



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

Data Mining and Visualisation

Traffic Dataset Analysis Using Data Mining Techniques and
Visualization Tools

Project Report

<i>Name:</i>	<i>Shahriar Khan</i>
<i>Student ID:</i>	<i>2507023</i>
<i>Programme:</i>	<i>Data Science and Analytics</i>
<i>Module:</i>	<i>Data Mining and Visualisation</i>
<i>Module Code:</i>	<i>ACCA7015</i>
<i>Lecturer:</i>	<i>Prof. Nik Whitehead</i>
<i>Internal Verifier:</i>	<i>Dr. Tim Bashford</i>
<i>Date of submission:</i>	<i>15th May 2026</i>

Contents

1	Abstract	3
2	Business Understanding	3
3	Dataset Description	4
3.1	Dataset Source	4
3.2	Purpose and Suitability of the Dataset	5
3.3	Key Attributes	6
4	Data Prepartion	6
4.1	Primary Data Preprocessing	6
4.1.1	Java Heap Memory Problem	6
4.1.2	Dataset Reduction	7
4.1.3	Attribute Selection and Cleaning	8
4.1.4	Creating the Traffic_Level Attribute	9
4.1.5	Replacing Missing Values	10
4.1.6	Converting String Attributes into Nominal Values	10
4.1.7	Discretisation of Numeric Attributes	11
4.2	Secondary Data Preprocessing	12
5	Data Analysis	12
5.1	Classification Analysis	12
5.1.1	J48 Decision Tree Classification	13
5.1.2	Naive Bayes Classification	14
5.1.3	Random Forest Classification	15
5.1.4	Classification Model Comparison	15
5.1.5	Critical Evaluation of Classification Models	16
5.2	Clustering Analysis	18
5.2.1	SimpleKMeans Clustering Results	18
5.2.2	Cluster Visualisation	20
5.3	Association Rule Mining	21
5.3.1	Apriori Association Results	22
6	Tableau Visualisation Analysis	23
6.1	Hourly Pattern	25
6.2	Scatter Relationship	25
6.3	Geographic Map	26
6.4	Dashboard	26

7	Key findings included	27
8	Deployment and Business Recommendations	28
9	Personal Reflection	29
10	Conclusion	30

1. Abstract

This project investigates UK road traffic data using data mining and visualization techniques implemented in WEKA and Tableau. The dataset was collected from the UK Department for Transport's road traffic statistics database and focused on traffic records from London and Wales between 2020 and 2024 [1]. The original dataset contained a very large number of records, which created processing and memory management problems during analysis. Therefore, several preprocessing techniques were applied to optimize the dataset before machine learning (ML) analysis. The project followed the CRISP-DM framework, which includes business understanding, data understanding, data preparation, modeling, evaluation, and deployment stages [2]. This structure ensured that the project remained focused on both technical analysis and practical business applications. Different preprocessing methods such as filtering, attribute reduction, missing value handling, discretization, and conversion of string attributes into nominal values were applied to prepare the dataset for analysis in WEKA. Classification analysis was conducted using J48 Decision Tree, Naive Bayes, and Random Forest algorithms to predict traffic levels. Clustering analysis was performed using SimpleKMeans to identify traffic behavior groups, while Apriori association rule mining was applied to discover relationships between traffic variables. Tableau visualizations were also developed to analyze traffic trends and traffic distribution patterns. Tableau used the original dataset because it was not creating any memory issues. The results demonstrated that the J48 algorithm achieved the highest classification accuracy of 88.42%, while clustering analysis identified clear traffic behavior groups. Association rule mining also revealed strong relationships between vehicle categories and traffic levels. The findings indicate that data mining techniques can support transportation planning, traffic management, and future smart city development.

2. Business Understanding

Traffic congestion has become a major challenge in urban environments due to increasing population growth, vehicle ownership, and transportation demands. High traffic density creates problems such as road congestion, travel delays, fuel wastage, environmental pollution, and increased accident risk [3]. Government organizations and transport authorities therefore collect large amounts of traffic data to monitor traffic conditions and improve transportation systems. The primary business objective of this project was to analyze UK road traffic data to identify traffic behavior patterns and evaluate how ML algorithms can classify traffic levels effectively. The project also aimed to investigate how data preprocessing affects ML performance in WEKA.

The project focused on the following business questions:

- Can ML algorithms accurately classify traffic levels?
- Can clustering techniques identify different traffic behavior groups?
- Are there strong relationships between vehicle categories and traffic conditions?
- How can traffic visualization support transportation planning?

3. Dataset Description

3.1 Dataset Source

The dataset used in this project was collected from the UK Department for Transport Road Traffic Statistics database [1]. The dataset is publicly available through the Department for Transport website and contains traffic count information from different UK regions.

The original raw dataset included:

- Vehicle count information
- Road categories
- Regional information
- Hourly traffic counts
- Heavy goods vehicle counts
- Pedal cycle counts
- Bus and coach counts

The raw CSV dataset was extremely large and unsuitable for direct analysis in WEKA because of memory limitations. Therefore, filtering and optimisation procedures were required before analysis. The dataset was filtered according to:

- Years: 2020–2024
- Regions: London and Wales

After filtering:

- Initial filtered dataset size: 22,897 records
- Second filtered dataset size: 16,966 records
- Final processed dataset size: 15,030 records

The final dataset satisfied the assignment requirement of at least 15,000 records and 12 attributes.

3.2 Purpose and Suitability of the Dataset

The UK Department for Transport road traffic dataset was selected because it provides real-world traffic count information collected from different road locations and regions across the UK [1]. In general, this type of dataset is used for traffic monitoring, transport planning, congestion analysis, infrastructure development, road safety studies, and public policy support. It contains practical variables such as year, hour, region, road category, vehicle type, and total motor vehicle counts, which make it suitable for both statistical analysis and ML tasks.

This dataset is useful for data mining because it contains a large number of records and mixed attribute types. It includes numeric attributes such as **cars_and_taxis**, **LGVs**, **all_HGVs**, and **all_motor_vehicles**, as well as categorical attributes such as **region_name** and **road_category**. This combination makes the dataset suitable for classification, clustering, association rule mining, and visualization. Data mining techniques are commonly applied to large datasets to discover patterns, relationships, and useful knowledge that may not be obvious from manual inspection [3], [4].

The dataset also supports meaningful business and social analysis. For example, transport authorities can use traffic pattern analysis to identify high-traffic areas, compare regional traffic behavior, understand peak travel periods, and support future road planning. In this project, the dataset was filtered to London and Wales for 2020–2024 so that the analysis remained focused, manageable, and relevant to recent traffic behaviour.

This report is useful because it shows how raw traffic data can be converted into meaningful insights through a complete data mining process. The report does not only present results; it also explains how the dataset was cleaned, reduced, transformed, analyzed, evaluated, and visualized. This makes the work practical and suitable for a real data mining workflow.

The report is strongly aligned with the **CRISP-DM** methodology because it follows the main stages of a structured data mining project [2]. First, the business problem was identified: understanding traffic behavior and classifying traffic levels. Second, the dataset was explored and described. Third, data preparation was completed using Excel and WEKA through filtering, attribute selection, missing value handling, string-to-nominal conversion, and discretisation. Fourth, modeling was performed using classification, clustering, and association rule mining. Fifth, results were evaluated using accuracy, confusion matrix, cluster interpretation, and association rules. Finally, Tableau dashboards were used to present the findings in a clear and practical format.

Therefore, this dataset and report are suitable for the assignment because they demonstrate the full data mining lifecycle, from raw data preparation to final interpretation. The project also shows realistic challenges, such as WEKA memory issues, and explains

how preprocessing solved them. This makes the report stronger because it reflects both technical analysis and practical problem-solving.

3.3 Key Attributes

Table 1: Final Selected Attributes

Attribute	Description
year	Traffic count year
hour	Hour of traffic measurement
region_name	Region location
road_category	Road classification
pedal_cycles	Number of pedal cycles
two_wheeled_motor_vehicles	Motorbike counts
cars_and_taxis	Cars and taxis count
buses_and_coaches	Bus and coach count
LGVs	Light goods vehicle count
all_HGVs	Heavy goods vehicle count
all_motor_vehicles	Total motor vehicle count

4. Data Prepartion

4.1 Primary Data Preprocessing

Data preprocessing is one of the most important stages in data mining because raw datasets frequently contain unnecessary information, missing values, duplicated records, and incompatible data types [4]. The preprocessing stage was completed using Microsoft Excel and WEKA.

4.1.1 Java Heap Memory Problem

When classification algorithms such as Random Forest were applied to the original dataset, WEKA generated Java heap memory errors because the dataset exceeded the available memory allocation [5].

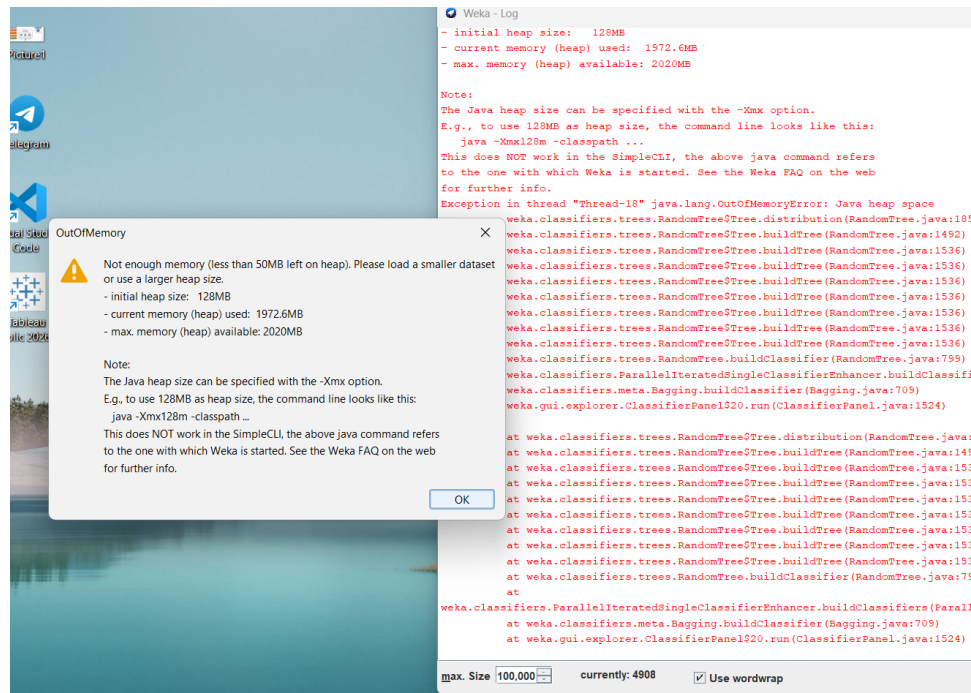


Figure 1: Java Heap Memory Error

This figure shows the Java heap memory error generated during WEKA processing. The error occurred because Random Forest required high memory resources while generating multiple decision trees simultaneously. The error message indicated that the available heap memory was insufficient for model training.

To solve this problem:

- Dataset size was reduced
- Unnecessary attributes were removed
- Less important records were filtered out
- Numeric attributes were discretised

These preprocessing techniques improved WEKA performance significantly.

4.1.2 Dataset Reduction

The original traffic dataset was too large for efficient WEKA analysis. Therefore, dataset filtering was performed in Microsoft Excel. The filtering criteria included:

- Selecting records only from 2020 to 2024
- Keeping records only from London and Wales

This significantly reduced the dataset size while preserving meaningful traffic behavior information.

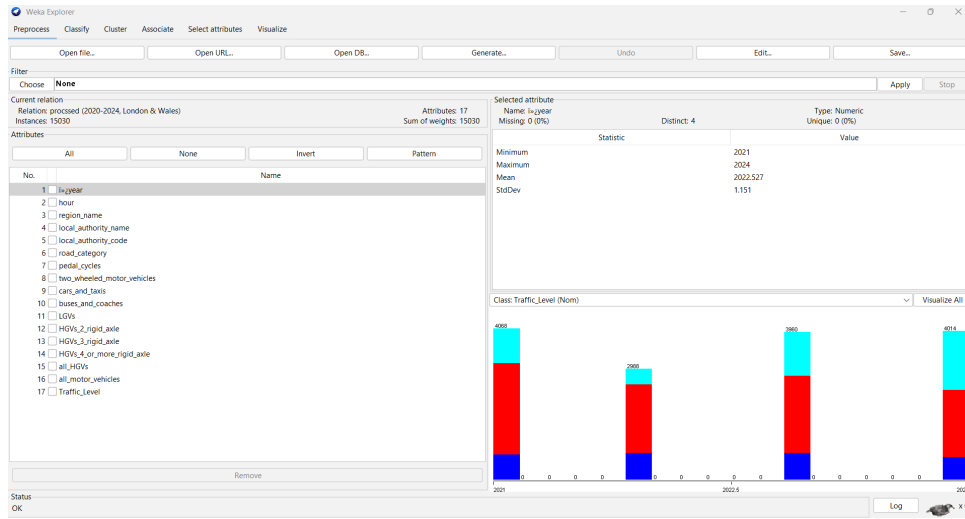


Figure 2: Dataset Loaded into WEKA

The visualization confirmed that the filtered dataset was successfully loaded into WEKA and ready for preprocessing and ML analysis. The dataset contains 15,030 records and 17 attributes before further preprocessing. WEKA automatically generated statistical summaries, including minimum values, maximum values, mean values, and standard deviation.

4.1.3 Attribute Selection and Cleaning

Several attributes in the dataset were unnecessary for ML analysis. Attributes such as local authority codes and duplicated identifiers did not contribute meaningful information for traffic prediction.

The following attributes were removed:

- local_authority_name
- local_authority_code
- HGVs_2_rigid_axle
- HGVs_3_rigid_axle
- HGVs_4_or_more_rigid_axle

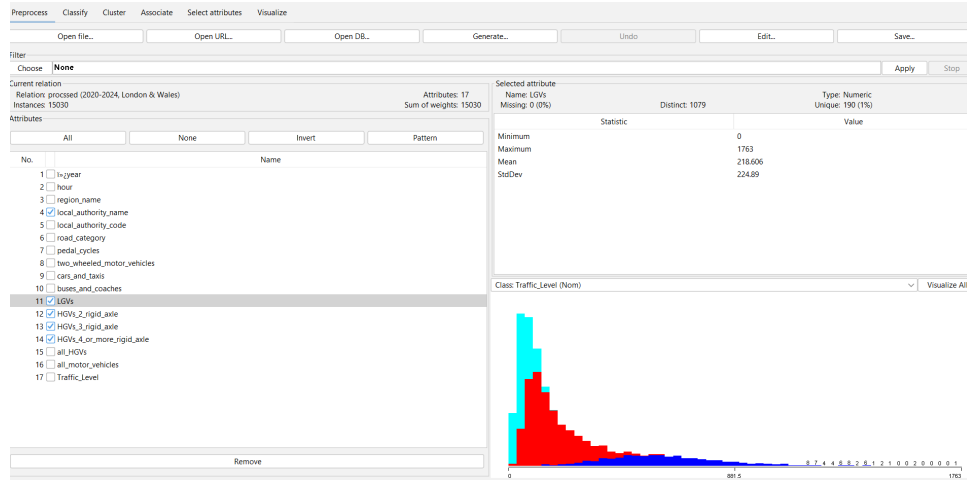


Figure 3: Attribute Selection and Cleaning

Here is the dataset after unnecessary attributes were removed. The attribute reduction process simplified the dataset and improved ML efficiency [4]. Removing irrelevant attributes also reduced overfitting risk and memory consumption during classification analysis.

The remaining attributes were selected because they directly represented traffic behaviour and vehicle activity.

4.1.4 Creating the Traffic_Level Attribute

A new classification attribute called **Traffic_Level** was created in Microsoft Excel to support supervised ML analysis.

The following Excel formula was used:

`=IF(P2<500,"Low",IF(P2<2000,"Medium","High"))`

This formula categorized traffic into low, medium, and high based on the **all_motor_vehicles** attribute.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	year	hour	region_nar	local_auth	local_auth	road_cate	pedal_cycl	two_wheel	cars_and	buses_and	LGVs	HGVs_2_ri	HGVs_3_ri	HGVs_4_o	all_HGVs	all_motor	Traffic_Level
2	2022	12	Wales	Newport	W0600002TM		0	10	2617	10	704	66	23	15	293	3634	High
3	2022	15	Wales	Newport	W0600002TM		0	11	3037	21	645	51	13	15	214	3928	High
4	2022	18	Wales	Newport	W0600002TM		0	21	2999	11	373	15	5	9	111	3515	High
5	2022	7	Wales	Newport	W0600002TM		0	5	2471	5	897	109	43	22	462	3840	High
6	2022	8	Wales	Newport	W0600002TM		0	12	2473	27	819	81	29	33	349	3680	High
7	2022	16	Wales	Newport	W0600002TM		0	9	3441	11	629	35	11	3	192	4282	High
8	2022	12	Wales	Newport	W0600002TM		0	10	2810	14	642	83	26	13	348	3824	High
9	2022	14	Wales	Newport	W0600002TM		0	13	2886	17	749	30	7	14	183	3848	High
10	2022	11	Wales	Newport	W0600002TM		0	12	2471	11	651	68	32	23	343	3488	High
11	2022	13	Wales	Newport	W0600002TM		0	8	2608	17	603	69	13	9	285	3521	High
12	2022	11	Wales	Newport	W0600002TM		0	15	2822	15	650	81	31	11	350	3852	High

Figure 4: Traffic_Level Attribute Creation

Here the dataset after adding the Traffic_Level attribute. The Traffic_Level column

became the target class attribute for classification analysis in WEKA. Creating this class attribute transformed the dataset into a supervised classification problem.

4.1.5 Replacing Missing Values

Some records contained incomplete or missing attribute values. Missing values can reduce ML accuracy and create processing problems during classification.

The **ReplaceMissingValues** filter was applied in WEKA.

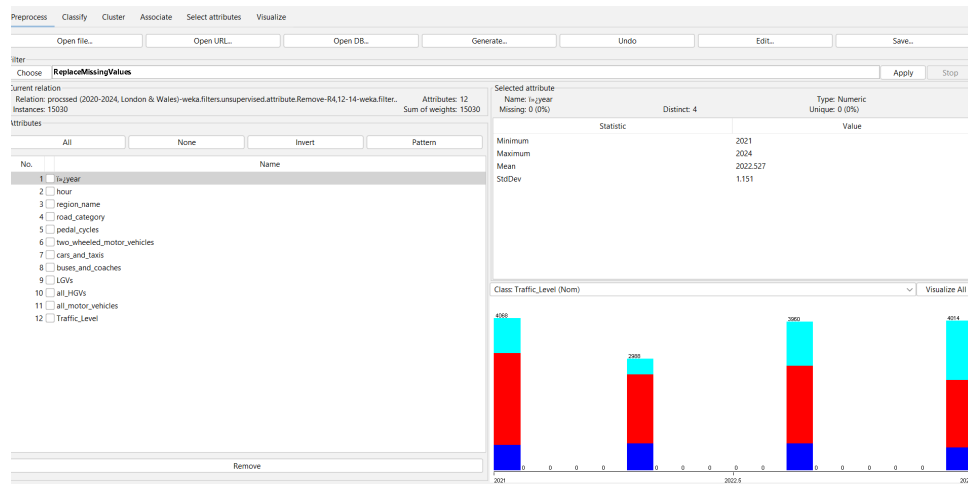


Figure 5: ReplaceMissingValues Filter

4.1.6 Converting String Attributes into Nominal Values

Some attributes such as **region_name** contained string data types. Several WEKA algorithms perform better when categorical values are stored as nominal data.

The **StringToNominal** filter was therefore applied.

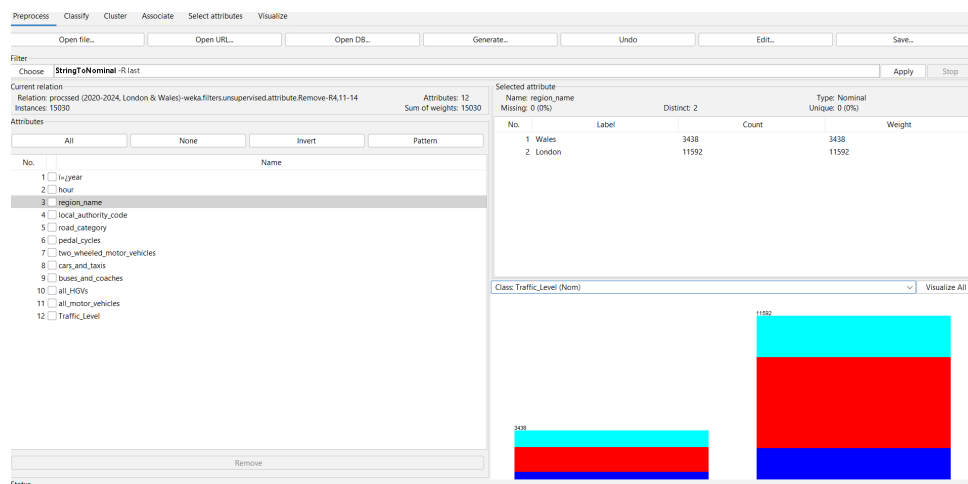


Figure 6: String to Nominal Conversion

After conversion, WEKA displayed distinct category counts for Wales and London. Converting string values into nominal format improved compatibility with classification and clustering algorithms.

4.1.7 Discretisation of Numeric Attributes

Association rule mining algorithms such as Apriori cannot directly process continuous numeric attributes efficiently. Therefore, discretization was applied to several traffic-related numeric variables.

The discretized attributes included

- cars_and_taxis
- buses_and_coaches
- LGVs
- all_motor_vehicles

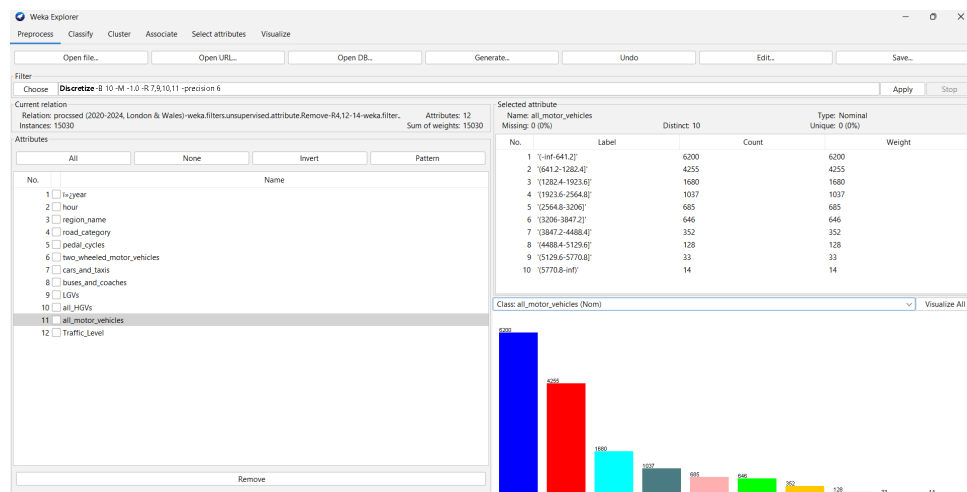


Figure 7: Discretisation Process

Numeric traffic attributes were transformed into interval-based categorical ranges by using a discretization filter.

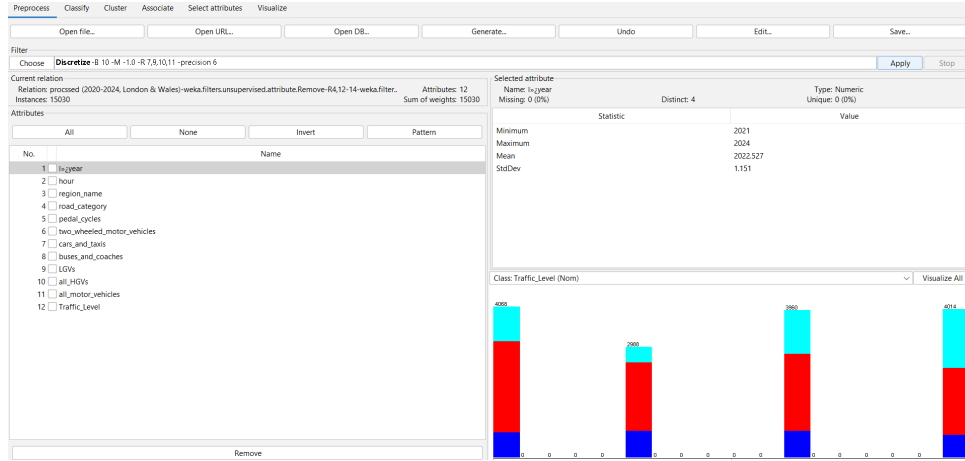


Figure 8: Discretised Attribute Output

This figure shows the discretized output for **Traffic_Level**. The numeric traffic values were divided into multiple categorical intervals, such as low, medium and high traffic ranges.

4.2 Secondary Data Preprocessing

Secondary preprocessing was performed after the initial analysis and modelling stages in order to further improve dataset quality and ML performance.

The secondary preprocessing process included:

- Less important features were reduced
- Attributes were adjusted for better accuracy
- Further discretisation supported Apriori

The additional preprocessing steps improved overall model quality, reduced unnecessary complexity, and increased compatibility with association rule mining algorithms.

This iterative approach aligns with CRISP-DM best practices because preprocessing and modelling stages are repeatedly refined until satisfactory analytical performance is achieved.

5. Data Analysis

5.1 Classification Analysis

Classification analysis was performed to predict traffic levels using traffic-related attributes. The **Traffic_Level** attribute was selected as the class variable.

The following algorithms were tested:

- J48 Decision Tree
- Naive Bayes
- Random Forest

Ten-fold cross-validation was used during training to improve model reliability [6].

5.1.1 J48 Decision Tree Classification

J48 is a decision tree algorithm based on the C4.5 classification method developed by Quinlan [7].

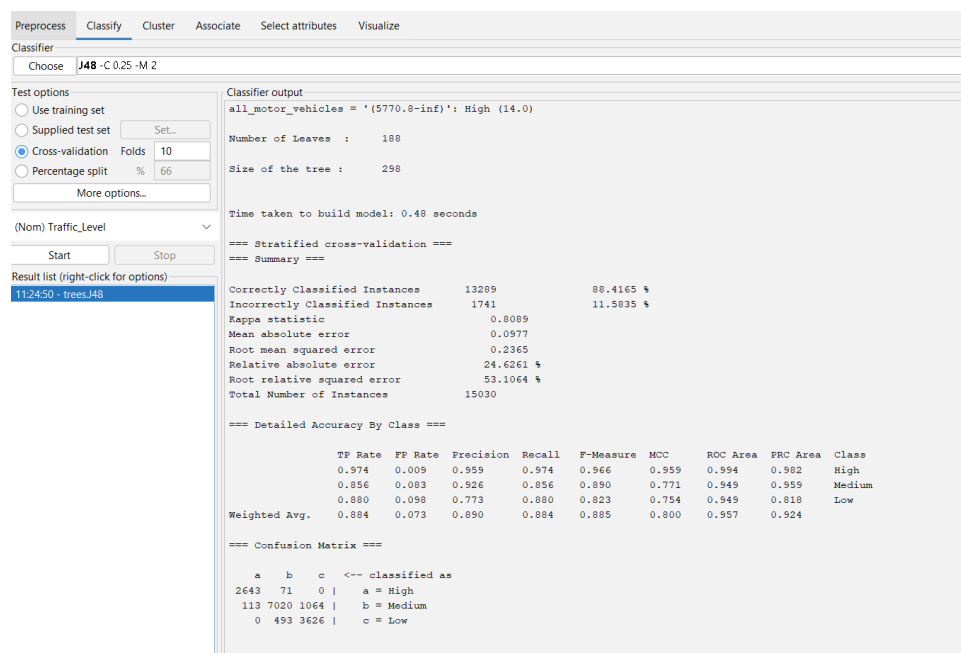


Figure 9: J48 Classification Results

It's shows the J48 classification output and confusion matrix. The algorithm successfully classified most traffic records correctly.

Results:

- Accuracy: 88.42%
- Kappa Statistic: 0.8089
- Weighted F-Measure: 0.885

The confusion matrix showed that high-traffic categories were classified particularly well. J48 produced the best overall performance because decision trees can effectively process mixed traffic attributes and categorical variables. Another advantage of J48 is interpretability. The decision-making process can easily be understood through branching

structures, which makes the algorithm suitable for business reporting and transportation analysis.

5.1.2 Naive Bayes Classification

Naive Bayes classification was also tested during the modelling stage.

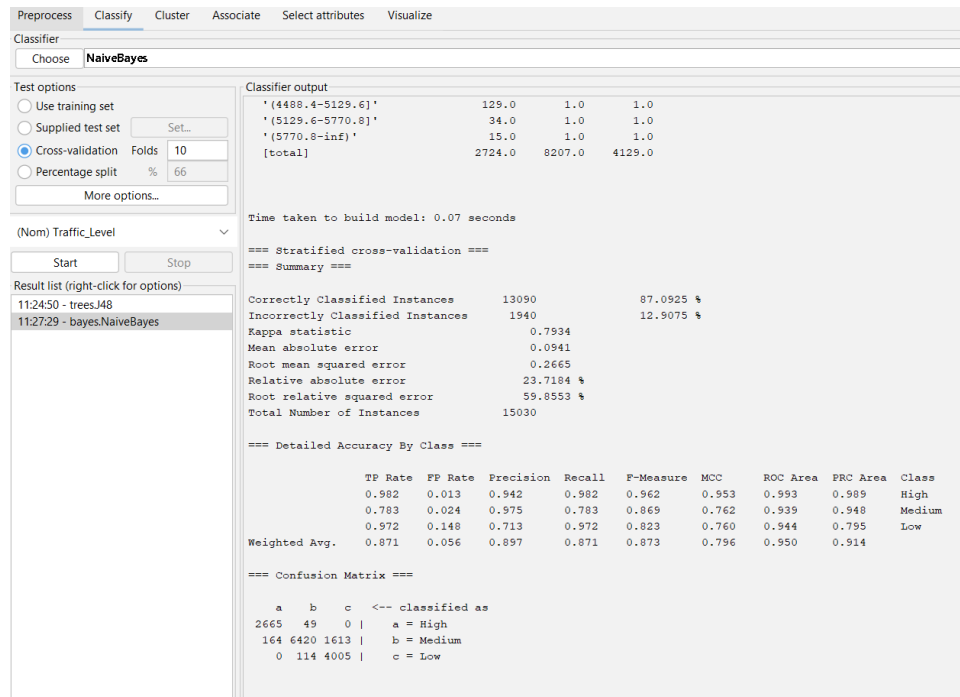


Figure 10: Naive Bayes Classification Results

The algorithm achieved slightly lower accuracy than J48 but processed the dataset very quickly.

Results:

- Accuracy: 87.09%
- Kappa Statistic: 0.7934
- Weighted F-Measure: 0.873

Naive Bayes assumes independence between attributes, which may not fully represent real traffic behavior because traffic variables are often related to one another.

Despite this limitation, the model still produced strong classification performance.

5.1.3 Random Forest Classification

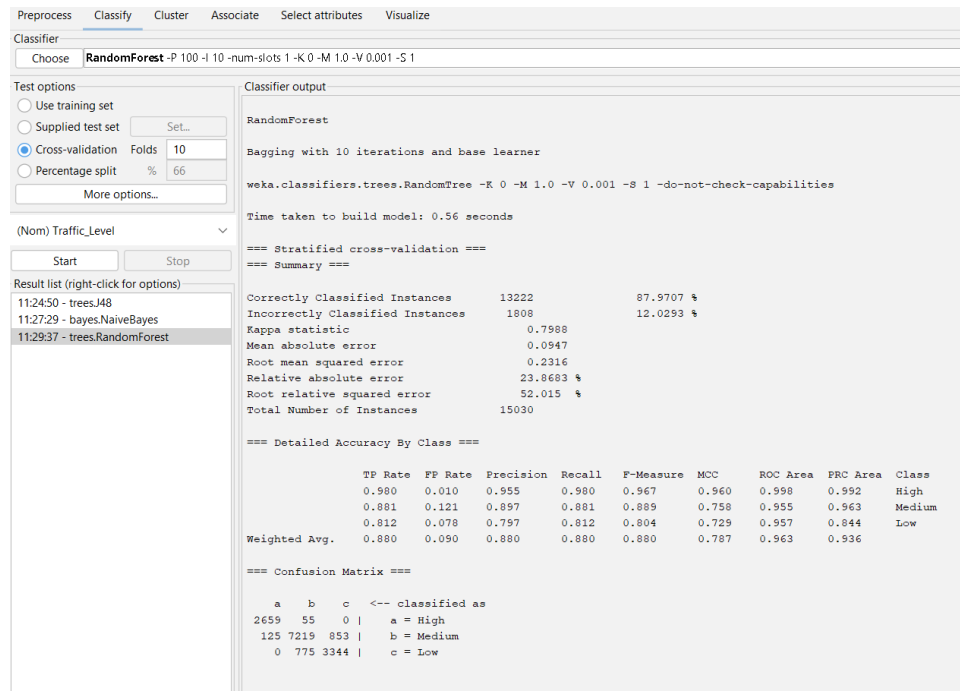


Figure 11: Random Forest Classification Results

Random Forest achieved high prediction accuracy but required significantly higher memory resources during training.

Results:

- Accuracy: 87.97%
- Kappa Statistic: 0.7988
- Weighted F-Measure: 0.880

Random Forest performed well because multiple decision trees were combined to improve prediction stability. However, the algorithm generated higher computational complexity compared to J48.

5.1.4 Classification Model Comparison

Table 2: Classification Algorithm Comparison

Algorithm	Accuracy	Kappa Statistic	Weighted F-Measure
J48	88.42%	0.8089	0.885
Naive Bayes	87.09%	0.7934	0.873
Random Forest	87.97%	0.7988	0.880

The comparison shows that J48 achieved the highest overall classification accuracy and F-measure.

The J48 model was selected as the best classifier because:

- Highest accuracy
- Strong confusion matrix performance
- Easier interpretation
- Lower memory consumption

5.1.5 Critical Evaluation of Classification Models

While J48, Random Forest, and Naïve Bayes all achieved relatively strong predictive performance, their suitability differs when aspects such as interpretability, computational efficiency, and underlying assumptions are considered.

Why J48 Performed Better Than Random Forest in This Dataset

Although Random Forest is widely recognised as a powerful ensemble learning technique, J48 demonstrated more practical effectiveness for this dataset beyond simple accuracy comparison. Firstly, the dataset contains a combination of categorical and discretised numerical attributes. Decision tree algorithms, such as J48, are inherently well suited to handling such mixed data types without requiring complex preprocessing. The model constructs a single tree using clear attribute splits, enabling efficient representation of structured traffic data. Secondly, computational constraints played a significant role. Random Forest builds multiple decision trees during training, which increases memory consumption and processing overhead. During experimentation, the algorithm generated Java heap memory errors, indicating that the model exceeded available computational resources. In contrast, J48 required minimal memory and successfully completed training without additional optimization. Ensemble methods are known to demand higher computational resources compared to single-tree approaches [8]. Another important factor is interpretability. J48 generates a transparent tree structure, allowing users to clearly understand the decision-making process. This characteristic is particularly valuable in transportation applications where stakeholders require explainable results. Random Forest, on the other hand, operates as an ensemble of multiple trees, making interpretation more difficult. Furthermore, the difference in predictive accuracy between J48 (88.42%) and Random Forest (87.97%) is marginal. This suggests that the additional complexity introduced by Random Forest does not provide significant benefits for this dataset. Previous research has shown that decision trees can perform competitively on structured datasets where relationships between variables are relatively clear [9].

Limitations of the Classification Models

Despite their effectiveness, each classification model has inherent limitations that must be acknowledged.

J48 Decision Tree:

- Susceptible to overfitting when the tree becomes overly complex
- Sensitive to small changes in the dataset
- Limited ability to model highly complex, non-linear relationships

Random Forest:

- High computational and memory requirements
- Reduced interpretability due to ensemble structure
- Increased training time, which may limit practical deployment

Naïve Bayes:

- Assumes independence between predictor variables
- This assumption is unrealistic in traffic data where attributes are often correlated
- May lead to reduced predictive performance despite computational efficiency

These limitations highlight the trade-offs between accuracy, interpretability, and computational efficiency, emphasising the importance of selecting models based on dataset characteristics and application needs [10].

As a result:

- Models may favour dominant classes such as “High” traffic levels
- Minority classes may be underrepresented, reducing classification reliability

The confusion matrix results suggest that high traffic categories were predicted more accurately than lower levels, indicating bias towards majority patterns.

Additionally, strong correlations between variables — such as **cars_and_taxis** and **all_motor_vehicles** — introduce redundancy, which can affect model generalization. This issue is particularly important for Naïve Bayes due to its independence assumption.

To mitigate these issues, future work could include:

- Applying resampling techniques such as SMOTE
- Performing feature selection to reduce redundancy

- Evaluating models on balanced datasets

Addressing bias is essential to ensure that predictive models perform consistently across different traffic conditions and regions.

Bias and Data Imbalance Discussion

The dataset shows evidence of imbalance, particularly due to the higher number of traffic records from London compared to Wales. This imbalance introduces potential bias during model training.

5.2 Clustering Analysis

Clustering analysis was performed using the **SimpleKMeans** algorithm.

The purpose of clustering was to identify natural traffic behaviour groups without using predefined class labels.

5.2.1 SimpleKMeans Clustering Results

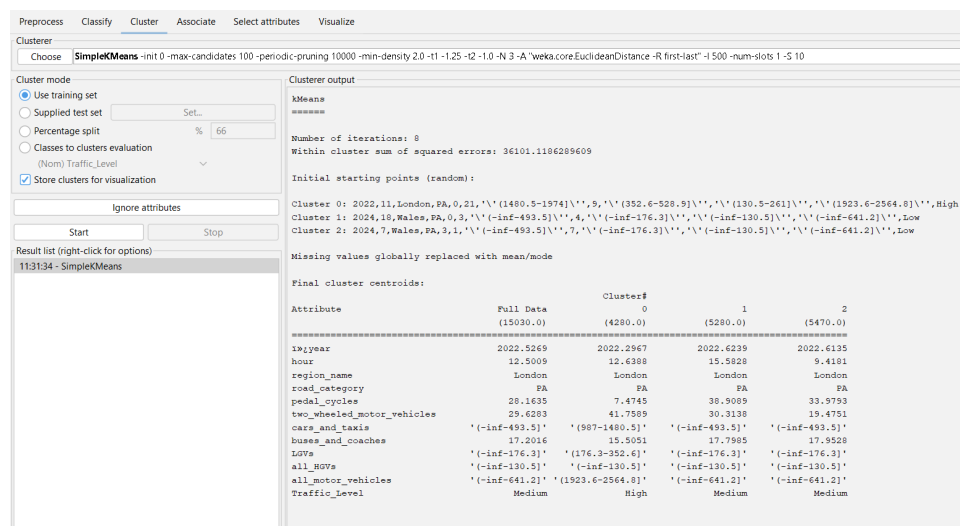


Figure 12: SimpleKMeans Cluster Output

This figure shows the initial SimpleKMeans clustering output.

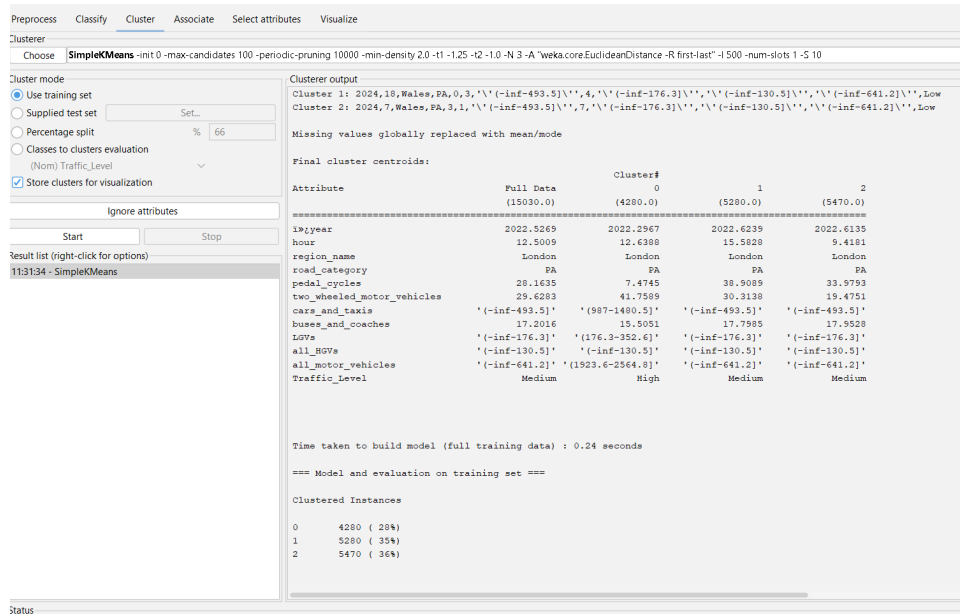


Figure 13: Cluster Centroid Information

Here is displayed centroid information for each cluster generated by the SimpleKMeans algorithm.

Cluster sizes:

- Cluster 0: 4280 records
- Cluster 1: 5280 records
- Cluster 2: 5470 records

The clustering analysis identified clear differences between traffic intensity groups.

5.2.2 Cluster Visualisation



Figure 14: Cluster Visualisation 1

It shows cluster grouping based on traffic distribution patterns.

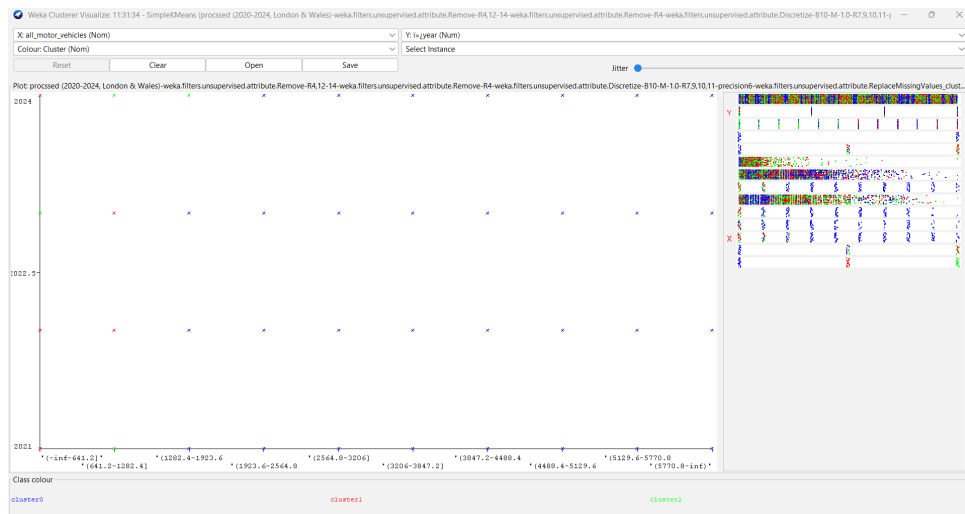


Figure 15: Cluster Visualisation 2

This visualisation demonstrates separation between low and high traffic behaviour groups.



Figure 16: Cluster Visualisation 3

The concentration of traffic records within different traffic ranges.

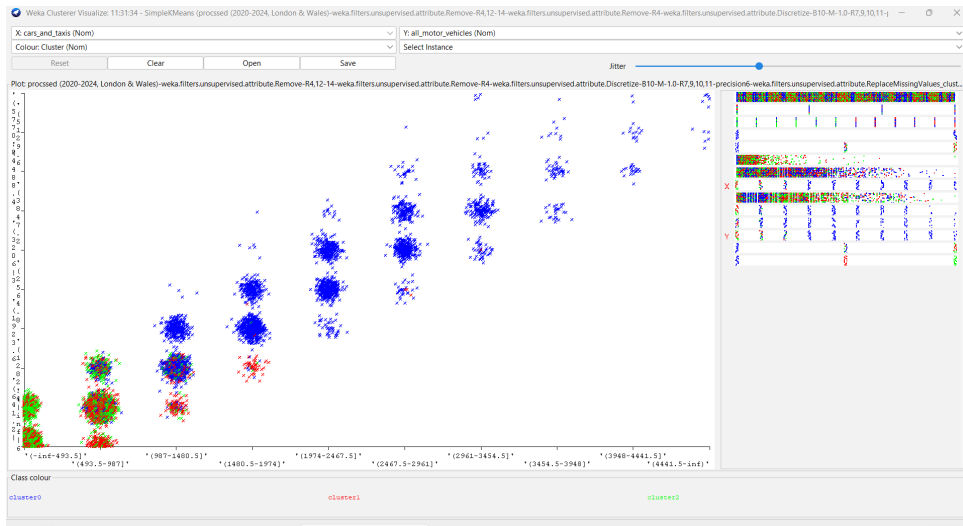


Figure 17: Cluster Visualisation 4

It demonstrates that high vehicle count records formed a separate traffic cluster.

The clustering visualisations confirmed that the dataset contained distinct traffic behaviour patterns.

5.3 Association Rule Mining

Association rule mining was performed using the Apriori algorithm to identify relationships between traffic-related variables. Apriori is a well-established algorithm that identifies frequent itemsets and generates association rules using support and confidence measures [a5]. In this study, continuous attributes were discretised into categorical ranges before

applying the algorithm, as Apriori is more effective with discrete data [a15]. The resulting rules revealed consistent relationships between vehicle categories and overall traffic levels. For example, low values of light goods vehicles were frequently associated with low heavy vehicle counts, while high total vehicle volumes strongly corresponded to high traffic conditions. These findings indicate that traffic behaviour is influenced by predictable interactions between different vehicle types, demonstrating the usefulness of association rule mining in transportation data analysis.

Association rule mining was performed using the **Apriori** algorithm.

5.3.1 Apriori Association Results

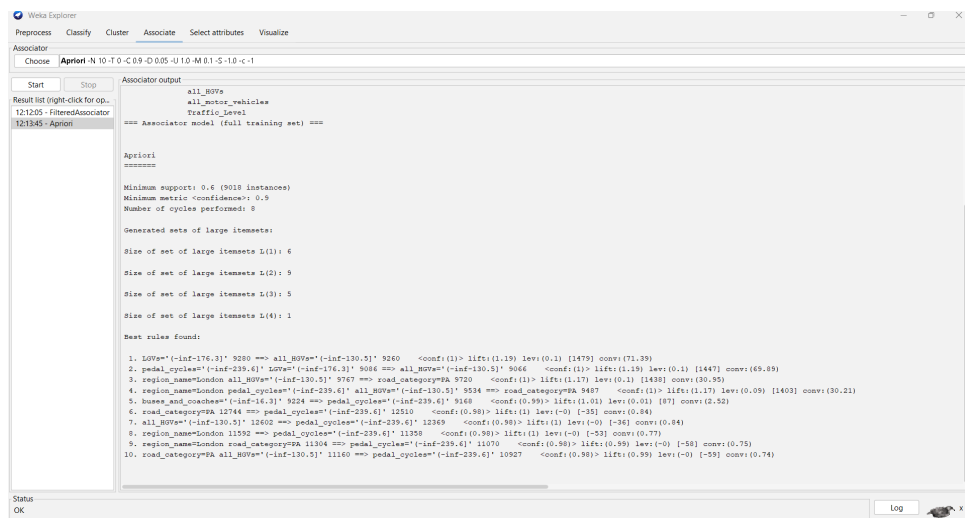


Figure 18: Apriori Association Rules

The Apriori association rule mining output generated in WEKA.

Several strong rules were identified:

- Low LGV traffic associated with low HGV traffic
- London traffic associated with specific road categories
- High motor vehicle counts associated with High traffic levels

The generated rules achieved high confidence values, indicating reliable relationships between traffic variables. Association analysis provided useful business insights because it identified relationships that may support transportation planning and traffic management.

6. Tableau Visualisation Analysis

Tableau provided more interactive and visually understandable analysis compared to WEKA histograms. Here, the original dataset is used because Tableau does not care about the dataset size or processing time of the dataset.

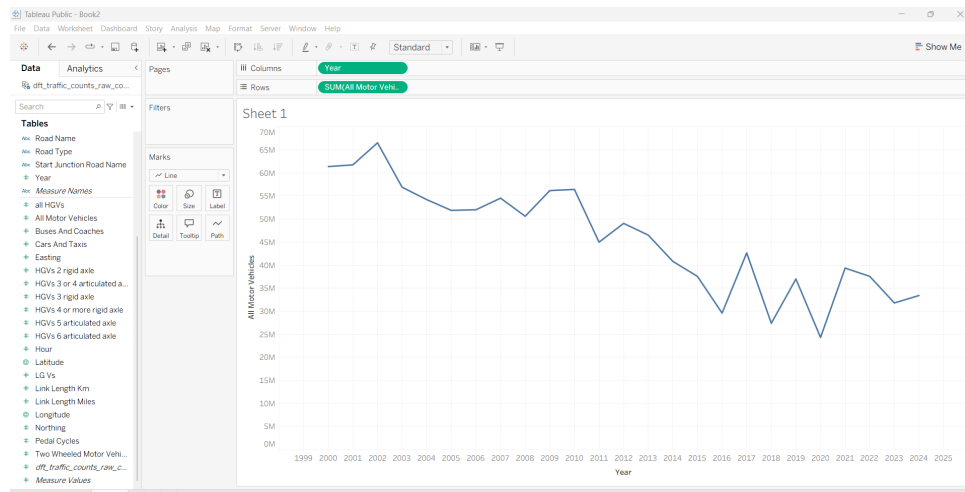


Figure 19: Year Distribution Visualisation

This figure shows traffic record distribution across years from 2000 to 2024. The visualization indicates relatively balanced traffic data collection across different years. Traffic records increased slightly during later years, particularly after recovery from pandemic-related transportation restrictions.

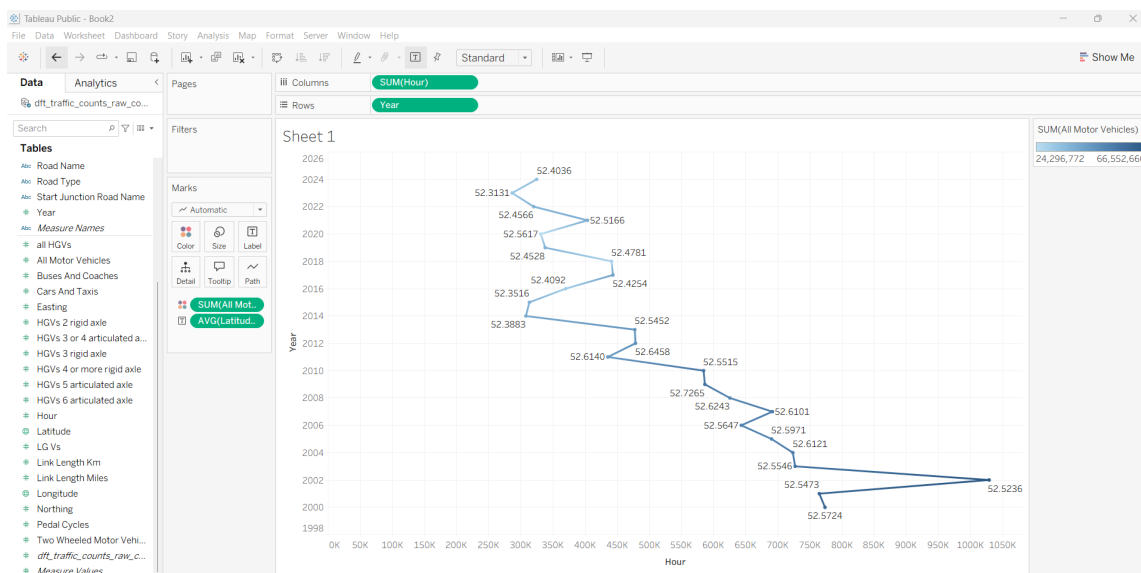


Figure 20: Yearly Traffic Pattern Visualisation in Tableau

The graph shows fluctuations in overall traffic behavior across different years, indicating

periods of increased and decreased traffic intensity. The visualization helps identify traffic pattern changes over time and supports comparative traffic analysis for transportation planning and congestion monitoring. The color intensity also represents variations in total motor vehicle counts, where darker shades indicate higher traffic volume, labeled by latitude.

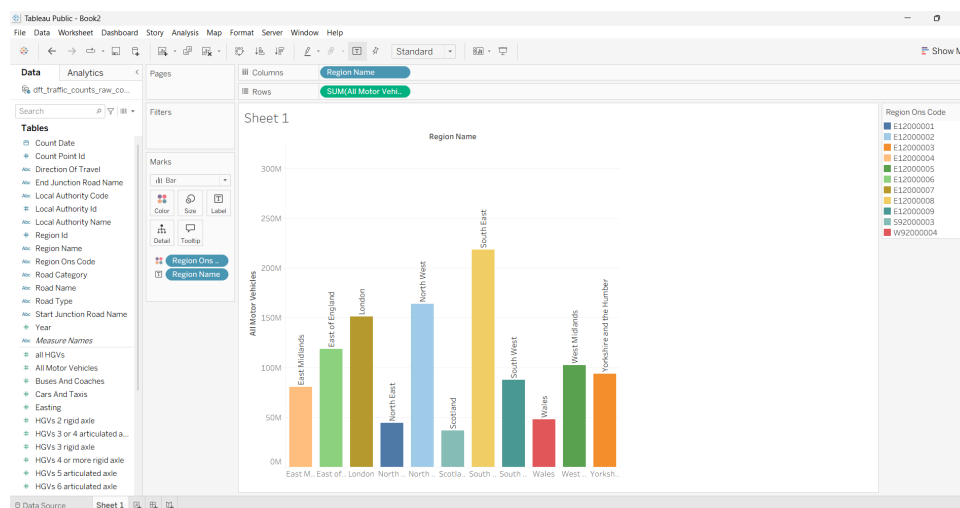


Figure 21: Traffic Distribution by Region

The visualization shows that regions such as the South East, North West, and London recorded significantly higher traffic volumes compared to Wales and Scotland. The chart helps identify regional differences in transportation activity and highlights areas with heavier road usage. This analysis is useful for traffic management, infrastructure planning, and understanding regional traffic distribution patterns.

6.1 Hourly Pattern

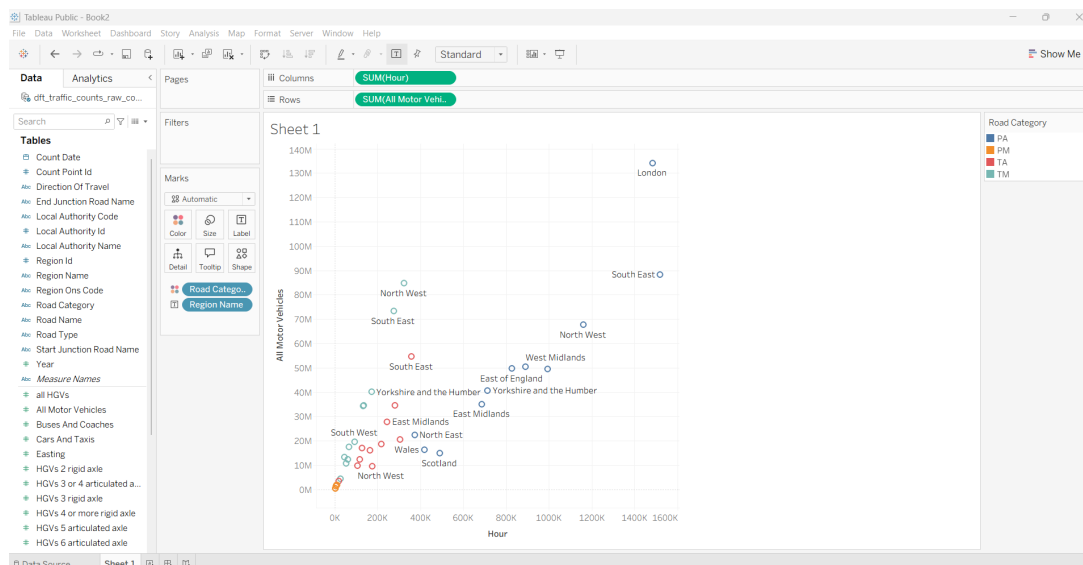


Figure 22: Hourly Traffic Distribution Across Regions

Here are mentioned hourly traffic distribution patterns across different regions. Traffic volumes increased significantly during daytime periods, particularly during morning and evening peak hours. London demonstrated consistently higher vehicle activity compared to Wales throughout most hourly intervals.

6.2 Scatter Relationship

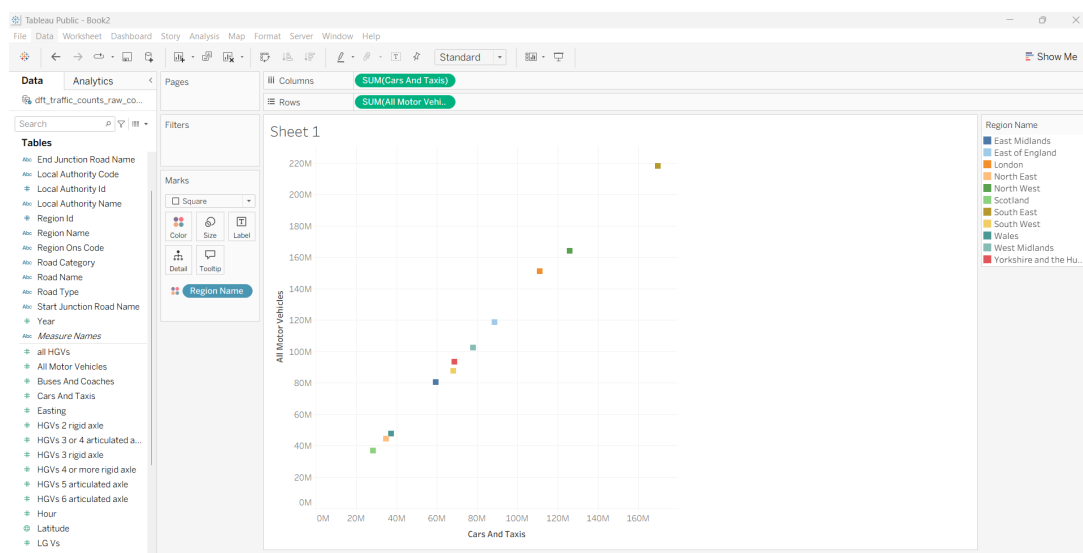


Figure 23: Relationship Between Cars/Taxis and Total Vehicles

It presents the relationship between cars/taxis and total motor vehicle counts. The scatter plot demonstrates a strong positive correlation between private vehicle activity and overall

traffic volume. Higher numbers of cars and taxis generally contributed to increased total traffic density.

6.3 Geographic Map

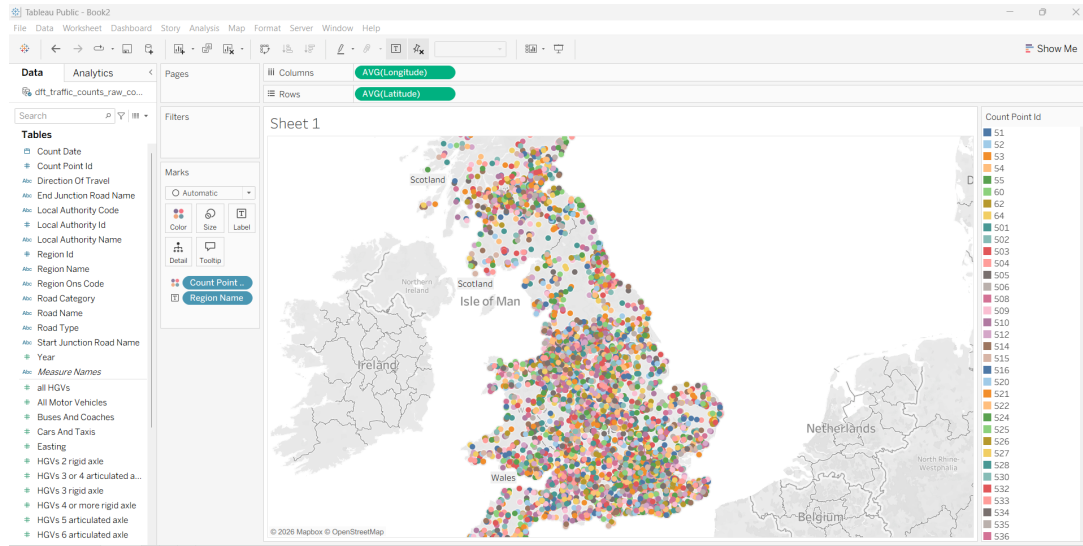


Figure 24: Geographic Distribution of Traffic Across the UK

This figure shows the geographic distribution of traffic activity across the UK. Metropolitan regions, particularly London, demonstrated significantly higher traffic density compared to rural regions. The map visualization helped identify regional traffic concentration patterns more effectively.

6.4 Dashboard

Tableau dashboards were created to visualize traffic trends more effectively.

The dashboards included:

- Regional traffic comparison
- Yearly traffic distribution
- Hourly Traffic Pattern
- Traffic level distribution

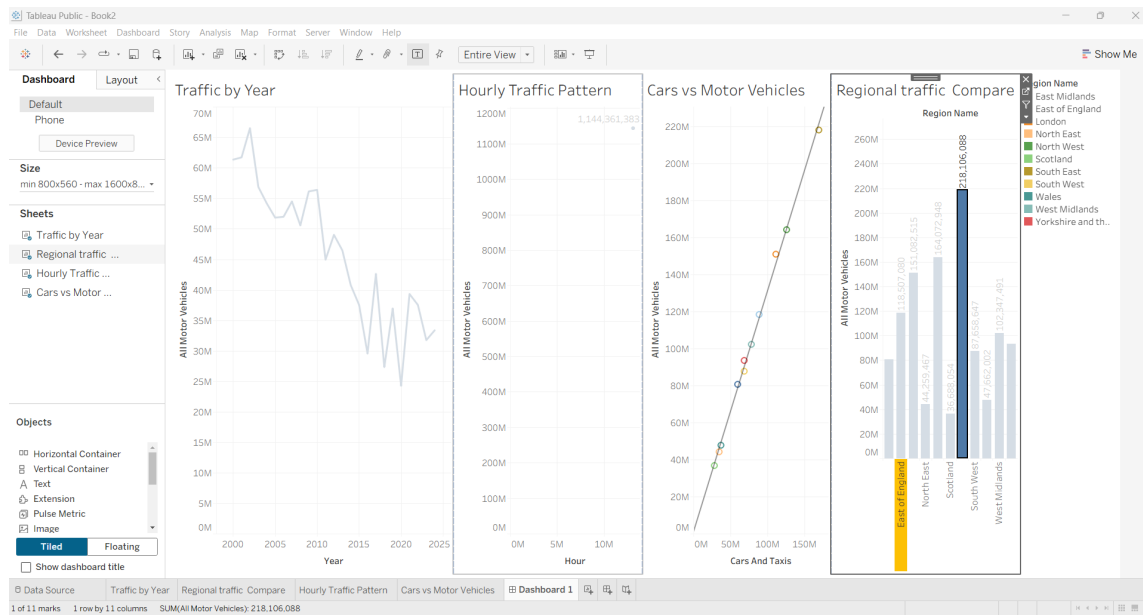


Figure 25: Integrated Tableau Dashboard Showing Multiple Traffic Patterns

The final integrated Tableau dashboard combining multiple traffic visualizations into a single interactive interface. The dashboard included regional comparison, yearly traffic distribution, vehicle category analysis, and traffic level visualizations. The dashboard improved overall understanding of traffic behavior patterns and supported clearer communication of analytical findings.

The dashboards demonstrated:

- London consistently had higher traffic volumes
- Vehicle traffic increased during daytime periods
- High traffic categories were concentrated in metropolitan regions

These visualisations improved communication of analytical findings to non-technical stakeholders. It also confirmed that regional traffic behaviour differences existed within the dataset.

7. Key findings included

The general finding overall:

- London contained significantly higher traffic density than Wales
- J48 achieved the highest classification performance
- Random Forest required higher computational resources
- Traffic variables formed clear clusters during clustering analysis

- Strong relationships existed between heavy vehicle counts and total traffic levels

The findings demonstrate that preprocessing strongly influences ML performance.

The project also showed that:

- Dataset reduction improved WEKA efficiency
- Discretisation improved Apriori performance
- Attribute selection reduced memory consumption
- Classification accuracy remained high after optimisation

These findings could support:

- Traffic monitoring systems
- Smart city transportation planning
- Congestion reduction strategies
- Infrastructure planning
- Public transportation management

8. Deployment and Business Recommendations

The developed models could potentially support transportation authorities in analyzing traffic behavior more efficiently.

Possible business applications include:

- Predicting high traffic conditions
- Monitoring congestion patterns
- Planning traffic control systems
- Identifying heavy traffic regions
- Supporting urban transportation planning

The J48 model appears suitable for practical deployment because:

- High accuracy
- Lower memory usage
- Easier interpretation
- Faster processing speed

Future improvements may include:

- Real-time traffic prediction
- Integration with weather datasets
- GPS-based traffic analysis
- Deep learning forecasting models
- IoT sensor integration

9. Personal Reflection

This project significantly improved my understanding of data mining, machine learning, and visualization techniques using WEKA and Tableau.

At the beginning of the project, I focused mainly on ML algorithms and expected classification accuracy to be the most important factor. However, during the preprocessing stage, I realized that data quality and preprocessing decisions strongly influence overall model performance.

One of the biggest challenges during the project was handling the large traffic dataset. When Random Forest classification was first attempted, WEKA generated Java heap memory errors because the dataset exceeded available memory limits. Solving this issue required reducing the dataset size, removing unnecessary attributes, and applying preprocessing techniques such as discretization and attribute selection.

This challenge helped me understand that preprocessing is often more important than the ML algorithm itself.

Another important learning experience involved algorithm comparison. Although Random Forest produced strong accuracy results, it consumed more memory resources compared to J48. J48 achieved slightly better accuracy while remaining easier to interpret and more efficient during processing.

The project also improved my ability to create visualizations in Tableau and interpret traffic behavior patterns. Before this project, I had limited experience with traffic data visualization. Creating dashboards and comparing regional traffic behavior improved my analytical and presentation skills.

I also learned the importance of reporting analytical findings clearly for business stakeholders. Technical outputs alone are not sufficient if decision-makers cannot understand the results or apply them practically.

Overall, this project improved my confidence in:

- Data preprocessing
- Machine learning analysis
- WEKA usage
- Tableau visualisation
- Business reporting
- Traffic pattern analysis

In future projects, I would like to improve my knowledge of advanced predictive analytics, deep learning techniques, and real-time traffic forecasting systems.

10. Conclusion

This project analyzed UK road traffic data using WEKA and Tableau through classification, clustering, association rule mining, and visualization techniques.

The dataset was successfully preprocessed and reduced from a very large raw dataset into an optimised dataset suitable for ML analysis. Several preprocessing methods including filtering, missing value replacement, discretization, and nominal conversion, improved system performance and reduced memory consumption.

Among the tested algorithms, J48 achieved the best overall classification accuracy of 88.42%. Clustering analysis identified clear traffic behaviour groups, while Apriori association analysis identified meaningful relationships between traffic variables.

The project demonstrated that data mining techniques can successfully analyze large traffic datasets and identify useful traffic behavior patterns for transportation management and smart city planning.

References

- [1] UK Department for Transport, “Road traffic statistics,” 2025. [Online]. Available: <https://roadtraffic.dft.gov.uk/downloads>.
- [2] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth, *et al.*, “Crisp-dm 1.0: Step-by-step data mining guide,” *SPSS inc*, vol. 9, no. 13, pp. 1–73, 2000.
- [3] M. Treiber and A. Kesting, *Traffic flow dynamics: data, models and simulation*. Springer Science & Business Media, 2012.
- [4] J. Han, M. Kamber, and J. Pei, “Data mining: Concepts and,” *Techniques*, Waltham: Morgan Kaufmann Publishers, pp. 2012–13, 2012.
- [5] Oracle Corporation, “Java heap space documentation.” <https://docs.oracle.com>.
- [6] Mitchell, “‘machine learning’, new york, ny, usa: Mcgraw-hill, inc,” 1997.
- [7] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.
- [8] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [9] D. Alekseeva, N. Stepanov, A. Veprev, A. Sharapova, E. S. Lohan, and A. Ometov, “Comparison of machine learning techniques applied to traffic prediction of real wireless network,” *IEEE access*, vol. 9, pp. 159495–159514, 2021.
- [10] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.