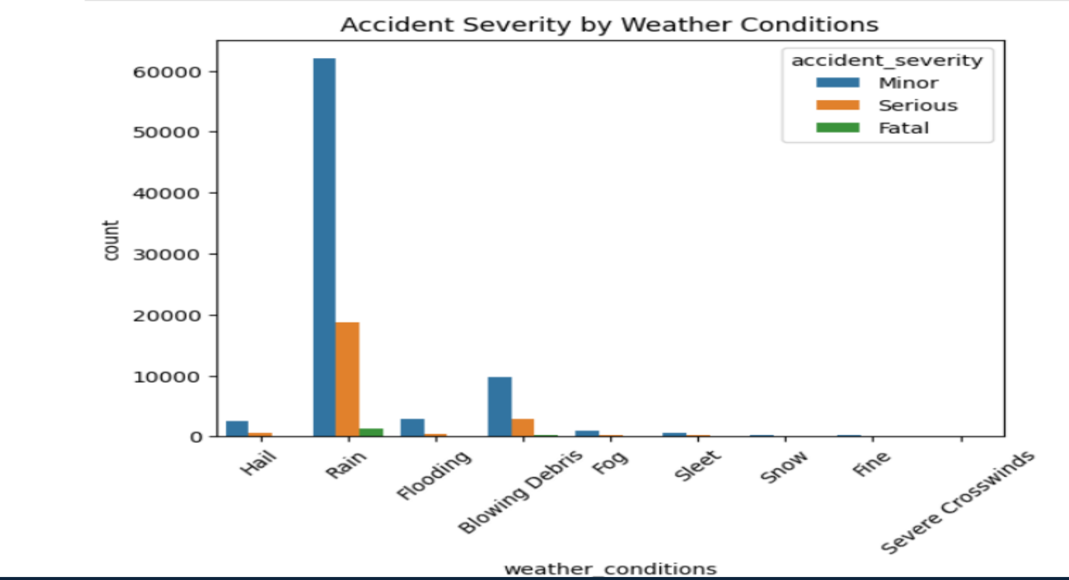
Most minor and serious accidents occur on wet/damp roads, with fewer on dry or icy surfaces.



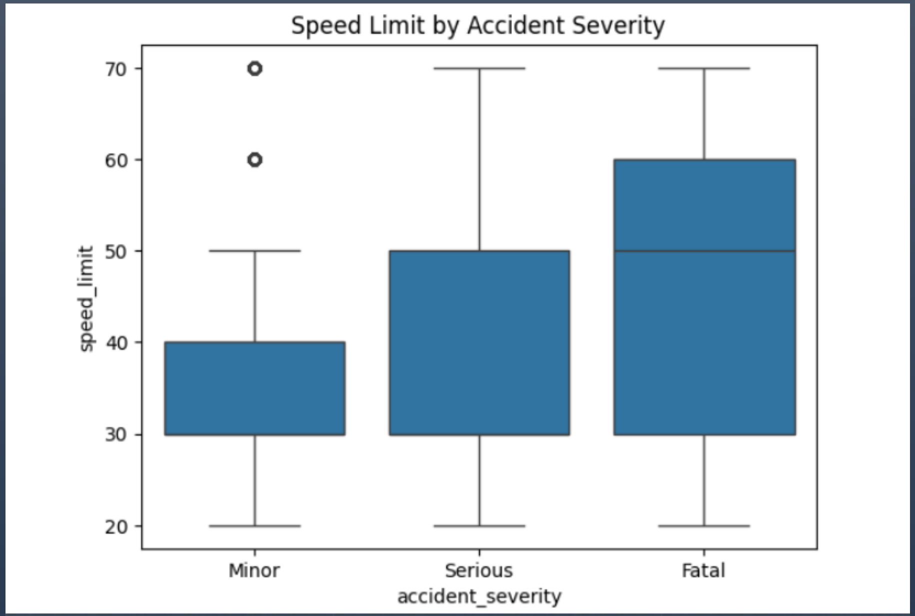
Accidents peak between 3 PM and 6 PM, with fewer incidents during early morning hours.



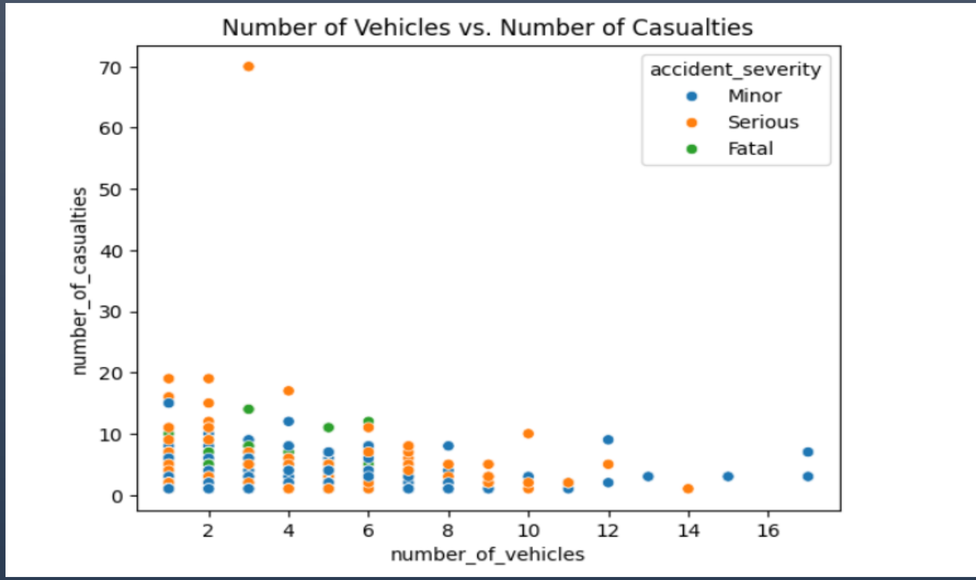
A geographic plot shows accident locations, densely clustered in central and southern regions.



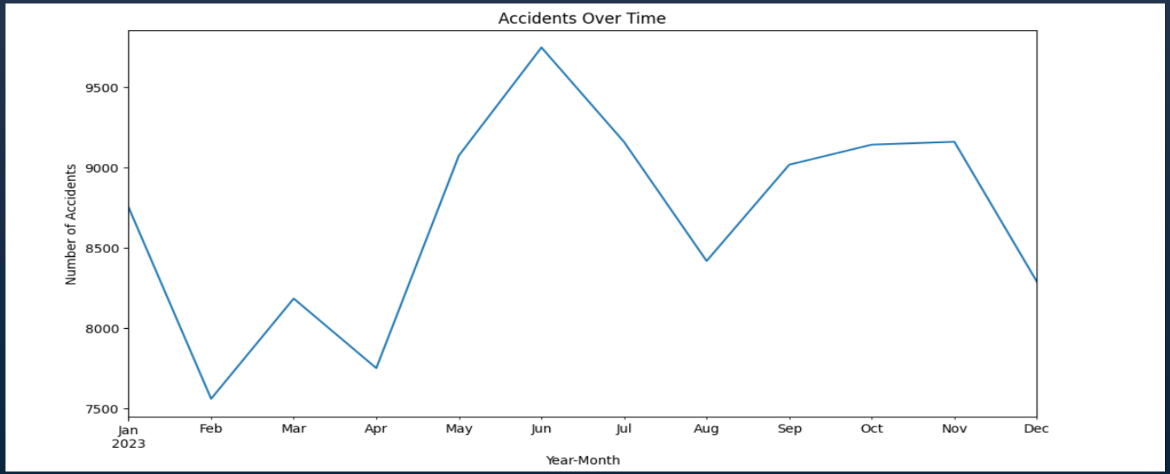
Most accidents happen during rain, with fewer incidents in severe weather like hail, fog, or snow.



Higher speed limits are linked to more severe accidents, with fatal accidents occurring at higher average speeds.



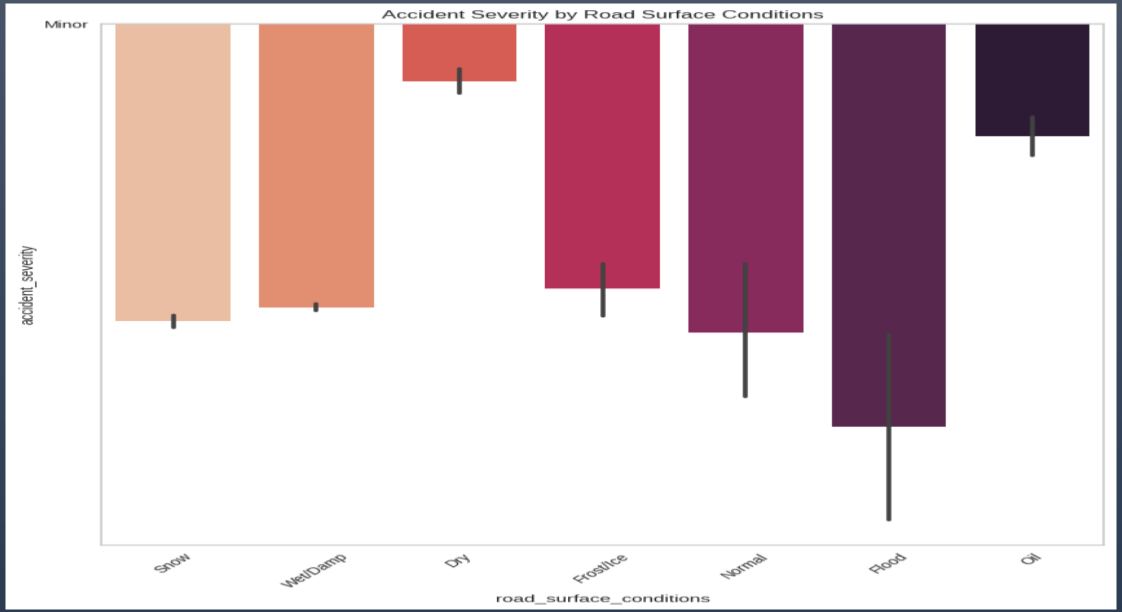
Accidents with more vehicles tend to have more casualties, though most incidents involve fewer than 6 vehicles.



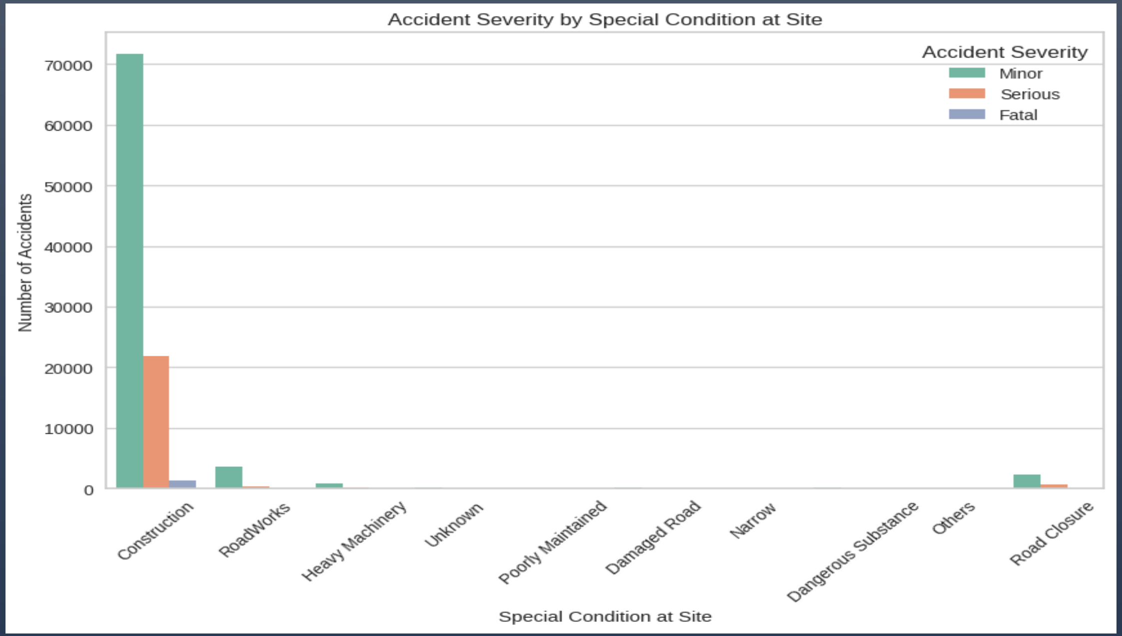
The number of accidents fluctuates throughout the year, peaking around June and dipping in February.

	Light_Conditions	Weather_Conditions	Accident_Severity
24803	Daylight	Rain	Minor
74405	Daylight	Rain	Minor
46472	Darkness without Streetlights	Rain	Minor
82604	Daylight	Rain	Minor
68654	Overcast	Rain	Minor
51875	Darkness without Streetlights	Rain	Minor

Most minor accidents happen during daylight or rain, with some occurring in darkness without streetlights or overcast conditions.

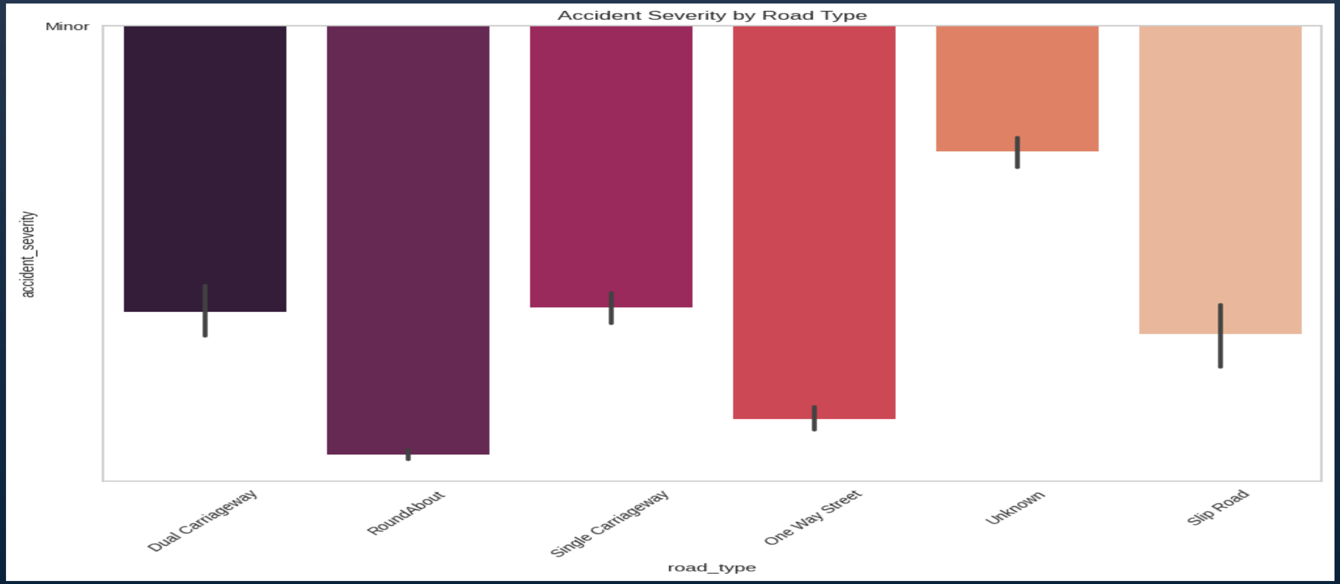


Accident Severity by Road Surface Conditions



Accident Severity by Special Condition at Site

Day_of_Week	Casualties_Count
Saturday	21802
Friday	19635
Thursday	19472
Wednesday	19175
Sunday	19018
Tuesday	18035
Monday	15827



Accident Severity by Road Type

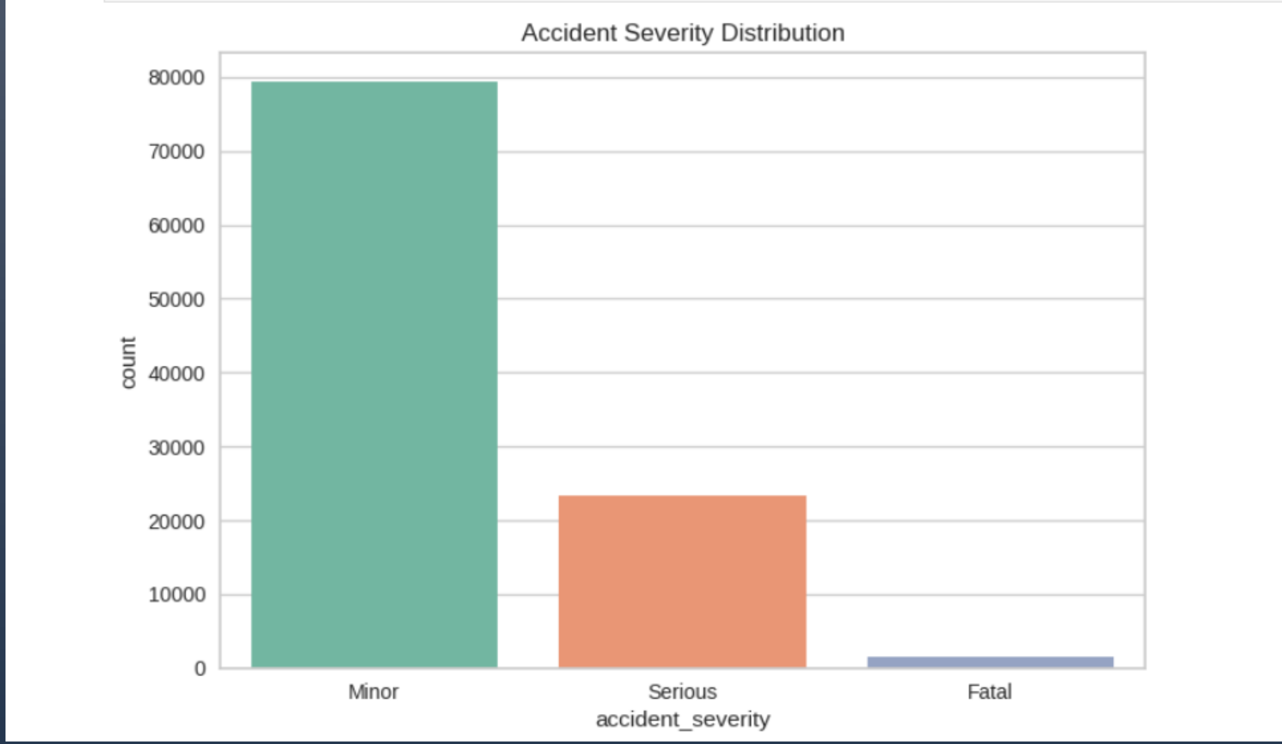
Data Imbalance

Class Imbalance and SMOTE:

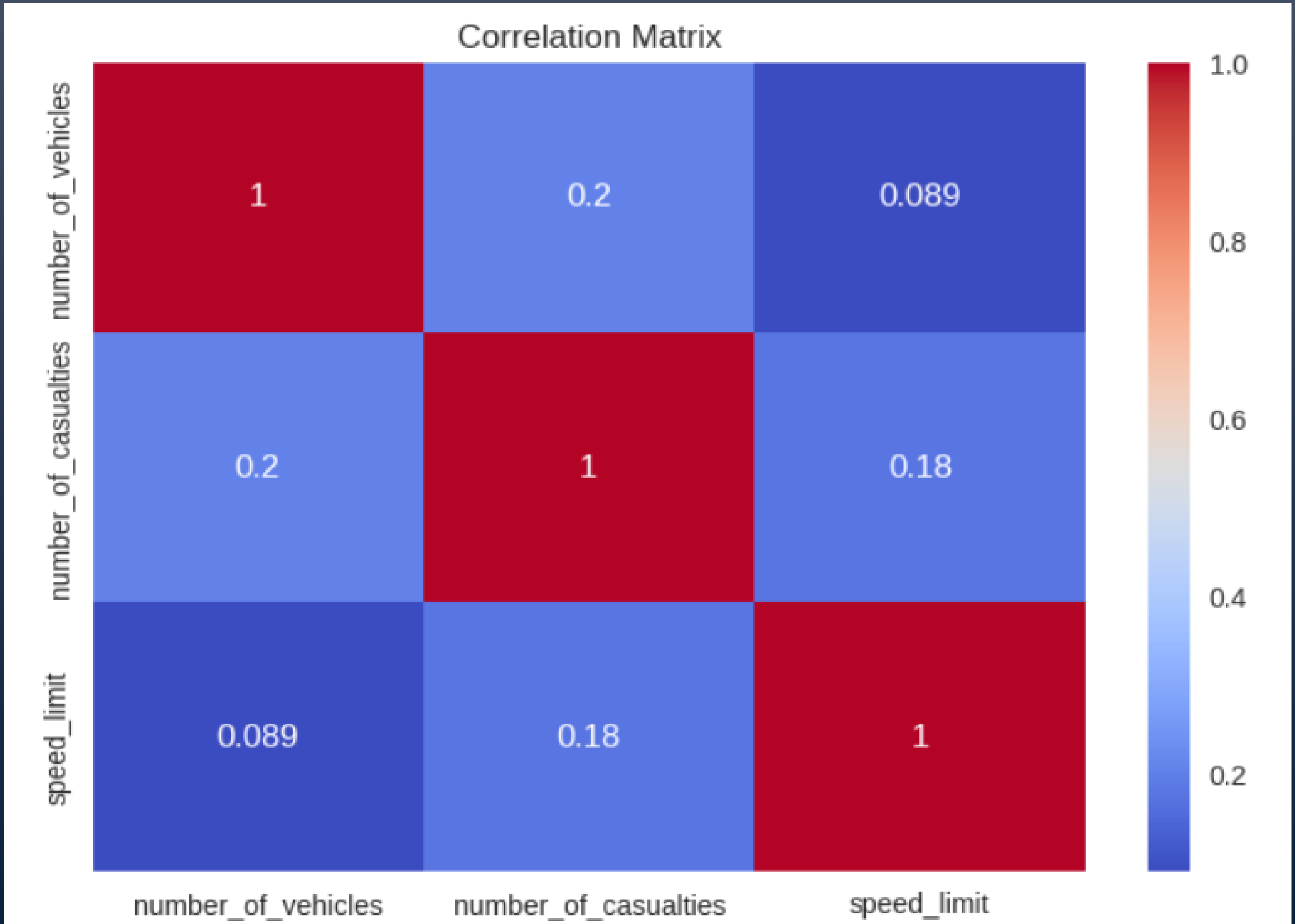
• The original dataset was **highly imbalanced**:
- **Minor**: 63,392
- **Serious**: 18,766
- **Fatal**: 1,238

• This could bias the model towards predicting accidents.

• **SMOTE** (Synthetic Minority Over-sampling Technique) 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.



CORELATION MATRIX



This shows that speed limit has minimal impact on number of vehicles and number of casualties. where number of vehicles and number casualties has a bit stronger relation that is 0.2 compared to speed limit.

Methodology

Chi-Square Test

- **Chi-Square Test on weather conditions and accident severity:**
- **p-value**: 1.3539774741973777e-74 (extremely small), indicating a strong statistical significance.

- **Chi-Square Test on light conditions and accident severity:**
- **p-value**: 1.1803413019381023e-132 (even smaller than the first), showing an even stronger statistical significance.

HYPOTHESIS TESTING FOR WEATHER AND LIGHT CONDITIONS

Hypothesis testing on weather conditions and accident severity:

- **Chi-square Statistic**: 397.28
- **Degrees of Freedom**: 16
- **P-value**: 0.00000
- **Hypothesis Decision**: Reject the null hypothesis (H_0) since p-value < 0.05.
- **Conclusion**: 🔴 Weather conditions significantly affect accident severity.

Hypothesis testing on light conditions and accident severity:

- **Chi-square Statistic**: 638.58
- **Degrees of Freedom**: 8
- **P-value**: 0.00000
- **Hypothesis Decision**: Reject the null hypothesis (H_0) since p-value < 0.05.
- **Conclusion**: 🔴 Light conditions significantly affect accident severity.

Feature Selection

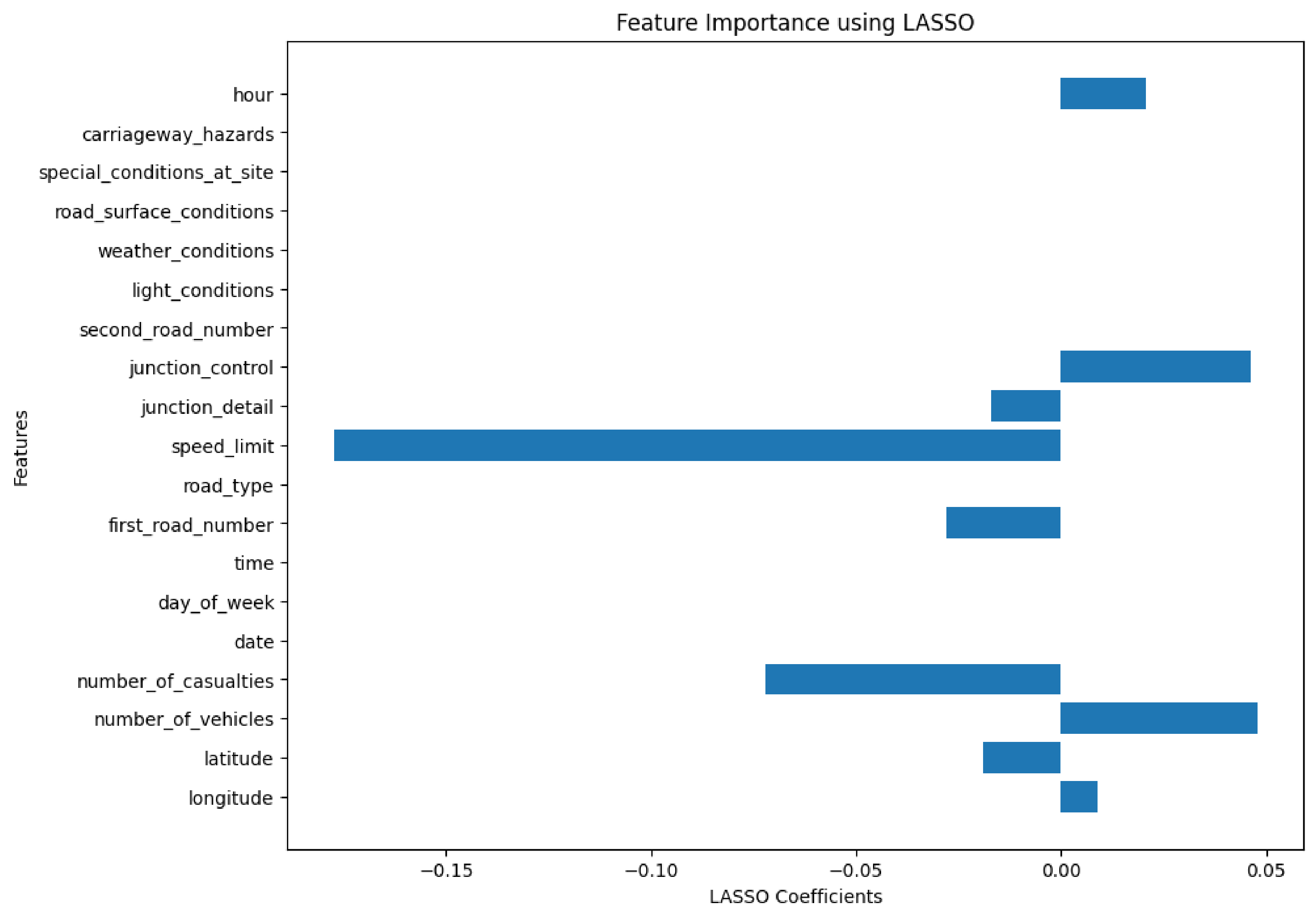
Filter Method

Top Features (Filter - MI):

- time,
- speed_limit,
- date,
- junction_control,
- first_road_number,
- junction_detail,
- latitude,
- number_of_vehicles,
- longitude,
- second_road_number]

EMBEDDED METHOD

LASSO



Results

'longitude’

'latitude'

'number_of_vehicles'

'number_of_casualties'

'first_road_number'

'speed_limit'

'junction_detail'

'junction_control' 'hour’

These are the top Features

Wrapper Method

Recursive feature Elimination

```
Selected Features (Wrapper - RFE): ['longitude', 'latitude', 'number_of_vehicles', 'date', 'time', 'first_road_number', 'speed_limit', 'junction_detail', 'junction_control', 'hour']
Model trained with selected features!
```

Results: Final Model

Using PyCaret

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
rf	Random Forest Classifier	0.6939	0.8553	0.6939	0.6861	0.6864	0.5408	0.5436	27.5760
et	Extra Trees Classifier	0.6886	0.8546	0.6886	0.6813	0.6826	0.5328	0.5347	16.9420
xgboost	Extreme Gradient Boosting	0.6762	0.8399	0.6762	0.6672	0.6574	0.5143	0.5250	4.0920
lightgbm	Light Gradient Boosting Machine	0.6682	0.8341	0.6682	0.6592	0.6470	0.5023	0.5146	8.6180
knn	K Neighbors Classifier	0.6595	0.8176	0.6595	0.6480	0.6466	0.4892	0.4946	3.0280
gbc	Gradient Boosting Classifier	0.6531	0.0000	0.6531	0.6418	0.6334	0.4796	0.4897	53.5300
dt	Decision Tree Classifier	0.6422	0.7317	0.6422	0.6406	0.6413	0.4633	0.4634	1.3010
ada	Ada Boost Classifier	0.6264	0.0000	0.6264	0.6158	0.6174	0.4396	0.4423	4.7780
ridge	Ridge Classifier	0.5349	0.0000	0.5349	0.5255	0.5224	0.3024	0.3067	0.4580
nb	Naive Bayes	0.5329	0.7100	0.5329	0.5292	0.5302	0.2993	0.2999	0.6370
lr	Logistic Regression	0.5322	0.0000	0.5322	0.5262	0.5274	0.2983	0.2993	11.5270
lda	Linear Discriminant Analysis	0.5310	0.0000	0.5310	0.5262	0.5275	0.2965	0.2972	0.6570
qda	Quadratic Discriminant Analysis	0.5300	0.0000	0.5300	0.5293	0.5282	0.2950	0.2959	0.5600
svm	SVM - Linear Kernel	0.4745	0.0000	0.4745	0.4982	0.4291	0.2117	0.2374	12.3950
dummy	Dummy Classifier	0.3333	0.5000	0.3333	0.1111	0.1667	0.0000	0.0000	0.6120

	Description	Value
0	Session id	123
1	Target	accident_severity
2	Target type	Multiclass
3	Target mapping	Fatal: 0, Minor: 1, Serious: 2
4	Original data shape	(190176, 7)
5	Transformed data shape	(190176, 7)
6	Transformed train set shape	(133123, 7)
7	Transformed test set shape	(57053, 7)
8	Numeric features	6
9	Preprocess	True
10	Imputation type	simple
11	Numeric imputation	mean
12	Categorical imputation	mode
13	Fold Generator	StratifiedKFold
14	Fold Number	10
15	CPU Jobs	-1
16	Use GPU	False
17	Log Experiment	False
18	Experiment Name	clf-default-name
19	USI	da8e

After running the models using PyCaret with our selected features, each group member picked one model to work on. We focused on tuning, hyperparameter optimization, and cross-validation for these models. The models we explored included Random Forest, Extra Trees, XGBoost, LightGBM, K-Nearest Neighbors, and Gradient Boosting Classifier. After comparing their performances, we found that Random Forest gave us the highest accuracy and consistent results across metrics, so we chose it as our final model."

Training Models and Evaluation Using Random Forest

The Random Forest classifier achieved an overall accuracy of **70.03%** in predicting accident severity. The model performed well for the **Fatal** (F1-score: 0.78) and **Minor** (F1-score: 0.75) classes, showing strong precision and recall. However, performance for the **Serious** class was weaker, with an F1-score of **0.55** and recall of only 50%, indicating that half of the serious cases were missed.

📌 Random Forest Accuracy: 0.7003

	precision	recall	f1-score	support
Fatal	0.74	0.82	0.78	12600
Minor	0.71	0.80	0.75	12539
Serious	0.63	0.50	0.55	12897
accuracy			0.70	38036
macro avg	0.69	0.70	0.69	38036
weighted avg	0.69	0.70	0.69	38036

Random Forest Accuracy after Hyperparameter Tuning

```
Fitting 3 folds for each of 10 candidates, totalling 30 fits
🎯 Best Parameters Found:  {'bootstrap': True, 'max_depth': None, 'max_features': 'log2', 'min_samples_leaf': 2, 'min_samples_s
plit': 3, 'n_estimators': 58}

📌 Best Random Forest Accuracy after Hyperparameter Tuning: 0.6994
              precision    recall  f1-score   support

       Fatal       0.74      0.80      0.77     12600
       Minor       0.71      0.82      0.76     12539
     Serious       0.63      0.48      0.55     12897

    accuracy                           0.70     38036
   macro avg       0.69      0.70      0.69     38036
weighted avg       0.69      0.70      0.69     38036
```

After applying hyperparameter tuning, the best Random Forest model achieved an accuracy of **69.94%**, which is comparable to the default model performance. The model continued to perform well on the **Fatal** (F1-score: 0.77) and **Minor** (F1-score: 0.76) classes, with high recall values of **0.80** and **0.82** respectively, indicating good sensitivity. However, the **Serious** class remains a challenge, with a lower recall of **0.48** and F1-score of **0.55**, suggesting that the model struggles to correctly identify many of the serious cases.

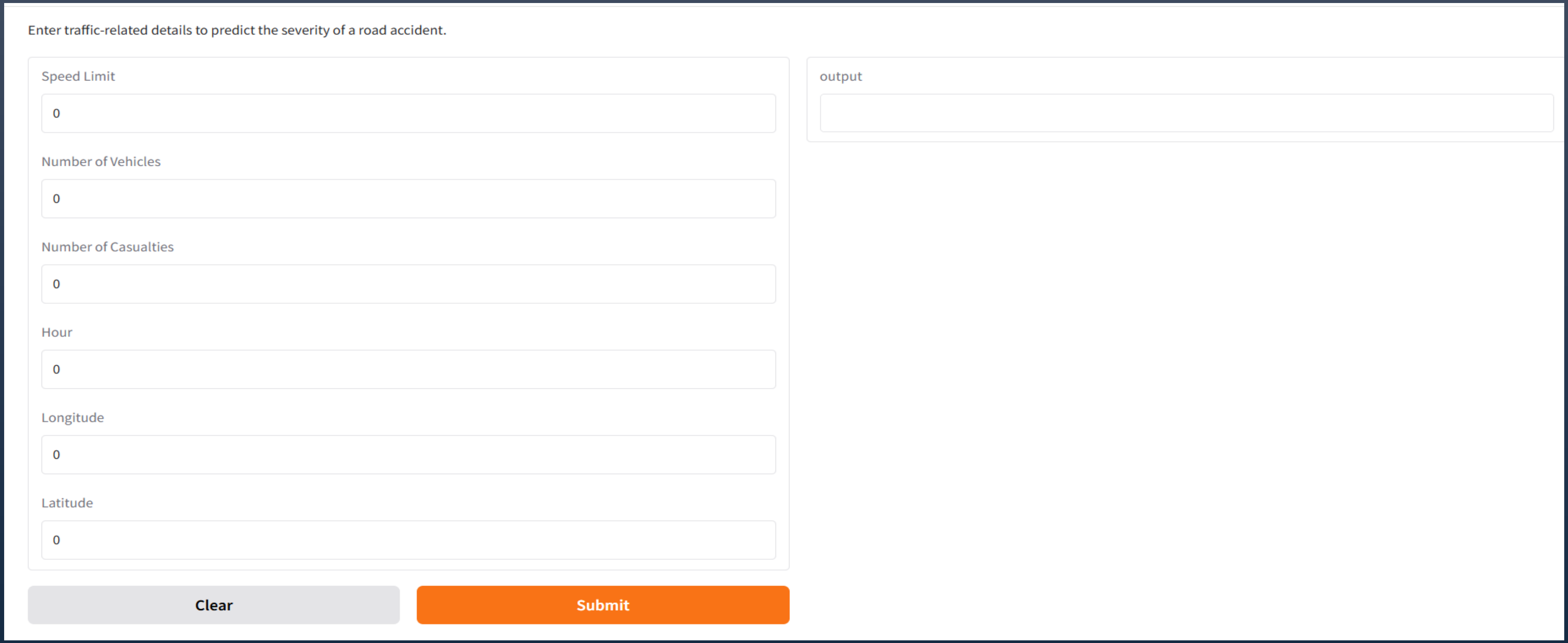
Despite tuning, the class imbalance and overlapping feature space may be limiting the model's ability to improve. To enhance performance further, additional steps such as **class balancing, feature selection, or alternative models** like **XGBoost or LightGBM** can be considered.

Cross-validation on the Random Forest

```
📊 Cross-Validation Accuracy Scores: [0.58652329 0.68450112 0.73193112 0.73106349 0.7335349 ]
✅ Mean CV Accuracy: 0.6935
📉 Standard Deviation: 0.0566
```

The Random Forest model was evaluated using 5-fold cross-validation, producing the following accuracy scores:
[0.5865, 0.6845, 0.7319, 0.7311, 0.7335]
The **mean cross-validation accuracy** is **69.35%**, with a **standard deviation of 0.0566**.
This indicates that the model is generally consistent in its performance across different data splits, though one fold (58.65%) showed notably lower accuracy, suggesting some variability possibly due to data imbalance or distribution differences.
To improve stability and performance, further tuning, stratified sampling, or addressing class imbalance could be beneficial.

Model Deployment



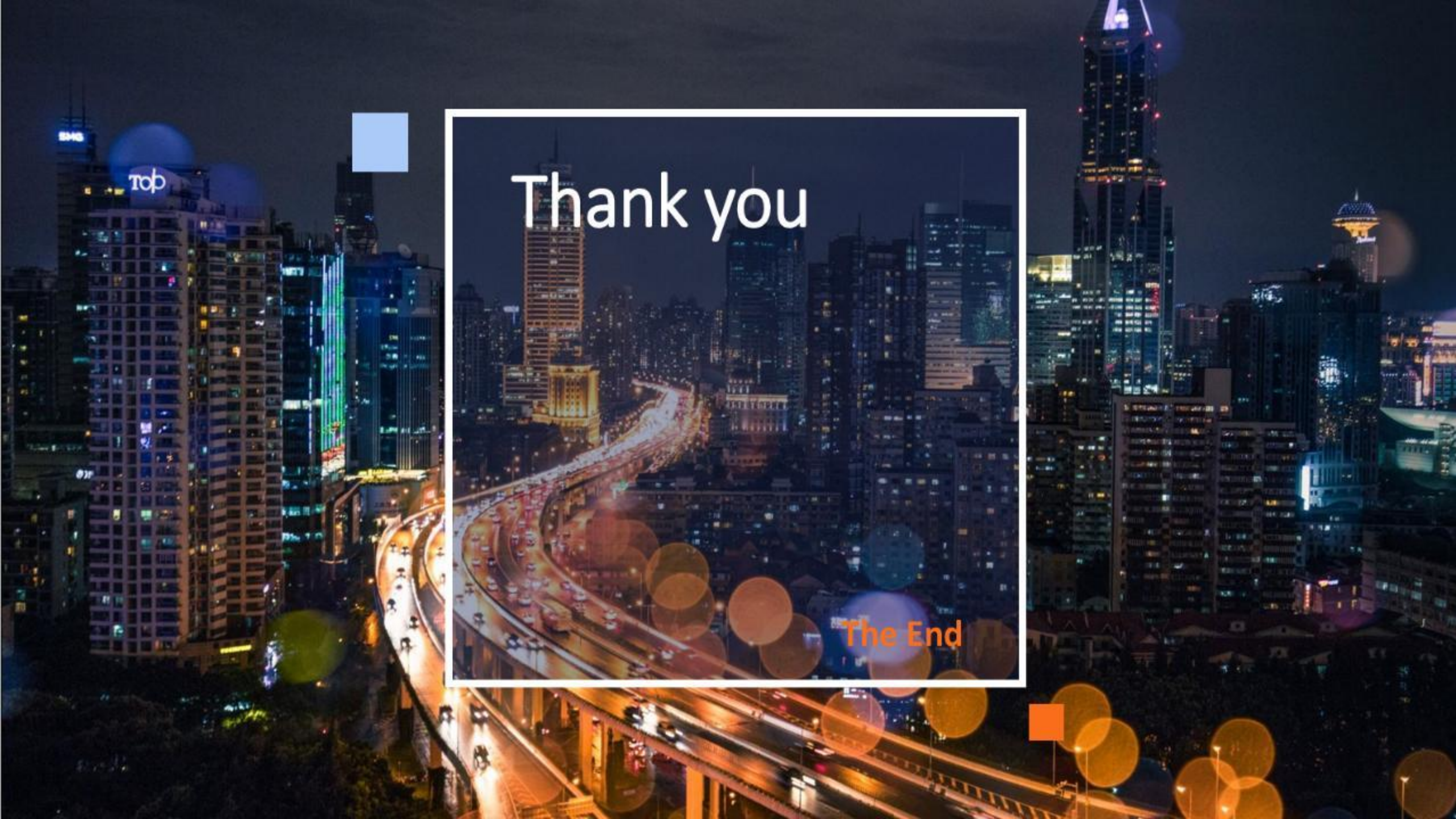
Gradio was used to create an interactive web interface for the UK road accident severity prediction model. It allows users to input accident-related features and instantly receive severity predictions from the trained Random Forest model.

Conclusions

The Random Forest model demonstrated satisfactory performance in predicting accident severity, achieving a maximum accuracy of **70.03%** before tuning and **69.94%** after hyperparameter tuning. Cross-validation results further confirmed model stability, with a **mean accuracy of 69.35%** and a standard deviation of **0.0566**.

The model performed well in identifying **Fatal** and **Minor** accidents, with high precision and recall. However, it struggled to accurately classify the **Serious** category, highlighting the impact of class imbalance and overlapping features. Despite tuning, gains were marginal, suggesting the need for further enhancements such as **advanced resampling techniques, feature engineering, or alternative ensemble models**.

Overall, the Random Forest model provides a solid baseline for accident severity prediction and can be effectively deployed with additional improvements for real-world use.


Thank you
The End