Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 10, 992-1004
2025
Publisher: Learning Gate
DOI: 10.55214/2576-8484.v9i10.10581

Predicting road accident risks using web data: A classification approach

Luan Sinanaj[1*], Erind Bedalli[2], Lejla Abazi-Bexheti[3]
[1,3]South East European University, North Macedonia; ls30441@seeu.edu.mk (L.S.) l.abazi@seeu.edu.mk (L.A.B.).
[2]University of Elbasan, Albania; erind.bedalli@uniel.edu.al (E.B.).

Abstract: With the rapidly increasing rates of vehicle usage during recent decades, road accidents have become a significant concern, posing not only risks of injuries but also ranking among the leading causes of fatalities for young and middle-aged individuals. Several factors influence the occurrence of accidents, including careless driving, atmospheric conditions, speeding, and driving under the influence. Understanding the circumstances that lead to a greater risk of road accidents is very helpful for their prevention. The primary goal of this work is to explore patterns in road accidents that have occurred within the state of Albania based on web data scraped from news portals and reports from governmental institutions. The data mining pipeline first involves an intensive data preprocessing phase where scraping, filtering, and refining techniques are employed. Subsequently, several classification models are built on the preprocessed data. These models are developed using various methodologies, including naïve Bayes, random forests, XGBoost, and LightGBM. The constructed classification models are evaluated based on training-test splitting of the preprocessed data using various performance measures. Finally, these models can be used to predict the likelihood of accidents based on certain circumstances.

Keywords: *Classification algorithms, Data preprocessing, Ensemble methods, Road accidents prediction, Web scraping.*

1. Introduction

Road accidents have been a serious problem in recent decades in every country of the world, and this has also occurred due to the increase in the use of vehicles. The created situation requires better management of the problem of road accidents to maximize the safety of citizens. The reasons why road accidents occur are among the most diverse, such as disobedience to road traffic rules by drivers or pedestrians, causes related to the weather and natural phenomena, or problems related to good or bad adaptation of the road infrastructure. In this context, automated tools for identifying key accident causes and geographic areas of elevated risk would be of significant utility in order to guide prevention strategies from responsible entities as well as to sensitize citizens.

While Albania faces a substantial burden of road traffic accidents, the data available from official institutions are, in general, limited to aggregated monthly statistics. The National Statistics Institute and the Ministry of the Interior of Albania only provide periodic reports summarizing statistics on road accidents, as well as the number of individuals injured or deceased each month and their respective genders [1, 2]. These published data on road accidents from the government are insufficient for conducting detailed studies on the patterns of accidents. To address this gap, this study leverages the use of online news portals as an important additional resource that provides more detailed information about accidents, capturing data about the cause, location, road type, and severity. An exploration of the literature concerning the state of Albania revealed that there are only a few studies on accidents utilizing news portals as a data source [3, 4]. This positions the current study among the first systematic efforts to use unstructured web data for structured accident data analytics and prediction in Albania.

In this research, the main goal has been the analysis and discovery of the main reasons that cause

History: Received: 30 July 2025; Revised: 17 September 2025; Accepted: 22 September 2025; Published: 16 October 2025
* Correspondence: ls30441@seeu.edu.mk

road accidents, as well as, if possible, the discovery of the combination of circumstances that increase the possibility of a road accident based on the data processed from online newspapers that publish articles about road accidents. Citizens can benefit from the results of this study, who can be informed about the causes or areas of road accidents more carefully, as well as responsible state institutes that can undertake awareness campaigns for the prevention of road accidents.

For the realization of this research, a methodology was followed, which is briefly described in the following steps:

- Implementing an algorithm in the Python programming language that reads and extracts articles discussing road accidents from online newspapers using scraping techniques, and saves them in a file.
- Implement another program in the Python programming language for processing the articles of the above step by filtering and finding data on road accidents and saving them in another file.
- The manual processing and control of the data mentioned above are essential for ensuring high-quality data for study and for applying it in the most effective way using machine learning algorithms.
- The last step is the application of classification methods such as k-nearest neighbors, naïve Bayes, and random forest.

The research questions raised for this case study are as follows:

- RQ1: How can web scraping be used to build a database of data from online newspaper articles?
- RQ2: How can data preprocessing techniques such as filtering, text normalization and data refinement be applied to prepare a cleaned dataset for applying classification algorithms?
- RQ3: How can simple and ensemble classification models be set to operate on preprocessed data?
- RQ4: Which of the classification models performs better in predicting the possibility of an accident on the basis of the circumstances?

This study contributes to the implementation of a multi-staged methodology that includes web scraping of accident data from Albanian online news portals, automated preprocessing associated with manual refinement of the extracted data, to build a structured dataset focused on road accidents. Two custom Albanian language corpora were developed to support the filtering of relevant content and the identification of accident-related features such as location, cause, and severity. The resulting dataset was then used to train and evaluate several classification models, including basic models such as naïve Bayes, and several ensemble models such as random forests, XGBoost, and LightGBM, in order to predict accident severity based on contextual variables.

This paper is structured as follows: in section two, an approach is provided to review the literature related to studies conducted on road accidents and related topics; in the next section (three), the methodology followed for this research is explained in detail; in the fourth section, the analysis of the results and findings of the study is included; and in the last part of the paper, the conclusions followed by the references used during this research are presented.

2. Literature Review

A good approach to scientific research is when previous studies and results are examined in relation to what we are interested in studying and developing further. For this reason, several more in-depth literature searches were conducted to identify previous scientific works related to the risk and prediction of road accidents, as well as the methods and techniques used to achieve the results. In the following, some information is provided from the literature review of previous works about the techniques used to find the causes or predict road accidents. A search of the literature revealed that in the state of Albania, few studies related to the prediction of road accidents exist, with the exception of the studies conducted by the authors of this paper. The only simple statistical information published in the state of Albania on road accidents is that of the Ministry of the Interior [1, 2].

Studies and scientific research on the prediction or analysis of road accidents have been conducted

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 10: 992-1004, 2025
DOI: 10.55214/2576-8484.v9i10.10581

since the 1980s and 1990s. A prominent work from this period was that conducted by Miaou [5] who proposed an alternative method for estimating basic roadside encroachment data without field collection [5]. This method explores the probabilistic relationships between encroachment events and run-off-road events, providing a wide range of data from conventional accident prediction models. In a subsequent study, Miaou and Lum [6] have examined the statistical properties of two conventional linear regression models and two Poisson regression models for vehicle accidents and highway geometric design relationships. It identifies limitations, such as distributional assumptions, estimation procedures, and sensitivity to short road sections [6].

Moreover, Furlonge [7] aimed to develop an economic model for estimating future road fatalities and determining the monetary value of fatal accidents via easily obtainable parameters [7]. In the early 2000s, many research studies on road accidents were conducted; for example, pattern recognition algorithms and direct diagnostic approaches were employed by Kononov [8] to model the prediction of traffic accidents and diagnose the causality of accidents [8]. Additionally, the study conducted by Larsen [9] employed a multidisciplinary approach and provided accurate information regarding the causes of traffic accidents [9].

In the last decade, various publications have focused on road accidents in different countries, including the use of different machine learning or deep learning algorithms.

In a notable study conducted by Lnenicka et al. [10], which has attracted our interest and sparked the inspiration for the current research work, the authors developed an application employing web scraping techniques to collect and extract crime-related news from online newspaper portals. The extracted information was subsequently processed, analyzed, and visualized, providing as an outcome both a geographical map and a statistical tabular form, allowing for the classification of regions according to their level of crime risk [10].

In the study by Biswas et al. [11], random forest regression was used to predict the number of road accidents and their casualties. The results of the study revealed that random forest regression is relatively effective for prediction and application in this field of study [11]. Additionally, Siddik et al. [12] have used four classification models: Decision Tree, K-Nearest Neighbors (KNN), Naïve Bayes, and Logistic Regression to predict deaths from road accidents. Their models were tested on traffic accident data taken from online newspapers. Among these methods, the decision tree algorithm produced the best accuracy, with an accuracy of 88%. These results will subsequently assist institutions in enhancing road safety [12]. A recent study conducted by researchers has developed nonlinear regression models to analyze and predict road traffic accidents in Albania using population projections and vehicle registration data as independent variables.

Chen and Wang [13] have analyzed traffic accident factors for different ages via road accident data. It proposes an adaptive boosting algorithm (AdaBoost) for data processing. The results revealed that weather and time of day significantly influence driving safety for senior drivers, with rainy and snowy days affecting older drivers the most [13]. Also, Zhao et al. [14] have proposed a vehicle accident risk prediction model via the trichotomy AdaBoost with SMOTE and the one-hot encoding algorithm. The model uses big data mining and real-life accident data analysis. The experimental dataset is reconstructed via synthetic minority oversampling, and weak classifiers are trained via the AdaBoost algorithm. The model considers a real-time area under the curve of 0.77, as demonstrated by extensive simulation results Zhao et al. [14]. AlMamlook et al. [15] have investigated accurate models for predicting traffic accident severity via machine learning techniques such as AdaBoost, logistic regression, naïve Bayes, and random forests. The RF model, with a 75.5% accuracy rate, outperforms other algorithms, such as LR, NB, and AdaBoost, demonstrating its potential for predicting injury severity in traffic accidents [15].

Three machine learning algorithms were tested against decision trees, random forests, and gradient-boosted trees via automobile crash data from the Maryland State Police in the research work by Elyassami et al. [16]. The study's conclusions demonstrated that the most significant variables in the predictive model of traffic accidents are disregard for stop signs and traffic signals, issues with road

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 10: 992-1004, 2025
DOI: 10.55214/2576-8484.v9i10.10581

design, low visibility, and unfavorable weather.

According to work by Chen and Chen [17], the random forest outperforms logistic regression and classification and regression trees in terms of accuracy and specificity in modeling the severity of traffic accidents. In line with earlier research findings, the random forest emerged as the most successful forecasting model among the three [17].

Yan and Shen [18] have employed a hybrid model that combines random forest and Bayesian optimization to obtain highly accurate results. Random forest was used as a basic predictive model, and Bayesian optimization was used to alter the random forest parameters. According to the experimental results, this model outperforms traditional algorithms in terms of accuracy [18]. The work by Karamanlis et al. [19] is another relevant contribution focusing on the identification of black spots (high-risk zones) in Northern Greece via classical statistical models and modern machine learning algorithms. Their research achieved promising accuracy and advocates hybrid approaches that integrate logistic regression and deep learning architectures as Karamanlis et al. [19].

Dong et al. [20] have proposed four boosting-based ensemble learning models for predicting road traffic injury severity, such as adaptive gradient boosting, natural gradient boosting, categorical gradient boosting, and light gradient boosting machines. The Light Gradient Boosting Machine (LightGBM) achieves the highest classification accuracy and precision. The study revealed that factors or variables such as the month of the year, driver age, cause of accident, and collision type significantly influence injury risk [20]. Combining LightGBM and SHAP could lead to the development of an interpretable model. In the study by Mahendra and Roopashree [21], four criteria are used to predict road accidents: the type of collision, the road type, the location, and the weather. Both a random forest and a decision tree regressor were included in their machine learning model, and the findings of the analysis demonstrated that the random forest regressor model outperforms the decision tree regressor model [21].

3. Methodology

This section explains the methodology followed for this research, the phases that have been completed, the data collection process, and the data mining models used for these data. For the realization of this project, the methodology was divided into two main parts: data preprocessing and classification models based on the Naive Bayes, Random Forests, XGBoost, and LightGBM algorithms.

3.1. Data Collection and Preprocessing

One of the key parts of this study was extracting and preparing the data, which were then utilized in the data mining models to meet the goals of this research work. This was an especially challenging endeavor, as only a limited number of papers and studies are available that target text processing in the Albanian language. During the preprocessing phase, the work was organized into four important steps: data extraction, filtering, refinement, and manual control. These steps were followed to provide the dataset for the study, as described in the following sections:

Step 1: Data extraction. As the primary resource for our data collection, several Albanian online news portals are utilized, from which the texts of various articles are extracted. This was achieved by implementing a web-scraping Python application based on the BeautifulSoup library [22]. This collection is designed to serve even for further research directions, as it will be a large base with articles from various topics. The BeautifulSoup Python module is designed to simplify the extraction of data from HTML or XML pages. Through this process, approximately 80,750 articles were extracted and stored in a CSV file, where the data were structured into a tabular format containing the following fields: number, news source, article date, and the article body.

Step 2: During the second stage, preprocessing on a large database of news articles is applied, filtering it to extract articles related to road accidents that have occurred within the state of Albania. During this stage, in addition to common data preprocessing libraries, such as Pandas and NumPy, extensive use of the text processing library NLTK was employed. NLTK is a widely used Python

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 10: 992-1004, 2025
DOI: 10.55214/2576-8484.v9i10.10581

toolkit for working with natural language processing, providing versatile tools for various language processing tasks, such as text segmentation, tokenization, stemming, and part-of-speech tagging. Despite being primarily designed for English, modifications were made to several of its components to support the Albanian language, owing to its modular and extensible nature [23]. To assist the NLP tasks in the Albanian language, we manually constructed two dedicated domain-specific corpora. The first corpus (CORPUS 1) included keywords and expressions in Albanian related to accident types, consequences, and severity (e.g., "aksident", "plagosje", "vdekje", etc.), while the second corpus (CORPUS 2) consisted of all known cities, towns, main villages, and roads of Albania, based on governmental administrative resources. Both corpora provided significant assistance for semantic filtering and geolocation tagging. A Python script utilizing these corpora processed the large base of news articles into a filtered set of 2,286 articles related to road accidents. The filtered database is structured into a tabular data frame with the following columns: number, date, region, location, injured, dead, number of involved vehicles, type of road, gender, cause, relevance score, and gravity of the accident. In addition to these natural attributes, an extra attribute was added to express the relevance score of each article. This score is calculated during the third step of preprocessing. The type of road is a categorical attribute with the following possible values: highway, interurban, urban, and rural. The primary cause of each accident is assessed based on text processing of the news article contents, aiming to classify it into one of the predefined causes: careless driving, atmospheric conditions, driving under the influence, wrong overtaking, speeding, running a red traffic light, or vehicle malfunctioning. The gender of the person causing the accident has been extracted through text processing, with values M (male), F (female), or N/A (not determined by the contents of the article). Gravity ranges from 1 (light) to 8 (severe), according to the number of people and vehicles involved in the accident and the number of casualties and injuries resulting from it.

Step 3: A further preprocessing step is applied to merge the articles referring to the same accident. During this stage, a relevance score is evaluated for each accident according to the number of news resources referring to it and the matches among these accidents. Furthermore, refinement is applied to mismatching data, using a weighted majority vote according to the credibility score of the news portals. The credibility score was manually assigned, varying from 5 (very well known) to 1 (less known) portals.

Step 4: After the application of the four aforementioned preprocessing stages, which operate in an automated way, a final stage of preprocessing is conducted, consisting of manual control and refinement of the obtained dataset. The control was primarily performed on data where mismatches occurred to confirm that the algorithm based on the majority vote score properly selected the most reliable version of the attributes. Furthermore, the dataset was refined to fill some cases of missing values via brief human inspection of the source news articles.

In summary, a set of 1854 instances was obtained at the end of the third stage, containing 183 articles where mismatching data occurred. Manual inspection identified 34 of these articles. Furthermore, there were 124 cases of missing values in the dataset, 40 of which were manually inspected.

Figure 1 schematically illustrates the workflow of the steps described above to obtain the final dataset related to road accidents in the state of Albania.

The program for web scraping, which reads articles from online portals and stores them in the Large CSV Articles Database file, is executed initially. The script then reads all articles from the Large CSV Article Database file, filters those related to road accidents within the state of Albania, and stores them in another CSV file named Accidents Raw. The script was trained with Albanian language data from the CORPUS1 and CORPUS2 files (which were described above).

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 10: 992-1004, 2025
DOI: 10.55214/2576-8484.v9i10.10581

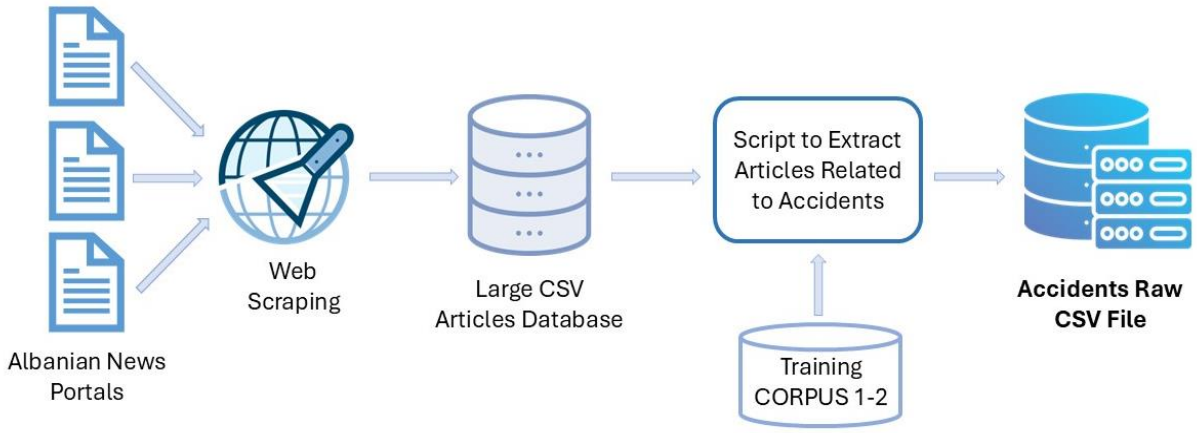

Figure 1.
Application Model for Data Preprocessing.

After preprocessing, the data contains 11 features, namely, number, date, region, address, type of road, age, gender, cause, relevance score, and gravity. More detailed descriptions of these fields are presented in Table 1. These data served as the input for the classification models.

Table 1.
Features after the preprocessing stage.

Feature	Description	Range
Date	The date of the publication of the road accident	01/01/2022 – 31/12/2023
Region	The region of the road accident	Albanian regions
Location	An approximate description of the address	Albanian locations according to CORPUS2
Injured	The number of injured persons in the accident	Nonnegative Integer
Dead	The number of deceased persons in the accident	Nonnegative Integer
Cars Involved	The number of vehicles involved in the accident	Nonnegative Integer
Type of Road	Types of roads in the Albanian state	Interurban, Urban, Rural, Highway
Cause	Categorizing the cause into one of the 7 foreseen categories (as given in the Range column)	Careless driving, Atmospheric conditions, Driving under the influence, Wrong overtaking, Speeding, running through a red traffic light, or vehicle malfunctioning.
Gender	The gender of the person causing the accident	M, F, N/A
Relevance Score	A numerical value based on the number of resources reporting the accident and the compatibility of their data	Nonnegative Integer
Gravity	A numerical value assessing the gravity of the accident	1 (light) to 8 (severe)

3.2. Classification Model

This model is applied to novel, unseen records, and on the basis of their features, their classes are determined. Classification has been successfully utilized in a wide variety of domains, including diagnostics in medicine, spam detection and sentiment analysis in natural language processing, fraud detection and risk assessment in finance, threat assessment in cybersecurity, and environmental monitoring.

The workflow of a classification task typically involves several key steps, such as data preprocessing, model selection, training, testing, evaluation, and prediction. During the preprocessing stage, the data are cleaned and transformed into a suitable format for training. Subsequently, the training model (i.e., the training algorithm) is selected and applied to the dataset to infer the decision

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 10: 992-1004, 2025
DOI: 10.55214/2576-8484.v9i10.10581

boundaries or rules for the separation of classes. Several options are available for the training model, including naïve Bayes classifiers, logistic regression, decision trees, support vector machines, ensemble techniques, or neural networks. The choice of training model generally depends on the nature of the problem and the complexity of the data, with multiple models often being applicable. The inferred model is then applied to test data to assess various performance evaluation metrics, such as accuracy, precision, recall, and F1 score. Once the performance evaluation indicates that the algorithm is capable of generalizing well to new data, the ultimate goal of classification is achieved: determining the classes of previously unseen data.

In this work, an extensive application of four classification methodologies is conducted: the naïve Bayes classifier, random forests, and two gradient-based boosting methodologies: the XGBoost and LightGBM algorithms. After the preprocessing stage, the training/testing split procedure is applied to the transformed training dataset with a ratio of 80% (training) - 20% (testing). The results of the classification model are compared with those of the original training set to evaluate the performance based on several performance metrics, such as accuracy, precision, recall, and logarithmic loss.

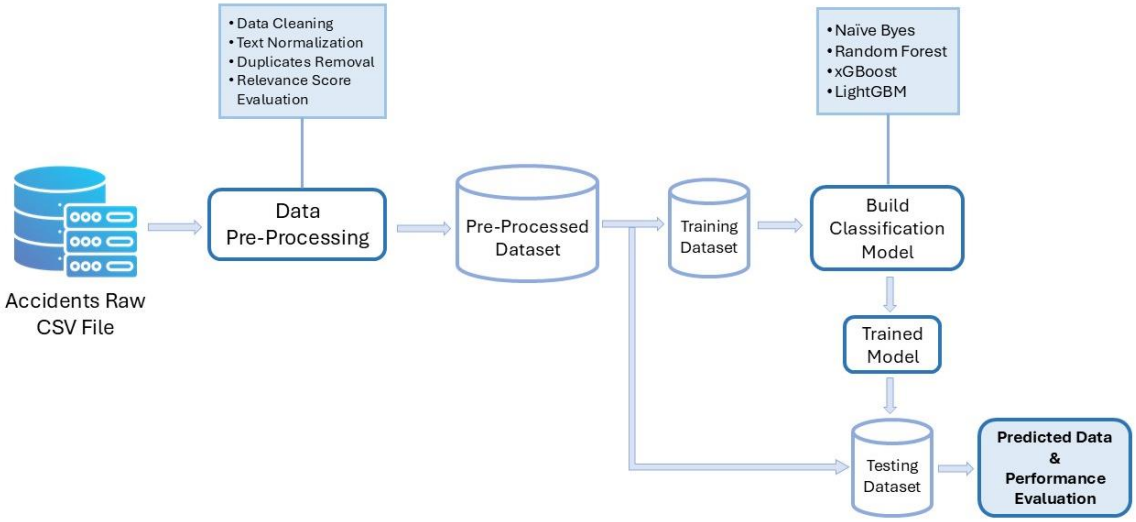

Figure 2.
Application of the classification models.

3.2.1. Naïve Bayes Classifier

The naïve Bayes classifier handles the classification problem via a probabilistic approach. The evaluation of the joint probability distribution values is significantly simplified by the 'naïve' assumption that the features are conditionally independent. This simplification makes the naïve Bayes classifier computationally efficient even for large and high-dimensional datasets. The conditional probabilities for the new entry belonging to each class are evaluated on the basis of Bayes' rule, and the maximal value among these probabilities determines the class assignment of the new entry [24]. More specifically, the evaluation of the conditional probabilities is handled as follows.

$$P(c = C_i | a_1 = v_1, a_2 = v_2, \dots, a_n = v_n) = \frac{P(a_1 = v_1, a_2 = v_2, \dots, a_n = v_n | c = C_i)}{P(a_1 = v_1, a_2 = v_2, \dots, a_n = v_n)} \quad (1)$$

Here V_1, V_2, …, V_n represent the values of the features of the new entry and the above calculation is repeated for each of the classes C_1, C_2, …, C_m of the training set. As the denominator in the above calculations is precisely the same for all classes and the objective is to find the maximum value, then the procedure can be carried out just by evaluating the numerators of these fractions. Despite being a fairly simple classification model, Naïve Bayes classification operates reasonably well in practice, making it a

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 10: 992-1004, 2025
DOI: 10.55214/2576-8484.v9i10.10581

significant tool in classification problems.

3.2.2. Random Forests

The central idea of ensemble methods is constructing multiple classification models simultaneously and yielding the final decision by aggregating the results of these classifiers. Each component classifier, known as a base learner, can be a simple model like a decision tree or a more complex model like a support vector machine. The results provided by the base learners can be combined into a single classification result by various methods, such as majority vote, weighted vote, averaging probabilities, etc. The rationale behind the ensemble methodology is that the aggregation of different methodologies eliminates errors caused by noise, biases, or variance in any particular base learner. Ensemble methods are used successfully in both classification and clustering problems, significantly increasing the stability of performance metrics, but at a higher computational cost [25].

The random forests classification method is an ensemble method that operates by constructing several decision trees simultaneously during the training phase. Each of these decision trees is built using a randomly selected subset of the original training set, via a procedure known as bootstrap aggregation (bagging). This randomness in data selection provides diversity in the generated decision trees. Additionally, the diversity is further enhanced as the splits during the construction of the trees are conducted by picking random subsets of features rather than the entire spectrum of features. The enhanced diversity enables the model to be not only more resistant to noise and biases but also much less susceptible to overfitting. Other significant strengths include the capability to operate on combinations of both categorical and numerical features, management of imbalanced datasets, and handling of missing values. The random forests method has been widely adopted in practice due to its ease of use, versatility supported by its set of strengths, and capability of dealing with both classification and regression problems [26].

3.2.3. XGBoost

Both XGBoost and LightGBM belong to the gradient boosting ensemble methods. The central idea here is still the construction of multiple base learners, but unlike bagging ensembles (e.g., Random Forests) where the trees are constructed independently, here, trees are constructed in an iterative manner, each focusing on rectifying the errors made by the previous one by optimizing a specific loss function.

Extreme Gradient Boosting (XGBoost) is a powerful gradient boosting method that uses a loss Function to guide the tree construction process. The loss function assesses the distances between predicted values and the actual values (known as residual errors), and the gradient boosting approach minimizes this loss iteratively. At each iteration, a new tree is constructed, which strives to reduce the residual errors manifested in the previous tree, thus, the overall performance of the model is gradually improved. This approach is characterized by very decent efficiency as it employs a histogram-based accommodation of the splits, which is faster than classical methods. Moreover, it is very well scalable, providing excellent support for parallel processing during the training phase. Another important feature of the XGBoost algorithm is the usage of regularization in two forms: L1 (Lasso) and L2 (Ridge), restraining the complexity of the trees to avoid overfitting the training data. The model includes several hyperparameters (like the number of trees, depth of trees, and learning rate), which allow for the adjustment of a balance among accuracy and computational complexity according to the priorities of the scenario [27, 28].

3.2.4. LightGBM

The LightGBM algorithm is another gradient boosting ensemble method that is devised focusing on being light in terms of memory usage and computational complexity, typically being faster compared to other ensemble methods. The central idea here is still the construction of multiple base learners, where the trees are constructed in an iterative way, each tree focusing on rectifying the errors made by

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 10: 992-1004, 2025
DOI: 10.55214/2576-8484.v9i10.10581

the previous one, via optimization of a specific loss function. The central mechanism is a histogram-based algorithm that splits the data more efficiently, reducing the memory usage and computation time [29].

One of the distinctive operational features of LightGMB is its leaf-wise growth approach. While other boosting algorithms grow the decision tree level by level, LightGBM grows leaf-wise, expanding the most promising leaves to reduce the loss function. Consequently, the generated trees have a larger depth and are typically more complex, improving performance measures such as accuracy. However, this approach involves an increased risk of overfitting, which is typically handled using parameters such as max_depth to control the growth of the tree by limiting the largest depth it can achieve [30].

For each of the aforementioned classification models, several parameters can be controlled. In our application, the GridSearchCV utility is employed to determine the best combination of the primary parameters. This utility is a hyperparameter tuning technique that systematically checks all the possible combinations from a predefined set of candidate values for the parameters. This technique practically trains the model and assesses its performance via cross-validation, taking the average performance among the folds as the final result. Table 2 summarizes the main parameters and their respective values for each classification model.

Table 2.
Main parameters for the classification models

Classification Model	Parameter 1 Name & Value	Parameter 2 Name & Value	Parameter 3 Name & Value
Naïve Bayes	alpha = 1.0	fit_prior = True	class_prior = None
Random Forest	n_estimators = 50	max_depth = 10	min_samples_leaf = 5
XGBoost	learning_rate = 0.1	n_estimators = 200	max_depth = 5
LightGBM	n_estimators = 100	num_leaves = 50	subsample = 0.7

4. Results

As described in the Methodology section, first, the data collection and data preprocessing stages are carried out. Data collection was implemented via web scraping, initially collecting a large base of news from Albanian news portals. Then, an intensive preprocessing stage involving data filtering, text processing, and data refinement was implemented. The outcome of this stage will be a cleaned tabular dataset, with each row corresponding to a particular accident containing features such as date, region, location, injured, dead, involved cars, cause, gender, road type, relevance score, and gravity. An overview of this table is provided in Figure 3.

After the second phase, four classification models are applied to the preprocessed data: naïve Bayes, random forests, XGBoost, and LightGBM. For each algorithm, the dataset is first randomly split into training and testing data, with the classification model being built relying solely on the training data. The target attribute of these classification models in all cases is the gravity scale of the accident. The generated classification models are then applied to the test data to assess their performance. The employed performance measures are accuracy, precision, recall, and logarithmic loss. The results of these evaluations are summarized in Table 3.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 10: 992-1004, 2025
DOI: 10.55214/2576-8484.v9i10.10581

	Date	Region	Location	Injured	Dead	Cars_Involved	Cause	Gender	Road_Type	Relevance_Score	Gravity
0	01/01/2022	Elbasan	rruga gostime mollas	1	0	2	Driving under inf...	M	Urban	3	3
1	01/02/2022	Kruje	aksi fushe kruje-thumane	1	0	1	Carelessness	M	Urban	3	3
2	01/02/2022	Lushnje	mbikalimi i savres	3	0	1	Wrong overtaking	M	Interurban	3	5
3	01/02/2022	Fier	autostrada vlore-fier	1	0	1	Atmospheric c...	M	Highway	3	3
4	01/03/2022	Sarande	xarres	1	0	1	Carelessness	M	Urban	2	3
...	...	...	...	...	...	...	...	...	...	...	...
1849	12/29/2023	Librazhd	aksi librazhd-stebleve	0	1	1	Carelessness	M	Urban	1	6
1850	12/29/2023	Fier	aksi fier-seman	0	0	2	Non-compliance...	NaN	Interurban	1	2
1851	12/29/2023	Tirane	afersi muzeu kombetar	1	0	1	Wrong overtaking	M	Urban	2	3
1852	12/30/2023	Lezhe	NaN	0	0	1	Speeding	M	Highway	3	1
1853	12/30/2023	Durres	aksi rini-shenavlash	0	0	1	Carelessness	M	Urban	3	1

1854 rows × 11 columns

Figure 3.
Overview of the preprocessing stage results.

Table 3.
Performance evaluation of the classification models.

Classification Model	Accuracy	Precision	Recall	Logarithmic Loss
Naïve Bayes	0.77	0.78	0.76	0.09
Random Forest	0.81	0.80	0.82	0.075
XGBoost	0.84	0.83	0.84	0.05
LightGBM	0.85	0.85	0.86	0.045

For each classification model, the preprocessed dataset was randomly split into training and testing datasets at a ratio of 80% training–20% testing, and this routine was repeated 200 times. The values of the performance measures (accuracy, precision, recall, and logarithmic loss) summarized in the given table represent the geometric averages of these 200 executions. Notably, the ensemble methods (Random Forests, XGBoost, and LightGBM) have demonstrated higher performance measures, with LightGBM being slightly superior to the other methods. On the other hand, the ensemble methods are characterized by higher computational costs, especially when large values of number of estimators are used. Nevertheless, the computational complexity was not the focus of this study.

5. Discussion

The findings of this study demonstrate the feasibility and utility of the methodology of data from online news portals with machine learning techniques to assess and predict the road accident risks in Albania. This was achieved through the implementation of a data mining pipeline starting with web scraping, proceeding with a complex sequence of preprocessing modules, and following with classification techniques and model evaluation.

BeautifulSoup served as a valuable tool for web scraping, providing a broad collection of accident-related content. However, significant challenges were encountered during the preprocessing phase due to the lack of natural language processing (NLP) tools designed for the Albanian language. In this context, the widely used English-language library NLTK was used with additional Python scripts customized to assist in recognition and text processing in the Albanian language. Furthermore, two domain-specific corpora were prepared to assist in semantic keywords detection and geographic location tagging. A hybrid approach combining automated preprocessing and final manual refinement improved data quality and enabled reliable feature extraction, including cause, gender, type of road, and severity level of the accidents. Considering the complexity of building structured datasets from unstructured online news text, especially in the context of the Albanian language, preprocessing represents a key methodological contribution of this work.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 10: 992-1004, 2025
DOI: 10.55214/2576-8484.v9i10.10581

In the classification phase, four distinct machine learning models were implemented, including the basic Naïve Bayes method and three ensemble methods, namely Random Forest, XGBoost, and LightGBM. The results demonstrated a better performance score of ensemble methods, especially LightGBM, which achieved high scores across all used evaluation metrics. These results indicate the effectiveness of boosting-based models in capturing complex, nonlinear relationships among the features of a dataset. Nevertheless, even the Naïve Bayes method, due to its simplicity and computational efficiency, remains a useful benchmark and a viable option for real-time systems with constrained resources.

LightGBM's superior performance is likely attributable to the nature of its histogram-based learning algorithm and leaf-wise tree growth strategy, which enable efficient capturing of complex attribute relationships and handling of imbalanced data. Additionally, its scalability and capacity to handle high-dimensional numerical and categorical characteristics make it particularly well-suited for diverse feature sets generated as the outcome of the preprocessing stage.

Building a classification model from web-scraped data to predict the likelihood of road accidents according to circumstances is a highly challenging procedure. Therefore, the performance metrics achieved by the applied classification algorithms, especially by ensemble classifiers, are regarded as not only promising but also technically robust.

Despite these promising results, there are some limitations that should be acknowledged. As the dataset is derived from media sources, it contains the potential for inherent biases in news reporting. Often, more severe or sensational accidents are more likely to be covered extensively by news portals, leading to the under-representation of minor or routine accidents, consequently skewing the distribution of variables.

From a practical perspective, this work demonstrates that it is possible to design early warning systems or dashboards that automatically track accident risk levels using real-time data from news portals. Such systems can potentially serve as support tools for road safety campaigns, public awareness initiatives, and infrastructure planning.

6. Conclusion

The work presented in this paper involved designing a multi-staged data mining pipeline for predicting the likelihood of road accidents based on specific circumstances. The pipeline comprised data collection, which was conducted through web scraping; data preprocessing, involving various tools such as filtering, text normalization, and data refinement; and a classification stage, executed using several algorithms, including naïve Bayes, random forests, XGBoost, and LightGBM. Multiple classification models were tested within this framework, and the performance evaluation metrics indicated that ensemble classifiers, particularly LightGBM, were the most effective models.

Although there have been similar works in various other languages, this is one of the few works on the Albanian language.

The outcome of this work may assist in the prevention of accidents by identifying areas with the highest risk in advance, thereby providing important information for both governmental institutions and citizens to enhance security at operational and strategic levels.

Transparency:

The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

Copyright:

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 10: 992-1004, 2025
DOI: 10.55214/2576-8484.v9i10.10581

References

[1] INSTAT, *Transportation, accidents and characteristics of road vehicles*. Tiranë, Albania: Instituti i Statistikave - Tiranë, 2024.

[2] Ministry of Interior, "Raporti mujor monthly report," Ministry of Interior in Albania, 2024. https://mb.gov.al/wp-content/uploads/2024/12/Buletini-Mars-2024.pdf

[3] L. Sinanaj and L. A. Bexheti, "Predicting road accidents with web scraping and machine learning techniques, in: Economic recovery, consolidation, and sustainable growth," in *ISCBE 2023.Proceedings of the 6th International Scientific Conference on Business and Economics (ISCBE 2023), Springer, Cham*, 2023.

[4] L. Sinanaj, E. Bedalli, and L. Abazi Bexheti, "A classification model for predicting road accidents using web data," *ENTerprise REsearch InNOVAtion*, vol. 9, no. 1, pp. 50-61, 2023. https://doi.org/10.54820/entrenova-2023-0006

[5] S.-P. Miaou, "Estimating vehicle roadside encroachment frequency using accident prediction models," presented at the Conference: 76. Annual Meeting of the Transportation Research board., Washington, DC (United States), 1996.

[6] S.-P. Miaou and H. Lum, "Modeling vehicle accidents and highway geometric design relationships," *Accident Analysis & Prevention*, vol. 25, no. 6, pp. 689-709, 1993. https://doi.org/10.1016/0001-4575(93)90034-T

[7] R. Furlonge, "Road fatality modelling in trinidad and tobago," *West Indian Journal of Engineering*, vol. 19, no. 1, pp. 10-15, 1996.

[8] J. Kononov, *Road accident prediction modeling and diagnostics of accident causality: A comprehensive methodology*. Denver: University of Colorado at Denver, 2002.

[9] L. Larsen, "Methods of multidisciplinary in-depth analyses of road traffic accidents," *Journal of Hazardous Materials*, vol. 111, no. 1-3, pp. 115-122, 2004. https://doi.org/10.1016/j.jhazmat.2004.02.019

[10] M. Lnenicka, J. Hovad, J. Komarkova, and M. Pasler, "A proposal of web data mining application for mapping crime areas in the Czech republic," presented at the International Joint Conference on Software Technologies (ICSOFT), Colmar, France, 2016.

[11] A. A. Biswas, M. J. Mia, and A. Majumder, "Forecasting the number of road accidents and casualties using random forest regression in the context of Bangladesh," presented at the International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, 2019.

[12] M. A. B. Siddik, M. S. Arman, A. Hasan, M. R. Jahan, M. Islam, and K. B. B. Biplob, "Predicting the death of road accidents in Bangladesh using machine learning algorithms," presented at the International Conference on Advances in Computing and Data Sciences, 2021.

[13] L. Chen and P. Wang, "Risk factor analysis of traffic accident for different age group based on adaptive boosting," presented at the International Conference on Transportation Information and Safety (ICTIS), 2017.

[14] H. Zhao, H. Yu, D. Li, T. Mao, and H. Zhu, "Vehicle accident risk prediction based on AdaBoost-SO in VANETs," *IEEE Access*, vol. 7, pp. 14549-14557, 2019. https://doi.org/10.1109/ACCESS.2019.2894176

[15] R. E. AlMamlook, K. M. Kwayu, M. R. Alkasisbeh, and A. A. Frefer, "Comparison of machine learning algorithms for predicting traffic accident severity," presented at the 2019 IEEE Jordan International joint Conference on Electrical Engineering and Information Technology, 2019.

[16] S. Elyassami, Y. Hamid, and T. Habuza, "Road crashes analysis and prediction using gradient boosted and random forest trees," presented at the IEEE Congress on Information Science and Technology (CiSt), 2021.

[17] M.-M. Chen and M.-C. Chen, "Modeling road accident severity with comparisons of logistic regression, decision tree and random forest," *Information*, vol. 11, no. 5, p. 270, 2020. https://doi.org/10.3390/info11050270

[18] M. Yan and Y. Shen, "Traffic accident severity prediction based on random forest," *Sustainability*, vol. 14, no. 3, p. 1729, 2022. https://doi.org/10.3390/su14031729

[19] I. Karamanlis, A. Kokkalis, V. Profillidis, G. Botzoris, and A. Galanis, "Identifying road accident black spots using classical and modern approaches," *WSEAS Transactions on Systems*, vol. 22, pp. 556-565, 2023.

[20] S. Dong, A. Khattak, I. Ullah, J. Zhou, and A. Hussain, "Predicting and analyzing road traffic injury severity using boosting-based ensemble learning models with SHAPley Additive exPlanations," *International journal of environmental research and public health*, vol. 19, no. 5, p. 2925, 2022. https://doi.org/10.3390/ijerph19052925

[21] G. Mahendra and H. R. Roopashree, "Prediction of road accidents in the different states of India using machine learning algorithms," presented at the 2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS), Raichur, India, 2023.

[22] L. Richardson, "Beautiful soup. Crummy," 2024. https://www.crummy.com/software/BeautifulSoup/

[23] nltk.org, "Natural language Toolkit," 2024. https://www.nltk.org/

[24] F. J. Yang, "An implementation of naive bayes classifier," presented at the 2018 International Conference on Computational Science and Computational Intelligence (CSCI), 2018.

[25] E. Bedalli, E. Mançellari, and O. Asilkan, "A heterogeneous cluster ensemble model for improving the stability of fuzzy cluster analysis," *Procedia Computer Science*, vol. 102, pp. 129-136, 2016. https://doi.org/10.1016/j.procs.2016.09.379

[26] Y. Ren, L. Zhang, and P. N. Suganthan, "Ensemble classification and regression-recent developments, applications and future directions," *IEEE Computational Intelligence Magazine*, vol. 11, no. 1, pp. 41-53, 2016. https://doi.org/10.1109/MCI.2015.2471235

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 10: 992-1004, 2025
DOI: 10.55214/2576-8484.v9i10.10581

[27] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM sigkdd International Conference on Knowledge Discovery and Data Mining*, 2016.

[28] P. Zhang, Y. Jia, and Y. Shang, "Research and application of XGBoost in imbalanced data," *International Journal of Distributed Sensor Networks*, vol. 18, no. 6, p. 15501329221106935, 2022. https://doi.org/10.1177/15501329221106935

[29] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA*, 2017.

[30] O. Alshboul, G. Almasabha, A. Shehadeh, and K. Al-Shboul, "A comparative study of LightGBM, XGBoost, and GEP models in shear strength management of SFRC-SBWS," *Structures*, vol. 61, p. 106009, 2024. https://doi.org/10.1016/j.istruc.2024.106009

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 10: 992-1004, 2025
DOI: 10.55214/2576-8484.v9i10.10581