Big Data and Data Mining

1. Introduction

A comprehensive analysis of traffic accidents that took place in the United Kingdom year 2020 is provided in this report. Utilizing the aid of a comprehensive data set that includes both documented and reported traffic events, the paper seeks to carry out two main goals:

- It looks to answer questions about the common conditions leading to accidents, the trends in accidents over time, and the distribution of accidents across different regions.
- It also displays the results of machine learning models designed to forecast the probability that an accident will result in a fatality.

The dataset analyzed for this analysis includes key features related to accident circumstances, fatalities, and incident severity. Report is divided into many sections that outline the procedures for getting the dataset ready for modeling and analysis, highlight important findings from data analysis,

2. Analysis

2.1 Data Cleaning

Before delving into data analysis and machine learning model development, a meticulous data cleaning process was executed, employing a combination of methods to ensure accuracy by addressing errors and outliers as needed. The dataset formed three key tables: accidents, casualties, and vehicles.

To begin, the data cleansing involved the removal of rows having null values where proper. Specifically, within the accidents table, 14 rows about location data coordinates—such as location easting & northing OSGR, latitude, and longitude—were eliminated. Due to the nature of location coordinates and the complexities involved in their imputation, addressing these null values proved challenging. However, considering the substantial size of the accidents table dataset (totaling 91,199 rows), the impact of removing these specific rows was considered negligible. Notably, the remaining tables did not show any null values upon inspection. Additionally, a thorough check for duplicates was conducted, resulting in the absence of any duplicate entries within the dataset.

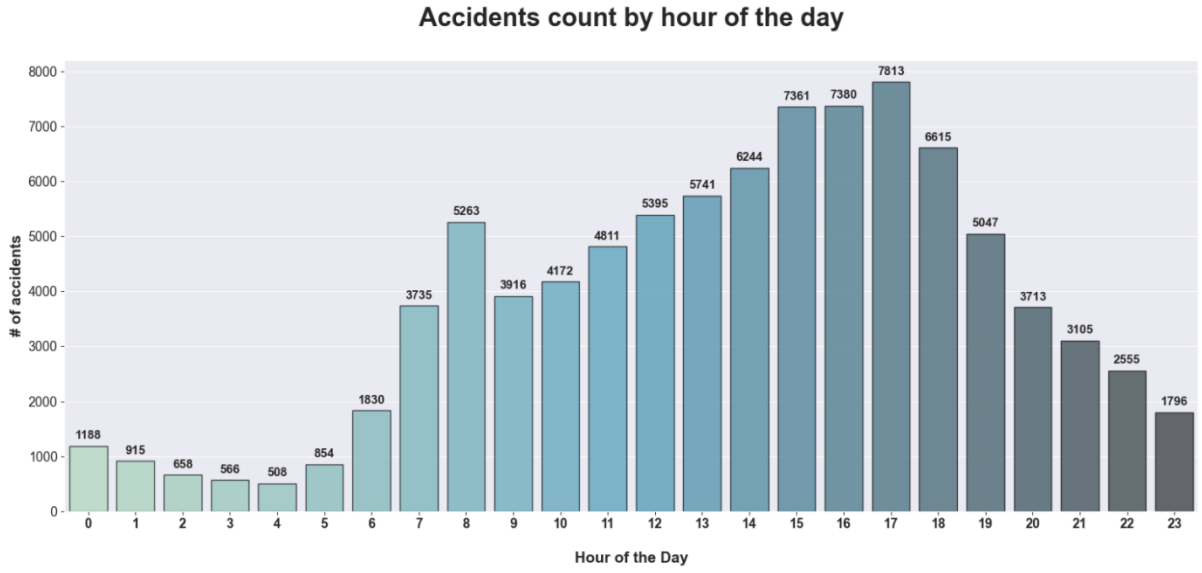
Further cleaning procedures involved confirming the appropriateness of values corresponding to distinctive features by aligning them with the guidelines outlined in the UK (United Kingdom) government's guidance document. This step ensured the accuracy and relevance of the dataset. Moreover, outlier detection was performed using boxplots, particularly focusing on numerical

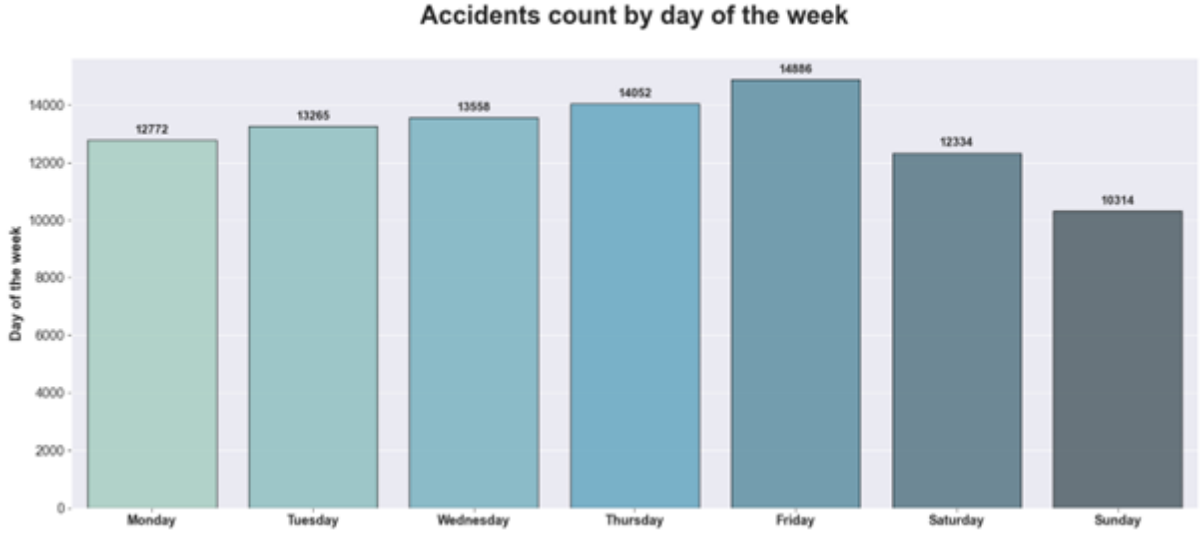
values across the dataset. Noteworthy outliers were detected in variables such as the number of casualties and engine ability. To keep data integrity, these showed outliers were meticulously removed from consideration in the next analysis.

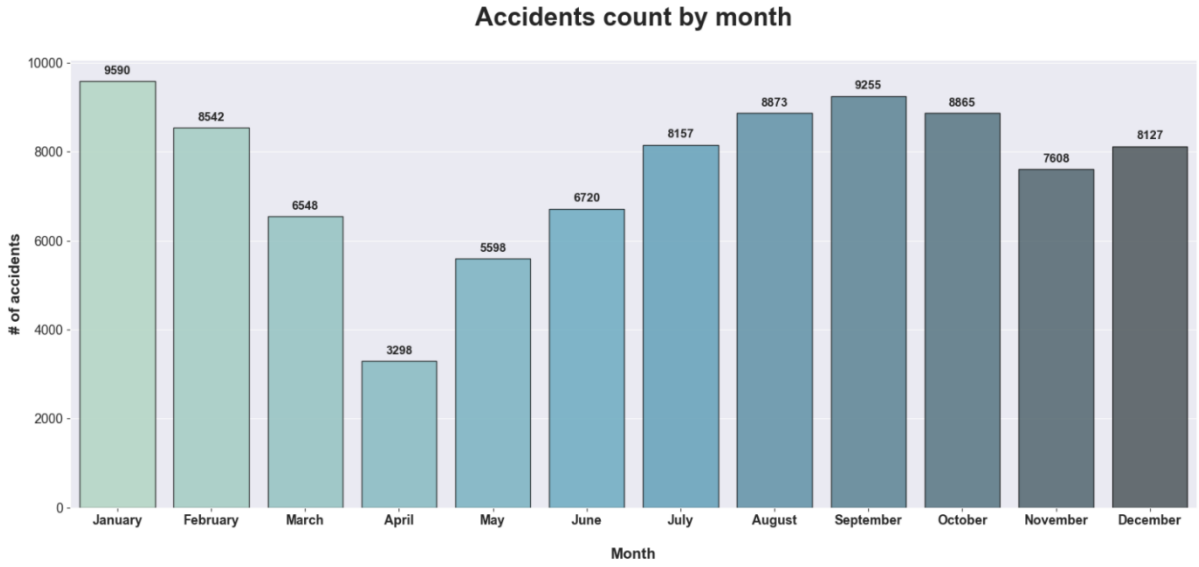
Overall, the rigorous data cleaning process involved the meticulous removal of null values where proper, adherence to proven guidelines for data accuracy, and the careful identification and elimination of outliers—culminating in a refined dataset conducive to robust analysis and model development.

2.2 What are the periods of the day and days of the week that accidents occur?

- In 2020, the most common hour of accident is between the hours of (17:00-18:00), with about 8.57% of all the accidents, with one-third happening between (15:00-19:00), the closest time to this is about (15:00-16:00) with about 8.07%and from (16:00-17:00) accounting for about 8.09%.
- With about 16.33% of all accidents with about 14,886 events, Friday have accounted for most of the accidents in 2020.
- January 2020 saw most of the accident with about 9,590 instances with about 10.52% of accident reported, while the most of April 2020 was the least of accident accounted for with about 3,298 cases with about 3.62% of accident recorded

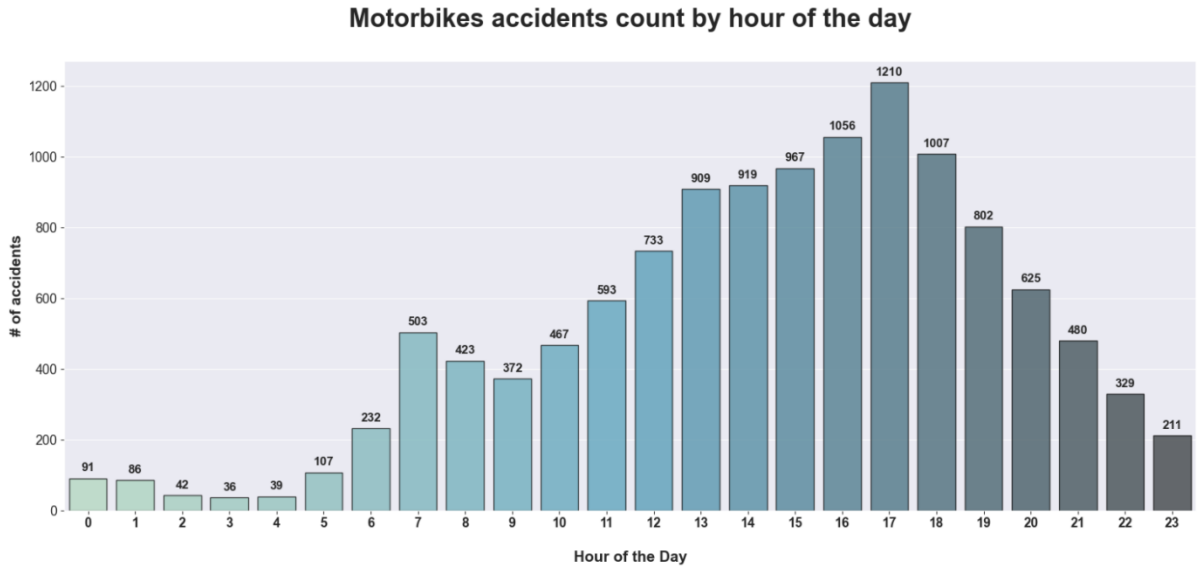

Accidents count by hour of the day
# of accidents
Hour of the Day
0 1000 2000 3000 4000 5000 6000 7000 8000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
1188 915 658 566 508 854 1830 3735 5263 3916 4172 4811 5395 5741 6244 7361 7380 7813 6615 5047 3713 3105 2555 1796

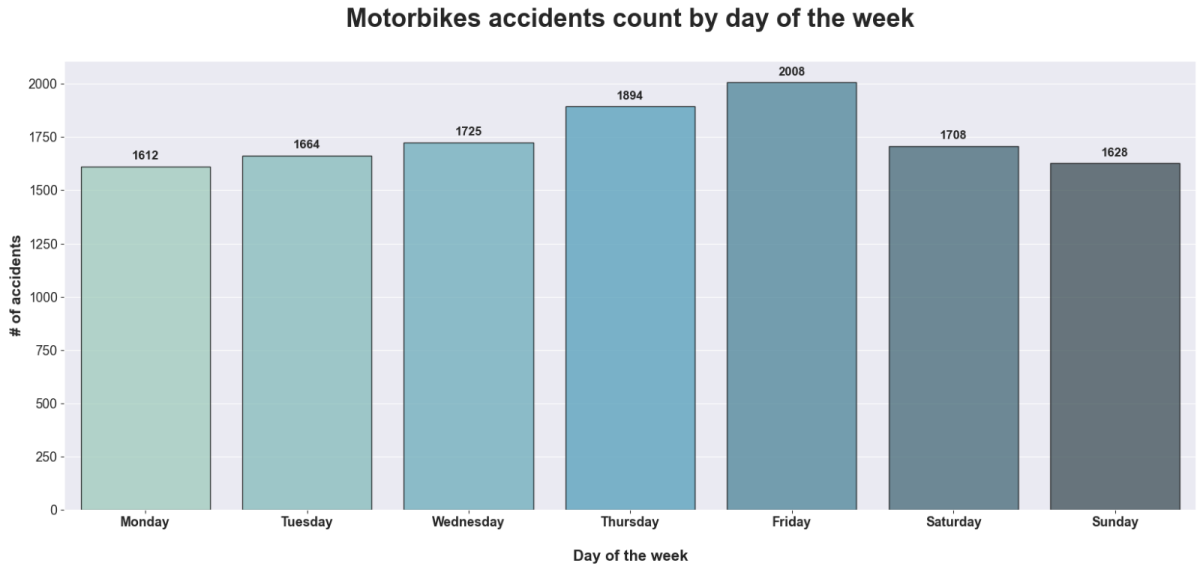

Accidents count by day of the week
Day of the week
0 2000 4000 6000 8000 10000 12000 14000
Monday Tuesday Wednesday Thursday Friday Saturday Sunday
12772 13265 13558 14052 14886 12334 10314



2.2.1 The exact moments of day and days of the week that motorbike accidents occur.

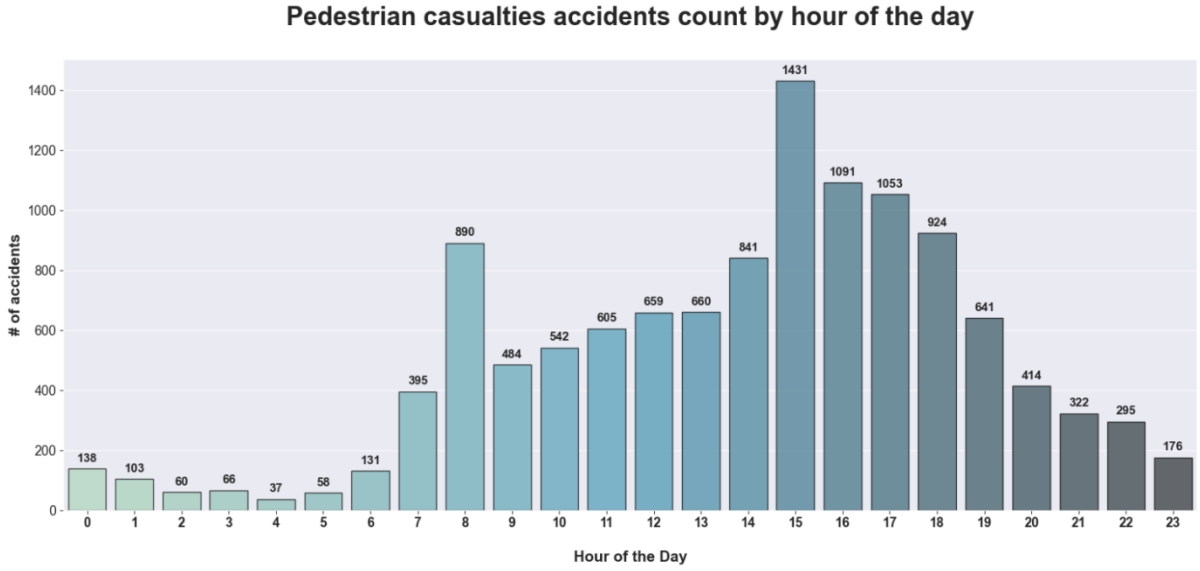
- In the year 2020, the largest percentage of accident give with about 34.65% of motorcycle accident happening between (15:00-19:00)
- In 2020, motorcycles will typically occur on Friday, with about 16.16.41% of all accidents that occurred each week





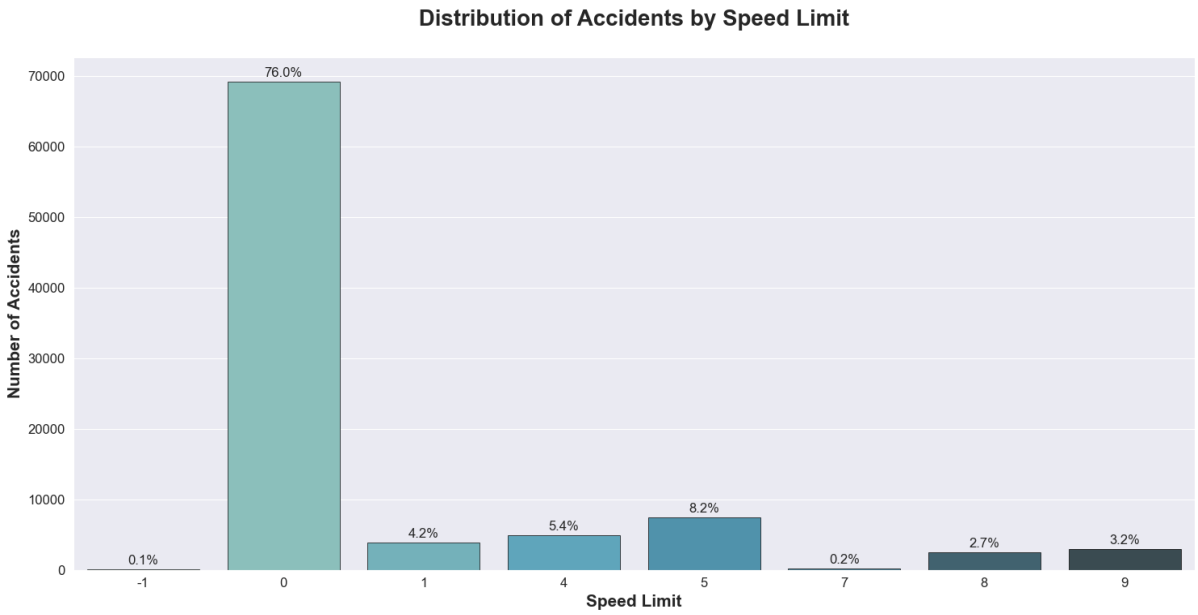
2.2.2 The times and days when pedestrian injuries occurred.

- Of all the pedestrian accidents, 7.41% occurred between 8:00 AM and 9:00 AM, and 11.91% occurred between 3:00 PM and 4:00 PM.
- surprisingly, the lowest number of pedestrian accidents happened on Sundays, in contrast to Fridays, which accounted for the highest incidence at 17.7%. Specifically, only 8.15% of accidents involving pedestrians occurred on Sundays.

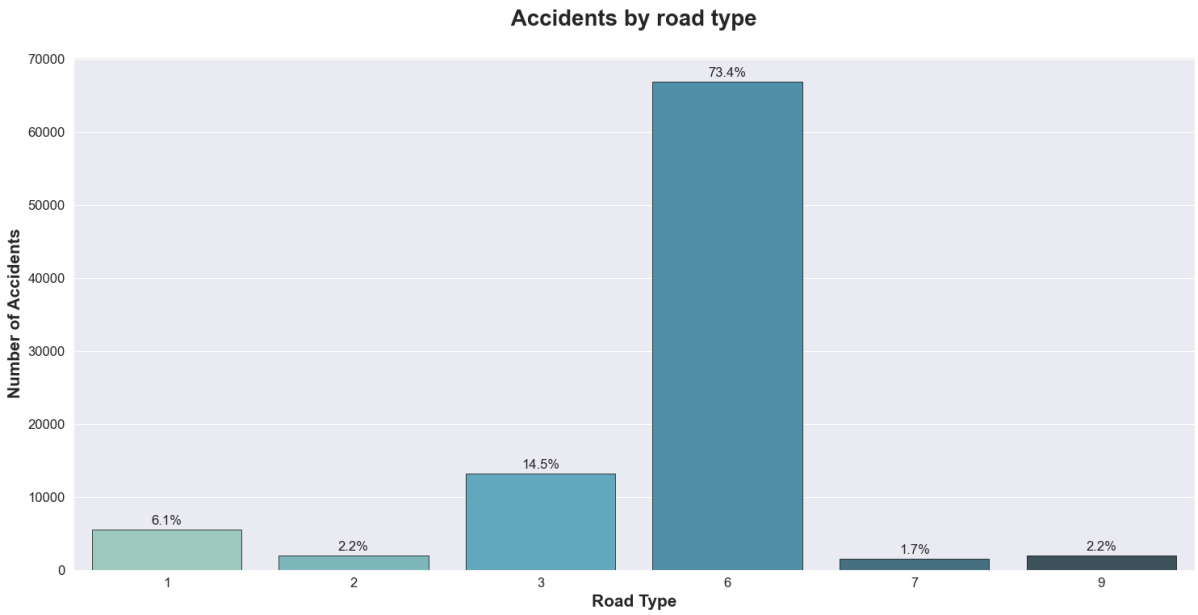


2.3 Where Accidents Happen

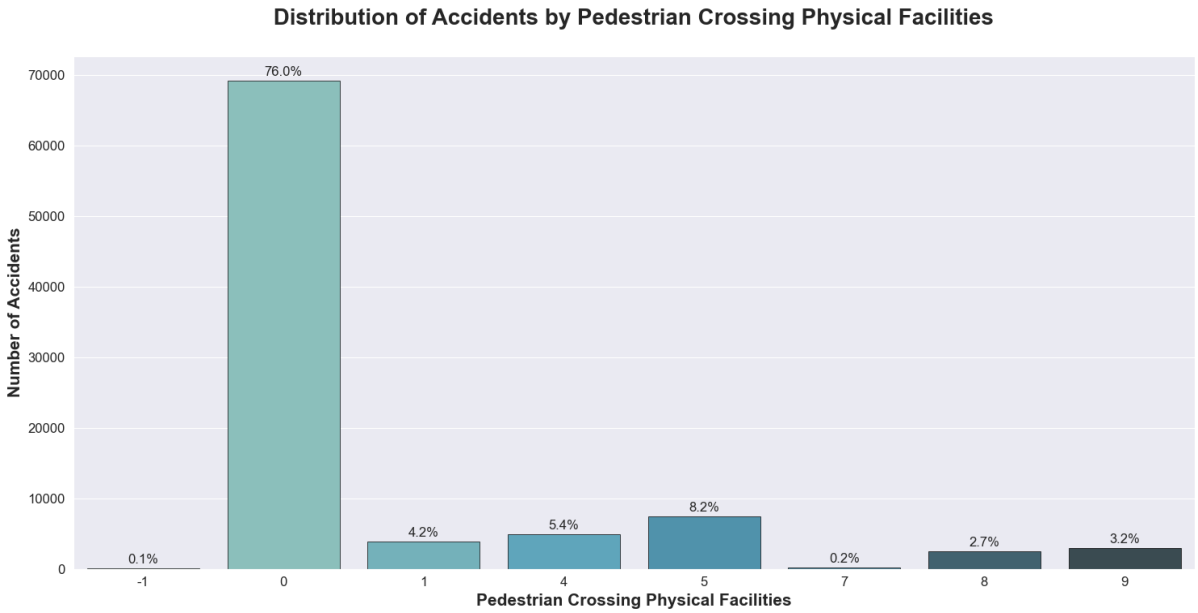
Interestingly, areas marked with a 30mph speed limit had a notably higher share of accidents, making up 57% of all incidents in 2020. This stands out considering that zones with 60 mph and 70 mph limitations on speeds just recording for 12.5% and 5.1% of impacts, respectively.



Single carriageways were the type of road that met the highest number of incidents in 2020. constituting 73.4% of all reported incidents.

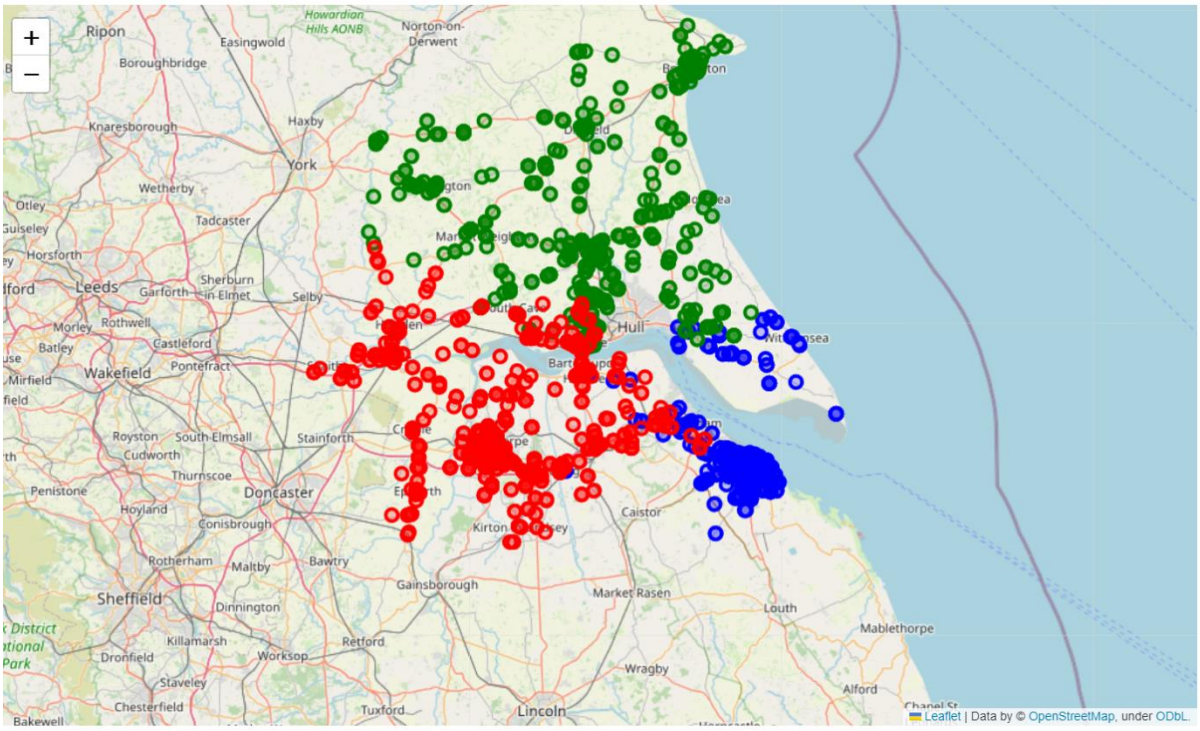


The data strongly recommends enhancing crossing facilities to curb accidents, considering that 76% of all reported incidents occur in these areas. Implementing such measures could significantly lower pedestrian casualties in accidents.



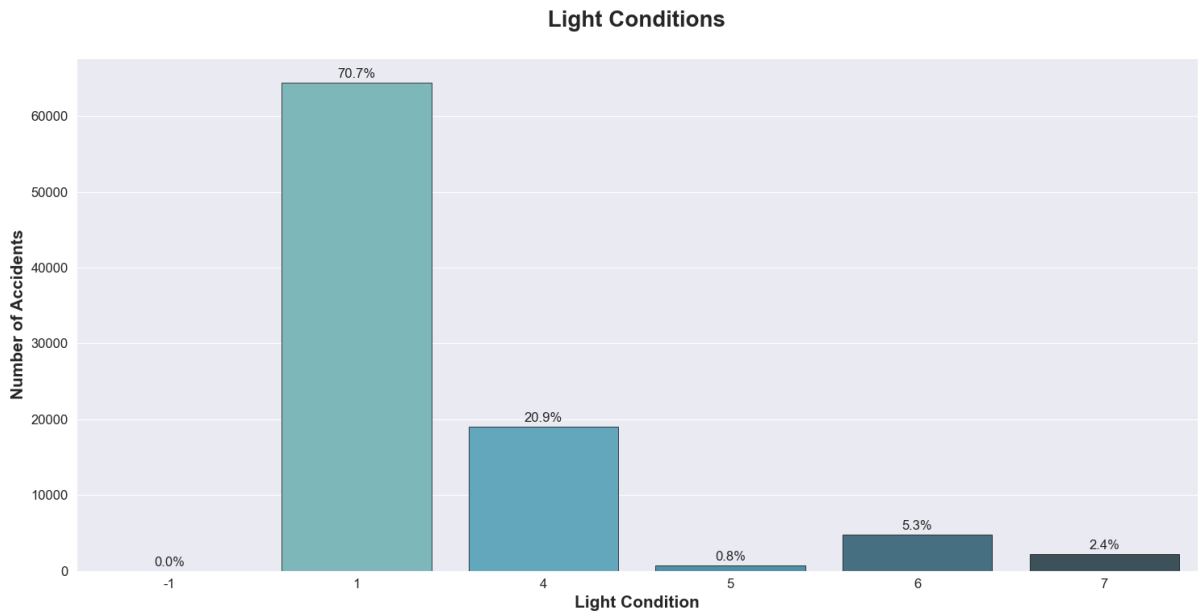
2.3.1 Accidents in the Humberside region

Utilizing local authority district codes, four key districts within the Humberside region were chosen: Kingston upon Hull, City of; East Riding of Yorkshire; North Lincolnshire; and Northeast Lincolnshire. Employing the K-means algorithm on longitude and latitude coordinates data revealed the presence of three distinct accident clusters across the region. These clusters, as illustrated in the figure below, encompass areas within Hull and East Riding, specific locations in North Lincolnshire, and areas within Northeast Lincolnshire.

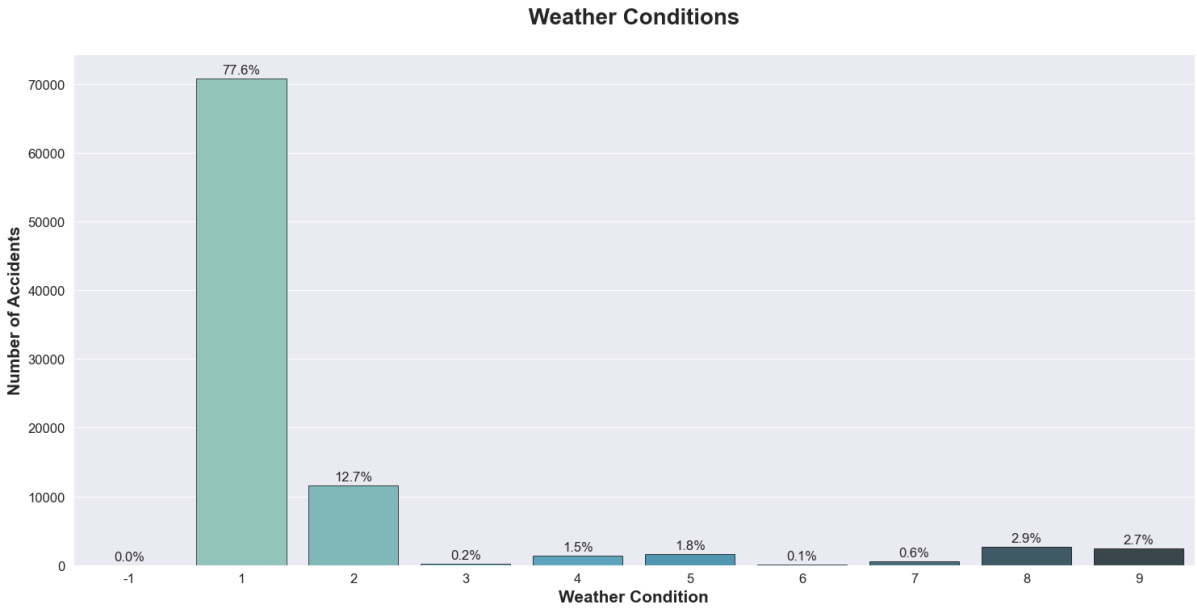


2.4 Situations when accidents happen.

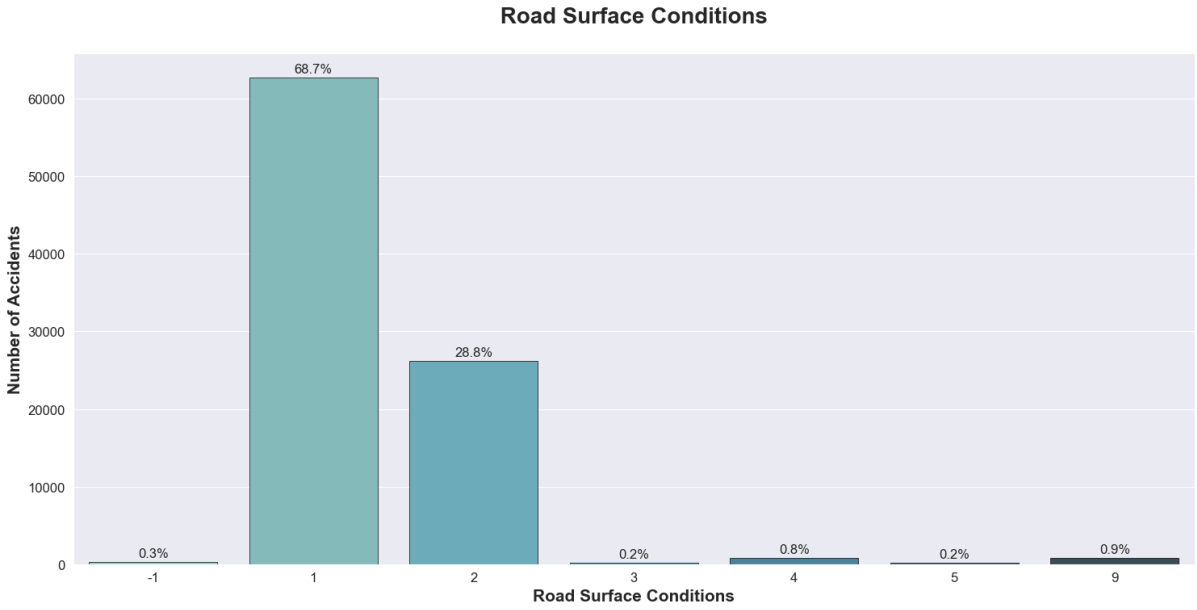
Seventy percent of all accidents occur during daylight conditions, while an added 20% of accidents occur in the darkness with lighting present.



Clear weather conditions, without high winds or rain, contributed to 90.3% of all accidents in 2020.



68.7% of accidents occurred on dry road surfaces, while 28.8% took place on wet or damp roads.



2.4.1 Utilizing the Apriori algorithm for association rules extraction.

The Apriori algorithm, a technique for uncovering hidden patterns, was used to uncover relationships within the dataset. This method explores how certain variables relate to each other by showing if/then rules based on their co-occurrence. It measures support, showing how often variables appear together, and confidence, measuring the strength of this association (Java Point, 2021).

In this analysis, only rules with a minimum support of 40% were considered. These criteria yielded 8 rules, one of which shows a strong association: "Daylight conditions have a 56% support with slight accident severity and a confidence score of 79%." This rule is highly relevant in the dataset, showing a substantial relationship between these variables.

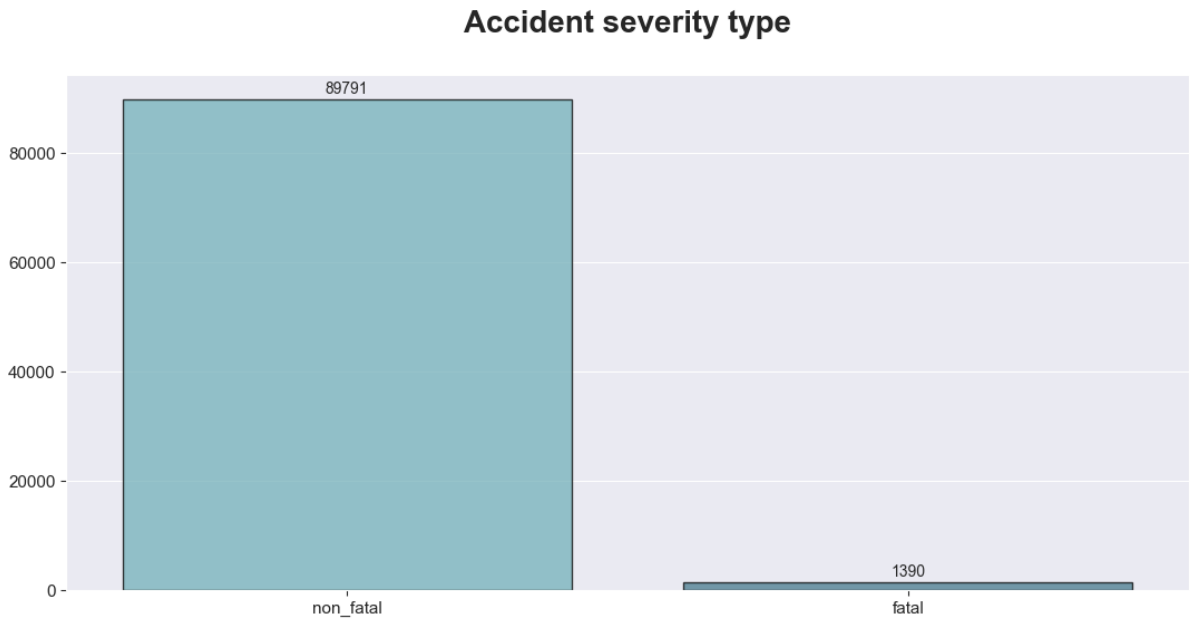
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
22	(light_conditions_1)	(accident_severity_3)	0.706781	0.783497	0.559349	0.791404	1.010092	0.005588	1.037906	0.034074
72	(special_conditions_at_site_0, light_condition...	(accident_severity_3)	0.677817	0.783497	0.535462	0.789981	1.008277	0.004395	1.030877	0.025478
8	(speed_limit_30)	(accident_severity_3)	0.573113	0.783497	0.460052	0.802725	1.024542	0.011020	1.097470	0.056113
54	(weather_conditions_1, light_conditions_1)	(accident_severity_3)	0.578684	0.783497	0.453845	0.784270	1.000987	0.000448	1.003585	0.002340
36	(special_conditions_at_site_0, speed_limit_30)	(accident_severity_3)	0.552374	0.783497	0.442362	0.800838	1.022133	0.009579	1.087071	0.048375
108	(weather_conditions_1, special_conditions_at_s...	(accident_severity_3)	0.559184	0.783497	0.438216	0.783670	1.000222	0.000097	1.000803	0.000503
66	(light_conditions_1, road_surface_conditions_1)	(accident_severity_3)	0.534607	0.783497	0.420855	0.787224	1.004757	0.001992	1.017516	0.010173
122	(special_conditions_at_site_0, light_condition...	(accident_severity_3)	0.517520	0.783497	0.406971	0.786386	1.003688	0.001496	1.013528	0.007616

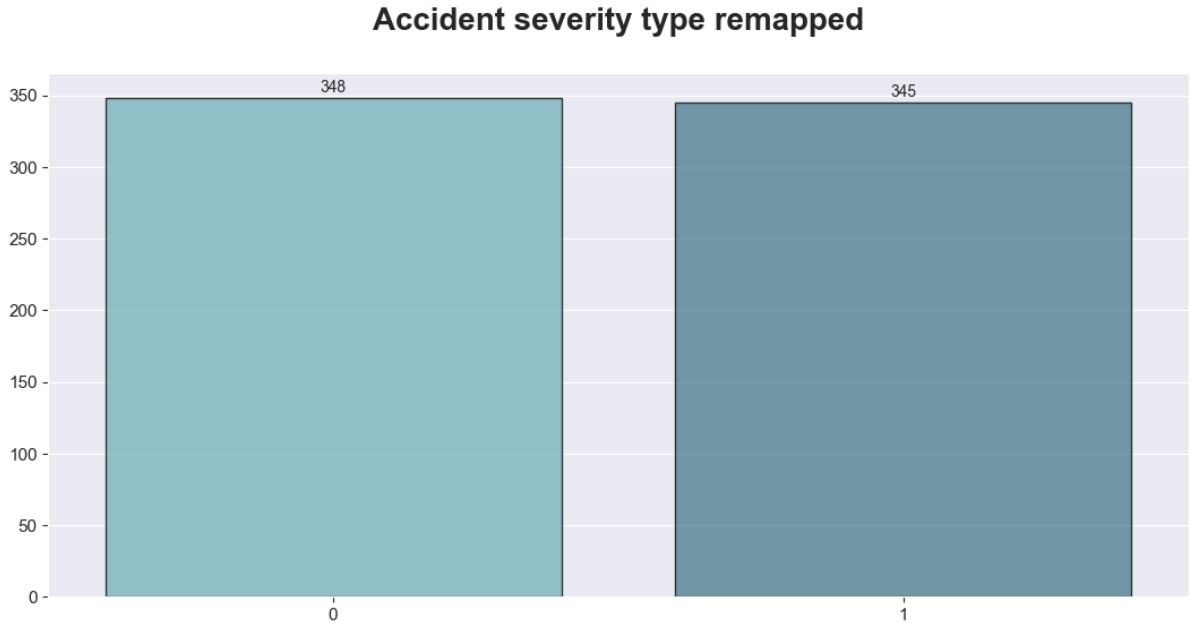
3. Predictions

This report aims to highlight the outcomes of machine learning models designed for accurately forecasting accident occurrences. Five models were employed: Random Forest, Decision Trees, KNN, Support Vector Machine, and XG Boost classifiers.

The dataset faced significant imbalance, with fatal accidents making up only 1.5% of the data after grouping serious and slight injuries into a single non-fatal class. To prepare the data for modeling:

- Dropped unknown, missing, or out-of-range values, ensuring only identifiable variables were used.
- Grouped slight and serious injuries into a single non-fatal class.
- Standardized numerical features and performed one-hot encoding for categorical variables.
- Addressed class imbalance by manually performing random under sampling, equating the number of observations in both majority and minority classes.
- Utilized Select K-Best to show top features influencing the target class.
- Optimized model performance by employing Randomized Search for hyperparameter tuning.
- These steps were crucial in preparing the dataset and perfecting the models for correct accident predictions despite the imbalanced nature of the data.





The models trained on the under sampled data performed impressively, showing high accuracy scores exceeding 98%. Metrics like precision, recall, and f1-score also displayed exceptional performance across all models.

Yet, when applying the best parameters from these models to predict the entire dataset, balancing the minority class using SMOTE, the results were disappointing. The models struggled, solely predicting the majority class without capturing the minority class effectively.

4 Recommendations

The analysis highlights a high occurrence of accidents in 30mph zones, constituting 57.3% of all incidents. It suggests improving speed limit adherence, possibly using speed display boards or transitioning more areas to 20mph zones, which have four times fewer accidents despite accounting for 12.3%.

Moreover, 76% of accidents occur where there are no nearby crossing facilities within 50 meters. This shows the need for more crossings, particularly in 20mph and 30mph zones, to enhance pedestrian safety and reduce vehicle accidents caused by pedestrians crossing indiscriminately.

5 References

Java Point (2021) Association Rule Learning. Available online: https://www.javatpoint.com/association-rule-learning [Accessed 11/08/2023].

Sanwoola, K. (2023) Big Data and Data mining. thesis.