UNIVERSITY OF SURREY

INDIVIDUAL COURSEWORK COVERSHEET

MSc. Business Analytics

Module - Data mining and text analytics with application in SAS

Assignment - Exploring Road Traffic Accident Data and Text Analytics Insights

Date - 7th Jan / 2024

Name - Mohamed Shahil Shahul Hameed

URN No - 6809743

Words - 4100

Introduction

This ongoing technical study begins with an introductory section that presents contextual information for the comprehensive analysis of Road Accident Surrey data utilising SAS Viya software. As a guide, it provides a clear explanation of the report's structured framework and offers an in-depth assessment of the primary actions that were accomplished. The objective of the study is to gain valuable insights from the information. The introduction highlights the three fundamental tasks that structure the subsequent study: data exploration, predictive modelling, and text analysis. Each project contributes to the overall goal of understanding and improving road safety in Surrey, with extreme attention to detail and careful thought.

The planned segmentation of the task into three segments reflects a systematic approach. The primary objective is to set the foundation by examining each aspect of the data, addressing issues related to data quality, and cleansing the dataset. The subsequent task focuses on predictive modelling, showcasing the ability of SAS Viya to precisely predict accident severity and provide valuable insights. The third task, which is to analyse tweets to capture the public sentiment and incorporate a community perspective into the investigation.

Task 1 – Data Exploration and Cleaning

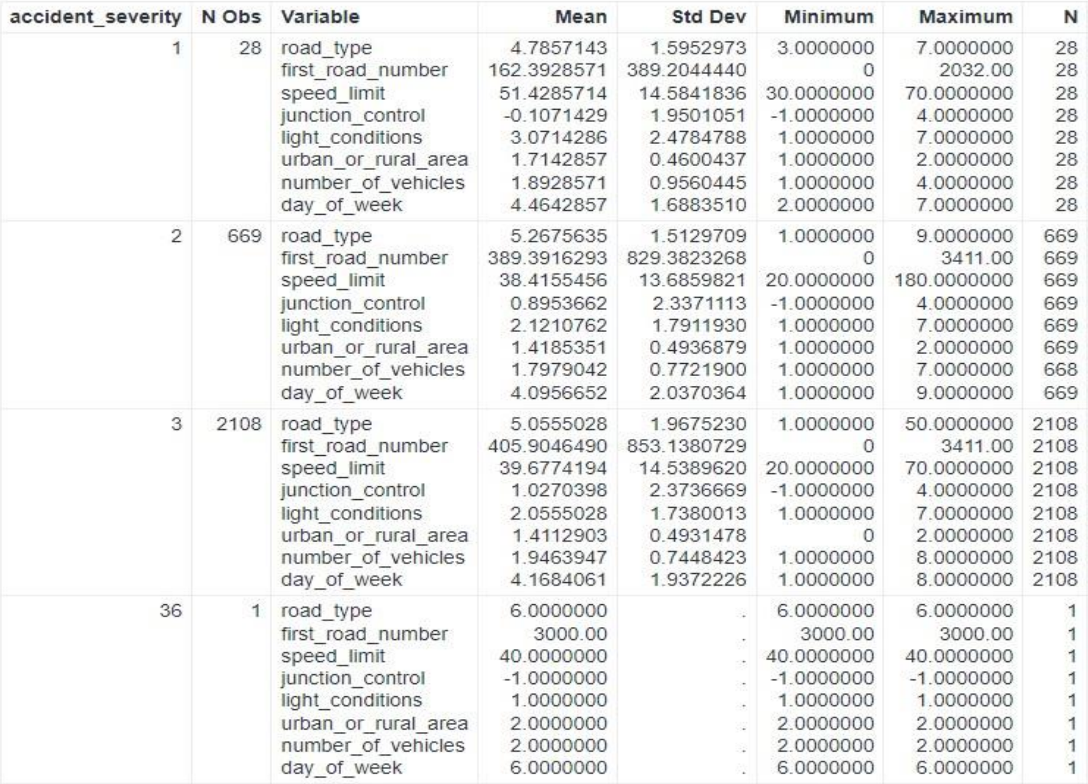
Exploratory data analysis

The dataset contains a variety of information related to road accidents. Here is a brief overview of some of the columns present:

- accident_reference: A unique identifier for each accident.
- location_easting_osgr, location_northing_osgr: The Easting and Northing coordinates of the accident location
- longitude, latitude: Geographical coordinates of the accident location.
- police_force: A code for the police force that attended the scene.
- accident_severity: A code indicating the severity of the accident.

- number_of_vehicles: The number of vehicles involved in the accident.
- number_of_casualties: The number of casualties in the accident.
- date: The date of the accident.
- day_of_week: The day of the week on which the accident occurred.
- time: The time at which the accident occurred.
- urban_or_rural_area: A code indicating whether the accident occurred in an urban or rural area.
- did_police_officer_attend_scene_of_accident: Indicates if a police officer attended the scene.

Summary Statistics



accident_severity	N Obs	Variable	Mean	Std Dev	Minimum	Maximum	N
1	28	road_type	4.7857143	1.5952973	3.0000000	7.0000000	28
		first_road_number	162.3928571	389.2044440	0	2032.00	28
		speed_limit	51.4285714	14.5841836	30.0000000	70.0000000	28
		junction_control	-0.1071429	1.9501051	-1.0000000	4.0000000	28
		light_conditions	3.0714286	2.4784788	1.0000000	7.0000000	28
		urban_or_rural_area	1.7142857	0.4600437	1.0000000	2.0000000	28
		number_of_vehicles	1.8928571	0.9560445	1.0000000	4.0000000	28
		day_of_week	4.4642857	1.6883510	2.0000000	7.0000000	28
2	669	road_type	5.2675635	1.5129709	1.0000000	9.0000000	669
		first_road_number	389.3916293	829.3823268	0	3411.00	669
		speed_limit	38.4155456	13.6859821	20.0000000	180.0000000	669
		junction_control	0.8953662	2.3371113	-1.0000000	4.0000000	669
		light_conditions	2.1210762	1.7911930	1.0000000	7.0000000	669
		urban_or_rural_area	1.4185351	0.4936879	1.0000000	2.0000000	669
		number_of_vehicles	1.7979042	0.7721900	1.0000000	7.0000000	668
		day_of_week	4.0956652	2.0370364	1.0000000	9.0000000	669
3	2108	road_type	5.0555028	1.9675230	1.0000000	50.0000000	2108
		first_road_number	405.9046490	853.1380729	0	3411.00	2108
		speed_limit	39.6774194	14.5389620	20.0000000	70.0000000	2108
		junction_control	1.0270398	2.3736669	-1.0000000	4.0000000	2108
		light_conditions	2.0555028	1.7380013	1.0000000	7.0000000	2108
		urban_or_rural_area	1.4112903	0.4931478	0	2.0000000	2108
		number_of_vehicles	1.9463947	0.7448423	1.0000000	8.0000000	2108
		day_of_week	4.1684061	1.9372226	1.0000000	8.0000000	2108
36	1	road_type	6.0000000	.	6.0000000	6.0000000	1
		first_road_number	3000.00	.	3000.00	3000.00	1
		speed_limit	40.0000000	.	40.0000000	40.0000000	1
		junction_control	-1.0000000	.	-1.0000000	-1.0000000	1
		light_conditions	1.0000000	.	1.0000000	1.0000000	1
		urban_or_rural_area	2.0000000	.	2.0000000	2.0000000	1
		number_of_vehicles	2.0000000	.	2.0000000	2.0000000	1
		day_of_week	6.0000000	.	6.0000000	6.0000000	1

Figure 1

Day of Week

• Central Tendency: The day_of_week variable represents the specific day of the week when accidents took place, with values ranging from 1 (Sunday) to 7 (Saturday). The average statistics for different levels of severity indicate that there may be fluctuations in the frequency of accidents during the week.

• Dispersion: The day_of_week variable has low standard deviation and range, as anticipated, due to its categorical nature with a restricted number of possible values.

First Road Number

• Central Tendency: The first_road_number is a numerical identifier for the specific road where the accident occurred. Mean values exhibit significant variation in relation to accident severity, indicating that specific roads may have an increased risk towards severe accidents.

• Dispersion: The standard deviation is significantly higher for less severe accidents, indicating a wider variety of road numbers involved in such instances.

Junction Control

Central Tendency and Dispersion: This variable categorises the type of junction control at the accident site. The presence of a negative mean in certain categories may suggest the presence of missing or encoded data, which should be examined more closely.

Light Conditions

• Central Tendency: The light conditions show that there are different average light conditions for different levels of accident severity. The conditions can be logically encoded to indicate various states such as daylight, darkness, and the presence of streetlights.

• Dispersion: The standard deviation numbers indicate that there is variation in light conditions across different levels of accident severity.

Number of Vehicles

• Central Tendency: This variable reflects the number of vehicles involved in accidents. The average number of vehicles involved in accidents usually varies from 1 to 2, with little consistency across different levels of severity.

• Dispersion: The low standard deviation indicates that the most of accidents involve an identical number of vehicles.

Road Type

• Central Tendency: road_type is likely a category variable with numeric coding. The median results for severity show that there may be a connection among different types of roads and the severity of accidents.

• Dispersion: Different standard deviations suggest that certain types of roads have a more consistent correlation with the severity of accidents compared to others.

Speed Limit

• Central Tendency: The speed_limit variable provides data regarding the legal speed of the road throughout the accident. The average speed limit displays variability.

• Dispersion: The quantitative measures of standard deviation and range for speed limit reveal substantial variability in the speed limitations of roads where accidents take place.

Urban or Rural Area

•Central Tendency: The variable "urban_or_rural_area" is presumably encoded with values 1 and 2 to differentiate between urban and rural areas, respectively, indicating the central tendency. The average numbers indicate that accidents are distributed across both areas.

• Dispersion: The small standard deviation suggests a noticeable disparity between accident locations in urban and rural areas.

Months

• Central Tendency and Dispersion: The variable "months" indicates the month where accidents took place. The data displays cyclical and category characteristics, which could suggest a correlation with seasons.

Data Quality Issues

• There is a potential outlier or error in the accident_severity variable, as it currently has a maximum value of 36, which comes outside the expected range.

• The count of number_of_vehicles are 2805, indicating that there is 1 missing value, as the other counts are 2806.

The variable "urban_or_rural_area" should have a minimum value of either 1 (urban) or 2 (rural), however it now has a value of 0. This could be a data stepping into mistake.

Pie chart - Urban or Rural Area Distribution

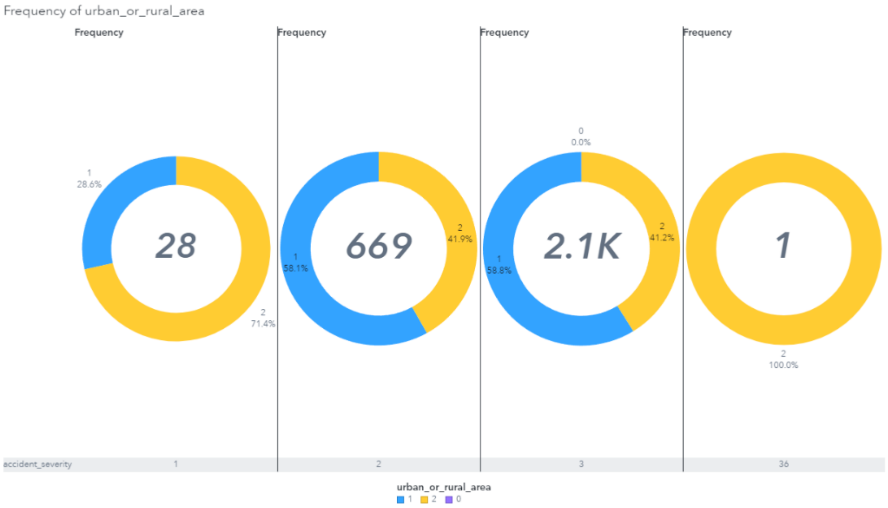


Figure 2

The above chart provides an in-depth assessment of accident severity, with severity levels ranging from 1 (most severe) to 3 (least severe), in rural as well as urban situations. The data indicates an evenly distributed distribution of accident severity in both urban and rural areas for severity levels 1 and 3. However, for accidents categorised as severity level 2, there is a little greater number of incidents taking place in towns and cities. The sample size for severity level 36 is insufficient to make any meaningful conclusions.

Bar Chart - Day of the Week Distribution

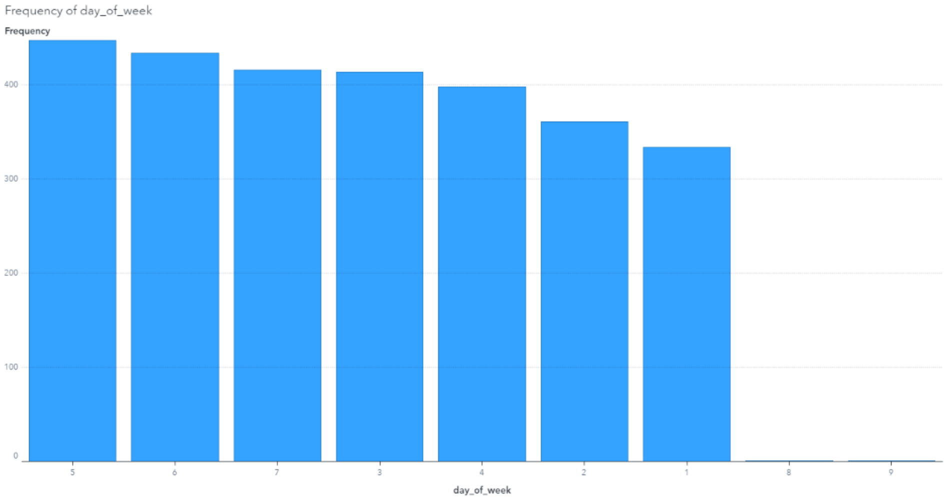


Figure 3

The bar chart displays the frequency of accidents depending on the day_of_week variable, assuming that the day codes follow to a traditional calendar sequence. Although the distribution is generally even, there is a minor rise in accidents throughout the middle of the week, reaching its highest point on day 5, which is likely to be a Friday, a day known for heightened travel and traffic.

Box Plot – Day of week by accident severity

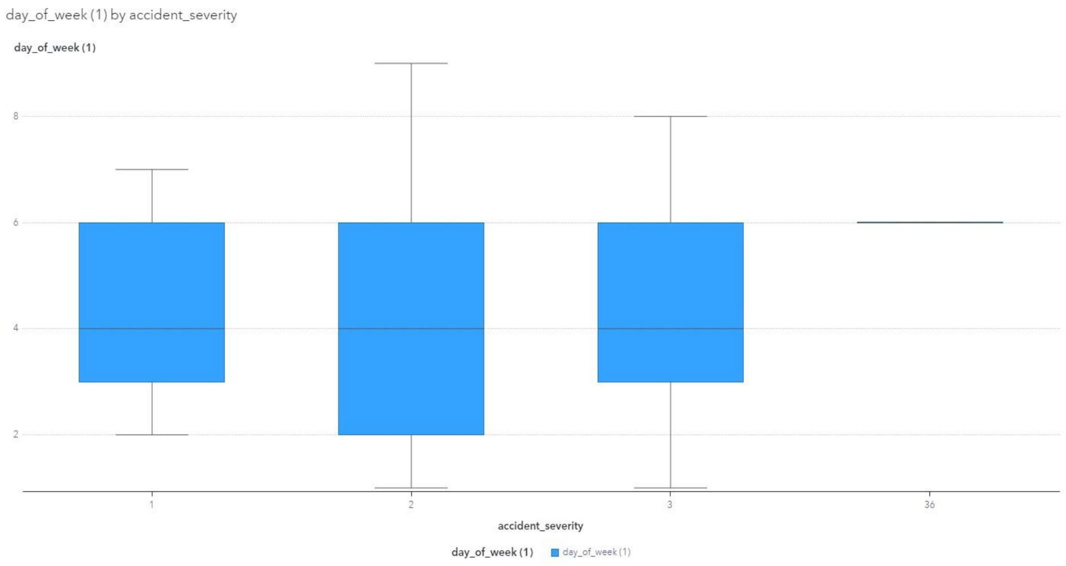


Figure 4

The box plot presents an analysis of accident severities categorized into three standard levels and one anomalous level. The median value across severities 1, 2, and 3 is consistently around 5, indicating a stable median accident count across these categories. The Interquartile Range (IQR) for these severities is relatively tight, which points to a consistent spread in the number of accidents occurring within each severity level; there are no wide variances or extreme values that would indicate irregularities in the data. The presence of a category '36' is not standard for accident severity and it is an outlier.

Scatter Plot - Correlation of Selected Measures

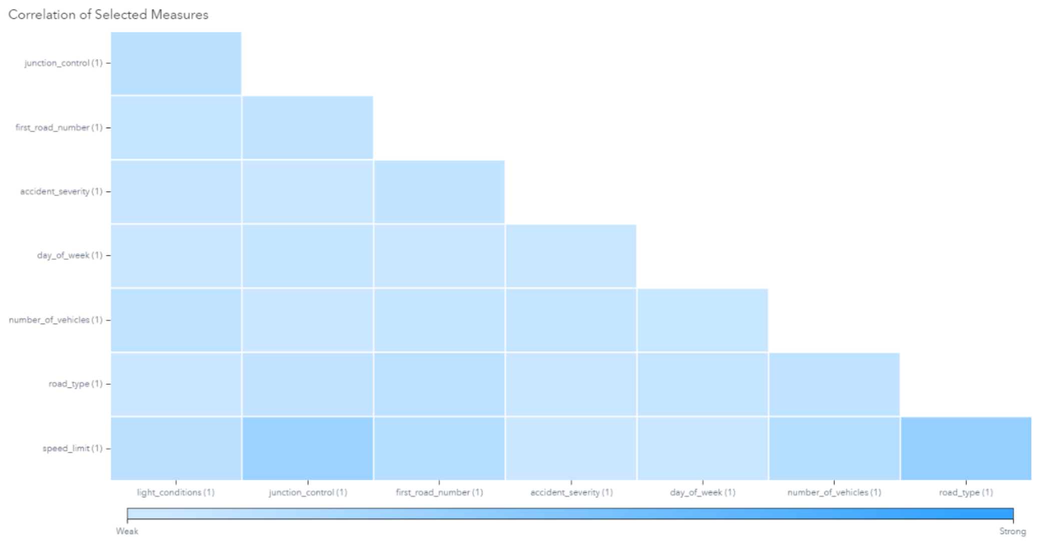


Figure 5

The correlation matrix provides a visual representation of the strength of the relationships between different variables. It appears like there is a relationship between the kind of road and the number of vehicles involved in an accident. This implies that some road types are more prone to multi-vehicle collisions.

Geographic Distribution of Accident Severity

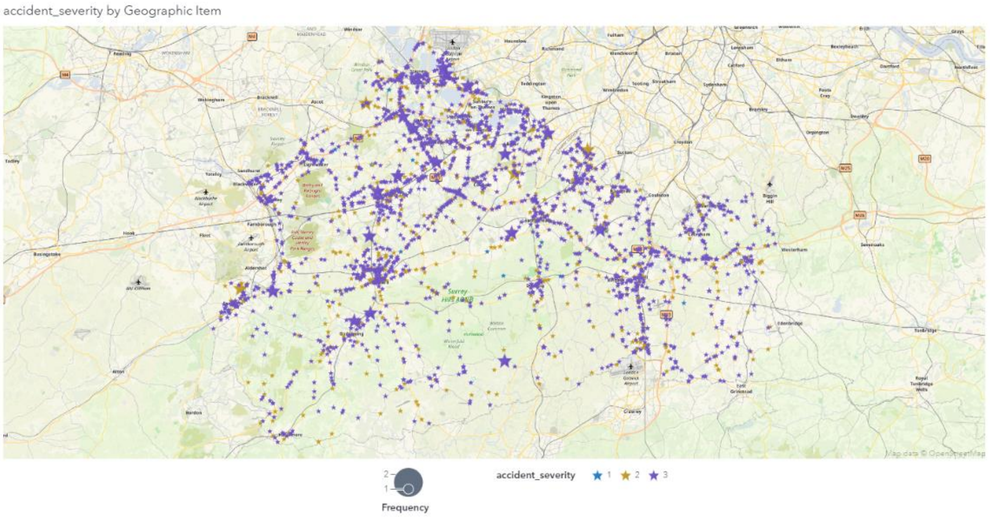


Figure 6

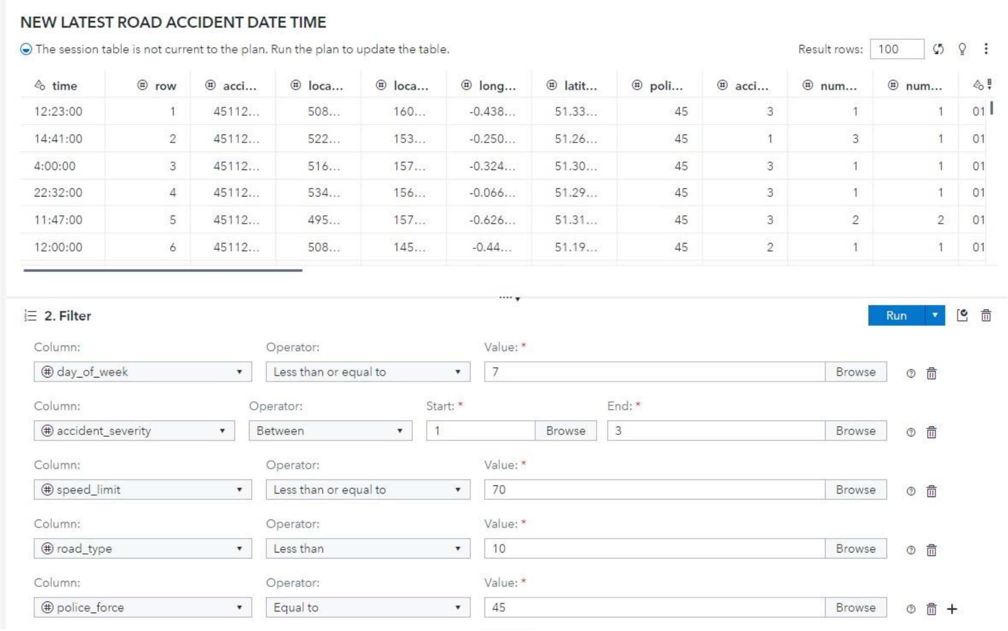
The spatial distribution of accidents, illustrated on a geo map, reveals clusters of accidents with greater severity. These clusters may indicate specific regions that necessitate focused safety actions or enhancements to infrastructure. Additionally, it enables the representation of accident trends in relation to geographical elements such as road intersections, metropolitan hubs, or recognised problematic areas.

Patterns and Relationships

The visualizations suggest the following patterns and relationships:

Urban areas show a greater incidence of accidents, which can be linked to multiple variables like increased traffic congestion, higher vehicle volume, and the complexity of metropolitan road networks. The day of the week does not appear to have a significant impact on the frequency of accidents, although there is an obvious rise towards the end of the week. This pattern may be associated with enhanced social and commercial endeavours. There is a significant association between the severity of accidents and the number of casualties, where more severe accidents lead to a higher number of casualties. The importance of safety initiatives in preventing severe accidents is made clear by this relationship.

Filtering of Data



NEW LATEST ROAD ACCIDENT DATE TIME

The session table is not current to the plan. Run the plan to update the table.

Result rows: 100

time	row	acci...	loca...	loca...	long...	latit...	poli...	acci...	num...	num...	
12:23:00	1	45112...	508...	160...	-0.438...	51.33...	45	3	1	1	01
14:41:00	2	45112...	522...	153...	-0.250...	51.26...	45	1	3	1	01
4:00:00	3	45112...	516...	157...	-0.324...	51.30...	45	3	1	1	01
22:32:00	4	45112...	534...	156...	-0.066...	51.29...	45	3	1	1	01
11:47:00	5	45112...	495...	157...	-0.626...	51.31...	45	3	2	2	01
12:00:00	6	45112...	508...	145...	-0.44...	51.19...	45	2	1	1	01

2. Filter

Column: day_of_week — Operator: Less than or equal to — Value: * 7

Column: accident_severity — Operator: Between — Start: * 1 — End: * 3

Column: speed_limit — Operator: Less than or equal to — Value: * 70

Column: road_type — Operator: Less than — Value: * 10

Column: police_force — Operator: Equal to — Value: * 45

Figure 7

The filter step is more detailed and appears to use multiple criteria for refining the dataset:

The 'day_of_week' filter is set up to include records with values less than or equal to '7', representing the days of the week, to eliminate any incorrect entries.

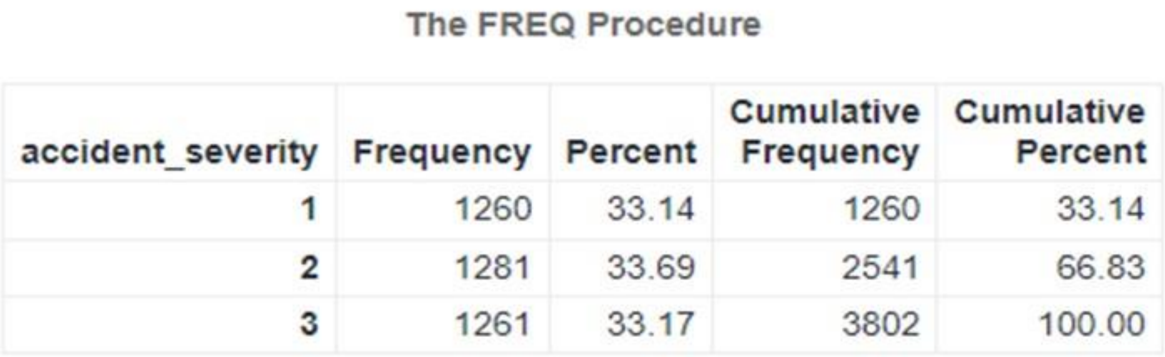
The ‘accident_severity’ is defined within the range of '1' to '3'. This likely refers to a categorization of accident severity, where only these specific levels are considered meaningful for the analysis.

The ‘speed_limit’ filter is activated for speed restrictions that are 70 or less. This stage may aim to concentrate on regions with speed restrictions falling within a specific range.

The ‘road_type’ filters the value less than '10' indicates that are encoded numerically and the analysis is restricted to specific categories.

The ‘police_force’ filter is implemented to select records where the police force is '45'. The analysis could be narrowed down to a particular jurisdiction.

Balancing Data



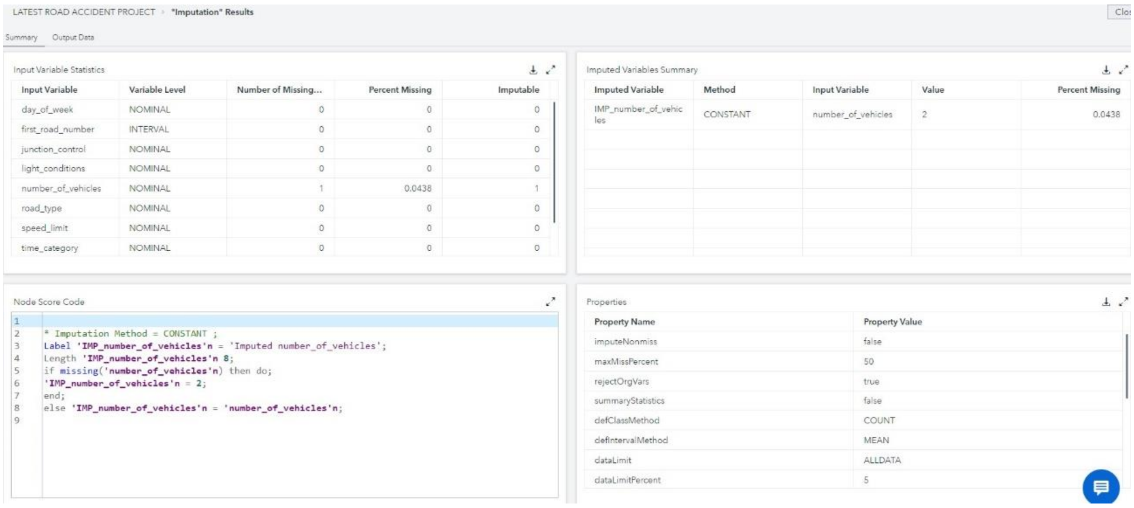
The FREQ Procedure

accident_severity	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1260	33.14	1260	33.14
2	1281	33.69	2541	66.83
3	1261	33.17	3802	100.00

Figure 8

The FREQ process in SAS Viya presents a well-balanced distribution across three accident severity categories, with each category representing approximately one-third of the dataset. The frequencies and percentages of severity levels 1, 2, and 3 are almost the same, with values of 1260, 1281, and 1261, and 33.14%, 33.69%, and 33.17% respectively. This indicates a uniform distribution of these severity levels. The equal distribution among severity levels in analytical models is useful as it prevents any of them from being ignored or outnumbered, thus preventing a possible error in the results.

Missing Values and Imputation Process



Input Variable Statistics

Input Variable	Variable Level	Number of Missing...	Percent Missing	Imputable
day_of_week	NOMINAL	0	0	0
first_road_number	INTERVAL	0	0	0
junction_control	NOMINAL	0	0	0
light_conditions	NOMINAL	0	0	0
number_of_vehicles	NOMINAL	1	0.0438	1
road_type	NOMINAL	0	0	0
speed_limit	NOMINAL	0	0	0
time_category	NOMINAL	0	0	0

Imputed Variables Summary

Imputed Variable	Method	Input Variable	Value	Percent Missing
IMP_number_of_vehicles	CONSTANT	number_of_vehicles	2	0.0438

Node Score Code

```

* Imputation Method = CONSTANT ;
Label 'IMP_number_of_vehicles'n = 'Imputed number_of_vehicles';
Length 'IMP_number_of_vehicles'n 8;
if missing('number_of_vehicles'n) then do;
'IMP_number_of_vehicles'n = 2;
end;
else 'IMP_number_of_vehicles'n = 'number_of_vehicles'n;

```

Properties

Property Name	Property Value
imputeNonmiss	false
maxMissPercent	50
rejectOrgVars	true
summaryStatistics	false
defClassMethod	COUNT
defIntervalMethod	MEAN
dataLimit	ALLDATA
dataLimitPercent	5

Figure 9

The variable number_of_vehicles has been identified to have a single missing entry, which accounts for 0.0438% of the dataset.

The problem of missing data for the variable "number_of_vehicles" has been resolved by employing a constant value imputation technique. This method replaces all null values in a variable with a predefined value. In this scenario, the chosen constant value is '2', which is the mode used for imputation.

Task 2 – Predicting Accident Severity

Scenario

During busy city traffic, a delivery truck rushing to a deadline didn't notice the red light because the driver was on the phone. The truck suddenly stopped, causing many cars behind to crash into each other.

Key variables for scenario

first_road_number , number_of_vehicles , day_of_week , junction_control , weather_conditions , speed_limit , urban_or_rural_area , road_type

light_conditions , time_caregory , _months_

Three Predictive Models – Neural Network, Logistic Regression, Decision Tree

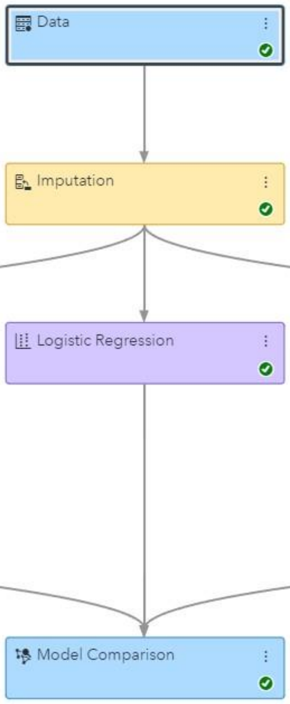
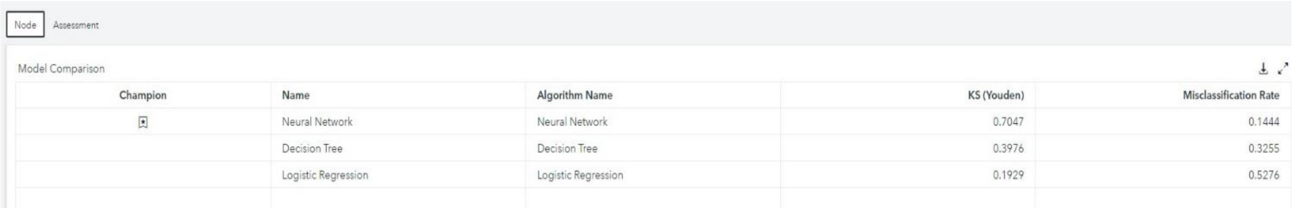


Figure 10

Champion Model



Model Comparison

Champion	Name	Algorithm Name	KS (Youden)	Misclassification Rate
★	Neural Network	Neural Network	0.7047	0.1444
	Decision Tree	Decision Tree	0.3976	0.3255
	Logistic Regression	Logistic Regression	0.1929	0.5276

Neural Networks:

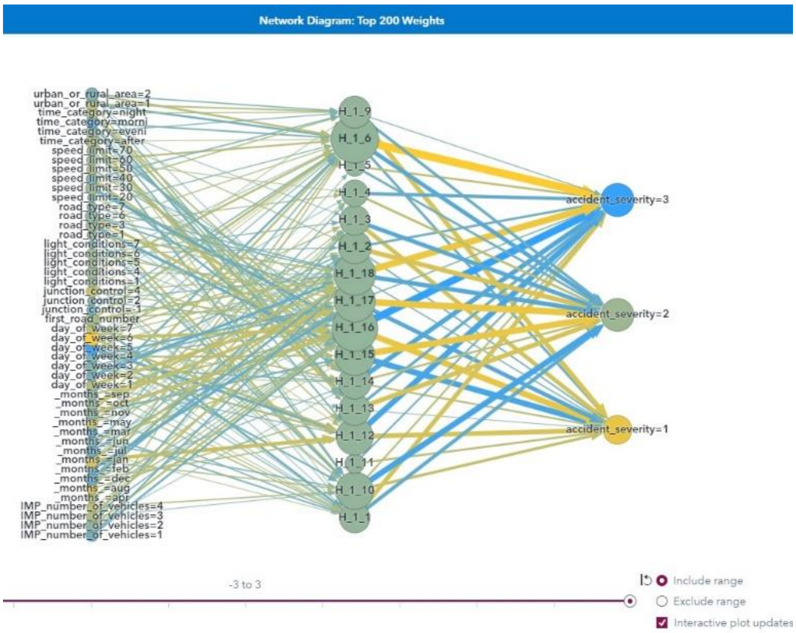


Figure 11

Neural networks are leading the way in predictive modelling, particularly when dealing with complex data interactions. These systems function by imitating the cognitive processes of the human brain, where information is processed by interconnected nodes, sometimes known as "neurons". When it comes to road accidents, a Neural Network can identify subtle patterns among extensive and diverse data sets, encompassing factors such as weather conditions and traffic patterns. The network diagram highlights the influential predictors with significant node connections, including light_conditions, road_type, and time_category. These indicators clearly have a significant impact on predicting the levels of accident severity.

Decision Trees:

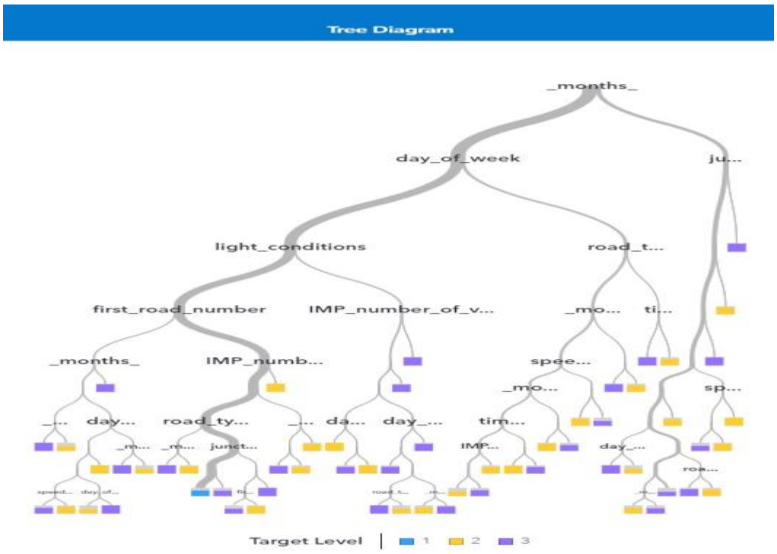
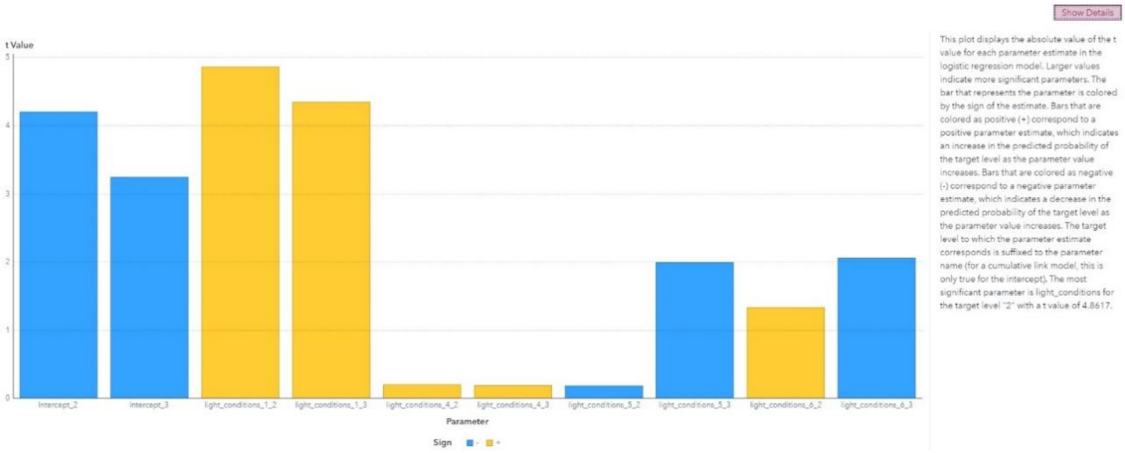


Figure 12

Decision Trees offer a distinct and contrasting methodology. The dataset has been separated into branches, resulting in unique and logical findings regarding the probable outcomes. Each node inside the tree indicates a point where a decision is made, whereas the branches indicate what comes out of that decision. The tree ends with a "leaf" node that predicts the conclusion. They identify the important factors that contribute to the risk of road accidents and allow for the implementation of targeted preventative measures with high efficiency.

Logistic Regression:



This plot displays the absolute value of the t value for each parameter estimate in the logistic regression model. Larger values indicate more significant parameters. The bar that represents the parameter is colored by the sign of the estimate. Bars that are colored as positive (+) correspond to a positive parameter estimate, which indicates an increase in the predicted probability of the target level as the parameter value increases. Bars that are colored as negative (-) correspond to a negative parameter estimate, which indicates a decrease in the predicted probability of the target level as the parameter value increases. The target level to which the parameter estimate corresponds is suffixed to the parameter name (for a cumulative link model, this is only true for the intercept). The most significant parameter is light_conditions for the target level "2" with a t value of 4.8617.

Figure 13

Logistic Regression is the most reliable and durable model for binary classification. When used to road safety, it can, for example, estimate the probability that specific conditions will end in serious accidents. Logistic Regression provides odds ratios, that offer valuable insights into the relative impact of many variables on the outcome. This is an essential tool for assessing risks and developing methods to mitigate them in urban planning and public safety projects**.**

Model Comparisons

Typically, these models are not employed separately. An effective analytical strategy utilises a comparative approach, taking advantage of the unique capabilities of each method. Neural Networks can detect intricate and non-linear connections that Logistic Regression cannot. On the other hand, Decision Trees offer clear and direct insights that can help improve and fine-tune the characteristics utilised in Neural Networks. By simultaneously executing different models and contrasting their forecasts, we may converge on the most precise results, leveraging the combined predictive capability of all three.

Model Effectiveness

The Neural Network, with its outstanding Kolmogorov-Smirnov statistic, has a consistent capability to differentiate across severity levels, suggesting a highly effective model for this prediction job. However, the misclassification rate of this organisation, although it is the lowest compared to the other two, needs careful consideration for possible improvement. The primary advantage of the Decision Tree is that it's able to be easily understand, which is particularly beneficial that want to comprehend the decision-making procedure. Nevertheless, its somewhat elevated misclassification rate implies that although it is valuable for acquiring insights, it may not be the most precise for prediction. Logistic Regression, while statistically informative, seems to be less effective in this scenario due to its higher misclassification rate in comparison to the Neural Network.

Conclusions and Recommendations

The Neural Network model stands out as the most effective in predicting accident severity due to its lower misclassification rate. However, its complexity requires careful consideration in execution. The Decision Tree offers valuable insights into the decision-making process, and its interpretability makes it a vital tool for communicating findings.

For enhancing road safety, the following recommendations are made:

- Improved lighting: Improving Street lighting and vehicle light visibility could minimise accident severity, given the importance of light conditions.

- Time and Categorical Emphasis: Given the importance of time_category and road_type, it may be beneficial to implement focused safety campaigns during high-traffic periods and on specific kinds of roads.

- Policy Implementation: Utilise the Decision Tree's explicit decision pathways to guide policy modifications, such as adjusting traffic lights and setting speed limit restrictions, based on the day of the week and speed limit characteristics.

Task 3 – Text Analysis of Tweets

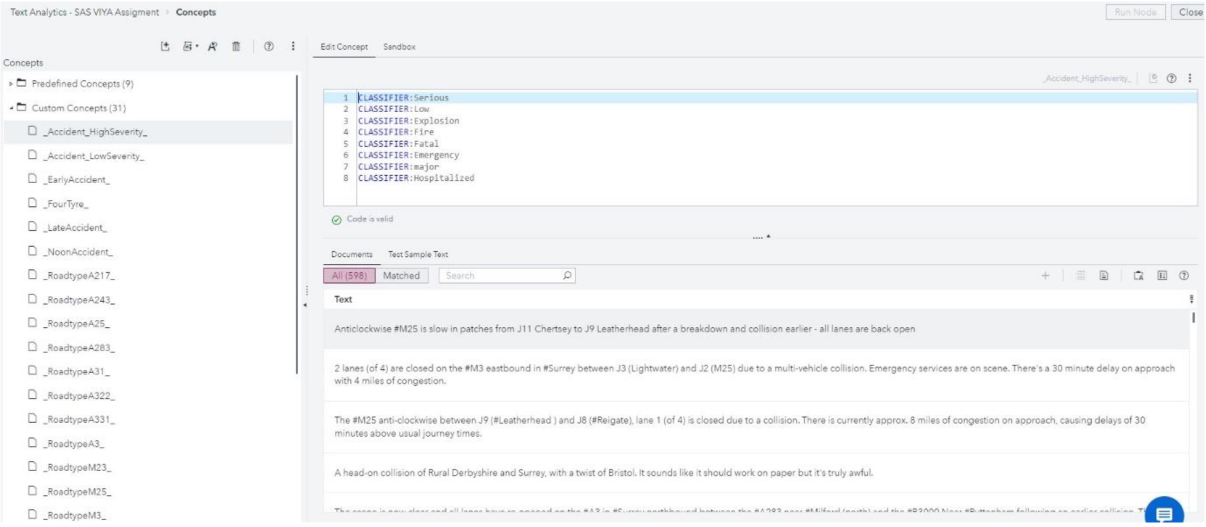
Exploring the Tweet Dataset

The 'Date' field uses a 'DD-MMM-YY' format that is clear and standard.

The 'Source' column highlights 'BBC Radio London Travel' and 'National Highways: South-East', indicating that the dataset combines information from various reliable traffic reporting agencies.

The 'Text' column contains precise updates related to traffic, including the names of roads like 'M25', descriptions of traffic conditions such as 'slow in patches' or 'lanes closed', and indications of locations using junction numbers like 'J11'.

CONCEPTS



```
CLASSIFIER:Serious
CLASSIFIER:Low
CLASSIFIER:Explosion
CLASSIFIER:Fire
CLASSIFIER:Fatal
CLASSIFIER:Emergency
CLASSIFIER:major
CLASSIFIER:Hospitalized
```

Anticlockwise #M25 is slow in patches from J11 Chertsey to J9 Leatherhead after a breakdown and collision earlier - all lanes are back open

2 lanes (of 4) are closed on the #M3 eastbound in #Surrey between J3 (Lightwater) and J2 (M25) due to a multi-vehicle collision. Emergency services are on scene. There's a 30 minute delay on approach with 4 miles of congestion.

The #M25 anti-clockwise between J9 (#Leatherhead) and J8 (#Reigate), lane 1 (of 4) is closed due to a collision. There is currently approx. 8 miles of congestion on approach, causing delays of 30 minutes above usual journey times.

A head-on collision of Rural Derbyshire and Surrey, with a twist of Bristol. It sounds like it should work on paper but it's truly awful.

The scene is now clear and all lanes have re-opened on the #A3 in #Surrey northbound between the #A283 near #Milford (north) and the #B3000 Near #Puttenham following an earlier collision. T

Figure 14

Predefined concepts were used to analyse the data and provide a broad understanding of the number of significant incidents compared to minor incidents. These categories are crucial for conducting in-depth trend research and immediately providing emergency response resources.

Custom concepts were evaluated, providing a more detailed perspective. For example, the 'Accident_HighSeverity' category indicated the most severe accidents that required quick care. Meanwhile, the concepts of 'EarlyAccident' and 'NoonAccident' provide insight into the timing of events, indicating possible patterns associated with traffic density and road usage at different times of the day. A sequence of classifiers, varying from 'Serious' to 'Hospitalised', that are used to categorise the severity of the incident. Each classifier functions as a filter, enabling the classification of incidents into pre-established severity categories. This facilitates a prompt evaluation of the urgency and characteristics of each event, hence facilitating the implementation of appropriate actions.

Word Clouds - Frequency of Keywords and Concept Names

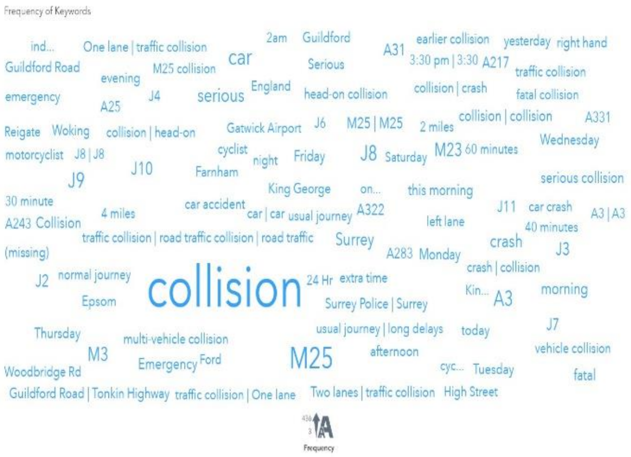


Figure 15

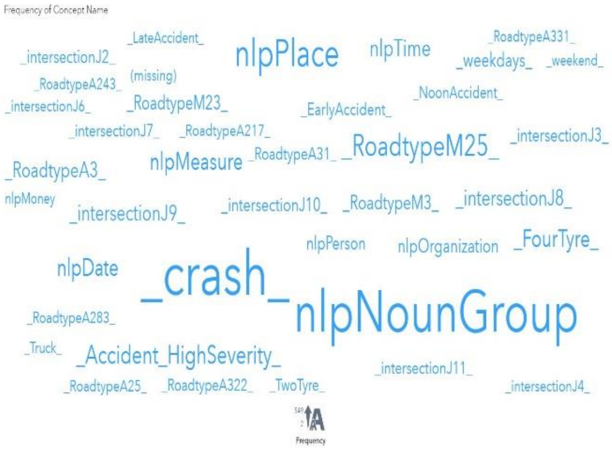


Figure 16

The above two-word cloud displays the frequency of keywords and concept names. Significant terms such as 'crash' and 'M25' showed up, indicating frequent sorts of incidents and their specific locations. Furthermore, the frequent reference to road designations and intersections such as 'A217' and 'J8' suggests the presence of areas that are prone to incidents.

Bar Chart - Concept Frequency

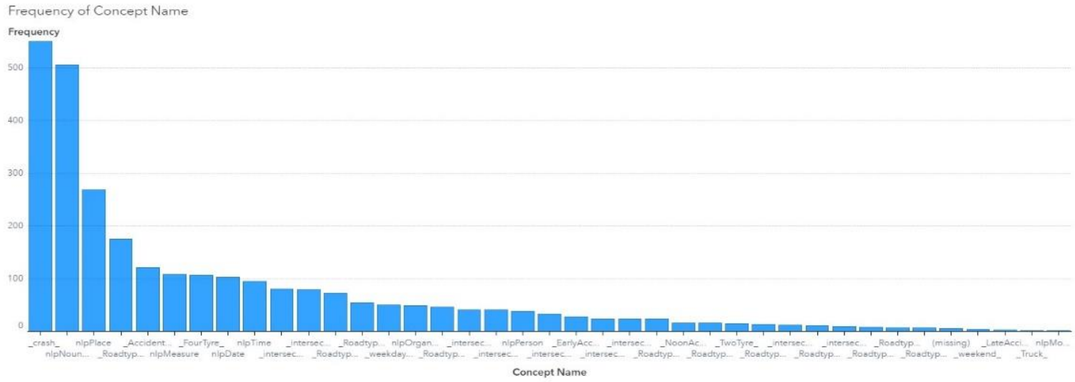
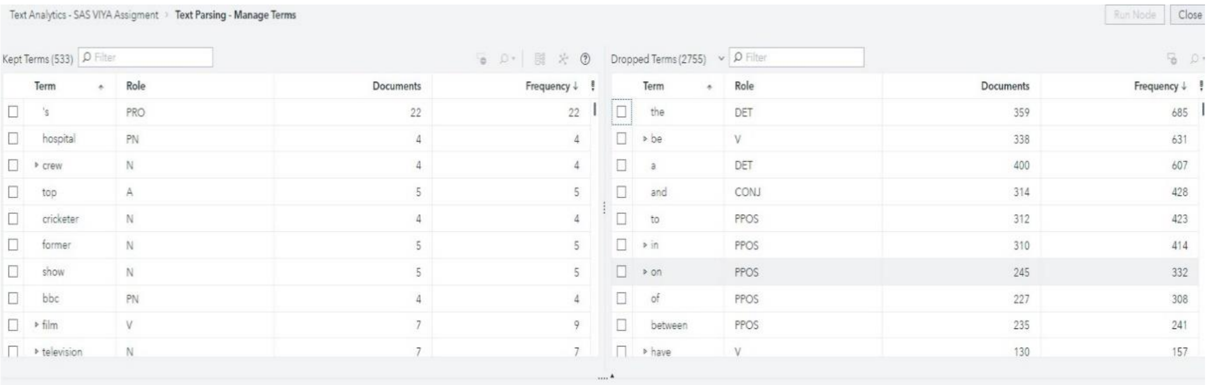


Figure 17

The bar chart depicts a trend study of concept frequencies, with 'crash' and 'nlpPlace' taking the lead. This data is crucial for identifying the main traffic concerns and their locations.

Text Preprocessing



Text Analytics - SAS VIYA Assigment › Text Parsing - Manage Terms
Run Node
Close

Kept Terms (533)

	Term	Role	Documents	Frequency ↓
□	's	PRO	22	22
□	hospital	PN	4	4
□	▸ crew	N	4	4
□	top	A	5	5
□	cricketer	N	4	4
□	former	N	5	5
□	show	N	5	5
□	bbc	PN	4	4
□	▸ film	V	7	9
□	▸ television	N	7	7

Dropped Terms (2755)

	Term	Role	Documents	Frequency ↓
□	the	DET	359	685
□	▸ be	V	338	631
□	a	DET	400	607
□	and	CONJ	314	428
□	to	PPOS	312	423
□	▸ in	PPOS	310	414
□	▸ on	PPOS	245	332
□	of	PPOS	227	308
□	between	PPOS	235	241
□	▸ have	V	130	157

Figure 18

Firstly, the text is often cleansed from special characters and punctuation. These elements can contribute noise into the data as they are typically irrelevant to the interpretation of the meaning of words and sentences. For instance, punctuation markings such as a dot at the end of a sentence or exclamation marks can be removed since they usually do not add to the understanding of the text's meaning.

Managing the inclusion and exclusion of initial and final stop words is equally important. Stop words are words that are excluded from text processing, either before or after. Typically, they are the most used words in a language, such as "the", "is", and "in", which contribute little to the overall meaning of a document. By eliminating these words, we may concentrate on the terms that are more likely to express the unique and fundamental nature of the text.

The interface categorises terms into 'Kept' and 'Dropped' groups, enabling the improvement of the dataset for more accurate analysis. 'Kept Terms' are recognised as potential conveyors of important meaning, while 'Dropped Terms' mostly include of stop words that are normally excluded prior to analysis due to their lack of informative value.

Bar Chart - Role by Frequency

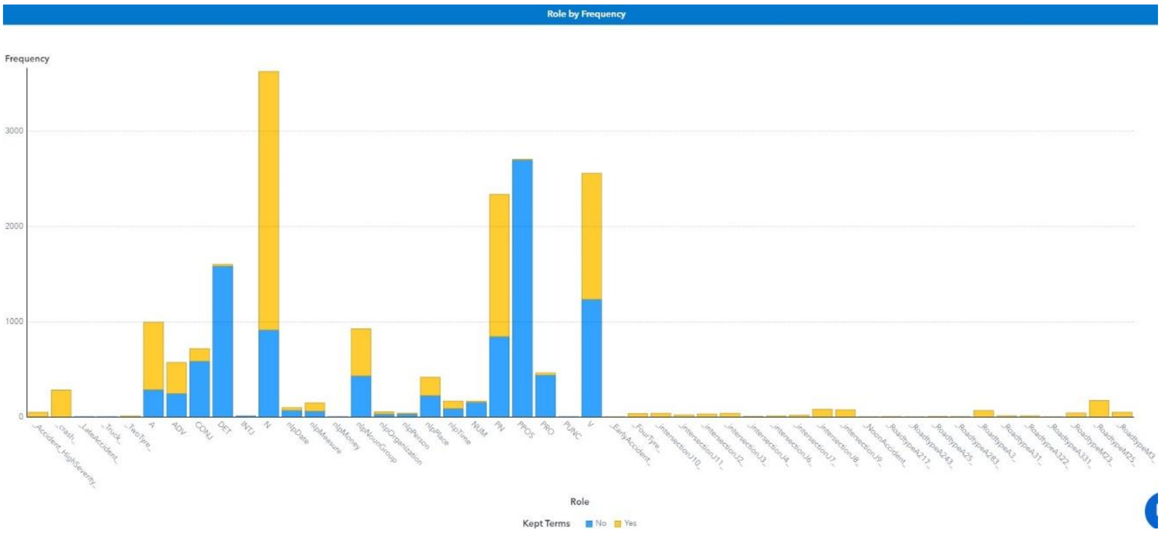
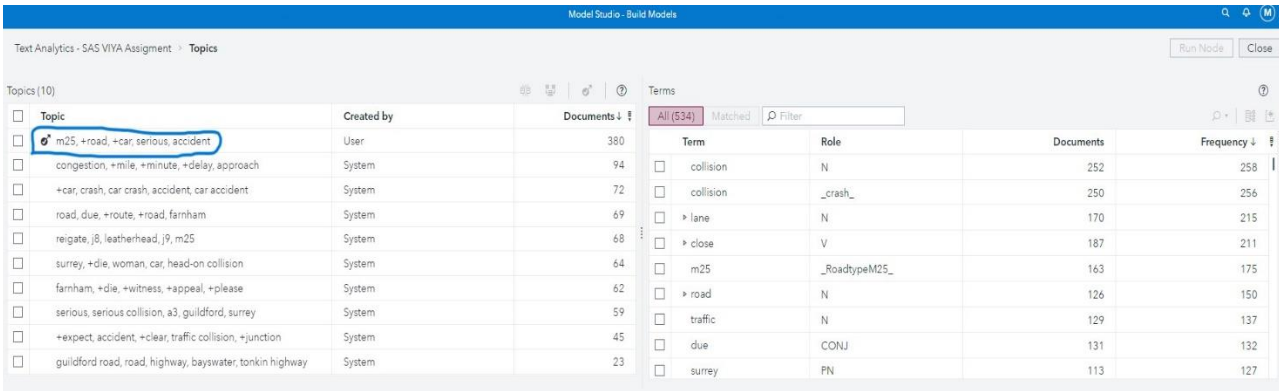


Figure 19

The frequency analysis graphically represents the presence of terms throughout the text. It emphasises the most used terms and their specific functions (such as pronouns, nouns, and verbs), enabling a quantitative evaluation of the data. A term's high frequency suggests its importance and relevance to the topic matter.

Sentiment Analysis & Topics



Text Analytics - SAS VIYA Assigment › Topics

Topics (10)

□	Topic	Created by	Documents ↓
□	m25, +road, +car, serious, accident	User	380
□	congestion, +mile, +minute, +delay, approach	System	94
□	+car, crash, car crash, accident, car accident	System	72
□	road, due, +route, +road, farnham	System	69
□	reigate, j8, leatherhead, j9, m25	System	68
□	surrey, +die, woman, car, head-on collision	System	64
□	farnham, +die, +witness, +appeal, +please	System	62
□	serious, serious collision, a3, guildford, surrey	System	59
□	+expect, accident, +clear, traffic collision, +junction	System	45
□	guildford road, road, highway, bayswater, tonkin highway	System	23

Terms

All (534) Matched Filter

	Term	Role	Documents	Frequency ↓
□	collision	N	252	258
□	collision	_crash_	250	256
□	▸ lane	N	170	215
□	▸ close	V	187	211
□	m25	_RoadtypeM25_	163	175
□	▸ road	N	126	150
□	traffic	N	129	137
□	due	CONJ	131	132
□	surrey	PN	113	127

Figure 20

My primary topic focuses on severe car accidents occurring on the M25. This issue is correlated with 380 documents, suggesting that much of the discussion concentrates on this topic. The model has generated nine additional subjects, each with a different number of documents associated with it. These include a variety of issues, including overall traffic congestion as well as individual occurrences like head-on crashes and road closures.

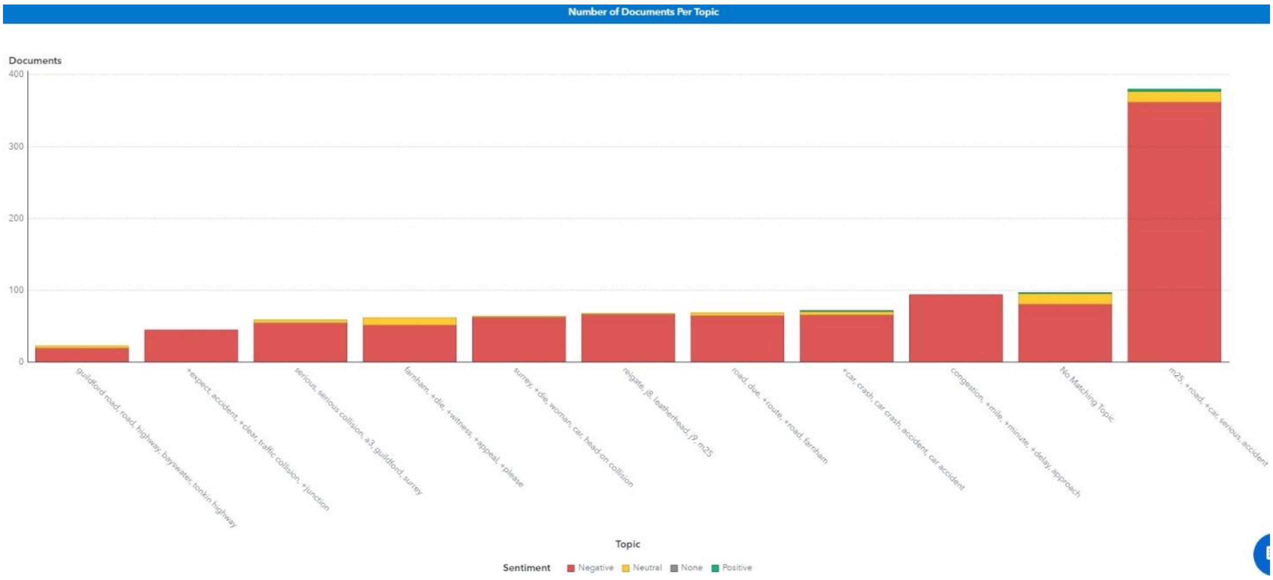


Figure 21

- Negative Sentiment: The majority of sentiments expressed in the documents were negative, which is consistent with the theme of traffic-related events.
- Neutral Sentiment: The majority of topics contain only a small fraction of texts with a neutral sentiment. This could denote objective reports or facts presented without any kind of emotional bias.
- Positive Sentiment: Only a handful of documents displays positive sentiment.

Figure 20 displays the outcome of sentiment analysis. The significant dominance of 'bad' sentiment, as shown by its disproportionately long bar, corresponds to the characteristics of the dataset centred around traffic incidents - a field where bad occurrences are frequently documented and deliberated over. The findings of 'Positive' and 'Neutral' attitudes are notably less frequent.

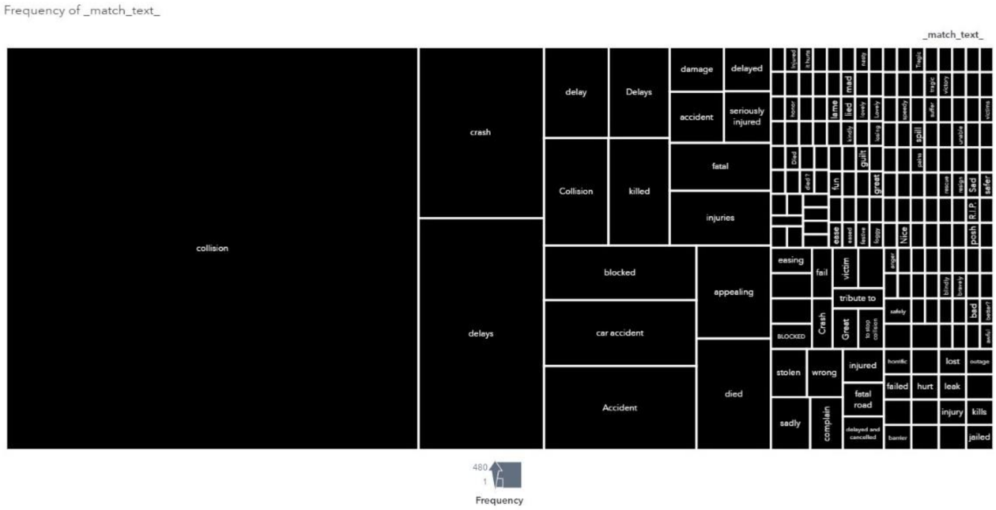
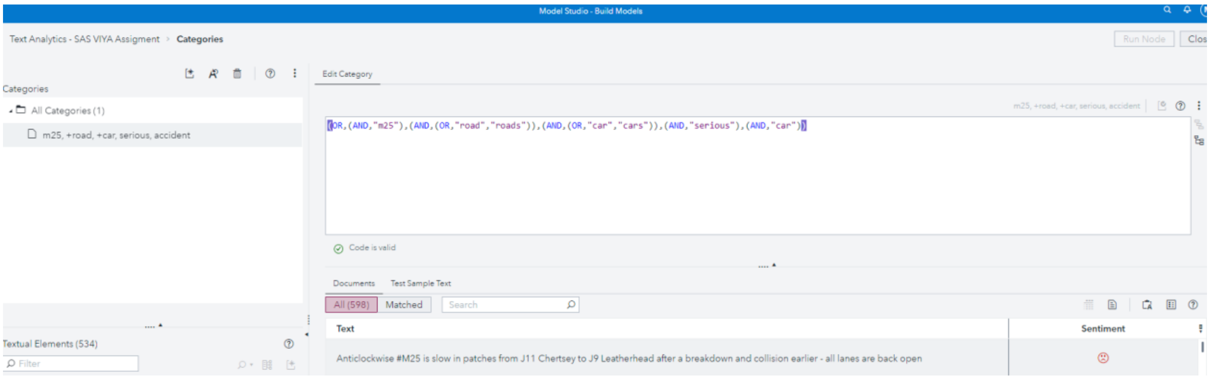


Figure 22

Categories



```
(OR,(AND,"m25"),(AND,(OR,"road","roads")),(AND,(OR,"car","cars")),(AND,"serious"),(AND,"car"))
```

Text	Sentiment
Anticlockwise #M25 is slow in patches from J11 Chertsey to J9 Leatherhead after a breakdown and collision earlier - all lanes are back open	☹

Figure 23

An evaluation was conducted on a category characterised by the terms 'M25,' 'road,' 'car,' 'severe,' and 'accident'. The findings indicate a significant degree of clarity, with an important number of correctly identified cases and a small fraction of incorrectly identified cases in both directions.

Bar Chart - Diagnostic Counts for Automatically Generated Categories

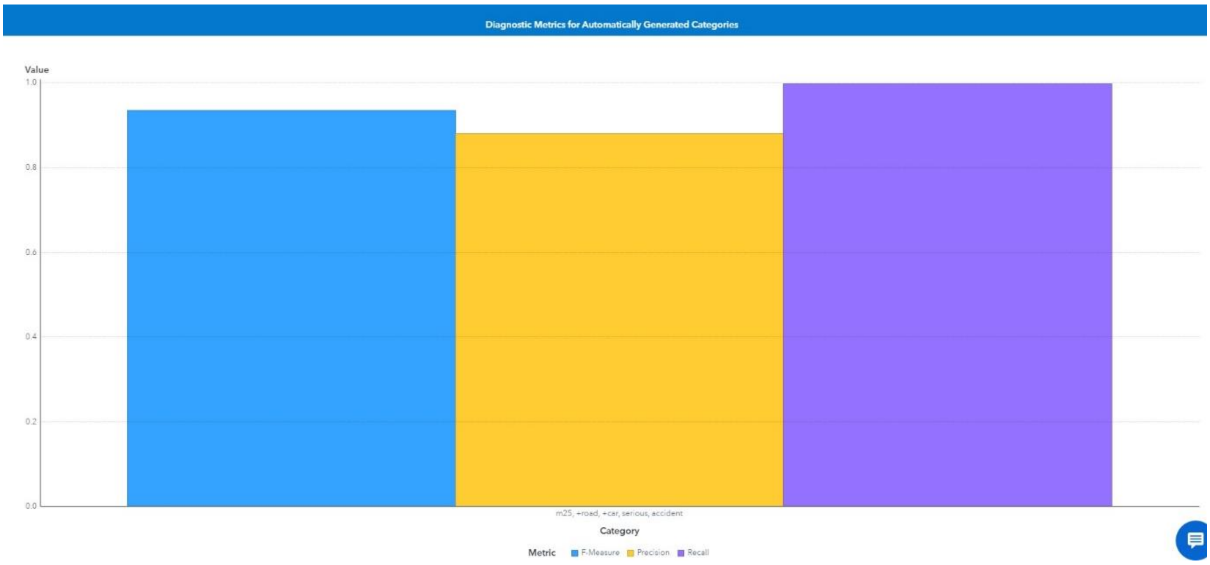


Figure 24

The above bar chart for Automatically Generated Categories" illustrates three metrics: F-Measure, Precision, and Recall. Each metric offers a unique lens into the performance:

F-Measure (Blue Bar): At near 1.0, this suggests a balanced harmonic mean between Precision and Recall, indicating a robust algorithm performance.

Precision (Yellow Bar): Lower than the F-Measure, Precision measures the accuracy of the positive predictions. The value suggests room for improvement in the specificity of the algorithm.

Recall (Purple Bar): This metric, also high, evaluates the algorithm's ability to find all relevant instances within a dataset. A high Recall indicates few missed relevant categories.

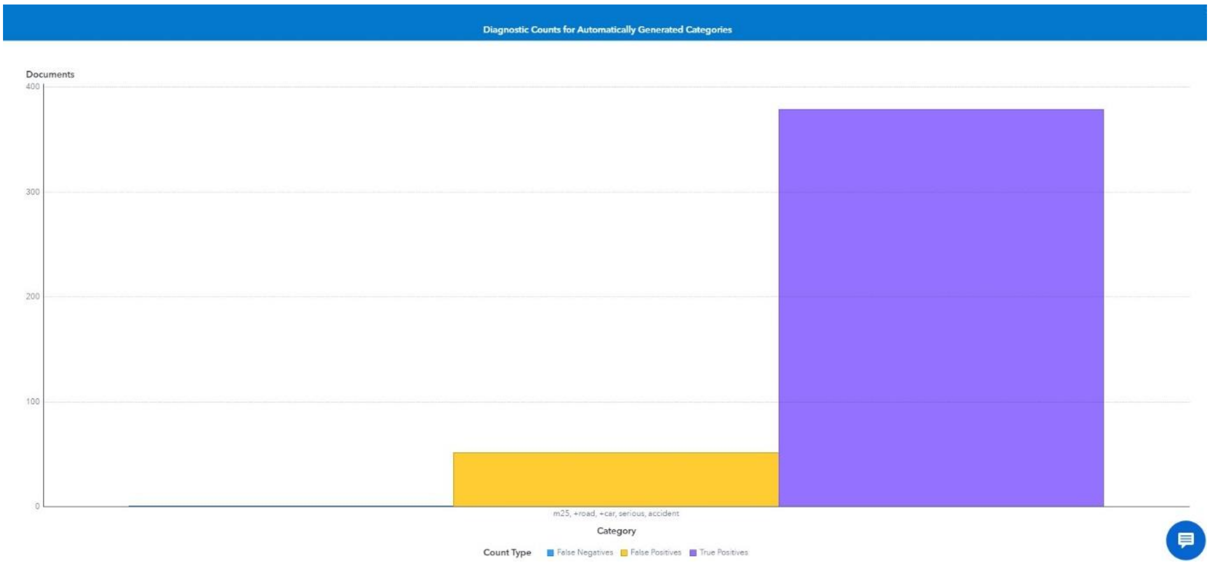


Figure 25

This bar chart quantifies the algorithm's output:

True Positives (Purple Bar): Significantly higher than the other counts, indicating most categories were correctly identified.

False Positives (Yellow Bar): Minimal, suggesting few irrelevant categories were incorrectly labelled as relevant.

False Negatives (Blue Bar): Non-existent in this dataset, indicating that the algorithm did not miss any relevant categories.

Key Insights

The analysis employs a range of classifiers, from 'Serious' to 'Hospitalised', serving as filters to systematically categorize incident severity.

Word clouds reveal the prominence of terms such as 'crash' and 'M25', highlighting recurrent incident types and pinpointing exact locations. The repeated mention of particular roads and junctions suggests certain areas are more susceptible to traffic incidents, necessitating targeted safety initiatives.

In the tokenization process, the text is stripped of noise-contributing elements like special characters and punctuation, allowing for a more focused analysis on meaningful terms. The exclusion of stop words further refines the dataset, directing attention to terms with significant semantic value.

The sentiment analysis unveiled a predominantly negative tone across the documents, aligning with the nature of traffic incidents.

While most topics had a minimal presence of neutral sentiment, indicative of objective reporting, positive sentiments were scarcely represented, reflecting the critical nature of the traffic situations.

Findings From Text Analysis

The findings from this comprehensive analysis underscore the criticality of high-severity accidents on the M25, a central theme across numerous documents.

Focused interventions should be directed at high-risk locations and times, utilizing the patterns observed in incident timing and severity (ESSWIDI, et al., 2023).

Continued refinement of data collection and analysis processes could significantly improve incident management, reduce traffic congestion, and elevate road safety measures (Choudhary, et al., 2023).

Task 4 – Decision-Maker's Summary and Recommendations

Overview of the Dataset

The dataset analysed provides comprehensive data on road accidents, including unique identifiers, location coordinates, police force codes, accident severity, vehicle and casualty numbers, date and time, and whether the accident occurred in an urban or rural area. It offers valuable insights into accident patterns and can be instrumental in developing targeted road safety strategies.

Key Insights and Analysis

Accident Distribution and Severity:

Accidents are evenly distributed between urban and rural areas.

Severity levels range from 1 (most severe) to 3 (least severe), with a notable proportion of level 2 severity accidents in urban areas.

An outlier in accident severity (level 36) suggests data quality issues.

Temporal and Spatial Patterns:

A slight increase in accidents towards the end of the week, peaking on Fridays.

Geographic distribution shows clusters of high-severity accidents, indicating potential high-risk zones.

Influencing Factors:

Road type and light conditions significantly impact accident severity.

Urban areas have higher accident rates, likely due to traffic congestion and complex road networks.

A correlation exists between the severity of accidents and the number of casualties.

Predictive Modelling:

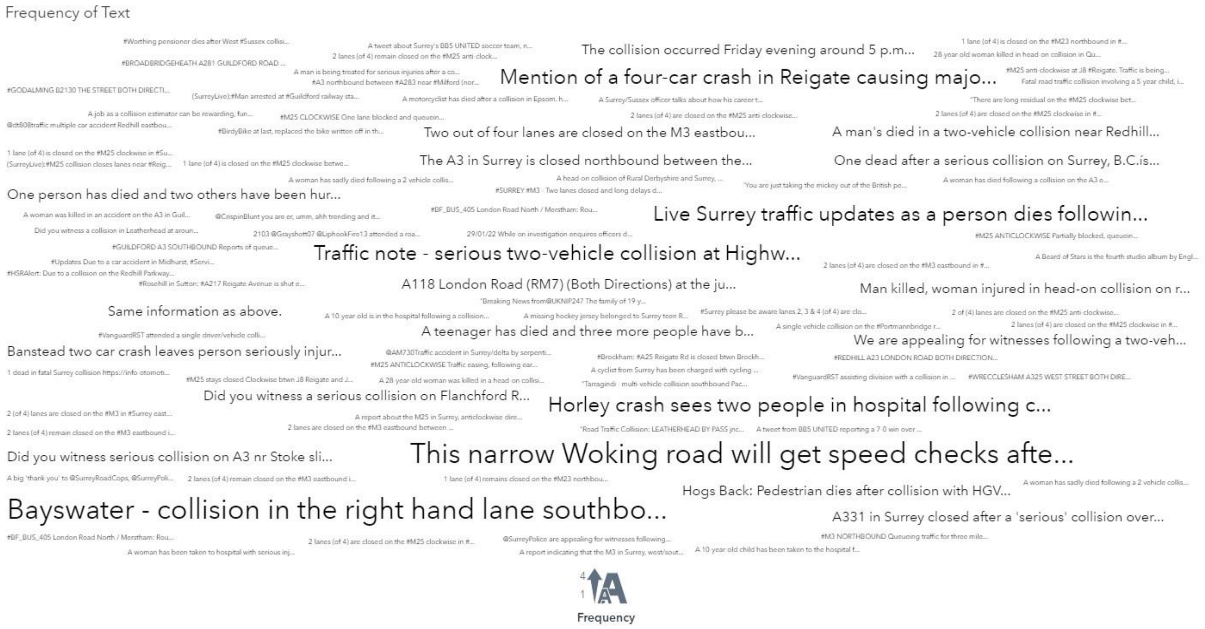
Neural Network, Logistic Regression, and Decision Tree models offer varied insights.

The Neural Network model is highly effective in predicting accident severity but complex in nature.

Decision Trees provide clear insights and are useful for policymaking.

Text Analysis of Tweets:

Focus on severe accidents on the M25, with frequent mentions of specific roads and junctions prone to incidents. Predominantly negative sentiment in traffic-related reports.



Recommendations

- ✓ Enhance street lighting and improve vehicle light visibility, particularly in recognised high-risk zones.
- ✓ Targeted Safety initiatives: Direct efforts towards high-traffic periods and certain road types for safety initiatives.
- ✓ Policy Implementation: Apply knowledge derived from Decision Trees to adapt traffic signal timing and establish speed restrictions according to the specific day of the week and type of road.
- ✓ Infrastructure Enhancements: Resolve areas of elevated risk determined through geographical analysis.
- ✓ Enhance public awareness and education by promoting knowledge on high-risk locations for accidents and promoting safe driving behaviours, particularly during periods of heavy traffic congestion.
- ✓ Continual Data Analysis: Consistently update and improve the process of collecting and analysing data to effectively manage incidents and promote road safety.

Conclusion

This investigation offers useful insights into the dynamics of road traffic accidents and emphasises the significance of data-driven decision-making in improving road safety. The recommendations, supported by the insights and predictive modelling of the dataset, seek to provide guidance for specific actions and policy modifications, thereby making an important contribution to the reduction of traffic accidents and betterment of public safety.

References

Abedi, M. a. S. E., 2024.. *A machine learning tool for collecting and analyzing subjective road safety data from Twitter..* [Online]
Available at: https://www.sciencedirect.com/science/article/pii/S0957417423030841

Agar, I. a. B. et al., 2023. The Essex risk-based policing initiative: evidence-based practices in problem analysis and crime prevention in the United Kingdom. *Justice Quarterly,* pp. 1--21.

Choudhary, J. K., Rayala, N., Kiasari, A. E. & Jafari, F., 2023. Road Safety in Great Britain: An Exploratory Data. *World Academy of Science, Engineering and Technology,* 17(7).

ESSWIDI, A., ARDCHIR, S. & A. D., 2023. SEVERITY PREDICTION FOR TRAFFIC ROAD ACCIDENTS. *Journal of Theoretical and Applied Information Technology ,* 101(8).