Road Accident Severity Classification

Karnik Ketan Kalani

Project Background

- Road accidents are a major public safety concern worldwide, resulting in significant human and economic costs.

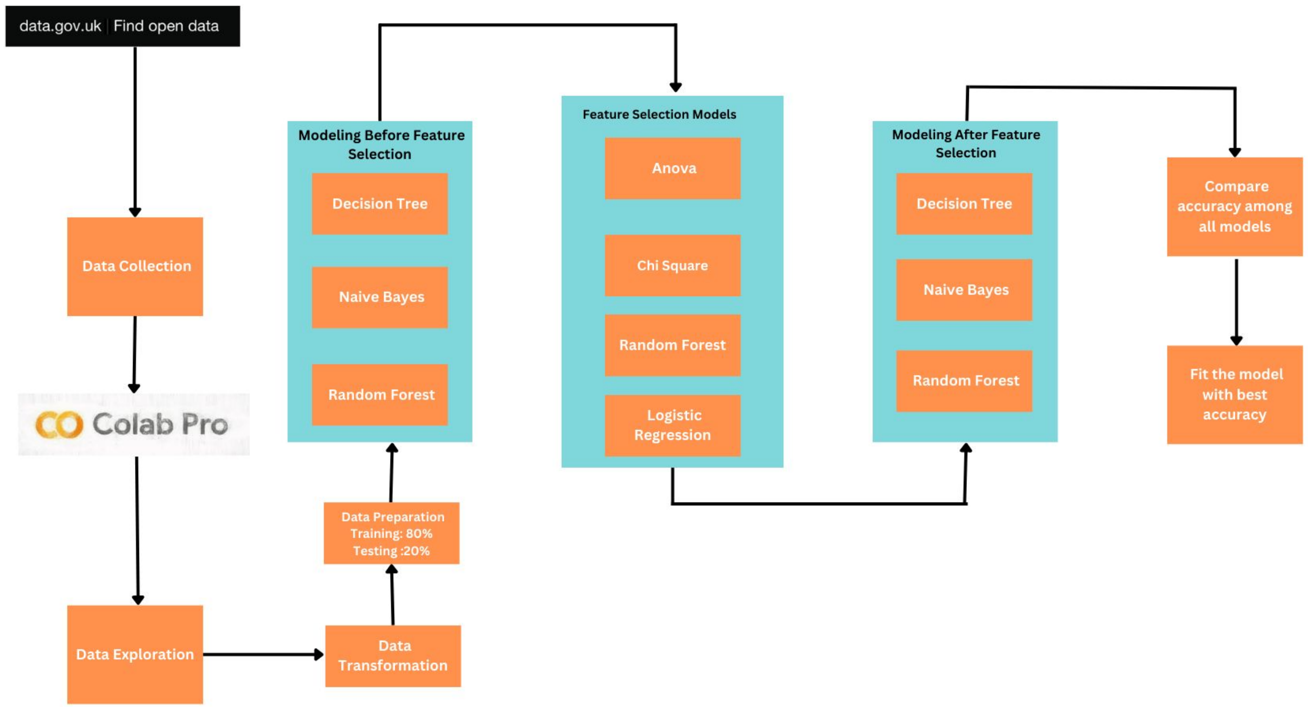
- The World Health Organization (WHO) estimates that road traffic injuries are the eighth leading cause of death globally, accounting for approximately 1.35 million fatalities annually, with an additional 50 million people sustaining non-fatal injuries.

- In addition to the loss of human life and physical injuries, road accidents can also have significant economic consequences, such as property damage, medical expenses, and lost productivity .

Motivation

- The main motivation behind the project is to reduce the number of accidents and to reduce the economic impact such as vehicle damages and the rise of insurance costs.
- Accidents can happen due to various factors such as bad lighting, bad weather, improper roads, speed limit, and so on.
- In this project, a Road accident classification model is built using machine learning models such as Decision Trees, Random Forest Classifier and Naive Bayes to predicting the factors that are responsible for catastrophic accidents.

Project Flow Diagram



Dataset Description

- The dataset is taken from data.gov.uk
- Data.gov.uk is a Uk Government project to make available non-personal UK government data as open data. It was launched in closed 30 September 2009.
- We merged three datasets based on the column Accident_index.
- Vehicles File(**3004425, 21)**
- Casualties File **(216720, 14)**
- Accidents File **(1780653, 31**)

```
casualties_df.head()
```

Accident_Index	Vehicle_Reference	Casualty_Reference	Casualty_Class	Sex_of_Casualty	Age_of_Casualty
200501BS00001	1	1	3	1	37
200501BS00002	1	1	2	1	37
200501BS00003	2	1	1	1	62
200501BS00004	1	1	3	1	30
200501BS00005	1	1	1	1	49

Dataset Description

- Accidents file

```
accidents_df.head()
```

Accident_Index	Location_Easting_OSGR	Location_Northing_OSGR	Longitude	Latitude	Police_Force	Accident_Severity
200501BS00001	525680.0	178240.0	-0.191170	51.489096	1	2
200501BS00002	524170.0	181650.0	-0.211708	51.520075	1	3
200501BS00003	524520.0	182240.0	-0.206458	51.525301	1	3
200501BS00004	526900.0	177530.0	-0.173862	51.482442	1	3
200501BS00005	528060.0	179040.0	-0.156618	51.495752	1	3

- Vehicle File

```
vehicles_df.head()
```

Accident_Index	Vehicle_Reference	Vehicle_Type	Towing_and_Articulation	Vehicle_Manoeuvre
200501BS00001	1	9	0	18
200501BS00002	1	11	0	4
200501BS00003	1	11	0	17
200501BS00003	2	9	0	2
200501BS00004	1	9	0	18

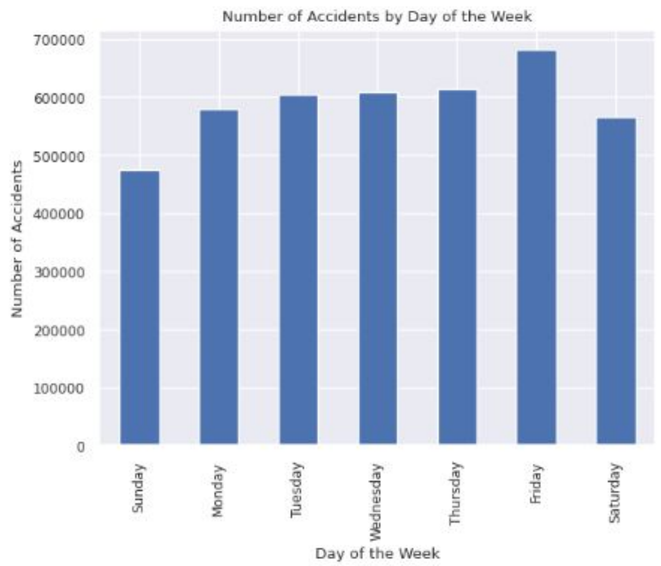
Data Exploration

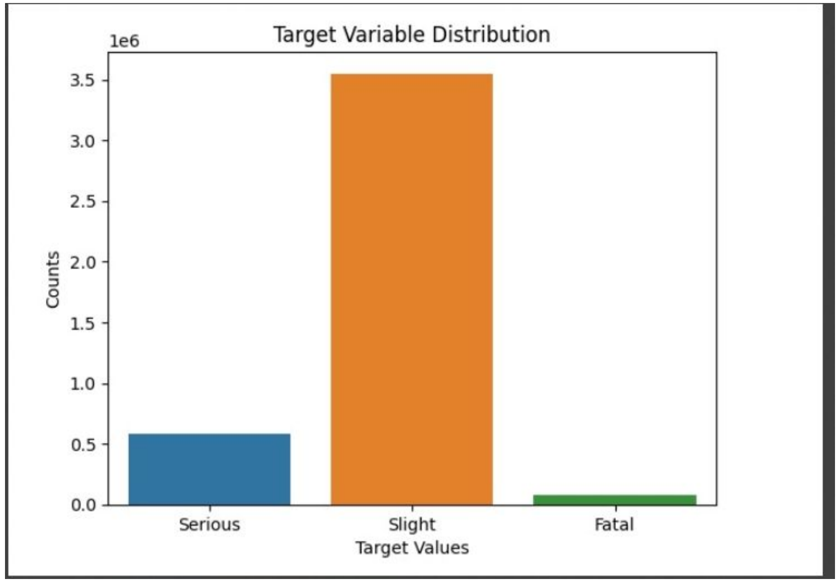
- The target Feature "accident_severity" has three accident types i.e., Fatal, slight and serious.

- There are imbalances in the values of "accident_severity" i.e., slight severity category has more data, this is resolved in the preprocessing step.

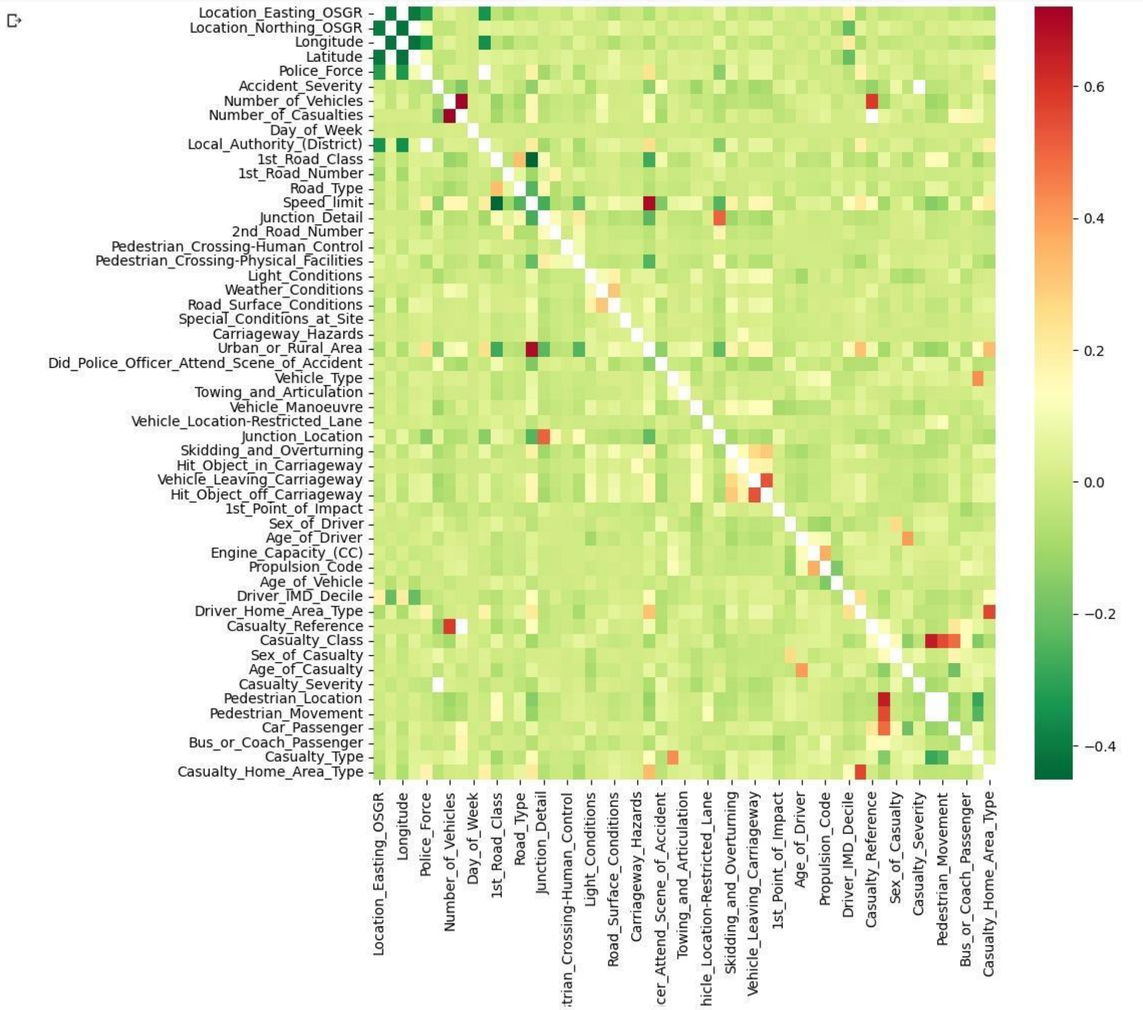
- Most of the accidents took place on Friday and the least ones on Sunday.

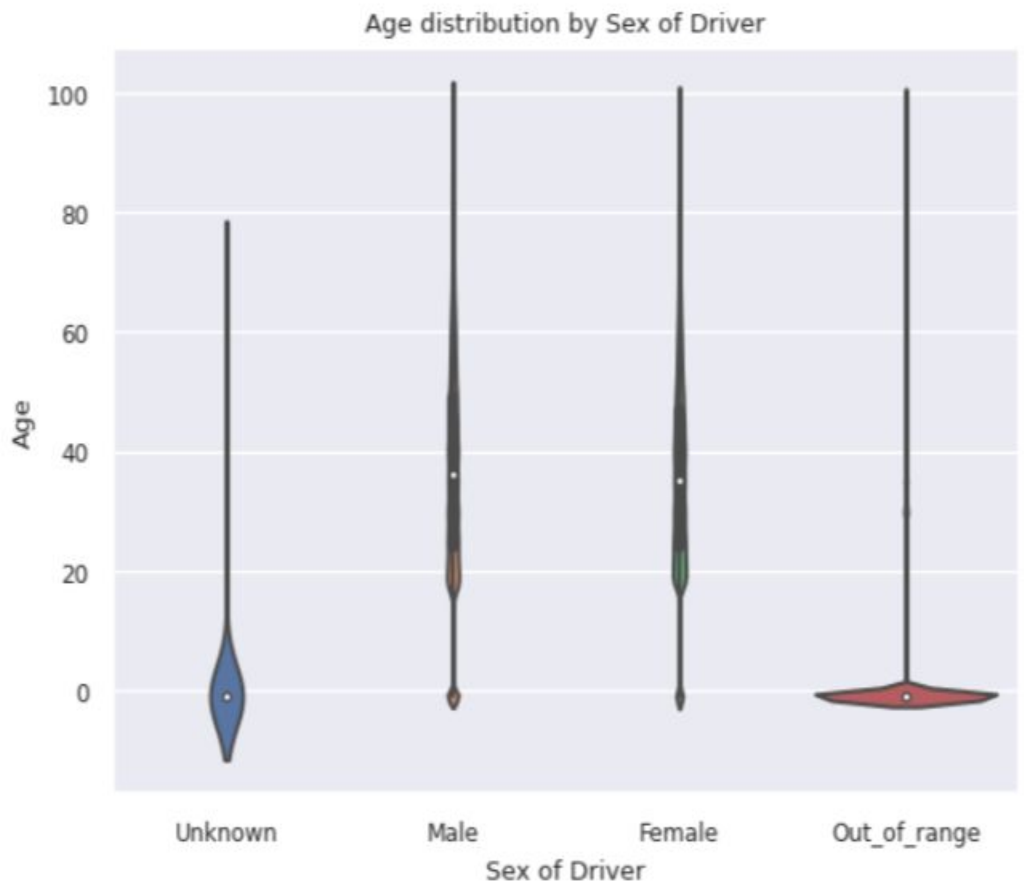
- There are many unknown and out of range values in the age distribution by sex of driver.





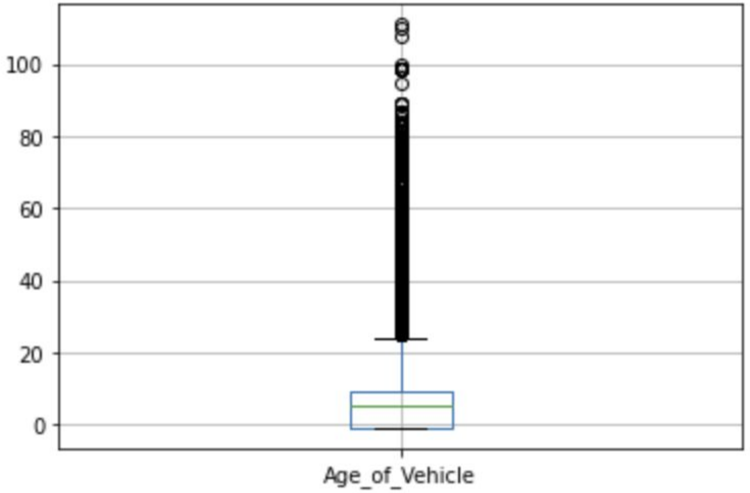
Data Exploration

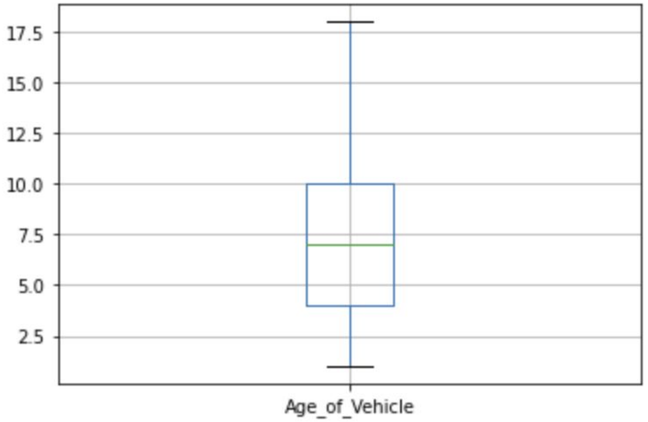




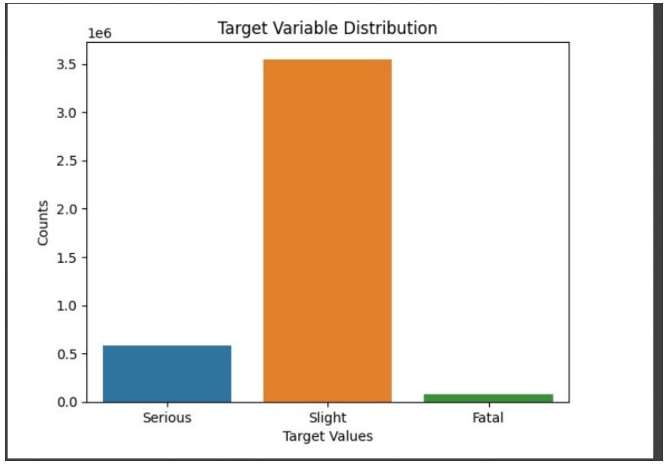
Data Preprocessing

- The merged dataset has many inconsistent values –1, so we replaced this NaN values and dropped the rows.
- Columns "Pedestrian_Road_Maintenance_Worker", "2nd_road_class",and "junction control" are removed as they have null values more than 30% in the dataset.
- ""Age_Band_of_Driver","Age_Band_of_Casualty","Was_Veh_Left_Hand_Driver", "Vehicle_Reference_y","Vehicle_Reference_x" columns are removed as they have no significance with the problem statement.
- "Age_of_vehicle", "Engine_Capacity", "Age_of_Driver", "Age_of_Casualty" have noticeable has significant number of outliers , which are replaced with median values of respective columns.
- The null values of categorical features are filled using mode of the respective columns.





Handling Target Class Imbalance

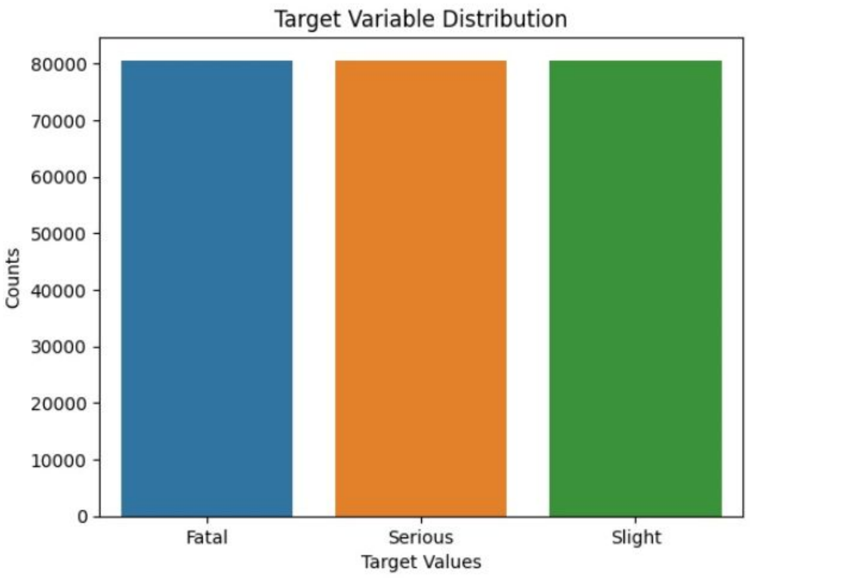


- In this dataset the target feature is not balanced.
- There are 50,000 serious injury, 350,000 slight injury, and 10,000 fatal injury values.

- Random under sampling was used to balance the training dataset and the target feature values.

Transformation And Modeling

- Date is converted into datetime objects in the format of mm/dd/yyyy.
- Converted the Time column to datetime format and extracted the time hours in a new column.
- Time is in string format and extracted the hours from time. All these were in object type and they were converted to int data type.
- Min-Max Scalar is used for data normalization to scale the features in a fixed range of 0 and 1.
- Random under sampling was performed to balance the target feature values in the dataset.
- The train and test datasets are divided in the ratio of 80:20.




```
# Apply RandomUnderSampler to balance the classes
rus = RandomUnderSampler(random_state=42)
X_res, y_res = rus.fit_resample(X_scaled, y)
print('Resampled dataset shape %s' % Counter(y_res))
```

```
Resampled dataset shape Counter({1: 80596, 2: 80596, 3: 80596})
```

Feature Selection Models

1. Anova Test

- Anova (Analysis of Variance) is a statistical method that can be used for feature selection.
- Anova helps to identify the features that have a significant impact on the target variable.
- Feature selection using Anova would help to identify which features have a significant impact on the "Accident_Severity " and should be included in the model.
- Anova is applied to the training set to rank the importance of the predictor variables based on their F-score value. The higher the F-score, the more significant the predictor variable can show the variation for target variable.
- ANOVA is best suited for continuous predictor variables and categorical target variables.

	feature	F-score
46	Casualty_Severity	89804.533300
12	Speed_limit	14731.762327
22	Urban_or_Rural_Area	12753.854820
23	Did_Police_Officer_Attend_Scene_of_Accident	8095.342602
27	Vehicle_Manoeuvre	5465.905830
9	1st_Road_Class	3848.771027
29	Junction_Location	3698.934785
13	Junction_Detail	3529.043288
4	Number_of_Casualties	3269.026282
32	Vehicle_Leaving_Carriageway	3168.233141
17	Light_Conditions	2994.139368
33	Hit_Object_off_Carriageway	2326.492588
30	Skidding_and_Overturning	2289.120226
42	Casualty_Reference	2240.736804
44	Sex_of_Casualty	2073.254726

2. **Chi-Square Test**

- Chi-Square can be used to test whether there is a significant association between a predictor variable and the target variable.
- The target variable is the variable that you want to predict, and the predictor variable is a variable that may have an impact on the target variable.
- If the p-value of the test is less than a predetermined significance level, then it can be concluded that there is a significant association between the two variables.

	feature	chi2-score	p-value
46	Casualty_Severity	21562.112132	0.0
23	Did_Police_Officer_Attend_Scene_of_Accident	7068.618337	0.0
22	Urban_or_Rural_Area	5219.310751	0.0
29	Junction_Location	4046.594809	0.0
32	Vehicle_Leaving_Carriageway	3822.887420	0.0
12	Speed_limit	3120.328119	0.0
33	Hit_Object_off_Carriageway	2982.142238	0.0
17	Light_Conditions	2824.564576	0.0
44	Sex_of_Casualty	2627.342966	0.0
13	Junction_Detail	2486.847687	0.0
30	Skidding_and_Overturning	2111.771641	0.0
41	Driver_Home_Area_Type	2015.826478	0.0
35	Sex_of_Driver	1776.268487	0.0
52	Casualty_Home_Area_Type	1611.983420	0.0
27	Vehicle_Manoeuvre	1524.775854	0.0

3. Random Forest

- Random Forest is a machine learning algorithm that uses an ensemble of decision trees to predict the target variable.

- Random Forest assigns an importance score to each feature based on its contribution to the accuracy of the model. The higher the importance score of a feature, the more important it is in predicting the target variable.

- In the given figure casualty severity is the most important feature as it has the highest value.

feature	importance
Casualty_Severity	0.350465
Number_of_Casualties	0.051915
Speed_limit	0.026162
Casualty_Reference	0.025996
Location_Easting_OSGR	0.023799
Location_Northing_OSGR	0.023767
Age_of_Casualty	0.022406
LSOA_of_Accident_Location	0.021827
Time	0.021476
1st_Road_Number	0.021326
ocal_Authority_(District)	0.020669
day	0.020487
Vehicle_Manoeuvre	0.020316
Casualty_Type	0.019180
Age_of_Driver	0.019142

4. Logistic Regression

- Logistic regression feature selection works by selecting a subset of relevant features from a larger set of candidate features.
- Logistic regression feature selection ranks the importance of each feature based on the magnitude of its coefficient.
- Logistic regression calculates coefficients during model fitting to minimize the error between predicted probabilities and actual labels.

feature	coefficient
Number_of_Casualties	15.241902
Casualty_Severity	11.549131
Number_of_Vehicles	3.849392
Casualty_Type	2.731302
Did_Police_Officer_Attend_Scene_of_Accident	2.667318
Police_Force	1.464642
Local_Authority_(District)	1.417717
Speed_limit	1.324297
Pedestrian_Movement	1.221576
Casualty_Reference	1.215717
Pedestrian_Location	1.143327
Propulsion_Code	1.068114
Road_Type	1.052925
Vehicle_Type	0.990601
Location_Northing_OSGR	0.615065

Model Proposal

Algorithm	Pros	Cons
Decision Tree	Easy to interpret, handles non-linear relationships, can handle mixed feature types	Prone to overfitting, sensitive to small variations in the data, may create biased trees
Random Forest	Reduces overfitting, handles non-linear relationships, can handle mixed feature types	Computationally expensive, difficult to interpret, may not perform well on imbalanced data
Naive Bayes	Simple and fast, handles high-dimensional data, performs well on small datasets	Assumes independence of features, may not perform well on highly correlated features, can be sensitive to outliers

Model Comparison

Accuracy	Decision Tree	Random Forest	Naïve Bayes
Before Feature Selection	0.83	0.89	0.687
After Feature Selection	0.85	0.854	0.62

Hyperparameter Tuning

- Hyperparameter tuning was performed by using Optuna library.
- Best parameters can be seen in the figure.
- After hyperparameter tuning Random Forest achieved an accuracy of 90%.

```
best_params
```

```
{'n_estimators': 827,
 'max_depth': 38,
 'max_features': 'sqrt',
 'min_samples_split': 7,
 'min_samples_leaf': 1,
 'bootstrap': False,
 'class_weight': 'balanced_subsample'}
```

```
Random Forest Classification Report (with best hyperparameters):
              precision    recall  f1-score   support

           1       0.95      0.93      0.94     15936
           2       0.90      0.85      0.87     15959
           3       0.87      0.93      0.90     16463

    accuracy                           0.90     48358
   macro avg       0.91      0.90      0.90     48358
weighted avg       0.90      0.90      0.90     48358

Accuracy: 0.9036353860788288
```

Columns Importance based on Domain Knowledge and Analysis

- The columns "journey_purpose_driver", "was_vehicle_left_hand_driven", "vehiclereference_x" and "vehiclereference_y" were removed from the dataset as they were not significant in classifying road accident severity.
- The columns "age_band_of_driver" and "age_band_of_casualty" were removed as age of driver and age of casualty features were already present in the dataset.
- "Speed_limit" is an important feature as accidents occurring at higher speeds tend to be fatal.
- "Casualty_Severity" was considered as it provides information on the severity of a person's condition after an accident.
- "Urban_rural_areas" and "Light_Conditions" were also considered as accidents in areas with poor roads, signage, or lighting can be more severe.
- "Vehicle_Type" and "Road_Type" were also considered as accidents involving larger vehicles can cause more severe casualties.

Columns Importance based on Domain Knowledge and Analysis

- The Latitude and Longitude columns were removed due to high correlation (> 0.9) with the local_authority_district column.
- Certain features were discovered to be equally important for prediction based on feature selection methods.
- "Did_Police_Officer_Attend_Scene_of_Accident" was found to be equally important as the absence of police officers could result in minor accidents.
- "Vehicle_Manoeuvre", "Junction_detail", "Junction_location", and "Pedestrian_location" were identified as important features since accidents involving maneuvers and junctions had a higher probability of being severe.
- "Number_of_Vehicles" and "Number_of_Casualties" were highly important features since higher values for both features corresponded to a greater severity of accidents.

Conclusion and Future Work

- Naive Bayes assumes that all features are independent of each other and that they have an equal impact on the classification. This assumption is not always true, hence model did not perform well with an accuracy of 62%.
- The decision tree has a better performance after feature selection as before feature selection the input features are large in number and the model is too complex and overfits the data , but after feature selection the input features will be less ,the model will be simple and reduces the risk of multicollinearity between features. So, the model performs well after feature selection .
- Random Forest is the best performer and hyper parameter tuning was performed on random forest and it got an accuracy of 90% .
- This project helps people and government organizations understand the severity of the situation and take the necessary steps to make the environment safer..
- As to the future scope, we would like to implement this process to on a wide range of data and provide real-time updates on accidents.