



DRAGANA RADOSAVLJEVIĆ

University of Priština
in Kosovska Mitrovica, Serbia

ORCID iD: 0000-0003-0243-2213

ALEKSANDRA PETROVIC

University of Priština
in Kosovska Mitrovica, Serbia

ORCID iD: 0000-0002-8062-9751

VLADIMIR MAKSIMOVIĆ

University of Priština
in Kosovska Mitrovica, Serbia

ORCID iD: 0000-0001-8479-0027

IWONA FLOREK

WSGE University of Applied Science
in Józefów, Poland

ORCID iD: 0000-0003-0194-3361

TRAFFIC SAFETY, SUSTAINABLE MOBILITY AND HUMAN RIGHTS: DATA-DRIVEN ANALYSIS OF TRAFFIC ACCIDENTS IN SERBIA

ABSTRACT

Goal: *This paper presents the application of data mining methods in the analysis of traffic accidents in the Republic of Serbia during the period 2022–2025, based on the open dataset available through the Open Data Portal of the Republic of Serbia. The aim of the study is to examine the extent to which spatial, temporal, and structural attributes enable the differentiation between accidents involving material damage only and accidents involving injured or deceased persons.*

Methods: *The analysis was conducted in the Weka environment using the J48, Random Forest, Naive Bayes, kNN, and Logistic algorithms, with evaluation based on Accuracy, Precision, Recall, F1-score, and ROC AUC measures.*

Results: *The results indicate that the Logistic algorithm achieved the best overall performance, while Naive Bayes demonstrated the highest recall for the class of accidents involving injured or deceased persons.*

Discussion: *The results indicate that the Logistic algorithm achieved the best overall performance, while Naive Bayes demonstrated the highest recall for the class of accidents involving injured or deceased persons.*

KEYWORDS: *data mining; traffic accidents; classification; Weka; sustainable mobility; public safety; human rights; Serbia.*

INTRODUCTION

Traffic accidents represent one of the major public risks of contemporary society, as their consequences affect not only property but also directly influence the lives, health, safety, and everyday mobility of citizens. According to the World Health Organization, approximately 1.19 million people lose their lives annually in road traffic, while traffic-related injuries are the leading cause of death among children and young people aged 5 to 29 years (World Health Organization, 2026). Therefore, road safety cannot be viewed solely as a technical issue of traffic management, but also as an important issue of public health, quality of life, and the protection of fundamental human rights. Road safety constitutes one of the fundamental areas of social life, enabling people to engage in everyday mobility and, consequently, to pursue development both individually and collectively. The development of motorisation, the use of modern technologies, and the increasing number of vehicles on the roads (Smith, 2016), as well as—equally importantly—changes in population density worldwide, bring not only obvious and desirable benefits

but also significant social and material losses. The importance and continuing relevance of this issue were highlighted by the World Health Organization in 2013 (World Health Organization, 2013, p. 3). The problem has also been recognised by the European Union. The European Commission's strategic action plan, *EU Road Safety Policy Framework 2021–2030*, sets the objective of reducing the number of road fatalities by 50%, as well as significantly reducing the number of serious injuries suffered by road users (Urban, Jurczak, Płocki, 2024, 124). This approach is consistent with Sustainable Development Goal 11.2, which aims to provide safe, affordable, accessible, and sustainable transport systems for all, while improving road safety (United Nations, 2015). In the context of the green transition, transport is not viewed only through the reduction of emissions and the implementation of cleaner technologies, but also through the safety, accessibility, and resilience of the system. A transport system that is not safe cannot be considered fully sustainable, even when it satisfies ecological or technological criteria.

The development of open data enables traffic accidents to be analyzed using data mining methods. Such methods allow the identification of patterns that are not always visible through conventional descriptive indicators. The subject of this paper is the application of data mining methods to an open dataset on traffic accidents in the Republic of Serbia, with particular emphasis on distinguishing accidents involving material damage from accidents involving injured or deceased persons.

Accordingly, the paper addresses three research questions:

- To what extent do available spatial, temporal, and structural attributes enable the differentiation between accidents involving material damage and accidents involving injured or deceased persons?
- Which of the applied classification algorithms achieves the best aggregated performance, and which most effectively recognizes the class of accidents involving human casualties?
- What is the influence of the attribute *accident_description* on classification results, and how does its post-hoc nature affect the interpretation of the obtained results?

The main contribution of the paper lies in connecting open data, data mining methods, and the broader social framework of road safety. The paper does not aim to establish causal relationships, nor to develop an operational system for early risk prediction before the occurrence of an accident. Since some available attributes describe an event that has already occurred, the results should be interpreted as a post-hoc classification and pattern analysis of recorded accidents.

THEORETICAL FRAMEWORK AND RELATED RESEARCH

In this paper, road safety is considered a public and public-health risk with consequences for the lives, health, safety, and everyday mobility of citizens. The consequences of traffic accidents may be associated with the right to life, the right to health, and the right to safe movement. This does not mean that every traffic accident in itself constitutes a violation of human rights, but rather that a high level of traffic risk may seriously affect the realization of these rights in everyday life.

Sustainable mobility encompasses more than the environmental dimension of transport. It implies that the transport system should be accessible, efficient, safe, and resilient. The European Green Deal and the Sustainable and Smart Mobility Strategy of the European Union identify transport as one of the key areas of the green and digital transition (European Commission, 2019; European Commission, 2020). Within this framework, the analysis of traffic accidents is important because it helps position safety as an integral part of a sustainable transport system.

Data mining methods make it possible to simultaneously examine relationships among multiple attributes. Unlike conventional statistical records, which most often describe the number of accidents and their distribution across categories, classification algorithms can assess the extent to which combinations of spatial, temporal, and structural characteristics contribute to distinguishing the severity of consequences. Attribute ranking, on the other hand, enables an assessment of which features carry the most information for separating the observed classes.

Related studies show that decision trees, Random Forest, Naive Bayes, kNN, Logistic Regression, and other models can be applied in classifying traffic accident severity and identifying factors associated with more severe outcomes (Adeliyi et al., 2023; Yan, Shen, 2022; Obasi, Benson, 2023). It is particularly important to note that the model with the highest overall accuracy is not necessarily the best at recognizing socially sensitive outcomes. Therefore, the results in this paper are interpreted not only through overall accuracy but also through recall for the class of accidents involving injured or deceased persons.

Methodologically, it is important to distinguish post-hoc analysis from operational prediction. If certain attributes describe events that have already occurred, classification results should be interpreted as an analysis of recorded patterns rather than as a model applicable before the occurrence of an accident. This observation is particularly relevant for the attribute *accident_description*, as it carries substantial informational value while potentially containing information indirectly related to the accident outcome.

DATASET AND DATA PREPARATION

The study used the open dataset *Data on Traffic Accidents by Police Administrations and Municipalities*, obtained from the Open Data Portal of the Republic of Serbia (Open Data Portal of the Republic of Serbia, n.d.). The data are organized in tabular form and contain basic information on traffic accidents: police administration, municipality, date and time of the accident, geolocation of the accident site, type of accident according to consequences, accident type, and accident description. A total of 132,118 instances for the period 2022–2025 were analyzed.

The number of instances by year is relatively balanced: 33,399 accidents were recorded in 2022, 32,974 in 2023, 32,330 in 2024, and 33,415 in 2025. Such a distribution allows the data to be treated as a joint multi-year dataset without pronounced dominance of a single year. The data were used as secondary administrative data without personal identification, which is important from the perspective of ethical acceptability of the analysis.

During data preparation, derived attributes *year*, *month*, *day_of_week*, and *part_of_day* were extracted from the field containing the date and time

of the accident. The *part_of_day* attribute was created for easier interpretation of temporal patterns and grouped into the categories *night*, *morning*, *day*, and *evening*. A particularly important step was the definition of the binary target variable *severity*, which distinguishes accidents involving material damage from accidents involving injured or deceased persons.

Table 1. Transformation rule from the original consequence category into the target severity class

Original Value	New Severity Class
Material damage only	material_damage
With injured persons	injured_or_deceased
With fatalities	injured_or_deceased

Source: own elaboration.

Table 2. Distribution of accidents according to the target severity class

Class	Number of Accidents	Percentage
Material damage only	78,115	59.13%
injured_or_deceased	54,003	40.87%
Total	132,118	100.00%

Source: own elaboration.

The original column from which the target class was derived was not used as an input attribute, as this would lead to data leakage. Data leakage occurs when information directly or indirectly associated with the target variable is introduced into the model and would not legitimately be available in real-world application of the model (Kaufman et al., 2012). Therefore, the variable *original_accident_type* was excluded from the set of input attributes.

Special attention was paid to the attribute *accident_description*. In this study, it was treated as a nominal, that is, categorical attribute rather than a free-text attribute. In the prepared dataset, this attribute contains 68 different categories with no missing values. Therefore, no textual transformation such as *StringToWordVector* was applied in the Weka environment. Instead, the attribute was used as a predefined categorical label describing a typical traffic accident description.

Table 3. *Groups of input attributes used in the analysis*

Attribute Group	Attributes Used	Role in the Analysis
Spatial	police_administration, municipality	description of the territorial context of the accident
Temporal	month, day_of_week, part_of_day	examination of seasonal, daily, and weekly patterns
Structural	accident_type, accident_description	description of the characteristics of the recorded event
Target class	severity	binary differentiation of consequence severity

Source: own elaboration.

RESEARCH METHODS

The methodological procedure included descriptive data analysis, target class formation, preparation of input attributes, application of classification algorithms, and evaluation of the obtained models. This approach is consistent with the general principles of the data mining process, according to which modeling is not viewed as an isolated step but as part of a broader procedure involving problem understanding, data understanding and preparation, method selection, evaluation, and interpretation of results (Shearer, 2000).

Classification was performed in the Weka environment, version 3.8.6, on the same prepared ARFF dataset (Hall et al., 2009). The applied algorithms were J48, Random Forest, Naive Bayes, kNN/IBk, and Logistic. These algorithms were selected because they represent different classification approaches: decision trees, ensemble methods, probabilistic models, nearest-neighbor methods, and linear statistical models.

Table 4. *Classification algorithms used and basic settings*

Algorithm	Approach	Basic Settings in the Weka Environment
J48	decision tree	pruned tree; confidenceFactor = 0.25; minNumObj = 2
Random Forest	ensemble of trees	numIterations = 100; seed = 1; maxDepth = 0
Naive Bayes	probabilistic model	default settings, without kernel estimator
kNN/IBk	nearest-neighbor method	k = 5
Logistic	logistic regression	ridge = 1.0E-8; maxIts = -1

Source: own elaboration.

Model evaluation was performed using 10-fold cross-validation. The performance measures used included Accuracy, Weighted Precision, Weighted Recall, Weighted F1-score, ROC AUC, and the confusion matrix. Accuracy

indicates the overall proportion of correctly classified records, but it is not sufficient when one type of error has greater practical significance than another. Therefore, special attention was paid to Recall for the class *injured_or_deceased*, since it shows the proportion of actual accidents involving human casualties that the model successfully recognizes.

In addition to the main experiment, a control experiment was conducted without the *accident_description* attribute. The objective was not to declare this attribute methodologically unacceptable, but to assess the extent to which the results depend on a post-hoc descriptive attribute and the extent to which classification can rely on the remaining spatial, temporal, and structural attributes.

RESULTS AND DISCUSSION

Descriptive analysis shows that the majority of traffic accidents belong to the class *material_damage*, while the class *injured_or_deceased* accounts for 40.87% of all records. This indicates that the analyzed problem is not merely material in nature but also has a clear safety, health, and social dimension. The number of accidents by month is not entirely evenly distributed: the highest numbers were recorded in October, December, September, and July, while the lowest numbers were recorded in February and January. However, these indicators should be interpreted as descriptive rather than causal.

Table 5 presents the results obtained using the J48, Random Forest, Naive Bayes, kNN, and Logistic algorithms. The displayed Precision, Recall, and F1-score values refer to weighted average measures from the Weka environment.

Table 5. *Comparison of classification algorithm results*

Algorithm	Accuracy	Weighted Precision	Weighted Recall	Weighted F1-score	ROC AUC
J48	0.747	0.744	0.747	0.743	0.807
Random Forest	0.711	0.709	0.711	0.710	0.788
Naive Bayes	0.743	0.744	0.743	0.743	0.823
kNN	0.730	0.728	0.730	0.725	0.805
Logistic	0.749	0.740	0.749	0.747	0.830

Source: own elaboration.

According to aggregated measures, Logistic achieved the best result, with Accuracy = 0.749, Weighted F1-score = 0.747, and ROC AUC = 0.830.

This indicates that Logistic is the most successful model in terms of overall classification performance. J48 achieved very similar overall accuracy and has the advantage of interpretability, while Naive Bayes demonstrated very good class separation capability. In this experiment, Random Forest did not achieve the best results, indicating that the success of ensemble methods depends on the structure of the specific dataset.

Table 6. *Model recall by class and number of misclassified severe accidents*

Algorithm	Recall for material damage	Recall for injured_or_deceased	Severe accidents incorrectly classified as material damage
J48	0.829	0.627	20,152
Random Forest	0.767	0.630	19,995
Naive Bayes	0.771	0.702	16,101
kNN	0.830	0.586	22,348
Logistic	0.817	0.651	18,871

Source: own elaboration.

Table 6 shows that Naive Bayes achieved the highest recall for the class *injured_or_deceased*, with a value of 0.702. This means that the model correctly recognized approximately 70.2% of accidents involving injured or deceased persons. From the perspective of public safety, this result is particularly important because failing to identify severe cases is practically more sensitive than incorrectly classifying some less severe accidents as severe. Therefore, selecting a model solely based on overall accuracy is not justified.

Table 7. *Attribute ranking using the Information Gain measure*

Rank	Attribute	Information Gain
1	accident_description	0.224241
2	accident_type	0.160628
3	municipality	0.068023
4	police_administration	0.057372
5	part_of_day	0.002846
6	month	0.002639
7	day_of_week	0.000022

Source: own elaboration.

The analysis of attribute importance shows that *accident_description* contributes the most to class separation, followed by *accident_type*, *municipality*, and *police_administration*. This finding should not be interpreted as evidence that accident description causes more severe consequences, but rather as

an indication that the categories of this attribute contain information associated with the already recorded outcome. Temporal attributes have significantly lower individual Information Gain values, which does not mean that time is unimportant for accident frequency, but rather that within this model it has limited informational value for distinguishing severity levels.

Table 8. *Classification results without the attribute `accident_description`*

Algorithm	Accuracy	Weighted Precision	Weighted Recall	Weighted F1-score	ROC AUC
J48	0.726	0.723	0.726	0.722	0.778
Random Forest	0.688	0.684	0.688	0.685	0.743
Naive Bayes	0.716	0.719	0.716	0.717	0.778
kNN	0.703	0.699	0.703	0.697	0.757
Logistic	0.724	0.721	0.724	0.720	0.794

Source: own elaboration.

The control experiment confirms that excluding the attribute *accident_description* reduces the performance of all models. Overall accuracy decreased by approximately 0.021 to 0.027, while ROC AUC values were also lower. Nevertheless, even without this attribute, the models retained usable classification capability. This indicates that classification is not based exclusively on a single attribute, but that the remaining spatial, temporal, and structural attributes also contain information necessary for distinguishing the observed classes.

Therefore, it is not justified to speak of a single absolutely best model. Logistic is the most successful according to overall measures, J48 is suitable when interpretability is important, and Naive Bayes has an advantage when the priority is recognizing accidents involving injured or deceased persons. This finding is important for the analysis of public risks because model selection should depend on the objective of the analysis rather than on a single aggregated measure.

The obtained results have practical value because they demonstrate that open data can be used not only for statistical reporting but also for analytical pattern recognition. From the perspective of sustainable mobility, the results confirm that safety must be an integral part of transport system planning. From the perspective of human rights, the value of such analysis lies not in allowing algorithms to make decisions independently, but in helping institutions identify patterns that deserve additional attention. Such analytical insights are most useful when combined with the operational preparedness of the services that

respond to road incidents, as illustrated by simulator-based training of police officers for their duties at the scene of a road accident (Pellowski W., Lubiewski P., Nepelski M., Klonowska I., Kuczyńska E., 2021).

RESEARCH LIMITATIONS AND DIRECTIONS FOR FUTURE WORK

The main limitation of the study arises from the structure of the available data. The dataset does not include important factors such as detailed weather conditions, road surface conditions, road type, speed limits, traffic intensity, driver age and experience, seat belt use, alcohol presence, or other behavioral factors. Consequently, the models cannot explain all causes of severe outcomes, but only identify patterns based on available attributes.

A second limitation concerns spatial resolution. The attributes *municipality* and *police_administration* allow analysis of the broader territorial context but do not enable precise identification of specific dangerous intersections, road segments, or micro-locations. Such analysis would require combining this dataset with more detailed geospatial data, road infrastructure data, and traffic intensity information.

A third limitation concerns the nature of the attribute *accident_description*. Although informative, this attribute describes an event that has already occurred, and therefore the results should be interpreted as a post-hoc analysis of recorded accidents. Future research should distinguish between two directions of analysis: first, post-hoc analysis using complete accident descriptions to understand patterns, and second, predictive models using only attributes available before the occurrence of an accident.

Future work may focus on including additional data sources, applying spatial analyses of high-risk locations, comparing multiple class-balancing strategies, and examining models that allow clearer explanations of individual predictions. It would also be useful to conduct specific analyses for vulnerable road users such as pedestrians, children, older persons, and persons with disabilities, because unsafe transport systems may more strongly affect the realization of their everyday rights.

CONCLUSIONS

This paper presented the application of data mining methods in the analysis of traffic accidents in the Republic of Serbia during the period 2022–2025. Based on the open dataset, a binary target class *severity* was formed, distinguishing accidents involving material damage from accidents involving injured or deceased persons. The J48, Random Forest, Naive Bayes, kNN, and Logistic algorithms were applied in the Weka environment.

The results show that Logistic achieved the best aggregated performance, while Naive Bayes achieved the highest recall for the class *injured_or_deceased*. This finding is particularly significant in the context of public safety, since traffic risk analysis should not focus only on overall model accuracy, but also on the ability to recognize accidents involving consequences for life and health.

The analysis of attribute importance showed that *accident_description*, *accident_type*, *municipality*, and *police_administration* contribute the most to class separation. The control experiment without the attribute *accident_description* confirmed that this post-hoc attribute significantly contributes to classification, while the models still retain usable classification capability based on the remaining attributes. Therefore, the results should be interpreted as an analysis of patterns in recorded accidents rather than as an early risk prediction system.

The scientific and practical contribution of the paper lies in connecting open data, data mining methods, and the broader framework of road safety, sustainable mobility, and the protection of fundamental human rights. The paper demonstrates that publicly available data can serve as a basis for analytical support in planning measures contributing to a safer and more sustainable transport system.

ACKNOWLEDGEMENT

This research paper has been supported by Erasmus+ Capacity Building in Higher Education project 101179358-Res-Uni-ERASMUS-EDU-2024-CBHE Research Universities (Res-Uni)

REFERENCES

- Adeliyi, T. T., Oluwadele, D., Igwe, K., Aroba, O. J. (2023). Analysis of Road Traffic Accidents Severity Using a Pruned Tree-Based Model, 7(2), 131–138. *International Journal of Transport Development and Integration*. <https://doi.org/10.18280/ijtidi.070208>
- European Commission. (2019). *The European Green Deal*. COM(2019) 640 final. Brussels.
- European Commission. (2020). *Sustainable and Smart Mobility Strategy – putting European transport on track for the future*. COM(2020) 789 final. Brussels.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H. (2009). The WEKA data mining software: an update, 11(1), 10–18. *SIGKDD Explorations*.
- Kaufman, S., Rosset, S., Perlich, C., Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance, 6(4), 1–21. *ACM Transactions on Knowledge Discovery from Data*.
- Obasi, I. C., Benson, C. (2023). Evaluating the effectiveness of machine learning techniques in forecasting the severity of traffic accidents, 9(8), e18812. *Heliyon*. <https://doi.org/10.1016/j.heliyon.2023.e18812>
- Open Data Portal of the Republic of Serbia. (n.d.). *Data on Traffic Accidents by Police Administrations and Municipalities*. Retrieved 17 May 2026 from: <https://data.gov.rs/sr/datasets/podatsi-o-saobratshajnim-nezgodama-po-politsijskim-upravama-i-opshtinama/>
- Pellowski W., Lubiewski P., Nepelski M., Klonowska I., Kuczyńska E. (2021). The Simulator for Training Policemen to Perform their Duties at the Scene of a Road Accident, *European Research Studies Journal* Volume XXIV Issue 4 – Part 1, 763–780
- Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining, 5(4), 13–22. *Journal of Data Warehousing*.
- United Nations. (2015). *Transforming our world: the 2030 Agenda for Sustainable Development*. A/RES/70/1. New York.
- Urban, A. C., Jurczak, J. & Płocki, R. (2024). Bezpieczeństwo pieszych na przejściach dla pieszych. *Journal of Modern Science*, 58(4), 122–137. <https://doi.org/10.13166/jms/192194>
- World Health Organization. (2013). *Pedestrian safety. A road safety manual for decision-makers and practitioners*.
- World Health Organization. (2026). *Road traffic injuries*. Fact sheet, 1 May 2026. Retrieved from: <https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries>
- Yan, M., Shen, Y. (2022). Traffic Accident Severity Prediction Based on Random Forest, 14(3), 1729. *Sustainability*. <https://doi.org/10.3390/su14031729>