Analysis of road accident factors using Decision Tree Algorithm: a case of study Algeria

Ouennoughi Nedjmedine
Mohamed Boudiaf University
M'sila, Algeria
nedjmedine.ouennoughi@univ-msila.dz

Mehenni Tahar
Mohamed Boudiaf University
M'sila, Algeria
tahar.mehenni@univ-msila.dz

***Abstract*—Road accidents become a worldwide health issue. With the enormous number of death and injuries, this problem pushes governments to create solutions to reduce those statistics. One of the solving ways is using machine learning algorithms, and with the data collected from road accidents, we can increase traffic safety. In this research, we use a decision tree model to analyze road accidents that happened in Algeria. Then, we do a comparison with some similar works using accuracy as a performance evaluation metric. This work can help government and traffic safety entities to improve road safety and minimize the number of accidents, also, it can help other researchers to develop other models in the analysis of traffic accidents in Algeria and other countries.**

Keywords—machine learning, road accident, decision tree, data analysis.

I. INTRODUCTION

Every day, many people die around the world for different reasons, some die from diseases like cancer, and others die from natural disasters like earthquakes. Thus, there are many reasons but the victim is one, a "human being", and one of the main causes of human death is road accidents.

The statistics show that road and traffic accidents (RTA) are one of the most fatal causes of death worldwide. According to the World Health Organization (WHO) [1], over 1.3 million people die from road traffic crashes with around 50 million people who suffer from non-fatal injuries. Those statistics push governments to preserve these accidents as a worldwide public health issue. To reduce those enormous numbers, many solutions had been proposed to predict and avoid those fatal crashes.

Many studies have been carried out by different researchers to propose and create prediction and analytic models using machine learning algorithms which provide the ability to analyze road accident factors to define patterns and give a real-time and accurate prediction model of any traffic road crash.

This paper aims to analyze road accident factors in Algeria using the decision tree algorithm on a dataset of road accidents that happened in 2021 and perform a comparison with similar works using the metric of accuracy and discuss the result obtained.

The skeleton of this paper contains 6 sections: Section 2 shows a review of the state of the art in the analysis and prediction of road accidents. Section 3 is about decision tree classification: what is a decision tree, how does it work, and what are the attributes selection measures of this technique. Section 4 discusses our proposed model: we start this section by defining the sources that provide data and the availability of each source in Algeria. Then, we describe the body of our dataset. After that, we explain how our work was built. Section 5 provides a discussion of the result obtained and a comparison with some similar works. Section 6 covered the conclusion of this paper with perspectives to emphasize the accuracy of the analysis methods.

II. RELATED WORK

The section below will mention some related work about analyzing road accidents using data mining techniques.

Camilo Gutierrez-Osorio and Cesar Pedraza [2] presented an overview of the state of art in the analysis and the perdition of road accident factors using machine learning algorithms and advanced techniques. Besides that, they discussed what are the analysis and prediction methods, what are the most used, and the aim of each one, and mentioned the different available data sources to collect data from. They conclude their article that combining two or more methods is the best way to get better results. As a perspective, they aim to inject heterogeneous data to improve the accuracy and precision of the different analysis and prediction algorithms.

Addi Ait Moulouk et al. [3, 4] proposed a technique based on an association rules model with the integration of the multi-criteria decision analysis system using ELECTRE TRI to analyze road accidents in Morocco. This technique gives decision-makers the ability to select their own decision rules among a large number of rules to prevent accidents and improve road safety as the main goal.

Sachin Kumar and Durga Toshniwal [5] suggested an approach based on two algorithms to characterize road accident locations, they used the k-means algorithm to group the accident locations into three categories depending on their frequency, then they used the association rules algorithm to detect the most revealed factor in each location.

Manoj Kushwaha and M. S. Abirami [6] suggested a comparative analysis on the prediction of road accident severity which helps researchers and the public work department (PWD) and is useful for the Indian government to identify and minimize road accidents. They used different machine learning (ML) algorithms with the severity of accidents as the main factor for this study, accuracy as a performance metric, and integrated the internet of things (IoT) with ML. They found that the random forest (RF) is the best algorithm.

Sumbal Malik et al. [7] developed a prediction framework and implemented six different ML algorithms, using a dataset published by the UK government. They found that Random Forest is the best in terms of all performance metrics. They also believed that the proposed framework and

2022 5th International Symposium on Informatics and its Applications (ISIA) | 978-1-6654-7473-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIA55826.2022.9993530

discovered patterns will help the authorities to improve the safety of roads and prevent accidents.

Buket Geyik and Medine Kara [8] suggested a severity prediction using Stats19, a UK dataset. By applying four classification methods on it with prior knowledge as Trainor data, they classified objects and compared the accuracy of each technique. They believe it is impossible to stop road accidents but their work will help to reduce traffic injuries.

Syed Ibrahim Kabeer [9] applied the Decision Tree, Naive Bayes, and Ensemble technique (Bayes Boosting) on the dataset of Leads in the UK. They compared the accuracy of these methods to determine the best one that can be used to predict road accidents.

Mohammad Rahmaninezhad Asil et al. [10] developed an analysis method based on the CART algorithm for vehicle-to-vehicle accidents with light conditions as the main factor. This study focused on the severity of accidents in Rasht city in Iran as a target variable.

Poojitha Shetty et al. [11] suggested an analysis of road accidents using association and classification rules to discover patterns in road accidents and predict the type of accidents for the existing and new roads.

We propose to develop an analysis model of traffic accidents in Algeria using a decision tree technique. We perform our study on a dataset of the accidents that happened in 2021 and focus on the fatality of the accident.

III. Decision Tree Algorithm

The main goal of a decision tree is to classify objects and situations by analyzing previous data with known classes[12].

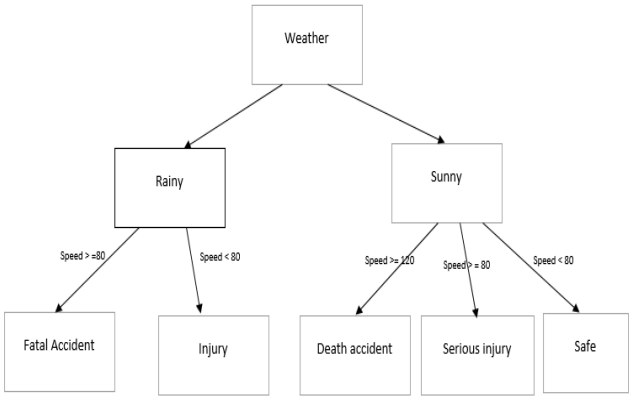
A decision tree **(DT)** is a supervised learning model that maps a data domain hierarchically to a response set. It is used in both **classification** and **regression** algorithms[13]. It is like a real inverted tree that contains sub-trees, each tree grows from top to down from the **root** down to the **leaves**. Besides **nodes** and **edges**, **root** and **leaves** define the elements of a decision tree (cf. Table I).

TABLE I. ELEMENTS OF DT

Element	Description
Root	is the main node where the tree begins from the top
Node	is where the split happens according to the value of a specific feature of the dataset.
Edge	It directs the outcome of a split between two nodes.
Leaf	is the final node up to down of DT which predicts the outcome.

The following example demonstrates a decision tree model of the fatality of road accidents based on speed and weather conditions. It is easy to define which are the factors that lead to a fatal accident by implementing this model with the dataset.

Fig. 1. Example of DT Model.



A. Attribute Selection Measure

The attribute selection measure (ASM) is a technique that helps to select the best attribute among all features of the dataset, as a root at the beginning of the decision tree building, then selects the best feature as the next node respectively until reaching the leaves of the tree finishing by the outcome. We can find two techniques of ASM in the literature [14]:

- **Information gain (entropy):** Information gain is the difference between a class entropy and the class conditional entropy and the selected feature.

$$Gain = E_{class} - E_{selected\ attribute} \quad (1)$$

Entropy measures the extent of impurity or randomness in a dataset where N is the number of classes and p_i is the probability that a tuple belongs to class C_i.

$$E = -\sum_{i=1}^{N} P_i \log_2 P_i \quad (2)$$

- **Gini index:** is defined as a measure of the purity of a specific class. It only creates binary splits, each attribute with a low Gini index value should be preferred. N and p_i are the same parameters as presented in entropy.

$$GINI = 1 - \sum_{i=1}^{N} p_i^2 \quad (3)$$

B. Algorithms of Decision Tree

According to[13], there are 4 different algorithms of DT:

- **ID3:** is a simple algorithm proposed in 1986 by SQUINLAN which uses information gain to decide the splitting features.
- **C4.5:** is an extension of **ID3** which both are used to classify data.
- **CART:** The classification and regression tree was introduced by BREIMAN et al. in 1984. CART is a binary tree algorithm where the outcome of each node is two edges.

- **CHAID:** Chi-square Automatic Interaction Detection (CHAID) is an algorithm developed to deal with nominal attributes.

IV. Our Work

In this section, we present how our work was built, but before that, we talk about what are the sources which researchers gather data from, their availability in Algeria, and what is the main source that we choose as a base for our algorithm. Then we close this section by performing a comparison with other similar works.

A. Data

Data is considered the main base of any ML algorithm. It is the most important item/piece for each data science work. Researchers use these data to classify the data itself or discuss and analyze it or implement it in tests to discover new data or patterns. Therefore, it is important to find and collect the data to develop different works in Machine Learning. There exist many sources that allowed researchers and developers to gather and collect data. Gutierrez-Osorio and Pedraza [2] defined five types of data sources, each source contains its privileges.

- **Government data:** Data made (generated, collected, preserved, and stored) by the government entities like traffic police and the national road safety directorate, and offered to the public with some restrictions (E.g., the confidentiality of people involved in road accidents).
- **Open data:** Data that can be found in catalogs that are stored and presented on websites to the public without restriction in a condition that data must comply with all legislation regarding privacy and confidentiality. The most common catalogs are the U.S data gov [15] and the U.K data gov [16].
  Both government data and open data are governmental sources, the difference between these two sources is in the restrictions and the support which are presented in.
- **Measurement technologies:** Data gathered from equipment that can be found in road infrastructures such as radar, cameras, and sensors.
- **Onboard equipment:** Data collected from any equipment/device installed on a vehicle that gives the ability to store and collect data about the vehicle, driver and even the road information.
- **Social media:** which are considered the newest sources to collect data from. The most famous are Google maps and Twitter streams.

Table II mentioned the strength and weaknesses of each source mentioned previously.

B. Data sources in Algeria

We discuss the availability of each of the previous sources in Algeria and which one we used as researchers to develop our work and why.

- **Government data:** The main source to collect data from is the government entities such as the police department and the National Delegation for Road Safety (NDRS) but, not anyone has the privilege to access it, it needs to present as a researcher for example and shows what are the intentions and purposes of using these data. Furthermore, it needs to write a worn declaration for not sharing this data on social media or the net.
- **Open data:** there are no online Algerian catalogs available to the public to collect data from, the only available website is [17] and it was shut down in 2019 (according to the last saved version)
- **Measurement technologies:** most road infrastructure in Algeria is not equipped with measurement technologies, and the only technology that exists like radar or video recorders is preserved by the government's entities like the police, and access to data of these technologies is barely possible.
- **Onboard equipment:** according to the law of road safety in Algeria, onboard equipment is illegally used in the country, and any used equipment without permission subjects the user to penalties that may reach imprisonment.
- **Social media:** Social media is an unreliable source in Alegria because they cannot verify and trust the origin of the information published, they are biased. Moreover, according to[18-20], Facebook is the most used platform in Algeria and unlike Twitter, Facebook does not allow users the ability to collect data like Twitter stream, so it is impossible to choose social media as a source of data.

C. Data description

1) Data source: the best source that was suitable for our needs to develop our work was the government data. Therefore, we collected the data from the National Delegation for Road Safety (NDRS).

2) Data: We have 23409 instances and 47 attributes distributed in three axes: Human factor, Vehicle condition, road infrastructure and atmospheric conditions. Table III shows the axes with some attributes.

TABLE II. Description of The Dataset

Axes	Some attributes
Human factor	Speeding
	Driving without a license
	Dangerous overtaking
	Dangerous maneuvers
Vehicle condition	Lack of lighting
	Mechanical defects
	Defective tires (blowout)
road infrastructure and atmospheric conditions	Defective road
	Lack of warning signs
	Obstacles on the road

TABLE III. STRENGTHS AND WEAKNESSES OF DIFFERENT ACCIDENT DATA SOURCES

	Government data	Open data	Measurement technologies	Onboard equipment	Social media
Strengths	-Historical (stocked data for decades) - Reliable (trustable source)	- Reliable - Easy to access (Public websites)	- reliable, - accurate - ease of use.	-Reliable -Accurate	- Quick - easy to collect - Divergence of data
Weaknesses	- Difficult to access - restrictions on data	- requires the handling of missing values and normalization and transformation of data	- Took a long time to collect data "field study" - Limited data (exp: loop detector "data related only to the vehicle") - Absence of the tech in some road infrastructure	- Took a long time to collect data "field study" - Illegal to use in some countries	- Unreliable - Limited sources - Difficult to interpret with

D. *Tools*

1) Python: is a very high-level open-source programming language to write software codes easily with the possibility of being maintained and reused. It is optimized for speed of development and designed to be integrated with other tools. Most Python programs run in every computer system, so it is not necessary to change the system. Besides that, Python is so powerful to deal with a huge amount of numerical data. All these reasons make it preferred and popular among all the other programming languages [21, 22].

2) Jupyter Notebook: an open-source web application that allowed the development of codes in over 40 programming languages with the advantage of the visualization option. It gives developers the possibility to divide code into blocks. Data scientists use Jupyter to analyze data and to create and share documents, equations, visualizations, and text.

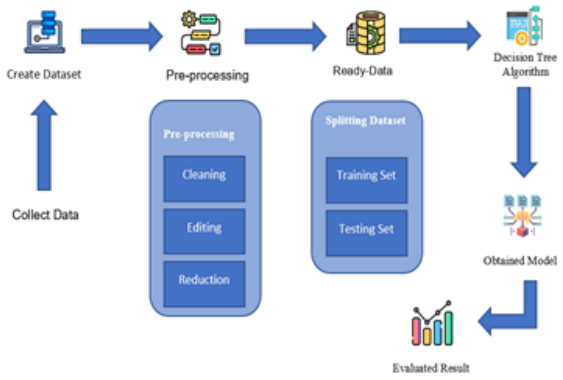
E. *Model*

We adopt a decision tree algorithm as our technique because, according to [2], DTs work well with large amounts of data, and their results can be presented graphically.

We used the ID3 algorithm with entropy as an attribute selection measure. We select the fatality attribute as a target feature with a fatal accident and no-fatal accident as its values. Figure 1 shows the steps of the proposed model.

We collect relevant data from the government source, the data was given as text saved in a Docx file, we transfer the data from the string to numeric values, and as a final result, we obtained a dataset in CSV format.

FIGURE 1. ROAD ACCIDENT ANALYSIS FRAMEWORK



F. *Preprocessing*

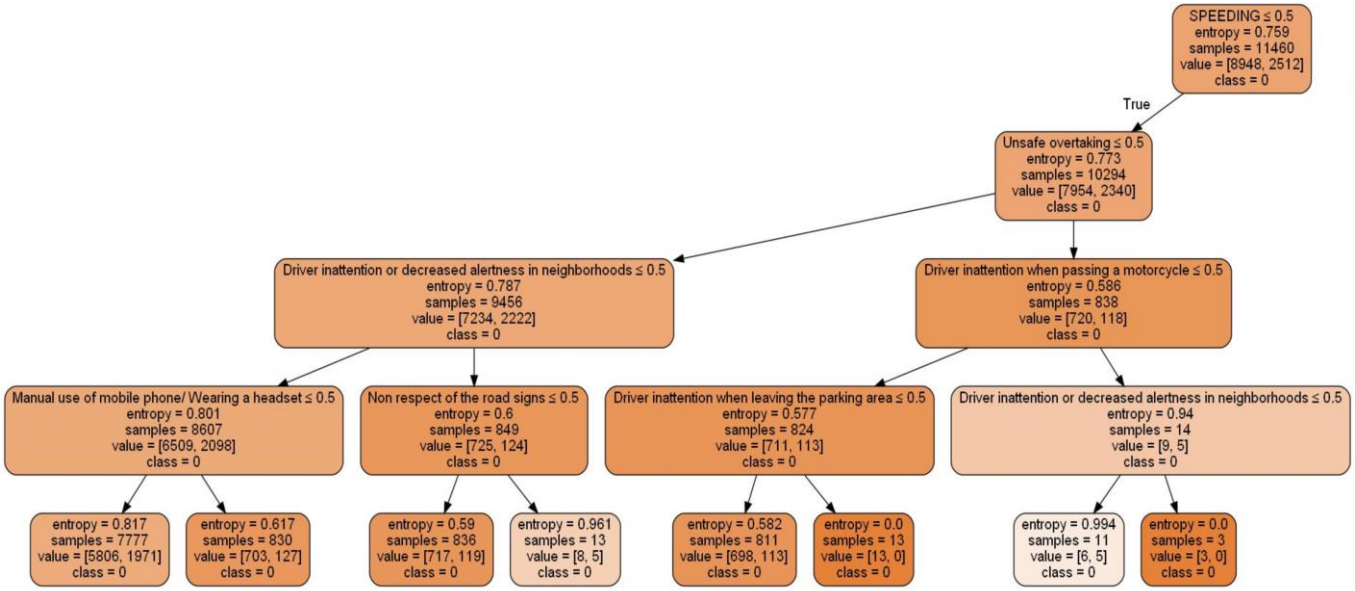
After we built the CSV file, we obtained a dataset with incomplete data, repeated and missing data. To have qualified data, and since we use python, importing useful python libraries is the second step of data processing: **NumPy, Pandas,** and **Matplotlib** are the most common libraries in the data science field. After that, we import the dataset that our work is based on, then, we need to apply the preprocessing and cleaning data which are the most important step of machine learning modeling by deleting incomplete data, duplicate data, and removing rows with null values.

After we handled the missing values, we check the categorial values and split the data into two sets: a **training set** and a **test set** with a ratio of 70:30, this means 70% of the data for the training set and 30% for the testing set.

V. DISCUSSION

In this section, we talk about the experiment in detail. We used Python as a programming language and the Jupyter notebook as an IDE. We realized this project on a computer with Windows 10 as an operating system, an Intel Xeon E3-1225 v5 as a processor, and with 16Mb RAM. We used accuracy as a performance evaluation metric which was

Figure 2. Visualization of our Decision Tree



calculated according to the values extracted from the confusion matrix where *TP* is true positive, *TN* is true negative, *FP* is false positive and *FN* false negative:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (4)$$

By applying the DT algorithm, we obtained a result with **78.125%** accuracy and **0.1037 s** as a run time. The whole process from creating the CSV file passing by the preprocessing to applying the DT took like **23 s** of run time.

Figure 2 shows a visualization of our Dt, each node contains the name of the feature, its entropy and sample which is the number of the apparition of the selected feature with the value of both classes of the target feature, those values refer to how many times the feature affected on the target attribute. In the end, we find in which class each feature belongs and whether it causes a fatal accident or not.

We performed a comparison with similar work found in the literature. As shown in table IV, it is evident that the result obtained compared to others is quite satisfying with a 78.125% of accuracy. On the other hand, our experiments were the fastest compared to the other works with a run time of 0.1037.

TABLE IV. COMPARISON RESULTS WITH SIMILAR STUDIES (N/A REFERS TO THE PARAMETERS NOT MENTIONED IN THE ARTICLE)

Algorithm	Accuracy (in %)	Run time (in seconds)
[7]	87.88	113.14
[8]	80.74	1.054
OUR WORK	**78.125**	**0.1037**
[23]	71.08	N/A
[9]	51.22	N/A

CONCLUSION

In this work, we built an analysis framework to analyze and predict the fatality of road accidents in Algeria. To realize this project, we used the decision tree algorithm.

First, we collect data from the government source which is the National Delegation for Road Safety (NDRS). After that, we made a preprocessing and cleaning of the data to get our final dataset. Then, we performed a comparison with similar works, the result obtained was quite satisfying.

This model can help the government to identify and minimize road accidents which is a worldwide health issue, and it can also help researchers to develop and create new projects by using this information.

As future work, we aim to use other performance evaluation metrics like precision, recall and F1-score. We also aim to develop hybrid models between the DT algorithm and another ML algorithm to increase the accuracy of our model.

REFERENCES

[1] World Health Organization. "Road traffic injuries." WHO. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries (accessed 30/11/2021.

[2] C. Gutierrez-Osorio and C. Pedraza, "Modern data sources and techniques for analysis and forecast of road accidents: A review," *Journal of Traffic and Transportation Engineering (English Edition),* vol. 7, no. 4, pp. 432-446, 2020/08/01/ 2020, doi: https://doi.org/10.1016/j.jtte.2020.05.002.

[3] A. Ait-Mlouk and T. Agouti, "DM-MCDA: A web-based platform for data mining and multiple criteria decision analysis: A case study on road accident," *SoftwareX,* vol. 10, p. 100323, 2019.

[4] A. Ait-Mlouk, F. Gharnati, and T. Agouti, "An improved approach for association rule mining using a multi-criteria decision support system: a case study in road safety," *European transport research review,* vol. 9, no. 3, pp. 1-13, 2017.

[5] S. Kumar and D. Toshniwal, "A data mining approach to characterize road accident locations," *Journal of Modern Transportation,* vol. 24, no. 1, pp. 62-72, 2016.
[6] M. Kushwaha and M. Abirami, "Comparative Analysis on the Prediction of Road Accident Severity Using Machine Learning Algorithms," in *Micro-Electronics and Telecommunication Engineering*: Springer, 2022, pp. 269-280.
[7] S. Malik, H. El Sayed, M. A. Khan, and M. J. Khan, "Road Accident Severity Prediction—A Comparative Analysis of Machine Learning Algorithms," in *2021 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT)*, 2021: IEEE, pp. 69-74.
[8] B. Geyik and M. Kara, "Severity prediction with machine learning methods," in *2020 international congress on human-computer interaction, optimization and robotic applications (HORA)*, 2020: IEEE, pp. 1-7.
[9] S. I. Kabeer, "Analysis of Road accident in Leeds," Dublin, National College of Ireland, 2016.
[10] M. R. Asil, H. Toroghi, and I. Bargegol, "Analysis of Factors Associated with Traffic Injury Severity on Urban Roads in Different Lighting Conditions," *Computational Research Progress in Applied Science & Engineering (CRPASE),* vol. 8, 2022.
[11] S. P. Poojitha Shetty, S. V. Kashyap, and V. Madi, "Analysis of road accidents using data mining techniques," *International Research Journal of Engineering and Technology (IRJET) e-ISSN,* pp. 2395-0056, 2017.
[12] J. R. Quinlan, "Learning decision tree classifiers," *ACM Computing Surveys (CSUR),* vol. 28, no. 1, pp. 71-72, 1996.
[13] S. Suthaharan, "Machine learning models and algorithms for big data classification," *Integr. Ser. Inf. Syst,* vol. 36, pp. 1-12, 2016.
[14] S. Tangirala, "Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm," *International Journal of Advanced Computer Science and Applications,* vol. 11, no. 2, pp. 612-619, 2020.
[15] T. T. S. U.S. General Services Administration. "Data Catalog " https://catalog.data.gov/dataset?q=traffic+accidents&sort=views_recent+desc&tags=crash&as_sfid=AAAAAAXHjZkDY7gFA5iMx_28NUE0FLt7GCD6A_wjSzainkj_rspLB-fqUew5h3LiHfKwq25Q1jllDf64k8tuEJ03xVdCKo4_qW6HRpHe_XBlCPYQhLUOwC0CkWT-WHXEHYKSTII%3D&as_fid=be93db12e7584b (accessed 07/05/2022.
[16] data.gov.uk. https://data.gov.uk/ (accessed 07/05/2022.
[17] cnpsr. http://www.cnpsr.org.dz/ (accessed 27/10/2021.
[18] SIMON KEMP. "DIGITAL 2022: ALGERIA." DataReportal. https://datareportal.com/reports/digital-2022-algeria (accessed 20/04/2022.
[19] statcounter. "Social Media Stats Algeria." StatcounterGlobalStats. https://gs.statcounter.com/social-media-stats/all/algeria (accessed 20/04/2022.
[20] Statista. "Number of social media users in Algeria as of 2021, by platform." Statista Research Department,. https://www.statista.com/statistics/1284853/number-of-social-media-users-in-algeria-by-platform/#:~:text=With%20around%2028%20million%20users,eight%20million%20Algerians%20used%20Instagram. (accessed 20/04/2022.
[21] W. Python, "Python," *Python Releases for Windows,* vol. 24, 2021.
[22] M. Lutz, *Programming python*. " O'Reilly Media, Inc.", 2001.
[23] S. Haynes, P. C. Estin, S. Lazarevski, M. Soosay, and A.-L. Kor, "Data analytics: Factors of traffic accidents in the uk," in *2019 10th International Conference on Dependable Systems, Services and Technologies (DESSERT)*, 2019: IEEE, pp. 120-126.