**Module Title : Data mining and text analytics With application in SAS**

**Student Number (ID): 6807428**

**Assessment Title : Individual Assignment - Exploring Road Traffic Accident Data and Text Analytics Insights**

**Table of Contents: Page No.**

1. Task 1
   1. Introduction 3
   2. Summary Statistics 3
   3. Box plot of Accident Severity 4
   4. Data cleaning 4
   5. Correlation Matrix 4
   6. Bar Graphs 5
   7. Line Graph 6
   8. List 6
   9. Heat Map 7

1.10Data Balancing 7

2. Task 2

2.1 Neural Network 8

2.2 Decision Tree 11

2.3 Conclusion 13

3. Task 3

3.1 Concepts Node 15

3.2 Text parsing Node 16

3.3 Sentiment Node 19

3.4 Topics Node 20

3.5 Categories Node 21

3.6 Conclusion 22

4. Task 4 23

5.References 25

**Task 1: Data Exploration and Cleaning**

*Introduction:*

This dataset contains most of the essential fields needed to analyse traffic crashes resulting in injury to people in the Surrey region in 2022. The data encompasses details about the specific locations of the occurrences, the total number of vehicles involved in the crashes, the type of roads involved, the lighting circumstances during the occurrences, date and time of accident occurrence, junction details, number of casualties, speed limit and the presence of road crossing amenities.

*Summary Statistics:*

A summary statistics (Table 1) of some variables has been generated to understand the central tendencies and dispersion of the data.

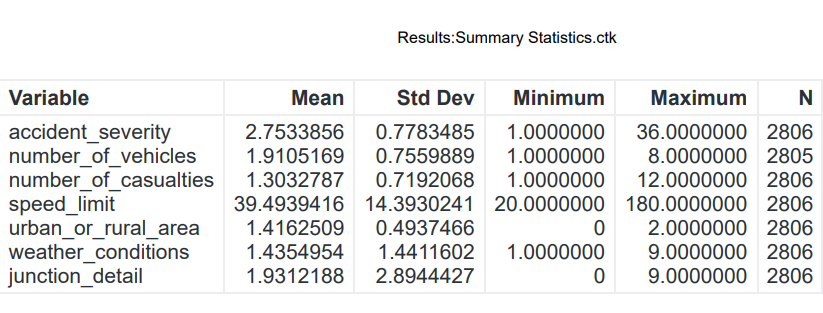


Table 1 : Summary Statistics

It can be inferred from the above table that most of the accidents which occurred are slight (mean accident severity ≈ 3). Similarly, average number of vehicles in the collision is 2 and number of causalities is approximately 1. Speed limit has a wide range, from 20 to 180, with an average speed limit of 40.

*Boxplot of Accident Severity:*

From the below Boxplot (Fig 1), it can be concluded that accident severity=36 is an outlier/undesired value which need to be eliminated. Also, day of the week has values 8 and 9, which also need to be filtered out.

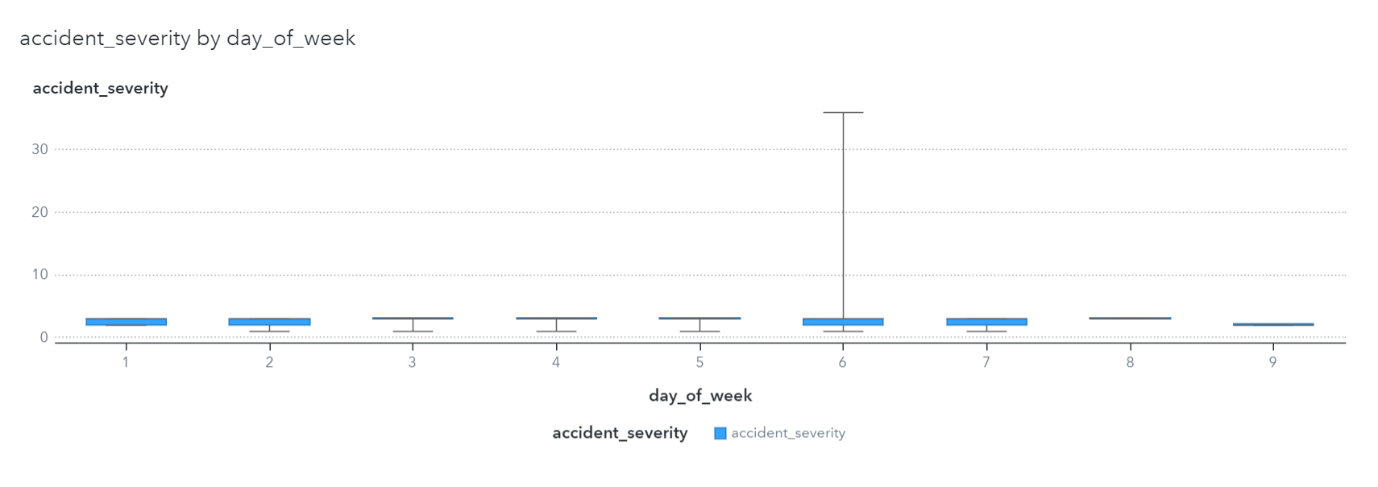


Fig 1 : Accident severity

*Data Cleaning:*

As part of data cleaning, some of the variable names have been changed through Prepare Data->Column Transforms->Manage Columns->Rename Columns. Also, the outliers and other undesired values has also been filtered out.

*Correlation Matrix*:

A correlation matrix (Fig 2) has been created to understand the relation between different variables and accident severity. There is a moderate negative correlation (-0.0371) between carriageway hazards and accident severity, suggesting its importance as a major factor. Similarly, trunk road flag, special conditions at sight, road surface condition, first road class, pedestrian crossing physical facilities, road type and light conditions also exhibit correlations with accident severity, ranging from -0.0282 to -0.0150. Since there are no missing values in the above chosen variables, no imputation needs to be done as part of data cleansing.

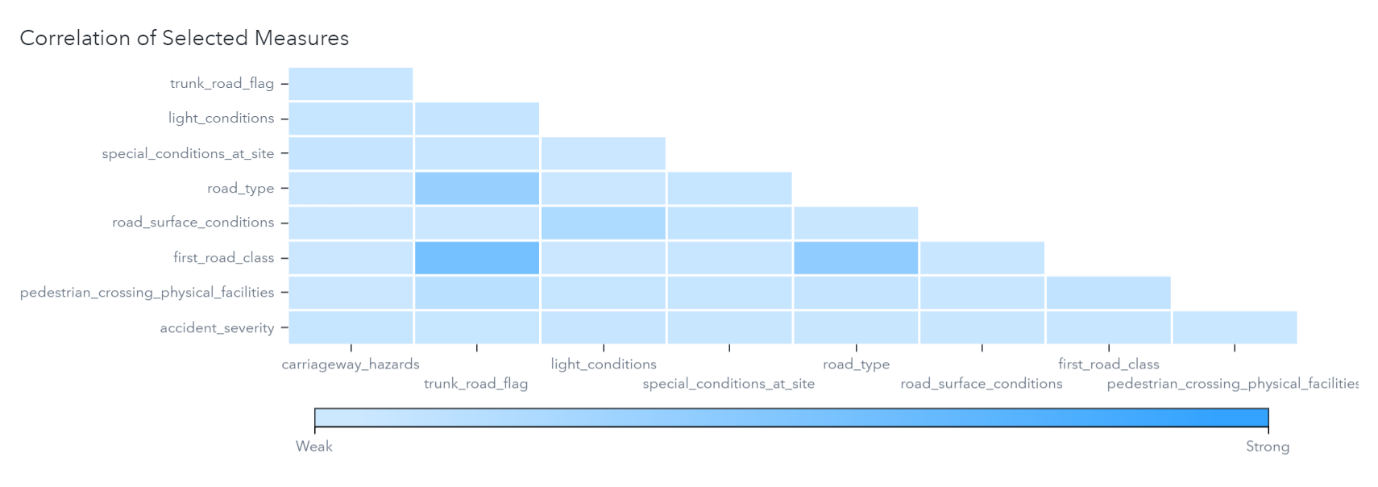


Fig 2 : Correlation Matrix

*Bar Graph:*

From the bar graph of accident frequency of each month (Fig 3), there is a notable difference in accident numbers between March, with 61% (1739) incidents, and December, with just 2.5% (72), highlights a significant disparity. Factors such as decreased traffic during holidays, heightened awareness, and extreme weather conditions in December might have contributed to this significant reduction in accidents compared to the busier and potentially more hazardous month of March.Top of Form

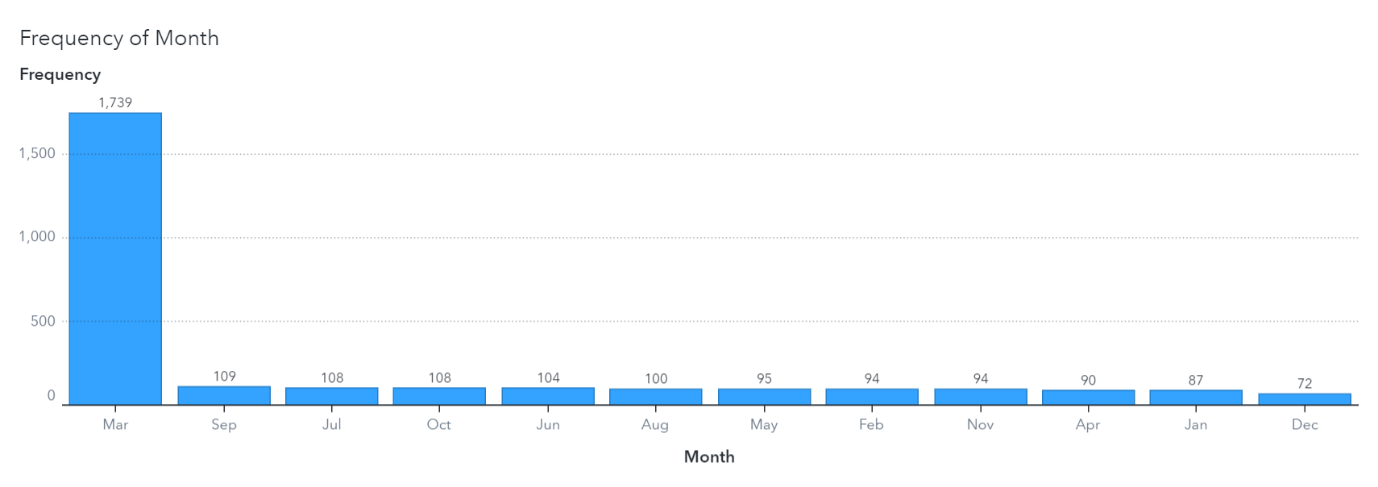


Fig 3 : Accident severity in each month

Considering the road surface conditions in Fig 4, dry road surface give way to the greatest number of accidents, with a distribution of 1537 (54%) slight, 492 (17%) serious, and 18 (0.6%) fatal incidents. Wet or damp road surfaces comprises of 522 slight (which is approximately 1/3rd of the former), 160 serious, and 7 fatal accidents. There are 39 slight, 16 serious, and 3 fatal accidents in frost or ice-covered roads, while deep flooded road surfaces being the roads with least accidents of only 6 slight accidents.

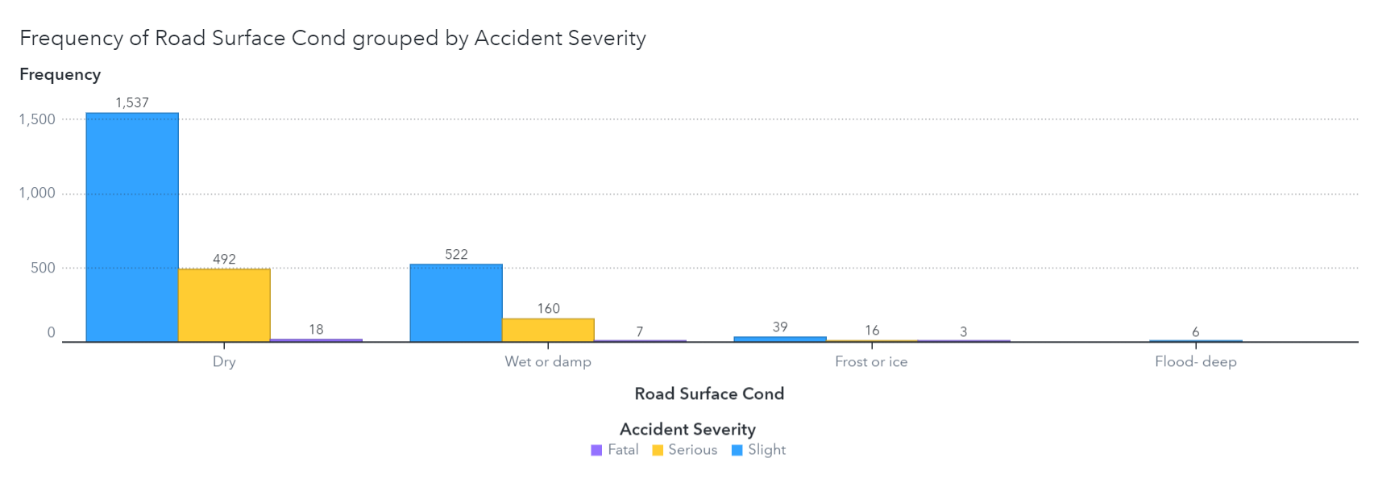


Fig 4 : Road surface conditions

*Line Graph:*

From the below line graph (Fig 5), it is understood that most accidents occurred on Thursdays, with 15% (446), followed by Friday with 433 while weekends had the least, accounting to around 10% (300) of accidents. This might be due to the difference in the mental stress of the drivers on end-of-week and weekends.

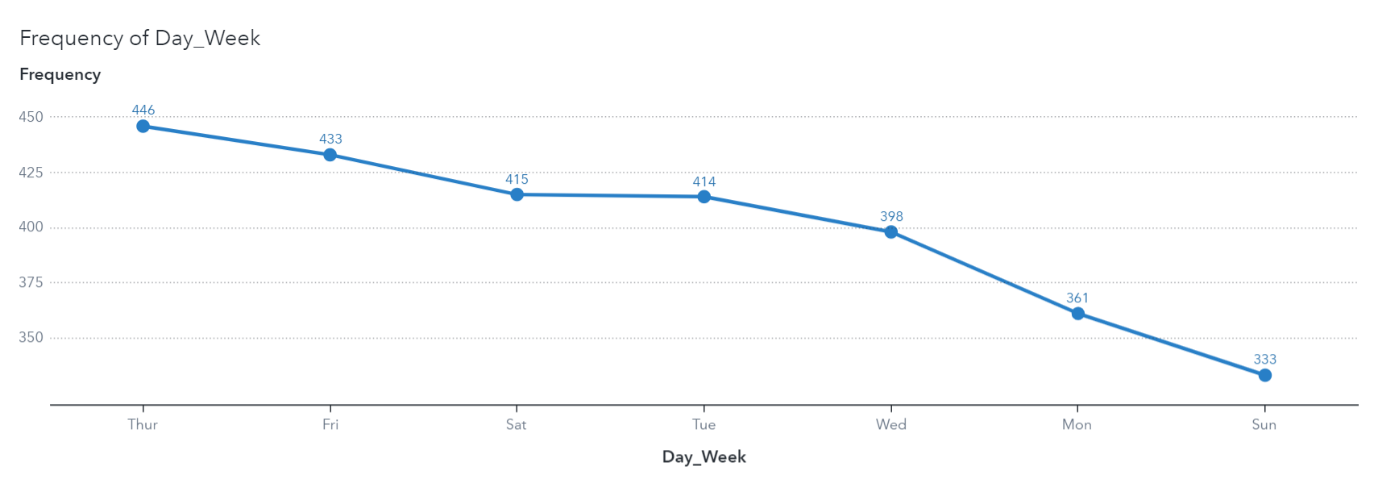


Fig 5 : Day of the week

*List:*

From the list of carriageway hazards (Table 2), it can be understood that around 70% (2000) of accidents had occurred on roads with no hazards, even though accident severity is very low. Drivers are more careful when there are vehicle loads on roads which had resulted in only 1 serious accident in the whole year. In general, accidents mostly occurred when there are no hazards on the carriageways. People tend to be more alert and careful when there is any other object, any previous accidents or animals on the roads.



Table 2 : Carriageway Hazards

*Heat Map:*

From the heat map of light conditions vs accident severity (Fig 6), it can be derived that around 50% (1497) accidents occurred in the presence of daylight (light condition=1) and all the accidents are fatal. Similarly, 16% of serious accidents also occurred during day time. Only 7% accidents happened in complete absence of light. When lights were lit in the dark, 15% fatal accidents took place.

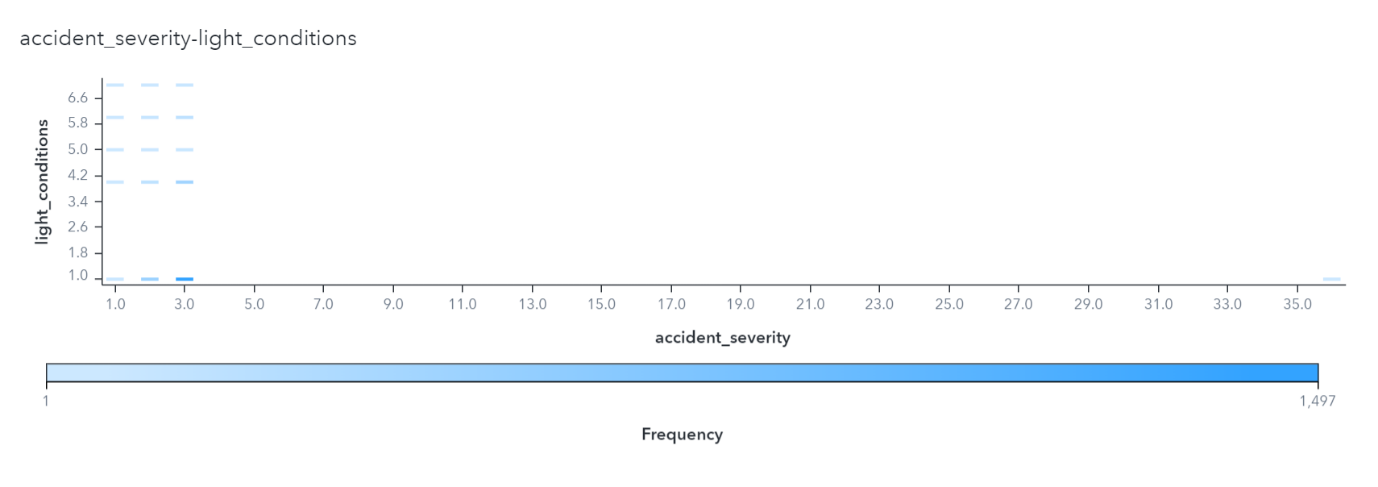


Fig 6 : Light Conditions

*Dataset Balancing:*

The dataset was quite imbalanced. So, a proper data balancing has been done before doing the prediction models. The data has been balanced to 33.32% for accident severity 1 and 2 and 33.35% for accident severity 3.

**Task 2: Predicting Accident Severity**

**2.1 Introduction:**

In today’s world, accident prediction world be immensely beneficial. Accidents can occur at any place, any time. But prediction of accident severity can be achieved to an extend by analysing different causes like junction details, light conditions of the place, weather conditions, etc. In this coursework, accident severity will be predicted by using the variables weather conditions, pedestrian crossing human control, pedestrian crossing physical facilities, junction control, carriageway hazards, time category, road surface conditions, first road class, second road class, special conditions at site, speed limit, trunk road flag, urban or rural area, road type, junction details, day of the week and light conditions. Prediction will be based on how the change in each of these variables affect the severity of the accident and suggesting measures to prevent accidents in future. The two models used in this coursework for accident severity prediction are Neural Networks and Decision Tree models.

**2.2 .Neural Network**

Neural network is a machine learning technique inspired by the human brain. It has a collection of interconnected nodes called neurons organized into layers. There is input layer which receives data, hidden layers which process it, and output layer which produces predictions. Weights (L1, L2) and biases are adjusted during training to optimize predictions.

In the neural network model used here, there are 34 neurons (which is double the number of input variables used) arranged in a single hidden layer. L1 weight of 0 and L2 weight of 0.1 is used, which helps in controlling model complexity. The optimization technique employed is LBFGS (Limited Memory Broyden-Fletcher-Goldfarb-Shanno).

After 240 iterations, the neural network model achieved promising results. The validation error stood at 0.202, which is a fairly good value, while the fit error was 0.124, indicating effective model fitting. The KS values reflected good discrimination, particularly evident in test data (0.5238), train data (0.7263), and validation data (0.5860).

*ASE Value:*

The Average Squared Error (ASE) factor is a common regression loss function, which computes the average of squared differences between neural network predicted values and actual target values. A lower value for ASE depicts close match between predicted and actual values, indicating good model performance in fitting the data. ASE values in this model (fig 7.) are favourable with 0.0981 for test data, 0.0670 for training data, and 0.0921 for validation data.

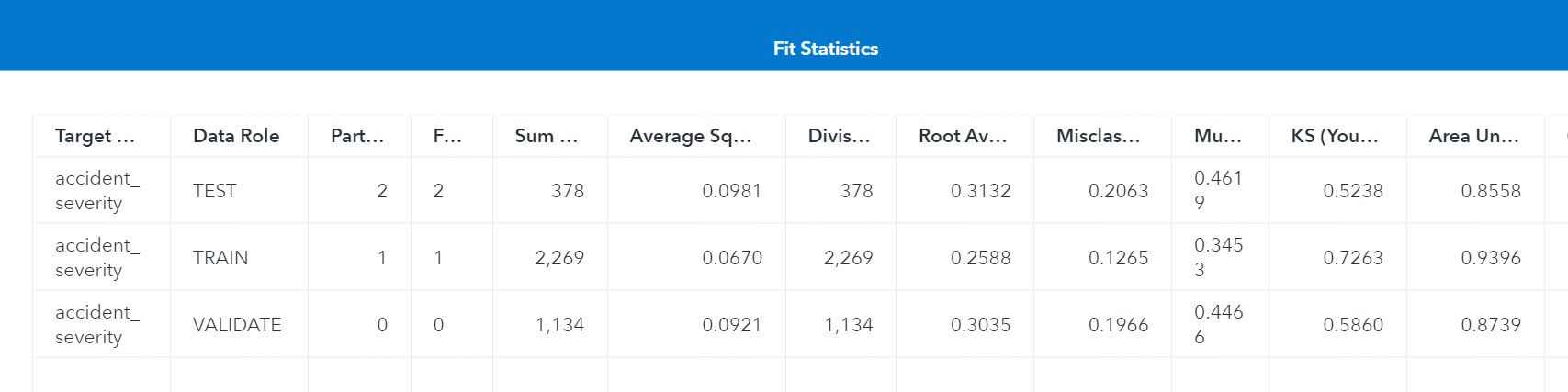


Fig 7. ASE and KS values for Neural Network model

*KS Value:*

The KS (Kolmogorov-Smirnov) value is a measure which is used to evaluate the discriminatory power of a classification model. In our model, the KS values are significant. In the train dataset, the KS value of 0.7263, indicates a considerable discriminatory ability. Similarly, the test dataset has a KS value of 0.5238, and the validation dataset with 0.5860. These values depicts that our model exhibits a strong discriminatory power across different datasets.

*Cumulative Lift:*

In the below graph (Fig 8), the validate, train, and test partitions has significant cumulative lift values of 2.65, 2.88, and 2.62, respectively, at the 10% quantile (depth of 10). These values suggests that events within the first two quantiles are approximately twice as frequent as expected by a random chance, accounting for 10% of the total number of events.

A graph of different colored lines

Description automatically generated

Fig 8 : Cumulative lift – Neural Network Results

*Receiver Operating Characteristic (ROC) Curve:*

The higher (closer to 1) the area under ROC curve, the more better the model is. In our model, ROC curve is depicted by fig 9. Area under ROC curve for validate data is 0.8739, train data is 0.9396 and test with 0.8558 , all are closer to 1 and thus the performance of our model is fair.

A graph of a graph

Description automatically generated with medium confidence

Fig 9 : ROC Curve – Neural Network Results

**2.3 .Decision Tree**

A decision tree is a machine learning algorithm for both classification and regression tasks. The tree structure includes nodes, where each internal node symbolizes a decision based on a feature, and each leaf node represents the predicted outcome.

The input values under Splitting options are as below:

* Maximum depth – 16
* Minimum leaf size – 15
* Number of interval bins – 100

All other values are default values. The above fields have been modified to improve the efficiency of the model.

*ASE Value:*

The lowest error for average squared error is for the training set (0.0990), indicating efficient model fitting to the training data. The validation data (fig 10) set also demonstrates a relatively low error (0.1201), depicting the model's ability to generalize well to a new dataset and the test dataset shows a slightly higher error of (0. 1261).

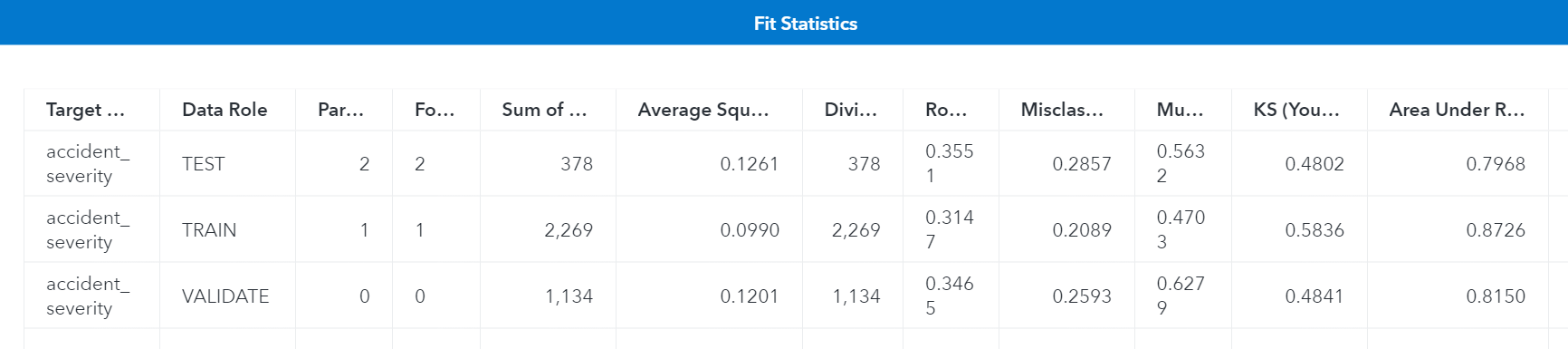
**

Fig 10. ASE and KS values for Decision Tree model

*KS Value:*

There is a significant improvement in the KS value in decision tree model compared to neural network model. The test data has a KS value of 0.4802, training data with 0.5836 and validate data with 0.4841 .

*Cumulative Lift:*

The Cumulative Lift (Fig 11) values for the validate data, train data and test data are 2.1, 2.52, and 1.72 respectively, indicating that the performance of model is efficient and better than random chance, particularly in the top 10% of predicted events. This means there are approximately two times more events happening in the predicted segments than expected randomly. A higher lift means the model is more efficient and accurate, especially in targeted actions.

A graph of different colored lines

Description automatically generated

Fig 11 : Cumulative Lift – Decision Tree Results

*Receiver Operating Characteristic (ROC) Curve:*

In the ROC curve (Fig 12) for decision tree analysis, there is a decrease in the area under the ROC curve, with 0.8726 for training data, 0.7968 for test data and 0.8150 for validate data, compared to neural network model.

A graph of a graph

Description automatically generated with medium confidence

Fig 12 : ROC Curve – Decision Tree Results

The figure below (Fig 13) demonstrates the importance of each input variables used in the analysis. Day of the weak has most priority while junction control has the least importance in predicting accident severity, suggesting that more care need to be taken on Thursdays and Fridays, when the most number of accidents occur. Some mitigations that can be taken are giving high priority to traffic control, improving pedestrian walking/crossing facilities, control on speed limit, usage of public transport, more ambulance services ready etc.

A screenshot of a computer

Description automatically generated

Fig 13 : Importance of Input Variables

**2.4 Conclusion:**

The neural network model beats decision tree model in several aspects (fig 14.). Firstly, the neural network provides lower ASE values for all datasets, suggesting a high level of accuracy in predicting target values compared to the decision tree. Secondly, the neural network exhibits higher KS values (0.5238) for all datasets, displaying a better discriminatory power of predictive model. Cumulative Lift which reflects model's prediction accuracy is greater in neural network model. Moreover, the neural network maintains a higher Area under ROC curve compared to decision tree, showing its ability to differentiate between classes and make appropriate predictions. Overall, the neural network signifies robust predictive power and effectiveness, making it a more advisable model over the decision tree based on lower ASE, higher KS values, superior Cumulative Lift, and consistently higher ROC AUC.

More police patrols should be there during high-risk hours and in areas prone to accidents. Improvements in non-trunk road infrastructure is also required.

A screenshot of a computer

Description automatically generated

Fig 14. Champion Model-Neural Networks

**Task 3: Text Analysis of Tweets**

The dataset contains textual information of 598 tweets discussing road accidents in the Surrey region. Data has been uploaded and the pipeline is created as below (fig 15.).

A screenshot of a computer screen

Description automatically generated

Fig 15. Pipeline of text analytics

*3.1 Concepts Node*

The concept node has some inbuild concepts made using advanced natural language processing techniques. Four new custom concepts-\_ARoads\_, \_Junction\_, \_Motorways\_, \_Places\_ have been created to better analyse the data. For -\_ARoads\_, there are 134 matches found. 175 matches for \_Junction\_, 195 matches for \_Motorways\_ and 299 for \_Places\_. These four concepts have been chosen as they are the most significant concepts from the tweets where information can be gained. Fig 16. represents the frequency of different concepts in documents.

A screenshot of a graph

Description automatically generated

Fig 16. Results from Concept node

A word cloud (fig 17.) is a visual representation where the frequency of the word is understood from the size of the font. The larger the font is, the more the occurrence of the word in the dataset.

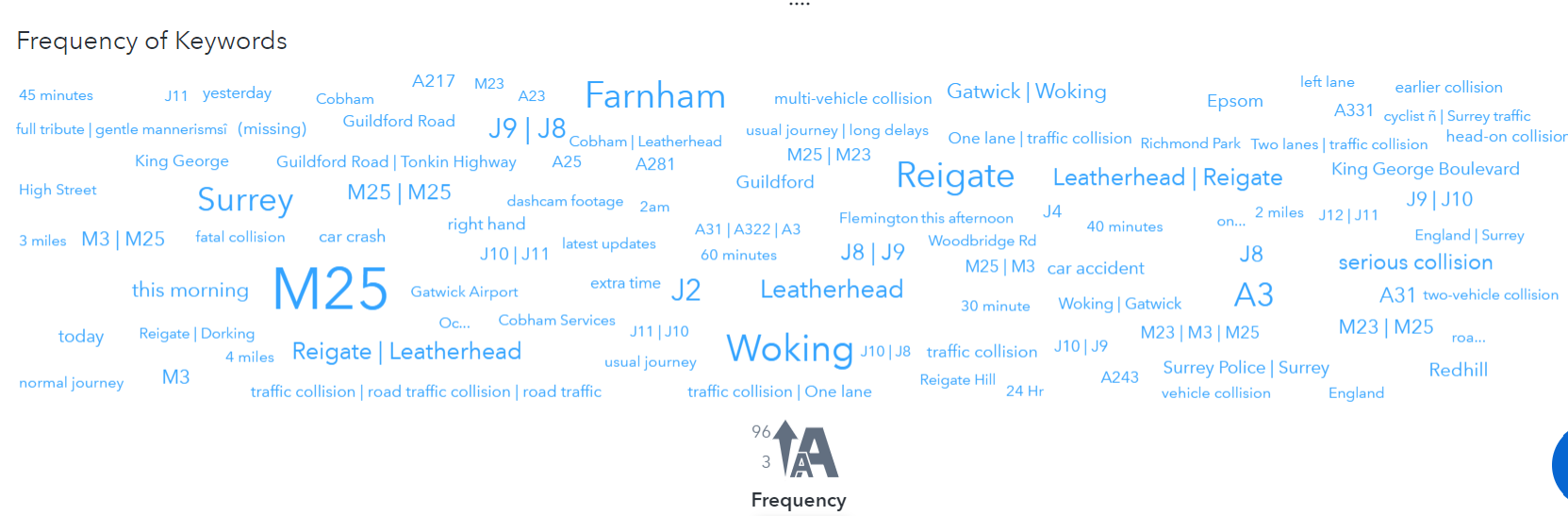


Fig 17. Word cloud of tweets

From the word cloud, M25 is the most occurring term in the dataset, with a frequency of 96. So, most of the tweets have the occurrence of the term M25 in it. Similarly, Woking, Farnham, Surrey and Reigate are some of the most commonly occurring places in the tweets. A23, M23, Surrey traffic etc are the rare words with a frequency of 3.

The Noun group concept has the highest frequency with 505 terms, followed by Places concepts with 299 (fig 18.) .

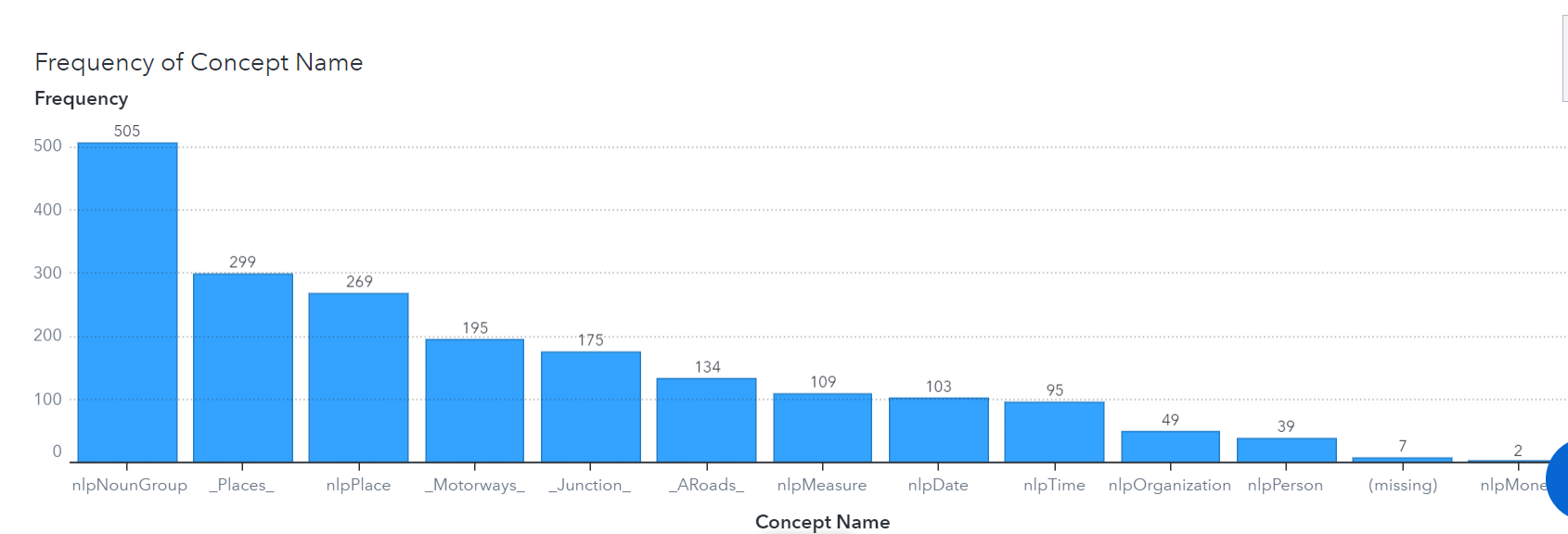


Fig 18. Frequency of Concepts

*3.2 Text Parsing Node*

The Text parsing node does the removal of special characters and punctuations, tokenization, stemming and part-of-speech tagging. The most frequent word in the tweets provided is ‘collision’, which occurs in 483 out of 598 tweets, with a frequency of 503. This is followed by ‘lane’ (215 times in 170 documents) and ‘close’ (211 times in 187 documents). Fig 19. depicts the ‘kept terms’ and ‘dropped terms’ in the decreasing order of their frequency.

A screenshot of a computer

Description automatically generated

Fig 19. Text Parsing Node

A term map (fig 20.) for the term ‘m25’ has been created inorder to identify and categorize the association within different terms. This term is strongly associated with the term ‘lane’, with an information gain of 203.45 . This is followed by junction term ‘j8’ with a gain of 186.11 .

A network diagram of blue dots and lines

Description automatically generated

Fig 20. Term map of ‘m25’

Using Similarity score (fig 21.), the terms close to the term ‘j8’ are found. ‘Reigate’ is closely related to j8, with a score of 0.951 . This is obvious as j8 is a junction at Reigate, Redhill. There are 88 matching documents for the same. Next highest similar term is j9 with a score of 0.869 with 78 matching documents. J9 is the junction Leatherhead, Epsom , which is close to j8 junction. Accidents occurring in j8 may cause traffic block in j9 or vice-versa.

A screenshot of a computer

Description automatically generated

Fig 21. Similarity scores with j8 term

*3.3 Sentiment Node*

Sentiment analysis is run to understand the sentiment expressed in the tweets-positive, neutral or negative. In our model, the Sentiment node is run and a graphical representation is present from the Explore and Visualise tab (fig 22.) .

A blue rectangles with black and white lines

Description automatically generated

Fig 22. Sentiment analysis of the tweets

The above bar graph (fig 22.) represents the frequency percent of different sentiments of the tweets- negative, positive and neutral. 92% of the tweets are composed of negative sentiments. The reason for this can be that the tweets are related to accidents, collision, block etc. 7% tweets have neutral sentiment while only 0.67% has positive sentiments.

*3.4 Topics Node*

In the Topics node, Maximum number of topics have been changed to 20. After analysing the term map for M25, a few new user-defined topics (fig 23.) have been created in order to increase precision in uncovering patterns in the dataset. The first topic, centered around keywords like "lane," "M25," "Reigate," "J8," and "J9," is associated with 225 matched documents. The second topic, characterized by terms like "M25," "J9," "Reigate," "Leatherhead," and "J8," is linked to 169 matched documents.

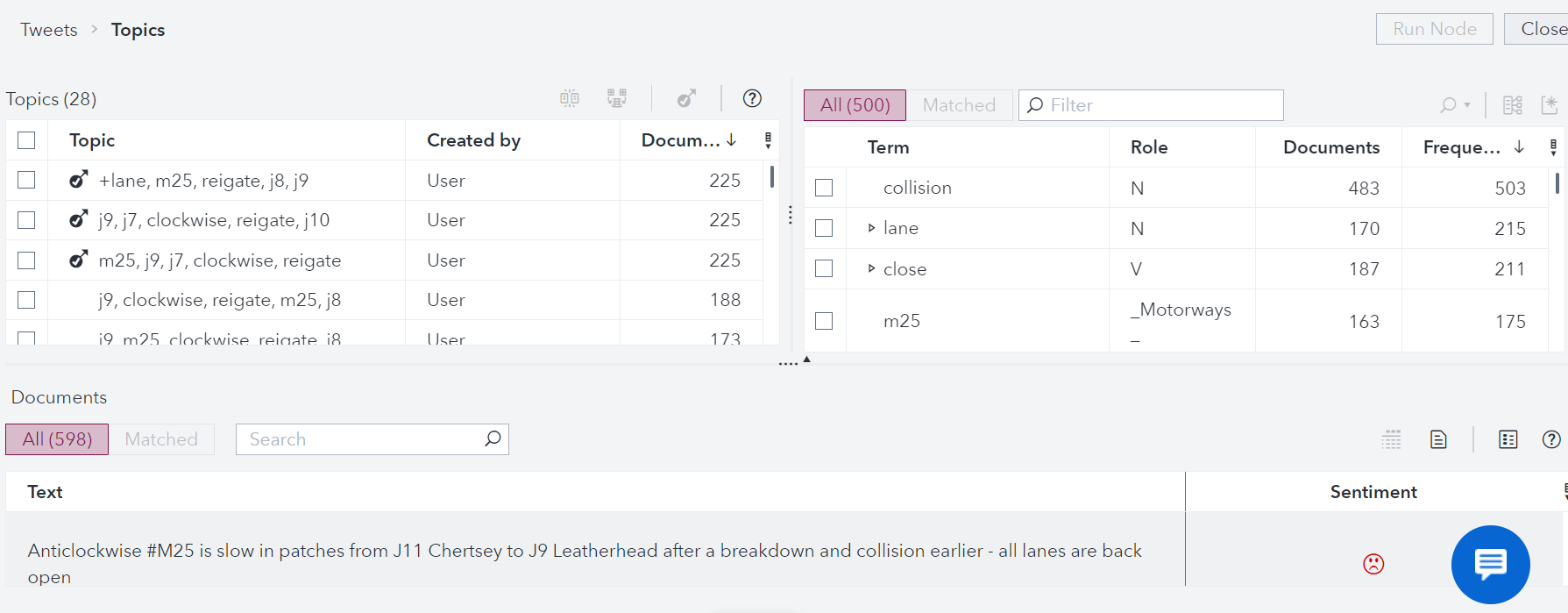


Fig 23. Topics node with new user-defined topics

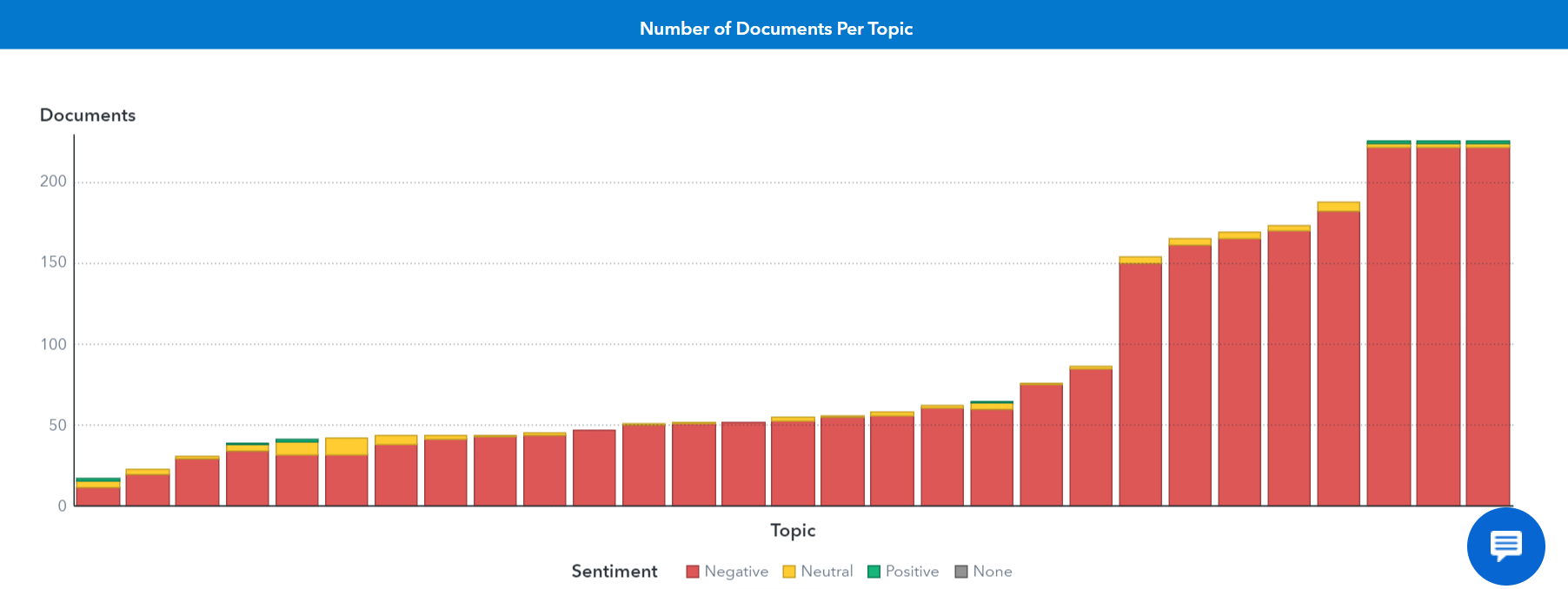


Fig 24. Sentiments for different topics

Fig 24. illustrates the sentiments on different topics-both user defined and system generated topics. From the above graph, it is obvious that majority of the topics has a negative sentiment.

*3.5 Categories Node*

Categories node helps to organize and categorize textual data. For the categories in our model, it is found from fig 25. that there is a large proportion of true positives compared to false positives and false negatives. 2223 documents match each of the 3 categories of true positives and only 32 and 2 documents match false positives and false negatives respectively.

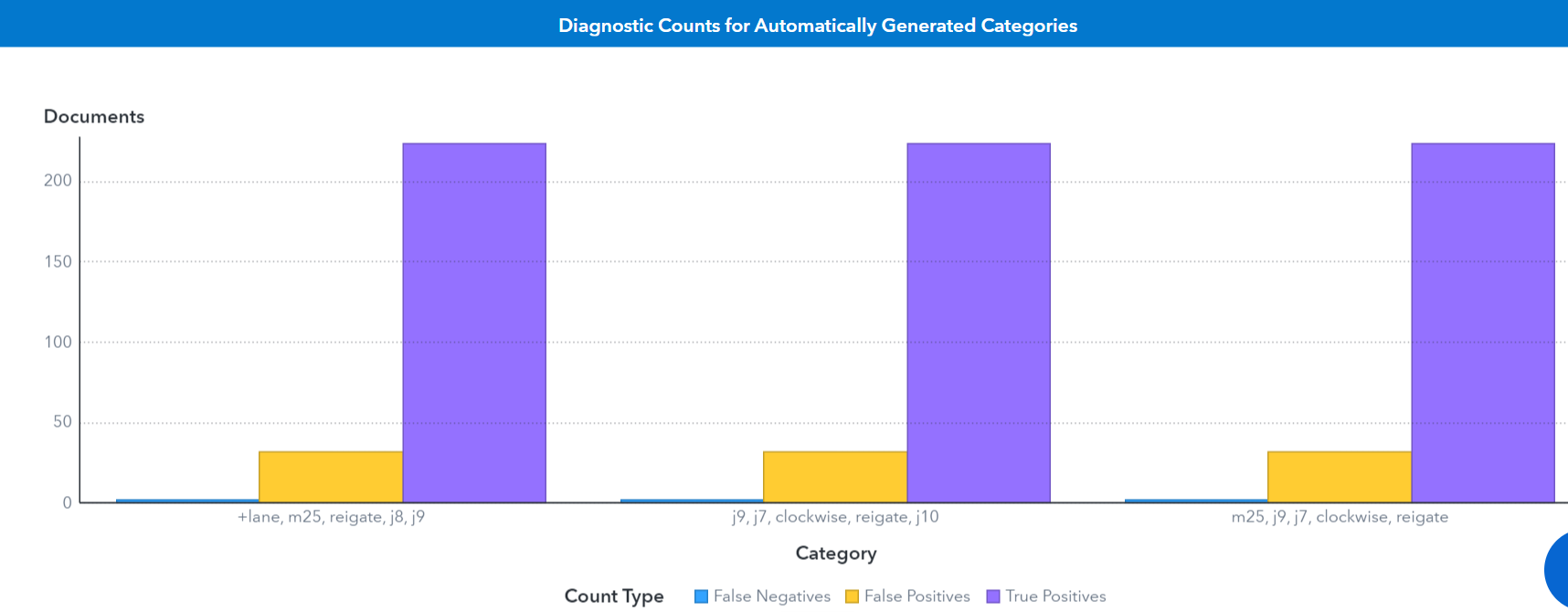


Fig 25. Diagnostic counts for categories

The recall, precision and F-measure for all the 3 categories are close to 1 (fig 26.), 0.991 for recall, 0.8745 for precision and 0.929 for F-measure, indicating the effectiveness of our model.

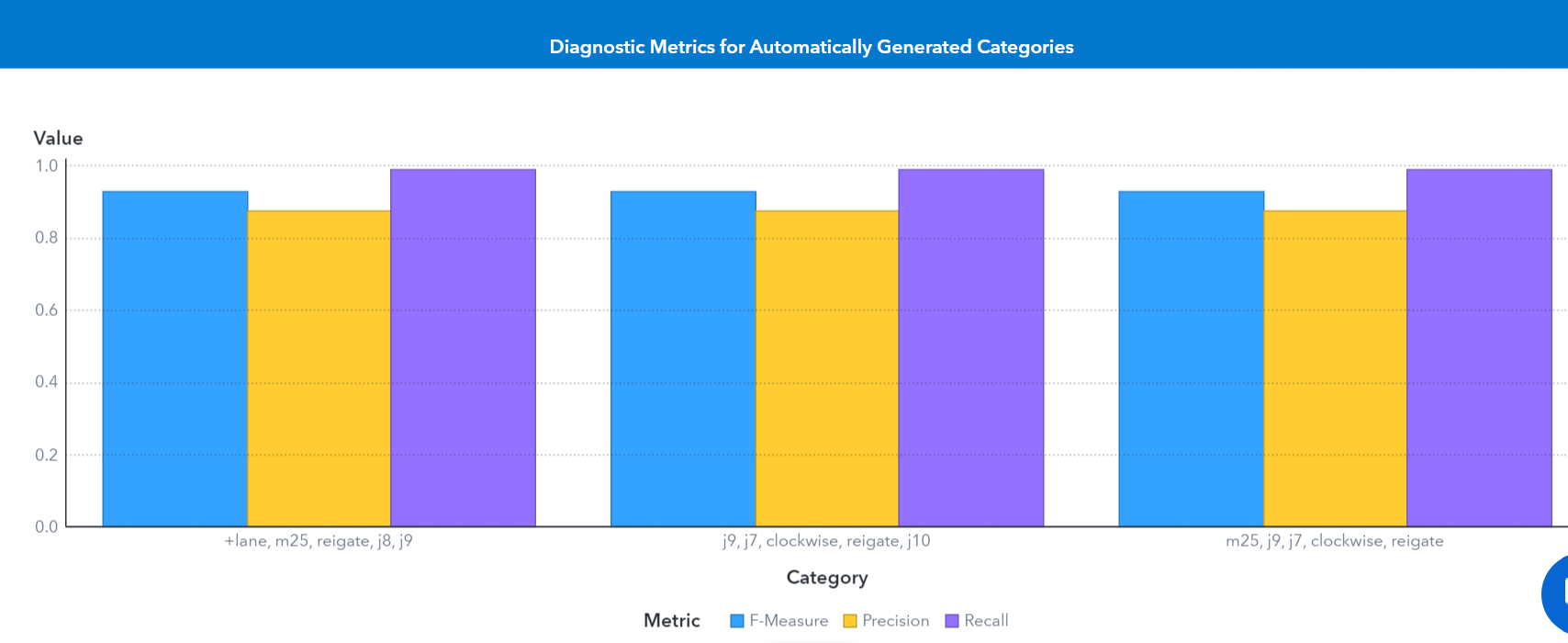


Fig 26. Diagnostic metrics for categories

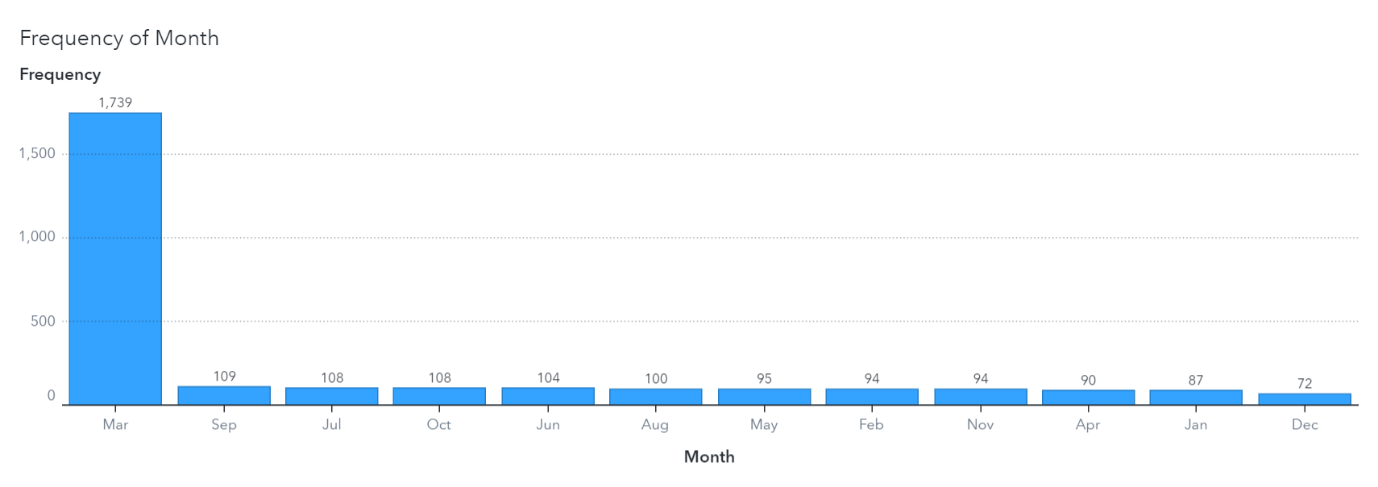
**Conclusion:**

To conclude, increased road safety measures are desperately needed, as evidenced by the frequent collisions that cause traffic jams between J9 Leatherhead and J8 Reigate on the M25. Reigate's accident concentration indicates a crucial area for intervention. The prompt response of emergency services emphasizes how crucial prompt assistances. It is imperative that public awareness campaigns discourage travel in these accident-prone areas during busy hours. Regrettably, deaths have happened, highlighting the necessity of thorough plans to stop more tragedies on Surrey roads. To solve this urgent problem and guarantee safer driving conditions, community members and authorities must work together to improve road safety and infrastructure.

**Task 4: Decision-Maker's Summary and Recommendations**

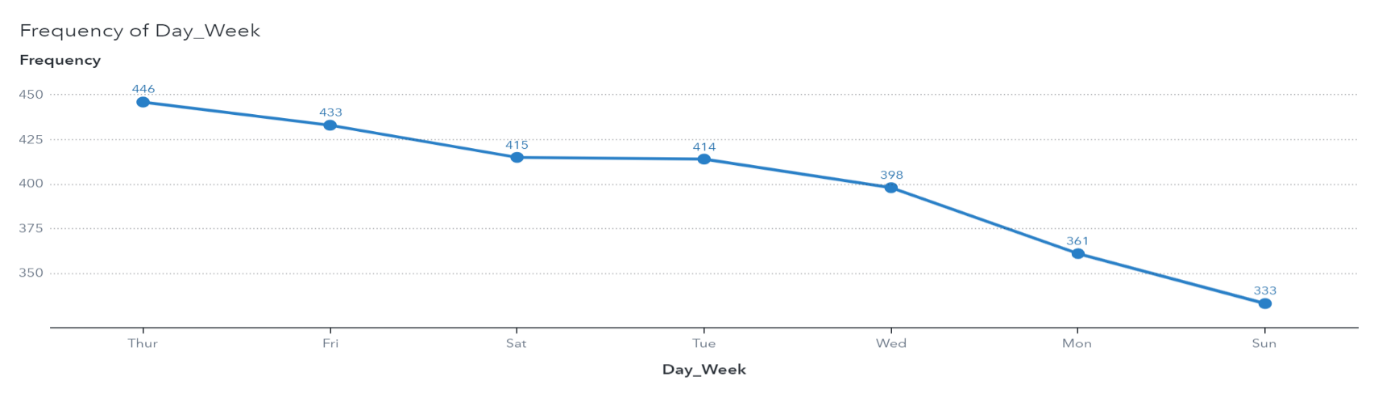
This report offers a detailed examination of the patterns observed in road accidents in the Surrey area throughout 2022.

It is noticeable that the highest number of accidents occur in March, comprising 61% of the total, while December reports the least, with only 2.5%. March is the transitional month with unpredictable weather like snowmelt, rain, start of spring etc.



Notably, 70% of accidents occur on non-trunk roads with single carriageways, highlighting the need for infrastructure enhancements in these areas.

Thursdays and Fridays mark peak accident days, representing 30% of the total cases, advocating for enhanced law enforcement initiatives during these specific days. The stress and tiredness of drivers towards weekend might impact their concentration and decision-making while driving. Nightclubs and other social activities during evenings can lead to an increase in the number of vehicles and people on the road, which can lead to more accidents.



The analysis reveals that a significant 40% of accidents occur during the afternoon, from 12 pm to 6 pm, indicating a significant peak in road accidents during this period. Morning hours, from 6 am to 12 pm, contribute to 27% of the accidents, while evenings, from 6 pm to 11 pm, account for 24% of the incidents.

**Measures to be taken:**

* Road design and infrastructure improvements such as more visibility, straight curves, blind spot elimination, warning signs for bends, intersections and pedestrian crossings, especially for non-trunk roads.
* Implement safe overtaking zones with clear road markings and signs.
* Schedules road maintenance on potholes, uneven surfaces and sign boards.
* Develop advanced warning systems for upcoming hazards, congestions and changing road conditions.
* Implement weather-responsive road maintenance strategies, which includes timely gritting and snow clearance during adverse weather conditions, especially in March.
* Installing proper road barriers and guardrails in critical areas which have good visibility
* Develop efficient emergency response plans and ensure quick response times to provide timely medical assistance and clear road blockages.
* Conduct random checks at different points, mainly on Thursdays and Fridays, to identify drivers under the influence of drugs and alcohol.
* Educate drivers about the importance of regular car maintenance such as tyre, brake and fluid checks in order to have safe driving. Additionally, mirrors and headlights should also be checked.
* Checking blind spots before lane changing and merging into highways need to be done on time using proper indicators.
* Drivers should put their mobile phones away during driving and need to take a break every 2 hours on a long journey, even in
* Local communities can be engaged to gather input on specific road safety challenges and potential solutions.
* Increasing the number of traffic calming methods like speed bumps, humps etc in local roads.

**References**:

* Sas Education (2019). *Exploring SAS Viya*.
* www.point-s.co.uk. (n.d.). *Road accident prevention: road safety with Point S!* [online] Available at: https://www.point-s.co.uk/advice/handy-tips/road-safety#:~:text=Road%20accidents%20are%20a%20serious