# IDENTIFICATION OF KEY RISK FACTORS FOR TRAFFIC ACCIDENTS USING MACHINE LEARNING

**Riste Ristov**
MSc, Teaching Assistant
Ss. Cyril and Methodius University in Skopje
Faculty of Civil Engineering Skopje
North Macedonia
ristov@gf.ukim.edu.mk
0000-0002-4996-0382

***Received: 27.11.2025***
***Accepted: 05.02.2026***
***Published: 06.02.2026***

***DOI: xxxx***

Road safety represents a significant challenge for the transport sector due to the severe consequences of traffic accidents and the need for timely identification of factors that increase risk. This study aims to quantify the influence of selected parameters on the weighted accident index and to establish a foundation for predictive models capable of identifying high-risk road sections. The analysis covers 161 sections of the national road network in the Republic of North Macedonia, with a total length of approximately 1,300 km, and includes 23 parameters. The assessment was performed using machine learning techniques, with model evaluation conducted on an 80/20 train–test split. The results reveal that road characteristics and traffic volume (AADT) exert the greatest influence on accident risk, whereas environmental factors have minimal impact. This approach enables more efficient planning of interventions and contributes to the overall improvement of road safety.

Keywords: accidents, machine learning, roads, safety, weighted index

## 1. INTRODUCTION

Traffic accidents represent one of the most serious global public-safety challenges, particularly among younger populations. According to the World Health Organization [1], road traffic injuries are the leading cause of death for individuals aged 5 to 29 years, claiming more than 1.19 million lives annually. Although developed countries have achieved considerable progress through systematic measures and infrastructure improvements, the average fatality rate in the European Union remains 45.5 deaths per million inhabitants [2].

In North Macedonia the situation is particularly concerning, with a rate of 69.5 fatalities per

million inhabitants [3], significantly above the European average. These figures underscore the urgent need for a proactive approach to road safety that does not rely solely on historical accident records but enables the identification and prediction of high-risk segments, allowing timely preventive interventions.

Modern analytical tools, such as machine learning and geospatial analysis, facilitate the detection of high-risk sections even before accidents occur. This proactive strategy substantially enhances the ability to plan targeted safety measures and manage risk across the road network more effectively.

The objective of this research is to determine the most influential factors affecting the weighted accident index (Wi) and to establish a foundation for predictive models that will support more efficient prioritization of interventions and national-level risk management.

## 2. LITERATURE REVIEW

The application of machine learning in the analysis of traffic accident risk factors has advanced significantly in recent years.

Wang [4] analyzed rural road accidents using Random Forest and SHAP values and identified road geometry and pavement condition (horizontal curves, longitudinal gradients, friction coefficient, sight distance, and signage) as dominant factors, parameters that strongly overlap with those in the present study.

Jiang [5] combined urban and national databases and confirmed that weather conditions and lighting have noticeable but clearly weaker influence than road characteristics and traffic volume (AADT).

Çelik [6] compared multiple techniques on large accident datasets and demonstrated that tree-based methods, including XGBoost, provide high accuracy and efficiency, supporting the models selected for this research.

These studies are fully consistent with the current findings: road geometry, pavement condition, and traffic volume are the primary risk drivers, while environmental factors play only a secondary role.

## 3. DATA AND MATERIALS

### 3.1 GENERAL INFORMATION ON THE ROAD NETWORK

The road network of the Republic of North Macedonia has a total length of 14,475 km and is classified into national, regional and local roads [7]. The national road network has a length of 897 km and represents a key segment of both national and trans-European transport infrastructure [8]. It comprises motorways, expressways, and two-lane roads that provide the main traffic connections within the country and with neighboring states.

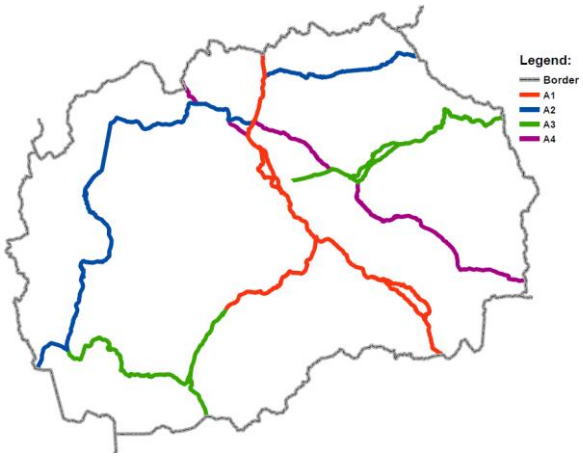


Figure 1. Overview map of Category A roads

The present research focuses on the national roads A1, A2, A3, and A4, which differ in technical characteristics and geometric elements. Although the official length of the national road network is 897 km, the analysis covers approximately 1,300 km due to the separate treatment of each carriageway on roads with divided lanes (motorways).

### 3.2 DESCRIPTION AND PROCESSING OF THE DATA

The data were processed using GIS tools, statistical techniques, and machine-learning methods, enabling the identification of spatial and temporal trends.

- Road characteristics. This category includes various geometric and functional parameters such as speed limits, horizontal curvature, curve radii, longitudinal gradients, and elevation. Additionally, lateral forces in curves, stopping sight distance, pavement evenness, rut depth, surface friction coefficient, and pavement condition index are analyzed. Data on the density of junctions, bridges, and viaducts, as well as

the condition of vertical and horizontal signage, are also incorporated [9][10].

- Traffic characteristics. Traffic intensity is expressed through annual average daily traffic (AADT), based on fixed and mobile automatic counts. This parameter provides insight into the influence of traffic volume on risk [11].

- Environmental characteristics. Climatic factors are represented by average and extreme annual values of precipitation and temperature, collected over a ten-year period from relevant meteorological stations. The data were processed using geospatial methods to achieve high resolution at the level of individual road sections [12].

- Accident data. The frequency and spatial distribution of traffic accidents were analyzed using a weighted accident index that takes into account both the number and severity of accidents. The index is based on a weighted scoring system according to injury severity and is normalized by the length of each road section. This index serves as the primary indicator for assessing and comparing the safety performance of different sections of the road network [13].

## 4. METHODOLOGY

The methodology investigates the factors influencing traffic accident risk using three machine-learning models: Bagging Regressor, CatBoost Regressor, and LightGBM Regressor. These models were selected to quantify the influence of 23 parameters (road geometry, traffic conditions, and environmental factors) on the weighted accident index (Wi).

The dataset comprising 161 road sections was divided into training 80% and test 20% subsets using stratified random sampling to preserve the distribution of the target variable. Model performance was evaluated on the test set using the coefficient of determination ($R^2$), mean absolute error (MAE), and root mean square error (RMSE).

The three models provide complementary strengths: Bagging Regressor reduces variance and ensures robustness, CatBoost Regressor efficiently handles categorical features and prevents overfitting, while LightGBM Regressor offers fast training and high accuracy through gradient boosting. Their combination yields stable and highly interpretable rankings of parameter importance.

## 5. RESULTS

### 5.1 BAGGING REGRESSOR

Bagging Regressor relies on the combination of multiple independent models to improve the stability and accuracy of predictions through the averaging of their outputs. Instead of depending on a single model, the method generates different subsets of the original dataset, each used to train a separate base learner, typically a decision tree. By aggregating their predictions, this technique produces more balanced and robust estimates, particularly when dealing with variable or complex data [14].

In the present study, the method was employed to predict the weighted accident index (Wi). The model consists of 500 decision trees, each trained on a different bootstrap sample, with their aggregation ensuring more uniform estimates and variance reduction.

The general prediction formula can be expressed as follows:

$$Y = (1/T)\sum(t = 1)^{T} Y_t \quad (1)$$

The final prediction Y is the average of all individual tree predictions, where $Y_t$ is the prediction of each individual model and T is the total number of models.

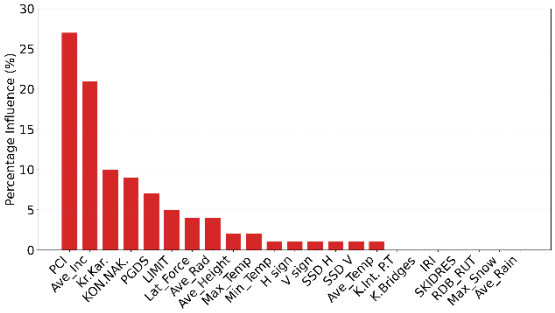


Figure 2. Percentage influence (Bagging Regressor)

As shown in the figure 2, the pavement condition index (PCI) and average longitudinal gradient (Ave_Inc) exert the strongest influence, traffic-related parameters exhibit moderate influence, while parameters associated with weather conditions have minimal impact.

The Bagging Regressor demonstrates solid predictive performance with $R^2$ = 0.522, MAE =

2.89, and RMSE = 3.78 on the test set, reflecting sensitivity to larger deviations.

## 5.2 LIGHTGBM REGRESSOR

With this technique, trees are added incrementally, growing along the most beneficial leaf (leaf-wise), and histogram-based thresholds are used for each feature to accelerate split selection [15]. This approach is well-suited to tabular data comprising section-wise indicators with the weighted accident index (Wi) as the target. The model deepens the leaf that yields the greatest local improvement, while feature histograms provide a limited set of candidate thresholds. This enables rapid model enhancement with moderate complexity and easily interpretable leaf rules.

In its simplest form, the update follows the gradient boosting scheme:

$$F_m(x) = F_{m-1}(x) + \gamma \cdot h_m(x) \quad (2)$$

Where $F_m(x)$ is the new predictive model after the m-th iteration, $F_{m-1}(x)$ is the previous model, $h_m(x)$ is the weak learner that corrects the previous errors and γ is the learning rate that controls the contribution of the new model. For selecting a split s (potential node split), the gain metric is used:

$$(s) = 0.5 * \left(G_L(s)^2/(H_L(s) + \lambda) + G_R(s)^2/ (H_R(s) + \lambda) - G(s)^2/(H(s) + \lambda)\right) - c \quad (3)$$

Where G and H are the sums of the first and second derivatives of the loss at the node and sides, λ is the L2 regularisation on complexity and γ is the fixed penalty for gain per split.

In this research, the method was applied to predict the weighted accident index (Wi). The model was optimised with 400 iterations, learning_rate ≈ 0.005, num_leaves = 7, and additional regularisation and subsampling parameters.

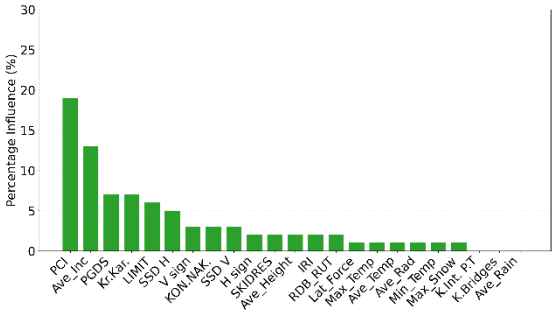


Figure 3. Percentage influence (LGBM)

As shown in the figure 3, road-related parameters (PCI, Ave_Inc, Kr.Kar.) and traffic-related parameters (PGDS, LIMIT) are clearly the most influential, while environmental parameters (precipitation, snow, temperature, elevation) have minimal influence.

The LightGBM Regressor achieved $R^2 = 0.514$, MAE = 2.94, and RMSE = 3.85 on the test set, confirming stable and accurate predictions.

## 5.3 CATBOOST REGRESSOR

CatBoost is a method that builds small symmetric trees and combines them into a stronger model. When categorical features are present, it encodes them using ordered target statistics, ensuring that each record uses information only from preceding rows and preventing information leakage from training to test data. It is particularly suitable for tabular data containing a mixture of numerical and categorical variables, requires no prior standardisation, and delivers stable estimates with simple and interpretable rules [16]. The model achieves progressive error reduction, with categorical features encoded in order (ordered encoding) without using information from future rows.

In its simplest form, the update follows the gradient boosting scheme:

$$F_m(x) = F_{m-1}(x) + \gamma \cdot h_m(x) \quad (4)$$

Where $F_m(x)$ is the new predictive model after the m-th iteration, $F_{m-1}(x)$ is the previous model, $h_m(x)$ is the weak learner that corrects the previous errors and γ is the learning rate. Categorical variables are transformed into numerical values using ordered statistics:

$$\mathrm{enc}(\mathrm{cat_i}) = \left(\Sigma_{j<i} y_j + a * \mathrm{prior}\right)/(n_{<i} + a) \quad (5)$$

Where $\mathrm{enc}(\mathrm{cat_i})$ is the encoded value for the category in the current row i, j < i – previous rows and prior is the initial value for stabilisation, a is the smoothing coefficient.

This ordered encoding between training and test sets is fully consistent with the 80/20 split and eliminates information overlap.

In this research, the method was applied to predict the weighted accident index (Wi). The model was configured with 1000 iterations, learning_rate = 0.05, and depth = 6.

As clearly visible from the figure 4, the dominant influence on the weighted accident index is exerted by road and traffic characteristics (PCI, Ave_Inc, PGDS, Kr.Kar., KON.NAK., LIMIT),

whereas the influence of environmental factors (precipitation, snow, temperature, and elevation) is negligible.

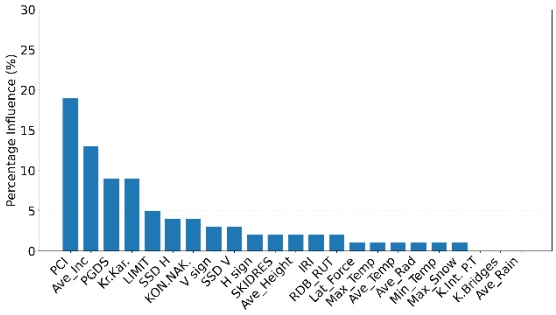


Figure 4. Percentage influence (CatBoost)

On the test set, CatBoost Regressor obtained $R^2$ = 0.520, MAE = 2.91, and RMSE = 3.81, confirming robust and predictive reliability.

## 5.4 AVERAGING AND NORMALISATION OF RESULTS FROM THE THREE METHODS

This section presents the combined results obtained from the three applied machine-learning models used to evaluate the influence of various factors on the weighted accident index (Wi).

To ensure a balanced evaluation, the importance of the parameters was averaged using weights proportional to the accuracy ($R^2$) of each model on the test subset (80/20):

- Bagging Regressor (33.7 %) – $R^2$ = 0.522
- CatBoost Regressor (33.5 %) – $R^2$ = 0.520
- LightGBM Regressor (32.8 %) – $R^2$ = 0.514

Based on these weights, the percentage influence values from the three methods were normalised so that the total sum across all parameters equals 100 %. This approach provides a more objective and comparable assessment of the significance of the factors.

Table 1. Percentage influence of parameters on the weighted accident index (Wi)

| No. | Abbr. | BR (%) | CB (%) | LGBM (%) | Res. (%) |
|---|---|---|---|---|---|
| 1 | PCI | 27.0 | 19.0 | 19.0 | **22.8** |
| 2 | Ave_Inc | 21.0 | 13.0 | 13.0 | **16.6** |
| 3 | Kr.Kar. | 10.0 | 9.0 | 7.0 | **9.0** |
| 4 | PGDS | 7.0 | 9.0 | 7.0 | **7.8** |
| 5 | KON.NAK. | 9.0 | 4.0 | 3.0 | **6.1** |
| 6 | LIMIT | 5.0 | 5.0 | 6.0 | **5.3** |
| 7 | RDB_RUT | 0.0 | 2.0 | 2.0 | **1.2** |
| 8 | K.Int. P.T | 0.0 | 0.0 | 0.0 | **0.0** |
| 9 | IRI | 0.0 | 2.0 | 2.0 | **1.2** |
| 10 | SSD H | 1.0 | 4.0 | 5.0 | **3.1** |
| 11 | K.Bridges | 0.0 | 0.0 | 0.0 | **0.0** |
| 12 | V sign | 1.0 | 3.0 | 3.0 | **2.2** |
| 13 | SSD V | 1.0 | 3.0 | 3.0 | **2.2** |
| 14 | Lat_Force | 4.0 | 1.0 | 1.0 | **2.3** |
| 15 | Ave_Temp | 1.0 | 1.0 | 1.0 | **1.0** |
| 16 | Ave_Rad | 4.0 | 1.0 | 1.0 | **2.3** |
| 17 | Ave_Rain | 0.0 | 0.0 | 0.0 | **0.0** |
| 18 | Max_Snow | 0.0 | 1.0 | 1.0 | **0.6** |
| 19 | Min_Temp | 1.0 | 1.0 | 1.0 | **1.0** |
| 20 | Ave_Height | 2.0 | 2.0 | 2.0 | **2.0** |
| 21 | SKIDRES | 0.0 | 2.0 | 2.0 | **1.2** |
| 22 | Max_Temp | 2.0 | 1.0 | 1.0 | **1.5** |
| 23 | H sign | 1.0 | 2.0 | 2.0 | **1.6** |

The resultant values, shown in Figure 5, indicate that the most influential factors are clearly the pavement condition index (PCI), the average longitudinal gradient (Ave_Inc), and the horizontal curvature characteristic (Kr.Kar.). Conversely, the lowest influence is exhibited by skid resistance (SKIDRES), maximum annual temperatures (Max_Temp), and horizontal signage (H sign).

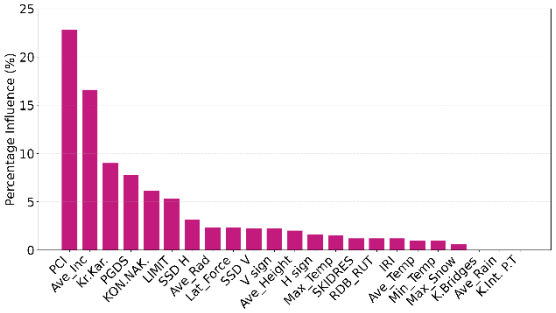


Figure 5. Normalised influence of parameters

The combined evaluation confirms the priority of measures aimed at pavement maintenance, management of longitudinal gradients, and improvement of horizontal alignment geometry. Factors with the lowest influence are useful as supplementary indicators and should be monitored contextually, without decisive impact on ranking.

## 6. DISCUSSION

The analysis reveals that the greatest influence on the weighted accident index is exerted by the geometric and structural characteristics of the road, particularly the pavement condition index (PCI), average longitudinal gradient (Ave_Inc), and horizontal curvature characteristic (Kr.Kar.). Together, these parameters account for 51.4 % of the total influence, highlighting their critical role in determining road safety. Parameters related to operational conditions exhibit moderate influence, whereas environmental factors contribute negligibly.

In a comparative context, the three models demonstrate a high degree of consistency.

Bagging Regressor provides the highest accuracy and robustness, CatBoost Regressor excels in interpretability and effective handling of categorical data, while LightGBM Regressor offers the fastest training with virtually identical rankings and only minimal deviations. This consistent ordering of the leading factors across all three models strongly reinforces the conclusions regarding priorities and supports the implementation of measures focused on improving pavement condition and managing longitudinal gradients and horizontal curves.

## 7. CONCLUSION

In summary, the results indicate that traffic-accident risk is primarily determined by pavement condition and road geometry. Indicators related to surface damage, longitudinal gradient, and horizontal curvature clearly dominate over all other variable groups. Parameters linked to operational conditions have a moderate contribution, while environmental factors prove negligible in the analyzed dataset.

The combination of three decision-tree-based machine-learning models (Bagging Regressor, CatBoost Regressor, and LightGBM Regressor) together with evaluation on an 80/20 train–test split provides a stable, robust, and highly interpretable assessment of the parameters, yielding consistent ranking of the key factors. These findings support the identification of high-risk sections and lay the foundation for predictive tools that will facilitate the planning and prioritisation of road-safety interventions.

Future research should focus on expanding the database and incorporating additional parameters (e.g. share of heavy vehicles, actual speeds, lane and shoulder widths), as well as introducing finer temporal resolution and external validation across different road categories. Such steps would strengthen the general applicability and increase the practical value of the results.

## REFERENCES


[1] World Health Organization. (2024). Road traffic injuries.
Available at: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries (accessed March 2025).

[2] European Commission. (2023). Road Safety Statistics 2023. Available at: https://road-safety.transport.ec.europa.eu/european-road-safety-observatory_en (accessed March 2025).

[3] State Statistical Office of the Republic of North Macedonia. (2025).
MakStat – Statistical database. Available at: https://makstat.stat.gov.mk/PXWeb/pxweb/mk/ (accessed March 2025).

[4] Wang, Y., Zhang, Y., Wu, J., & Xu, C. (2023). Analyzing the Risk Factors of Traffic Accident Severity Using Machine Learning and Association Rules. International Journal of Environmental Research and Public Health, 20(1), 345. https://doi.org/10.3390/ijerph20010345

[5] Jiang, Y., Li, S., Zhao, Z., & Chen, H. (2024). Machine Learning-Based Prediction Analysis of Potential Factors Influencing Traffic Accident Severity. Sustainability, 16(2), 1125. https://doi.org/10.3390/su16021125

[6] Çelik, E., & Sevli, D. (2022). Predicting Road Traffic Accident Severity Using Machine Learning Techniques. Applied Sciences, 12(15), 7485. https://doi.org/10.3390/app12157485

[7] Public Enterprise for State Roads. Web-GIS platform for spatial analysis and visualisation. Available at: http://62.77.137.99/pesr/webgis/#/map (accessed March 2025).

[8] Ministry of Local Self-Government. (2021). Programme for Development of the Planning Regions for the period 2021–2026. Government of the Republic of North Macedonia, Skopje.

[9] Doncheva, R., Ognjenovic, S. (2024). Road Design. Ss. Cyril and Methodius University – Faculty of Civil Engineering, Skopje. ISBN: 978-608-4510-60-4.

[10] Tobias, P., de León Izeppi, E., Flintsch, G., Katicha, S., McCarthy, R. (2023). Pavement Friction for Road Safety: Primer on Friction Measurement and Management Methods. Federal Highway Administration (FHWA), Report No. FHWA-SA-23-007.

[11] Public Enterprise for State Roads. Web-GIS platform for spatial analysis and visualisation. Available at: http://tdps.roads.org.mk/ (accessed March 2025).

[12] Gjeshovska, V., Taseski, G., Ilioski, B. (2024). Intense precipitation in the Republic of North Macedonia. Ss. Cyril and Methodius University – Faculty of Civil Engineering, Skopje. ISBN: 978-608-4510-56-7.

[13] Government of the Republic of North Macedonia, Ministry of Transport, Project Unit. (2024). Black Spot Management Handbook (BSM). Skopje, July 2024.

[14] Patil, P., Du, J.-H., Kuchibhotla, A. K. (2022). Bagging in overparameterized learning: Risk characterization and risk monotonization. arXiv preprint,arXiv:2210.11445.https://doi.org/10.48550/arXiv.2210.11445

[15] Ke, G., Meng, Q., Finley, T. et al. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems, 30. Available at: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[16] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. Advances in Neural Information Processing Systems,31.https://doi.org/10.48550/arXiv.1706.09516