India Road Accident Hotspot Area Prediction Using J48 Algorithm

Aina Farzana Binti Zulkifli, 2021817142
Nur Eisyatin Radhiah Binti Annuar, 2021610238
Siti Mukhlisah Binti Muskamal, 2022815898 &
Yasmin Nabila Binti Othman, 2021619506
Faculty of Computer and Mathematical Science
Universiti Teknologi MARA (UiTM)
21080 Kuala Terengganu, Terengganu, Malaysia
2021817142@student.uitm.edu.my, 2021610238@student.uitm.edu.my, 2022815898@student.uitm.edu.my &
2021619506@student.uitm.edu.my

I. Introduction

Road accidents pose a significant global concern, resulting in daily fatalities, injuries, and casualties on roads worldwide (Bahiru et al., 2023). According to the World Health Organization, road traffic injuries dominantly be the cause of death especially for children and young adults. By December 2023, approximately 1.19 million people die due to road accidents each year. Road accidents in India have become increasingly prevalent each year, posing a serious and escalating concern. The rise in the annual occurrence of India's road accident cases can be attributed to several key factors, including weather conditions and the state of the roads. Recognising the key contributing factors that elevate the risk of experiencing severe and fatal road injuries is crucial in mitigating the impact of this significant public health issue (Ghandour et al., 2020). In this study, the J48 classification technique is implemented to build a model for the specific purpose of predicting India's road accident hotspot areas.

II. Related Work

J48 is a classification algorithm that analyses a set of training examples and builds a classifier that must possess the ability to accurately categorize both training and test cases. As mentioned by (Bahiru et. al, 2018), the advantages of using the J48 classifier algorithm are it produces an accurate result with less memory required, less model-built time and has a short searching time that technically makes it faster. In the study of predicting road traffic accident severity written by (Bahiru et. al, 2018), the J48 approach has been used among other classifiers like the ID3, CART, and Naïve Bayes algorithm. Compared to the other classifiers, the J48 has been proven to have the highest accuracy with 96.30%.

(Posonia et. al, 2020) proposed a J48-based system to classify diabetes types to be gestational diabetic or non-gestational diabetic. As mentioned in the study, diabetes is now a typical disease that highly influences unpleasant side effects towards pregnant women. Using the J48 classifier algorithm, this system successfully achieved high prediction accuracy along with less processing time. Additionally, it is mentioned that the proposed work has achieved 91.2% efficiency from the experimental results.

In a study done by (Viswanath et. al, 2021), it is found that it is indeed important to forecast the number of traffic accidents for the purpose of making scientific transportation decisions. The researchers have developed an accident prediction model using the Apriori algorithm and Support Vector Machine (SVM). The result is aimed to be an initiative to prove the need for better-designed roads and vehicles. This project was successful in creating such an application that can help in the efficient prediction of road accidents based on factors such as types of vehicles, age of the driver, age of the vehicle, weather conditions and road structure and so on. This model was implemented by making use of several data mining and machine learning algorithms applied over a dataset for Bangalore and has been successfully used to predict the risk probability of accidents over different areas with high accuracy.

A study conducted regarding vehicular accidents has been done by (Kurika et. al, 2023) to identify factors of road accidents by using machine learning algorithm. A dataset containing seven years of road accidents in Ethiopia has been used and a few machine-learning algorithms have been applied which include the J48 decision tree, Random Forest tree, Naive Bayes and Bayesian network classifiers. The identification of common victims, common vehicles involved in accidents and black spot areas for accident occurrences was consequently done. The overall result shows that J48 Tree and Rep tree are two significant best models by performance compared to the other classifiers. Random forest and Random tree are poor classifiers and were not portraying good results for the given dataset.

The occurrence of road accidents has markedly escalated in conjunction with the rapid growth of the transportation sector (Paul et. al, 2022). Numerous factors contribute to the occurrence of road accidents, yet not all factors carry equal

responsibility for the severity of accidents. In the study done by (Bahiru, 2023), a few classification techniques are employed to construct the model for predicting accident severities including the J48, Support Vector Machine (SVM) and Naïve Bayes (NB). The J48 classifier has achieved a classification accuracy of 95.78%, surpassing the classification accuracy of the other classifiers. Naïve Bayes is the classifier that exhibited lower performance compared to others, but it demonstrated good efficacy in classifying fatal accident severities as opposed to serious and slight accident severities.

III. Data understanding

Data understanding is the first step in the process of analysing data. This stage requires examining and understanding the features of the dataset, which includes data on traffic accidents in India. Researchers and analysts analyse the dataset's structure, determine important characteristics, evaluate the data's quality, and obtain an understanding of the information's nature during the data understanding process. In addition to guaranteeing that the data is appropriate for analysis and setting the stage for later processes like data cleaning, preprocessing, and model development using the J48 algorithm, this fundamental step is essential for making well-informed decisions throughout the project.

The "Road Accident Severity in India" dataset is a manual compilation of records detailing road traffic accidents in the nation. It was obtained from Kaggle and covers the years 2017 to 2022. Each accident's specific details, including location, time, weather, kind of road, and vehicle information, are included in this dataset. The dataset's objective is to investigate and forecast the severity of traffic accidents, possibly using statistical or machine-learning techniques. Through the examination of this data, specific road safety interventions and measures can be formulated in India with the help of insights gained into the temporal patterns, geographical hotspots, and major factors contributing to accident severity.

Table 1. *Hotspot Areas for Road Accidents in India Classification*

Task	Algorithm	Purpose	Dataset Used
Classification	J48	Create a model to predict hotspot areas for road accidents in India	Road Accident Severity in India

In this study, we implemented a classification task using the J48 classifier algorithm. As mentioned in Table 1, we use the dataset “Road Accident Severity in India” along with the J48 classifier algorithm to create a model to predict hotspot areas for road accidents in India. We gathered both nominal and numerical data in our dataset. It has 32 attributes and 12316 instances, or data points. The data declaration for nominal and numerical data is shown in Table 2 below.

Table 2. *Numeric Data Description*

No.	Attribute Name	Description
1	Number_of_vehicles_involved	The count of vehicles involved in the accident
2	Number_of_casualties	The count of casualties in the accident

Table 3. *Nominal Data Description*

No.	Attribute Name	Description
1	Time	Indicates the time of accident occurred
2	Day_of_week	Indicates the day of accident occurred
3	Age_band_of_driver	Indicates the age group of the driver
4	Sex_of_driver	Indicates the gender of the driver
5	Educational_level	Indicates the educational level of the driver
6	Vehicle_driver_relation	Indicates the relationship between the vehicle and the driver
7	Driving_experience	Indicates number of years of driving experience of the driver
8	Type_of_vehicle	Categorization of the type of vehicle involved in the accident
9	Owner_of_vehicle	Indicates the owner of the vehicle involved in the accident
10	Service_year_of_vehicle	Indicates the number of years the vehicle has been in service
11	Defect_of_vehicle	Identification of any defects with the vehicle involved in the accident
12	Area_accident_occured	Indicates the location or area where the accident occurred
13	Lanes_or_Medians	Identification of the road configuration where the accident occurred
14	Road_allignment	The directional orientation or alignment of the road

No.	Attribute Name	Description
15	Types_of_Junction	Indicates the type of road junction where the accident occurred
16	Road_surface_type	The type of surface material of the road where the accident occurred
17	Road_surface_conditions	The conditions of the road surface where the accident occurred
18	Light_conditions	The lighting conditions at the time of the accident.
19	Weather_conditions	The weather conditions at the time of the accident
20	Type_of_collision	The collision type in which vehicles collided during the accident
21	Vehicle_movement	Indicates the movement or direction of the vehicle involved in the accident
22	Casualty_class	Indicates the role or category of individuals involved in the accident
23	Sex_of_casualty	The gender of the individual involved as a casualty in the accident
24	Age_band_of_casualty	Categorization of the age group of the individual involved as a casualty in the accident
25	Casualty_severity	Categorization of the severity of the injury sustained by the casualty in the accident
26	Work_of_casuality	The occupation or type of work of the casualty involved in the accident
27	Fitness_of_casuality	Categorization of the fitness level or condition of the casualty involved in the accident
28	Pedestrian_movement	Description of the movement or actions of the pedestrian involved in the accident
29	Cause_of_accident	The factors identified as the primary cause of the accident
30	Accident_severity	Categorization of the severity of the accident

This dataset consists of 30 nominal attributes and 2 numerical attributes. To study nominal attributes, one common approach is to examine the frequency or count of each distinct value within each attribute. This involves identifying the number of occurrences of each category and providing insights into the distribution of data across the various groups.

Studying numerical attributes involves calculating the mean (average), minimum, maximum values, and standard deviation of the numerical attribute. Table 4 shows a way to study the attribute for this dataset.

Table 4. *Ways to Study Attributes*

Nominal Attributes	Properties	Numerical Attributes
30	**Total Attributes**	2
Looking at the count of each value in each attribute.	**Way to Study**	Identify mean, min & max and its standard deviation

Based on Figure 1, we identified that the counts for both male and female drivers are 11437 and 701 respectively for nominal attributes.

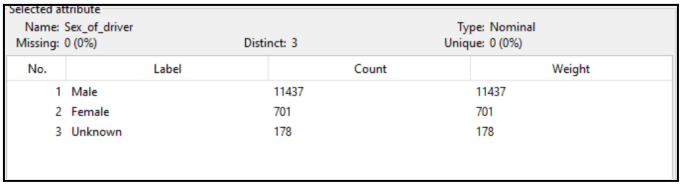


Figure 1. *Example of Nominal Attribute*

Based on Figure 2, using the attribute number_of_vehicles_involved as a numerical attribute as an example, the mean is 2.041, minimum and maximum value are 1 and 7 respectively. Lastly, the standard deviation for the attribute is 0.689.

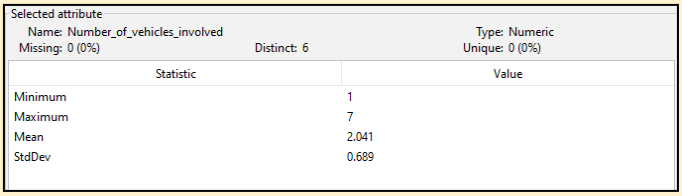


Figure 2. *Example of Numerical Attribute*

IV. DATA PREPARATION

According to the Amazon Web Services (AWS) official website, data preparation is needed to prepare the raw data so it will be suitable for further processing and analysis. In this study, data preparation has been done using WEKA and Microsoft Excel. Data preparation is performed in two parts which are data cleaning and data transformation as shown in Figure 3.

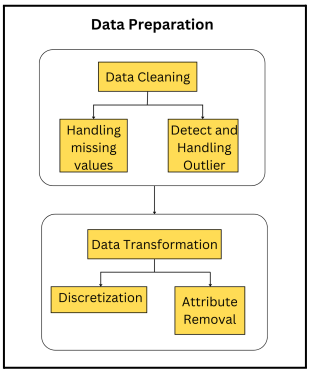


Figure 3. *Data preparation that has been done in this study*

Data cleaning is the vital process of removing incorrect, incomplete, and inaccurate data from the dataset (Bahiru et al., 2023) which includes handling missing values and addressing outliers. Managing missing values ensures dataset completeness and representativeness while detecting and handling outliers enhances model robustness by reducing the impact of anomalies.

To handle missing values, the "ReplaceMissingValues" filter in WEKA is employed. This filter is commonly used for handling missing values due to its simplicity and flexibility. It automates the process, replacing missing values in both nominal and numeric attributes with modes and means derived from the training data. This standardized approach ensures dataset consistency, making it suitable for reliable analysis and J48 model training. Table 5 shows a comparison of the "Educational_level" attribute before and after performing this filter.

Table 5. *Before and After "ReplaceMissingValues"*

BEFORE

Selected attribute
Name: Educational_level Type: Nominal
Missing: 741 (6%) Distinct: 7 Unique: 0 (0%)

No.	Label	Count	Weight
1	Above high sch...	362	362
2	Junior high sch...	7619	7619
3	Elementary sch...	2163	2163
4	High school	1110	1110
5	Unknown	100	100
6	Illiterate	45	45
7	Writing & read...	176	176

AFTER

Selected attribute
Name: Educational_level Type: Nominal
Missing: 0 (0%) Distinct: 7 Unique: 0 (0%)

No.	Label	Count	Weight
1	Above high sch...	362	362
2	Junior high sch...	8360	8360
3	Elementary sch...	2163	2163
4	High school	1110	1110
5	Unknown	100	100
6	Illiterate	45	45
7	Writing & read...	176	176

To detect and handle outliers, there are two filters performed which are "InterquartileRange" and "RemoveWithValues" respectively. By using the "InterquartileRange" filter in WEKA, the outlier can then be identified. Figure 4 and Figure 5 show the outlier attribute and extreme value that indicates the count for "Yes" which are 119 and 3976 respectively.

Selected attribute
Name: Outlier Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	no	12197	12197
2	yes	119	119

Figure 4. *Outlier Detection*

Selected attribute
Name: ExtremeValue Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	no	8340	8340
2	yes	3976	3976

Figure 5. *Extreme Value Detection*

To address outliers and extreme values, the "RemoveWithValues" filter is employed to eliminate anomalies. Before applying this filter, it is essential to identify the attribute index and nominal indices that will be considered. The attribute index specifies the column or feature in the dataset, while nominal indices identify columns containing nominal attributes. These indices are crucial for accurately applying the filter, ensuring that only the specified attributes are considered during the removal of outliers. Table 6 shows the example of RemoveWithValues setting in the outlier attribute. For outliers, the attribute index changes from last to 33 while the nominal indices change from first-last to last.

Table 6. *Comparison for RemoveWithValues Setting*

BEFORE

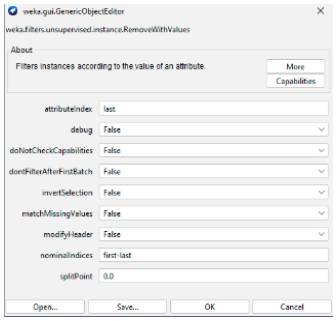


Table 6. *Continued*

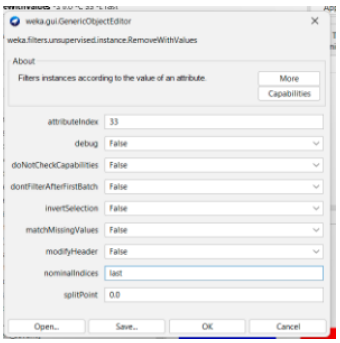
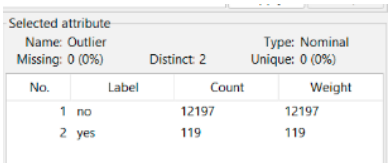
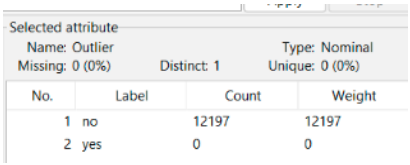
AFTER


Table 7 provides a comparison of the dataset before and after applying the "RemoveWithValues" filter. After applying this filter, the value of "Yes" is 0, which means there is no outlier.

Table 7. *Before and After "RemoveWithValues"*

BEFORE

AFTER


The Second part of the data preparation phase is data transformation which includes discretization and attribute removal. Data transformation is the process of transforming one type of data into another format that is suitable for analysis tasks (Bahiru et al., 2023). In this study, the dataset is not suitable for performing normalization. The reasons found are the values in the numeric attributes are on a similar scale and J48 algorithm is less sensitive to the scale of numeric attribute. According to the article "Towards AI", normalization is not always necessary when the values are already small and have similar ranges (Towards AI, 2019).

According to the Predictive Analytics Application with WEKA (Mutalib, 2021) e-book, discretization means the process of transforming numeric attributes to nominal. For the discretization process, only "Number_of_casualties" has been performed using the filter "discretize" in WEKA. In this process, there are some alterations made to the properties section that include specifying attribute indices and the number of bins. For example, for the mentioned attribute, the attribute indices are 22 and the number of bins is 3 (refer to Figure 6). The number of bins in discretization refers to the count of intervals or categories into which the continuous data is divided. Table 8 shows a comparison before and after applying the "discretize" filter.

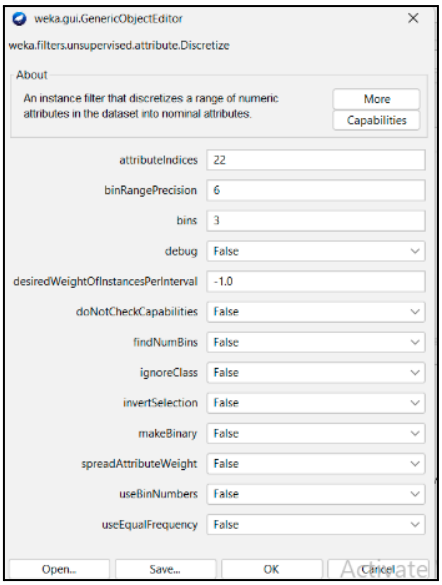
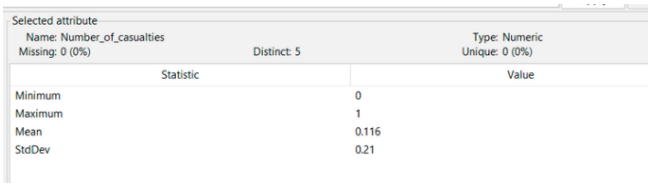
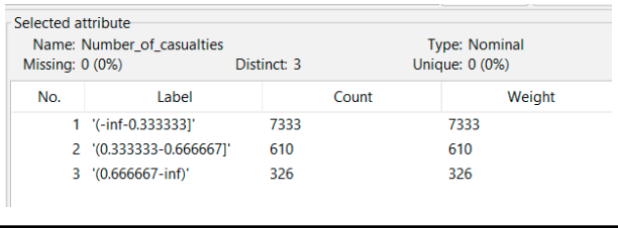


Figure 6. *Discretize Setting*

Table 8. *Before and After "Discretize"*

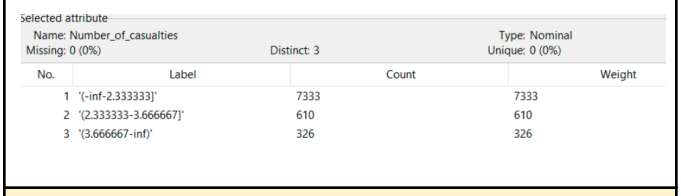
BEFORE

AFTER


Next, rename the category using the "Find and Replace All" feature in Microsoft Excel. The categories were renamed as low, intermediate and high referring to the level of casualties. Consequently, it develops a better understanding of the

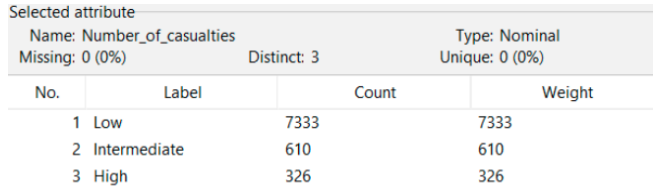
number of casualties and a clearer view of the statistic graph or chart based on the level of casualties. Table 9 shows the comparison new category name with the previous data.

Table 9. *Before and After Rename Category*

BEFORE



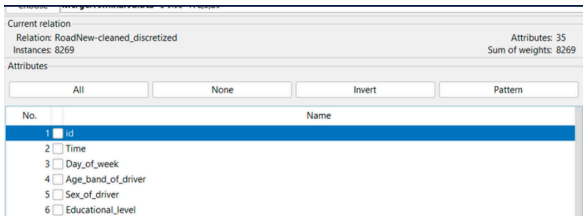
AFTER



The last process is attribute removal. Attribute removal refers to the removal of any unnecessary attributes from the dataset that do not correlate with the aim of the study. 4 attributes have been removed which are "ID", "Educational_level", "Fitness_of_casualties", and Pedestrian_movements. This process does not require any filter to be used but simply tick the desired attribute and click on the remove button at the bottom. Table 10 shows the number of attributes before and after the attribute removal process which is 35 and 31 attributes respectively.

Table 10. *Before and After Attribute Removal Process*

BEFORE



AFTER

V. MODEL ANALYSIS & EVALUATION

The process done on the dataset is classification. The classifier algorithm chosen is the J48 algorithm, also known as the decision tree algorithm. J48 can handle mixed data types with both numerical and nominal type of attributes. In addition, J48 provides decision rules, which can be used to identify combinations of factors that are high-risk conditions. In this phase, the model evaluation is done before and after the reduction of data. The analysis and evaluations done are

1. Cross-validation folds, k =10 and k = 20
2. The percentage split with the percentage of 70 and 80
3. Generate the tree visualizer and choose the tree with the best accuracy.

Cross-validation involves dividing the dataset into k folds and testing the model using a different fold for evaluation each time. Table 11 shows the accuracy results for 10 and 20 folds before reduction.

Table 11*. Cross-validation Before Reduction*

Cross-validation Model, k	Accuracy (%)
k = 10	72.7035
k = 20	72.3766

From the result shown, the 10-fold cross-validation shows higher accuracy with an accuracy percentage of 72.3766%. Next, for percentage split, the dataset is split according to a specified percentage. For example, for a percentage split of 70:30, 70% of the dataset is used to train the model, while the remainder is used to test the model. Below is Table 12, the accuracy results before the reduction for percentage split are shown

Table 12. *Percentage Split Before Reduction*

Percentage Split (%)	Accuracy (%)
70:30	71.6776
80:20	72.3856

Based on the result shown, the percentage split of 80:20 has higher accuracy, 72.3856% compared with 70:30, 71.6776%.

In Figures 7 and 8, the tree visualizer for 10-fold cross-validation and percentage split 80% is shown respectively.

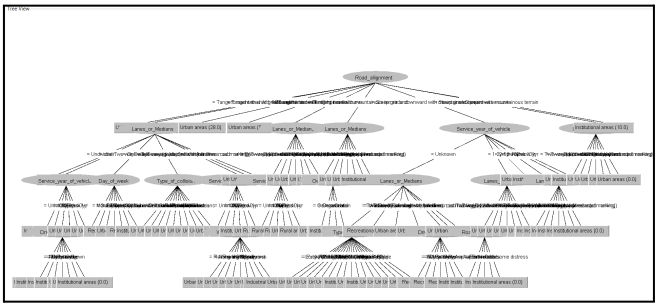


***Figure** 7. Tree Visualizer For Best Accuracy in Cross-validation Before Reduction*

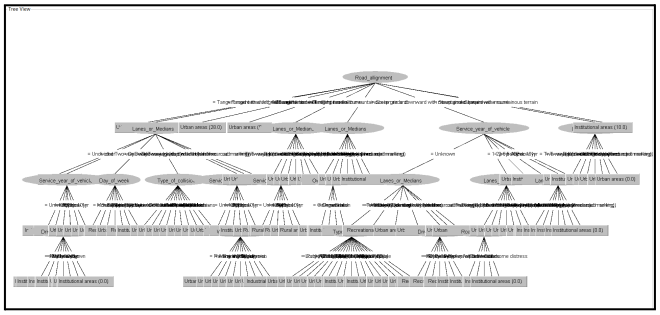


***Figure** 8. Tree Visualizer For Best Accuracy in Percentage Split Before Reduction*

The next step in this phase is the reduction of attributes. This is achieved by ranking and selecting the most important attributes in the dataset. This is important as it reduces overfitting, improves accuracy, and reduces the training time of the model. For this step, the attribute evaluator chosen is InfoGainAttributeEval which evaluates the worth of an attribute by measuring the information gain for the class. Then, the Ranker method would be automatically chosen. The number of attributes selected has to be set in the Ranker properties at the numToSelect field, which is set to 15 and 20. Figure 9 shows the 15 ranked attributes.

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 29 Area_accident_occured):
        Information Gain Ranking Filter

Ranked attributes:
 0.47553     1 Time
 0.0702     12 Road_allignment
 0.04896     9 Service_year_of_vehicle
 0.03467    11 Lanes_or_Medians
 0.01821    27 Cause_of_accident
 0.01599     8 Owner_of_vehicle
 0.0143     21 Vehicle_movement
 0.01381    18 Type_of_collision
 0.0136      7 Type_of_vehicle
 0.01266    13 Types_of_Junction
 0.00781     2 Day_of_week
 0.00694    14 Road_surface_type
 0.00663    17 Weather_conditions
 0.00526    26 Work_of_casuality
 0.0047      6 Driving_experience

Selected attributes: 1,12,9,11,27,8,21,18,7,13,2,14,17,26,6 : 15
```

***Figure 9**. List of 15 Ranked Attributes*

Figure 10 shows the list of 20 ranked attributes.

```
Attribute selection output
 0.034673   11 Lanes_or_Medians
 0.018207   27 Cause_of_accident
 0.015992    8 Owner_of_vehicle
 0.014303   21 Vehicle_movement
 0.013808   18 Type_of_collision
 0.013599    7 Type_of_vehicle
 0.012659   13 Types_of_Junction
 0.00781     2 Day_of_week
 0.006945   14 Road_surface_type
 0.00663    17 Weather_conditions
 0.005258   26 Work_of_casuality
 0.004699    6 Driving_experience
 0.004171   24 Age_band_of_casualty
 0.004012    5 Vehicle_driver_relation
 0.003125    4 Sex_of_driver
 0.00303    28 Accident_severity
 0.002252   10 Defect_of_vehicle
 0.00181    15 Road_surface_conditions
 0.001732    3 Age_band_of_driver
 0.001651   22 Casualty_class
 0.001588   16 Light_conditions
 0.001105   23 Sex_of_casualty
 0.000892   20 Number_of_casualties
 0          25 Casualty_severity
 0          19 Number_of_vehicles_involved

Selected attributes: 1,12,9,11,27,8,21,18,7,13,2,14,17,26,6,24,5,4,28,10,15,3,22,16,23,20,25,19 : 28
```

***Figure** 10. List of 20 Ranked Attributes.*

After the attributes have been ranked, the reduced data is saved. In the reduced dataset, only the 15 and 20 ranked attributes and the class attributes are saved. The evaluation was done again using both of the reduced datasets. Table 13 shows the accuracy of cross-validation and percentage split for the 15 ranked attributes.

***Table 13**. Accuracy Results For 15 Ranked Attributes*

Evaluation	Accuracy (%)
10-fold Cross-Validation	72.54
20-fold Cross-Validation	72.5074
Percentage Split of 70%	72.3312
Percentage Split of 80%	72.7124

Then, the same evaluations are done to the 20 ranked attributes. Table 14 shows the accuracy results.

Table 14. *Accuracy Results For 20 Ranked Attributes*

Evaluation	Accuracy (%)
10-fold Cross-Validation	72.5727
20-fold Cross-Validation	72.5074
Percentage Split of 70%	72.4401
Percentage Split of 80%	72.8758

Then, a tree visualizer was generated for the best results for each validation. Figures 11 and 12 below show the tree visualizer generated for 15 ranked attributes, both validation respectively. For cross-validation, the 10-fold shows higher accuracy compared with the 20-fold. The percentage split 80% shows higher accuracy compared with 70%. This is the same case for the 20 ranked attributes.

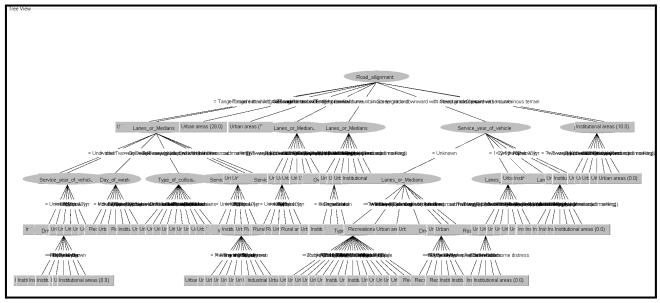


Figure *11. Tree Visualizer For Best Accuracy For 15 Ranked Attributes in Cross-validation*

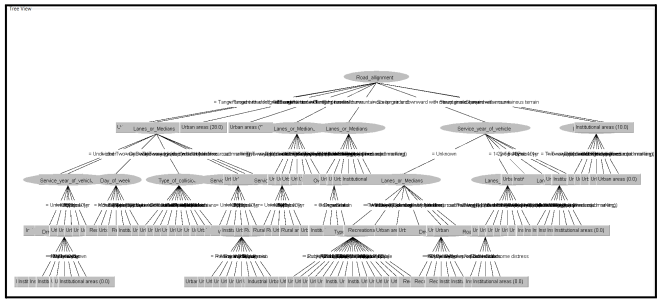


Figure *12. Tree Visualizer For Best Accuracy For 15 Ranked Attributes in Percentage Split*

Next figures 13 and 14 below show the tree visualizer generated for 20 ranked attributes, both validation respectively with the best accuracy.

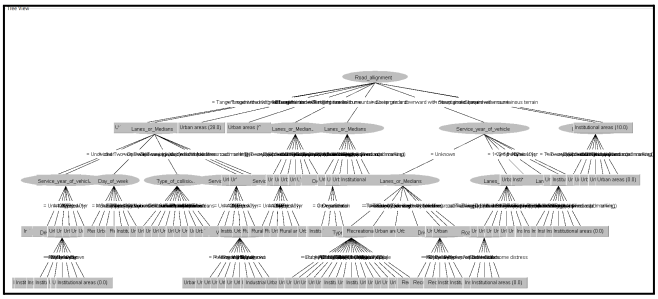


Figure *13. Tree Visualizer For Best Accuracy For 20 Ranked Attributes in Cross-validation*

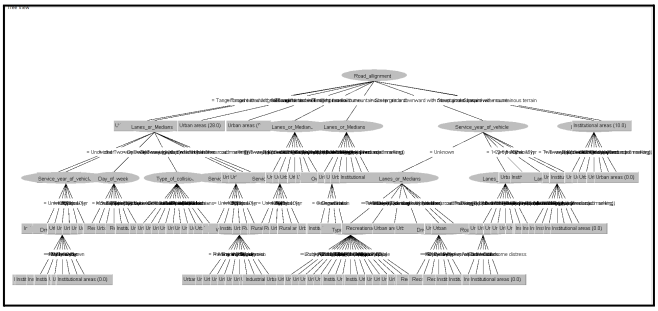


Figure *14. Tree Visualizer For Best Accuracy For 20 Ranked Attributes in Percentage Split*

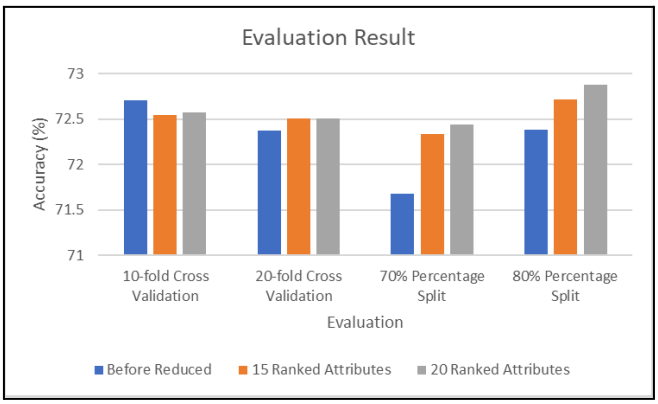


Figure *15. Bar Graph Evaluation Result*

Based on the figure above, it is shown that the accuracy result for 10-fold cross-validation, before the data was reduced, shows the highest accuracy. However, for other evaluations, the accuracy for the before-reduced dataset was the lowest. In addition, the accuracy for the dataset after reducing the attributes to 20 was similar with the attributes reduced to 10, and the highest for the other two evaluations. This shows that the 20 attributes are the most suitable for modelling. Other than that, the figure also shows for percentage split, the accuracy increases when the number of attributes is reduced.

I. CONCLUSION

In conclusion, this study harnesses the power of machine-learning methodologies to conduct a comprehensive analysis and prediction of road accident hotspot areas in India. Through the implementation of the J48 classifier algorithm, it

successfully identified and forecasted regions with elevated instances of road accidents, showcasing the algorithm's remarkable accuracy. Additionally, it is found that the accuracy becomes even higher after the dataset has been reduced to only containing 20 attributes. The findings underscore the efficacy of employing advanced data analytics tools, specifically the J48 classifier, in enhancing the understanding of road safety dynamics, thereby contributing valuable insights for informed decision-making and targeted interventions to mitigate the frequency of accidents in identified high-risk areas.

References

Bahiru, T. K., Manjula, V. S., Akele, T. B., Tesfaw, E. A., & Belay, T. D. (2023). Mining Road Traffic Accident Data for Prediction of Accident Severity. 2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), 606–612. https://doi.org/10.1109/IDCIoT56793.2023.10053409

Bahiru, T. K., Manjula, V. S., Akele, T. B., Tesfaw, E. A., & Belay, T. D. (2023). Mining Road Traffic Accident Data for Prediction of Accident Severity. *IDCIoT 2023 - International Conference on Intelligent Data Communication Technologies and Internet of Things, Proceedings*, 606–612. https://doi.org/10.1109/IDCIoT56793.2023.10053409

Ghandour, A. J., Hammoud, H., & Al-Hajj, S. (2020). Analyzing factors associated with fatal road crashes: A machine learning approach. International Journal of Environmental Research and Public Health, 17(11). https://doi.org/10.3390/ijerph17114111

How, When, and Why Should You Normalize / Standardize / Rescale… – Towards AI. (2019). TOWARD AI. https://towardsai.net/p/data-science/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff

Kurika, A. E., Ganie, I. A., Kadir, Y., Cerna, P. D., & Desei, F. L. (2020). Predicting Factors of Vehicular Accidents using Machine Learning Algorithm. International Journal of Emerging Trends in Engineering Research, 8(9), 5171–5176. https://doi.org/10.30534/ijeter/2020/46892020

Mutalib, S. A. R. &. (2021). Predictive Analytics Applications with WEKA. https://play.google.com/store/books/details?id=cG9JEAAAQBAJ

Paul, A. K., Boni, P. K., & Islam, M. Z. (2022). A Data-Driven Study to Investigate the Causes of Severity of Road Accidents. 2022 13th International Conference on Computing Communication and Networki

Posonia, A. M., Vigneshwari, S., & Rani, D. J. (2020). Machine learning based diabetes prediction using decision tree J48. Proceedings of the 3rd International Conference on Intelligent Sustainable Systems, ICISS 2020, 498–502. https://doi.org/10.1109/ICISS49785.2020.9316001

Viswanath, D., Preethi, K., Nandini, R., & Bhuvaneshwari, R. (2021). A Road Accident Prediction Model Using Data Mining Techniques. Proceedings - 5th International Conference on Computing Methodologies and Communication, ICCMC 2021, 1618–1623. https://doi.org/10.1109/ICCMC51019.2021.9418336

What is Data Preparation - Data Preparation Explained - AWS. (n.d.). Amazon Web Services. Retrieved January 11, 2024, from https://aws.amazon.com/what-is/data-preparation//