Proceedings of the 35th European Safety and Reliability & the 33rd Society for Risk Analysis Europe Conference
Edited by Eirik Bjorheim Abrahamsen, Terje Aven, Frederic Bouder, Roger Flage, Marja Ylönen

doi: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P2972-cd

A methodology to find the importance of winter road characteristics on winter road accidents

Mahshid Hatamzad

Bane NOR, Oslo, Norway. E-mail: mahshid.hatamzad@banenor.no
UiT/ The Arctic University of Norway

Various factors such as road geometry, precipitation, freezing temperature, and ice on the road surface increase the risk of different types of road accidents in winter. This study proposes a methodology to classify and model winter road accidents and determine the importance of each input variable (locational characteristics, road characteristics, and winter weather characteristics). This methodology utilizes a machine learning method for multi-class classification and then applies an approach to identify important input variables affecting classification of winter road accidents. The methodology has seven main stages and starts with data analysis, which gives a general overview of the dataset and is a major stage in using machine learning algorithms. Next, four different classes regarding personal injuries are defined for road accidents. After dividing the dataset into training and testing sets, categorical variables need to be transformed into numerical variables to be understood by machine learning algorithms. Then, different models for multi-classification need to be trained and tested to find the model with the best performance based on various evaluation metrics and plotting the process of learning and testing the model. Finally, the recursive feature addition method can be used to rank the importance of input variables on classifying severe road accidents in winter.

Keywords: machine learning, multi-class classification, winter road accidents, personal injuries, recursive feature addition

1. Introduction

Winter road conditions can have a significant impact on increasing the risk of road accidents in clod-climate areas due to snow, icy road surfaces, and bad visibility. Critical winter roads and weather conditions in cold-climate areas can easily lead to road accidents with personal injuries and fatalities. Therefore, identifying the impact of winter road characteristics on winter road accidents plays an important role in maximizing road traffic safety. Many studies have proposed different methods and methodologies to determine the effect of winter road and weather characteristics on road traffic safety in winter. Andreescu and Frost (1998) used a quantitative method and a two-year dataset to analyse the impact of average temperature (monthly, yearly, and two years), rain, and snow on vehicle incidents in Montreal, Canada. The results showed that snow played a significant role in road accidents in winter, and the number of road accidents increased when the weather was snowy. Eisenberg (2004) identified a link between precipitation and road crashes in the USA from the year 1975 to 2000. He utilized a negative binomial regression method, and the results demonstrated a strong relationship between monthly precipitation and road accidents that led to fatalities. Golob and Recker (2003) applied both linear and non-linear statistical analysis to identify the influential variables on road accidents. The results depicted that median traffic speed, lane (right or left), and wet roads affect the severity of the road accident. Yu and Abdel-Aty (2014) utilized a random forest model to find the important variables of road incidents that led to injuries. In addition, they used a support vector machine to explore the non-linear relationship between crashes and the variables. The results investigated that temperature, snow, speed deviation, and steepness are the most critical variables. Thordarson et al. (2021) identified a relationship between winter road accidents and weather conditions. The results showed that slippery road surface conditions and wind power had an impact on road accidents in winter. Eboli et al. (2020) utilized logistic regression to investigate the influential factors (road, environment, and driver) on different types of vehicle collisions. Shaik et al. (2021) published a review paper for studies that applied a neural network algorithm to determine the factors, which increase the risk of road accidents with fatalities. Lee et al. (2019) compared traditional statistical methods and machine learning algorithms to explore the advantages and disadvantages of these methods. They used three different datasets (road geometry, road incidents, and precipitation) recorded for nine years in Seoul, South Korea. The results showed that machine learning techniques were able to handle complex relationships, however, they were in danger of overfitting. In addition, curve length, precipitation, driver gender, and road surface condition were important parameters in the severity of road accidents. While numerous studies focused on determining of impact of winter weather and road characteristics on road accidents, there is no study to propose a multi-class imbalanced classification using machine learning algorithms. Hence, this study focuses on proposing a methodology to classify and model four different types of road accidents in winter using various types of input variables (location, road, and weather characteristics), and utilizing the recursive feature addition method to find the importance of each variable.

2. Dataset

A dataset to apply the proposed methodology consisted of different types of road accidents in winter in cold-climate areas and various types of input variables. The input variables were divided into three categories:

- Locational characteristics of the accident, such as altitude from the sea level.
- Road characteristics of the accident, such as lane (right, middle, left), presence of turn, bump, crossing, giveaway, junction, roundabout, railway, level crossing, and station.
- Weather characteristics when the accident occurred, such as temperature, precipitation, wind speed, wind direction, and pressure.

Different open-source datasets include these variables and can be used by researchers (Moosavi et al. 2019 a), (Moosavi et al., 2019b).

3. Methodology

Fig. 1. shows the general view of the methodology. The methodology consists of seven stages, and each stage is explained in this section.

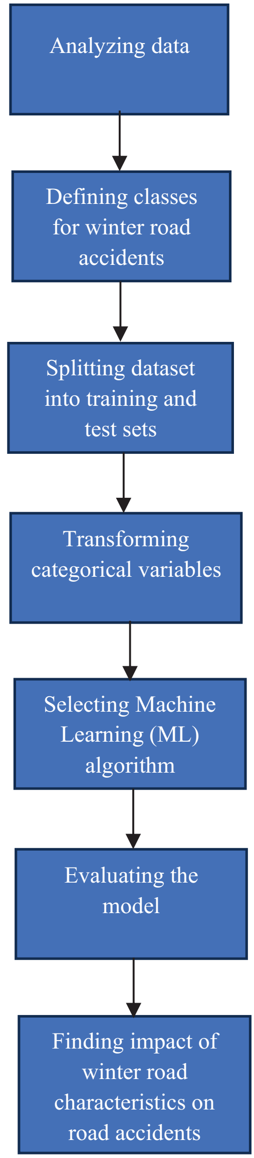

Fig. 1. Methodology Framework

3.1. *Stage 1: Analyzing the dataset*

Dataset analysis helps to get a general overview of different variables. For example, which lane is more prone to accidents, what type of accident occurs more often, what is the impact of weather on the type of road accidents in winter, what is the impact of precipitation and temperature on the type of accidents, which month has the highest accident rate, etc. Additionally, what type of classification problem do we face, whether it is a balanced multi-class classification or an imbalanced multi-class classification? Moreover, this stage helps to see the number of missing values in the dataset. If removing the number of missing values does not change the distributions of the variables, missing values can be removed, otherwise, missing values need to be dealt with by a method such as imputation techniques, forward/backward fill, etc. Fig. 2. shows the schematic view of Stage 1.

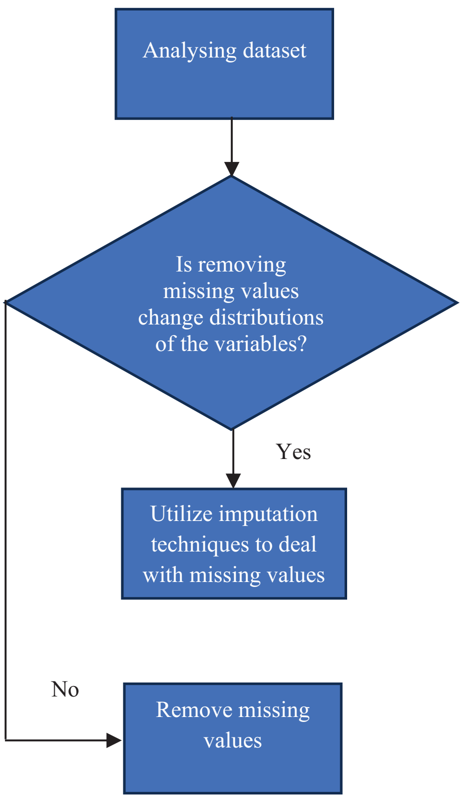

Fig. 2. Schematic view of Stage 1

3.2. *Stage 2: Defining accident classes*

Winter road accidents can be classified into 4 classes:

- Class 1: winter road accidents with no personal injuries
- Class 2: winter road accidents with personal injuries treated by first aid
- Class 3: winter road accidents with severe personal injuries
- Class 4: winter road accidents resulting in fatalities

Fig. 3. shows the schematic view of Stage 2.

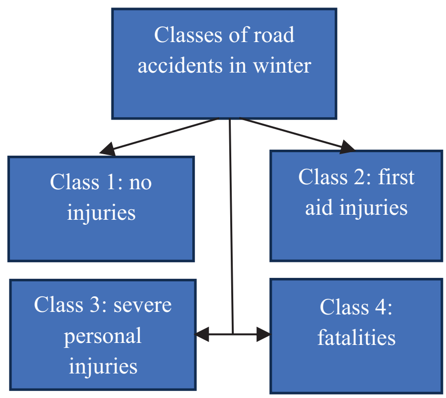

Fig. 3. Schematic view of Stage 2

3.3. *Stage 3: Splitting the dataset into training and testing sets*

First, the dataset must be split into training and testing sets. For instance, 70% or 60% of the data points can be considered for training, and 30% or 40% can be considered for testing the model. Fig. 4. shows the schematic view of Stage 3.

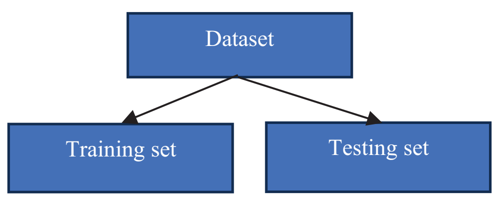

Fig. 4. Schematic view of Stage 3

3.4. *Stage 4: Transformation of categorical variables*

Categorical variables such as wind direction, month, and some road characteristics need to be transformed into numbers to be understood by ML algorithms. There are different methods, which can be used for this objective, such as dummy encoding and label encoding. Fig. 5 shows the schematic view of Stage 4.

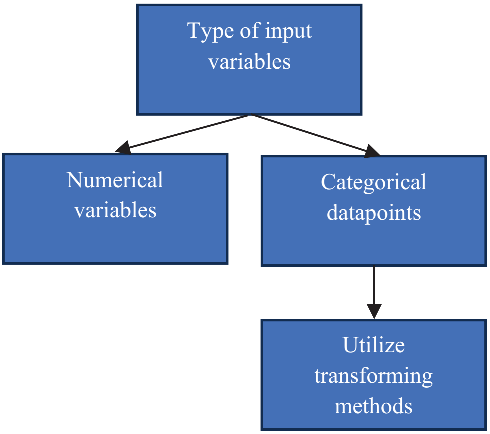

Fig. 5. Schematic view of Stage 4

3.5. *Stage 5: Dealing with classification problems and selection of the ML algorithm*

If the number of winter road accidents in different classes is almost the same, we deal with a balanced multi-class classification problem. However, it is not common to have the same number of road accidents in each class. Therefore, if the number of winter road accidents is very different in each road accident class (for example, class 1: 2000, class 2: 500, class 3: 100, class 4: 20), then, we need to use: 1) resampling methods such as synthetic minority oversampling technique (SMOTE) or under-sampling the majority class to achieve a better balance in the accident classes, and 2) utilizing ML algorithms to train and test the model for multi-class classification. Various algorithms can be used, such as decision tree, xgboost, random forest, and support vector machine. Fig. 6. shows the schematic view of Stage 5.

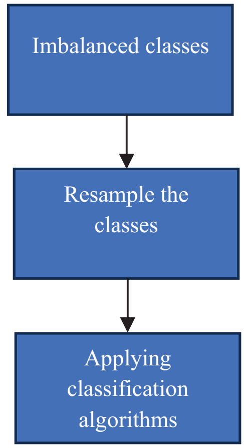

Fig. 6. Schematic view of Stage 5

3.6. *Evaluating different classification algorithms*

The confusion matrix shows precision, recall, and F1-score for each class. In addition, it demonstrates how many data points are incorrectly or correctly classified by the algorithms. Moreover, the weighted average can be calculated by the confusion matrix. However, numerical metrics are not enough to evaluate the model's performance. These metrics should be accompanied by graphical evaluation metrics such as learning curves and ROC-AUC (receiver operation characteristics and the area under the curve). The graphical model evaluation helps to understand if the model is overfitted or underfitted. In this stage, different algorithms need to be tested to find the algorithm with the best performance. Additionally, the performance of the selected model can be improved by tuning the hyperparameters. This means that it is possible to utilize different methods such as the methods proposed by grid-search (Zhang et al, 2022), and AI optimization techniques such as ant colony, genetic algorithm, and particle swarm optimization algorithms (Hatamzad et al. 2021). Fig. 7. shows the schematic view of Stage 6.

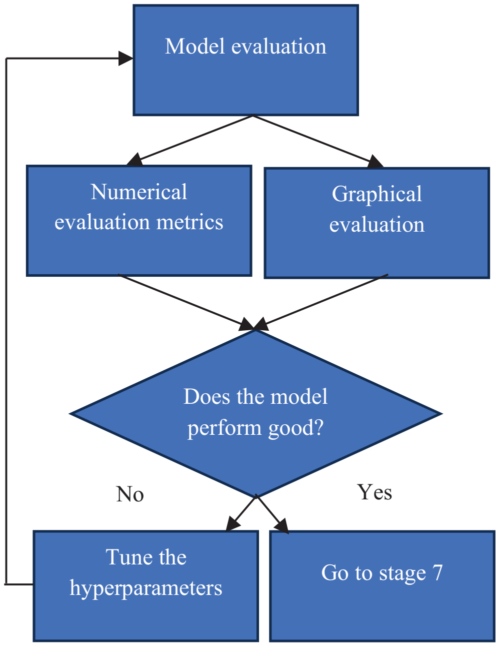

Fig. 7. Schematic view of Stage 6

3.7.*Stage 7: Impact of winter road characteristics on road accidents*

The selected algorithm (from Stage 6) is used in this stage to find the impact of winter road characteristics on classifying road traffic safety. The impact of each variable can be identified in two ways: 1) each winter road characteristic (location, road information, and weather) can be considered in the model separately, and 2) all the input variables can be considered in the model together. If the goal is only to explore the impact of one characteristic type such as winter weather characteristics on road accidents, the model can be run by weather characteristics. However, if the goal is to determine the impact of all the input variables, the model can be run to identify the important input variables on winter road traffic accidents.

In this stage, the recursive variable addition approach is used to determine the importance score for each variable (Github, 2022). The following steps show the procedure for using this method:

- Step 1: Setting a threshold
- Step 2: Constructing the ML classification algorithm using one variable and calculating the ROC-AUC value
- Step 3: Adding one more variable and calculate the new ROC-AUC value
- Step 4: If the ROC-AUC value increases more than the threshold in Step 1, it shows the variable's importance, which should be kept. Otherwise, the variable should be removed.
- Step 5: Steps 2 to 4 need to be repeated until all the input variables are evaluated.

Fig. 8. shows the schematic view of Stage 7.

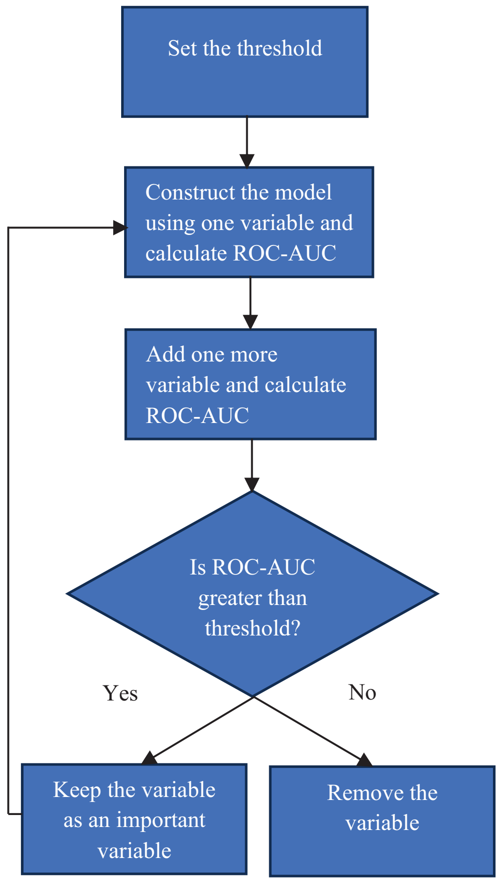

Fig. 8. Schematic view of Stage 7

4. Application to a real dataset

Hatamzad and Gudmestad (2025) have applied a slightly modified version of the methodology to a large dataset (Moosavi, 2021) using various machine-learning algorithms. The methodology is used for the coldest states in the USA. First, general statistical information from the original dataset is analyzed. Next, the dataset is filtered, according to the coldest states, cold months, and accident classes. Third, the dataset is split into training and test sets, and a transformation of categorical variables is performed. An imbalanced binary classification is dealt with because the distributions of accident classes (class 0 and class 1) are not equal (we used a ratio for an imbalanced classification). Machine learning algorithms for classification are compared. The classification algorithms are logistic regression, random forest, support vector machine, extreme gradient boosting, and decision tree. The weather factors affecting the classification of severity of winter accidents in the three coldest states in the USA with the maximum number of winter road accidents are determined using the feature importance approach.

4.1. *Results and discussion*

Severe road accidents in wintertime in cold climate areas were successfully classified and modeled. In addition, the most significant weather variables influencing classification of severe winter accidents were identified using the proposed methodology and applying different machine learning techniques for classification. The dataset (Moosavi, 2021), containing observations from the USA was pre-processed by extracting relevant variables and filtering for the coldest states and severe incidents.

The results demonstrated that the Extreme Gradient Boosting (xgboost) model outperformed other models for the coldest states in the US (achieving an F1-score of 87% for classifying serious incidents and 53% for very serious incidents). The xgboost model was applied to identify important variables influencing accident severity classification to derive future insights, which contributes to data-driven decision-making for preventing severe road accidents in winter. The most important variable was the atmospheric pressure in all the selected cold climate states in the US.

Further analysis was performed for three specific locations (states) in the US. In the first and second locations, pressure was a key factor; however, in the third location, wind speed was the most significant variable. The findings showed that although pressure was the most important feature in most of the locations, it cannot be concluded that it applies to all different locations, since the most important feature in the third location was wind speed. Therefore, it is important to classify and model each location separately.

5. Conclusion and future work

This paper proposed a methodology to classify and model multi-class road accidents in winter using a machine learning algorithm and then utilizing a recursive variable addition approach to find the importance of each input variable on classifying winter road accidents. The methodology has seven main stages. The first stage is analyzing data to get a general overview of the dataset. This stage also helps to find the presence of missing values and how to deal with them. In the second stage, four classes of road accidents in winter were defined: 1) no injuries, 2) first aid treatment, 3) severe personal injuries, and 4) fatalities. In the third step, the dataset is split into training and testing sets. In the fourth stage, feature engineering such as transforming categorical data is done. In the fifth step, imbalanced classes need to be resampled, and then a machine learning algorithm for multi-classes needs to be selected. In the sixth stage, the selected model is evaluated based on graphical and numerical metrics to avoid overfitting. In the seventh stage, recursive feature addition is used to find the importance of each variable. For future research, the proposed methodology should be applied to more real case studies. In addition, the driver's behavior data can be considered in the model.

Appendix A. Definition of important terms

An imbalanced classification for a multi-class problem contains an unequal distribution of observations (data points) in the multi-classes (4 classes). To make sure that algorithms properly classify and model the observations, evaluation metrics need to be calculated. There are various numerical metrics to measure the ability of the algorithms in imbalanced multi-classification.

A confusion matrix is a table that summarizes model performance using different metrics. A confusion matrix for 4 classes has 4 dimensions, which are actual values, and predicted values. Figure A1 shows a confusion matrix for a 4-class classification. The following letters describe the different cells in the confusion matrix:

- A: class 1 predicted as class 1
- B: class 2 predicted as class 1
- C: class 3 predicted as class 1
- D: class 4 predicted as class 1
- E: class 1 predicted as class 2
- F: class 2 predicted as class 2
- G: class 3 predicted as class 2
- H: class 4 predicted as class 2
- I: class 1 predicted as class 3
- G: class 2 predicted as class 3
- K: class 3 predicted as class 3
- L: class 4 predicted as class 3
- M: class 1 predicted as class 4
- N: class 2 predicted as class 4
- O: class 3 predicted as class 4
- P: class 4 predicted as class 4

	Real class				
Predicted class	Classes	Class 1	Class 2	Class 3	Class 4
	Class 1	A	B	C	D
	Class 2	E	F	G	H
	Class 3	I	J	K	L
	Class 4	M	N	O	P

Fig. A1. Confusion matrix for classifying multi-classes

The performance of the model can be evaluated by the F1-score for each class and learning curves. The metric F1-score for multi-class classification combines recall and precision for each class, and the learning curve plots the learning process and model performance to see if the model diagnoses with underfitting or overfitting (Hatamzad et al. 2021).

The equations to calculate the F1-score are as follows:

Precision (P_i):

Class i: P_i: $TP_i /(TP_i+FP_i)$ (1)

Recall (R_i):

Class i: R_i: $TP_i /(TP_i+FN_i)$ (2)

F1 -score:

Class i: F_i: $(2*P_i*R_i) /(P_i+R_i)$ (3)

Where, for example, TP_1 shows data points that the algorithm can truly predict as Class 1 or A in Fig. A1. FP_1 is data points that the algorithm falsely predicts Class 2, Class 3, and Class 4 as Class 1 such as B, C, and D in Fig. A1. FN_1 shows the data points that algorithm falsely predicts Class 1 as Class 2 or Class 3 or Class 4, such as E, I, and M in Fig. A1 (Rajesh, 2024).

References

Andreescu, M. P., & Frost, D. B. (1998). Weather and traffic accidents in Montreal, Canada. *Climate Research, 9(3),* 225-230.

Eboli, L., Forciniti, C., & Mazzulla, G. (2020). Factors influencing accident severity: an analysis by road accident type. *Transportation Research Procedia*, 47, 449-456.

Eisenberg, D. (2004). The mixed effects of precipitation on traffic crashes. *Accident Analysis & Prevention*, *36*(4), 637-647.

Github.com (2022). Feature selection for machine learning. https://github.com/solegalli/feature-selection-for-machine-learning/blob/main/11-Hybrid-methods/11.3-Recursive-feature-addition.ipynb

Golob, T. F., & Recker, W. W. (2003). Relationships among urban freeway accidents, traffic flow, weather, and lighting conditions. *Journal of Transportation Engineering, 129(4),* 342-353.

Hatamzad, M. & Gudmestad, O.T. (2025 Modeling the importance of weather variables on classifying severe winter road incidents in cold-climate areas. To be submitted for publication.

Hatamzad, M., Pinerez, G. P., & Casselgren, J. (2021). Using slightly imbalanced binary classification to predict the efficiency of winter road maintenance. *IEEE Access*, *9*, 160048-160063.

Lee, J., Yoon, T., Kwon, S., & Lee, J. (2019). Model evaluation for forecasting traffic accident severity in rainy seasons using machine learning algorithms: Seoul city study. *Applied Sciences*, *10(1),* 129.

Moosavi, S., Samavatian, M. H., Parthasarathy, S., & Ramnath, R. (2019a). A countrywide traffic accident dataset. arXiv preprint arXiv:1906.05409.

Moosavi, S., Samavatian, M. H., Parthasarathy, S., Teodorescu, R., & Ramnath, R. (2019b, November). Accident risk prediction based on heterogeneous sparse data: New dataset and insights. In *Proceedings of the 27th ACM SIGSPATIAL international conference on advances in geographic information systems* (pp. 33-42).

Moosavi, S., (2021, January), US-Accidents: A Countrywide Traffic Accident Dataset, Smoosavi, https://smoosavi.org/datasets/us_accidents

Shaik, M. E., Islam, M. M., & Hossain, Q. S. (2021). A review on neural network techniques for the prediction of road traffic accident severity. *Asian Transport Studies, 7*, 100040.

Rajesh, S. (2024). Confusion Matrix for Multiclass Classification, Medium. https://medium.com/@sreenilarajesh/confusion-matrix-for-multiclass-classification-fe7b6901a541

Thordarson, S., & Olafsson, B. (2008). Weather-induced road accidents, winter maintenance, and user information. *Transport Research Arena Europe*, 2008, 72.

Yu, R., & Abdel-Aty, M. (2014). Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data. *Safety Science*, 63, 50-56.

Zhang, L., Sun, B., Geng, R., Li, S., & Shen, T. (2022, October). Boosting Machine Learning Algorithms with Grid-Search for Transport IoT Data Prediction. In *International Conference on Computer Engineering and Networks* (pp. 903-910). Singapore: Springer Nature Singapore.