Analysis of Hidden Patterns in Road Accident Dataset Using Clustering Techniques

Henry O. Velesaca[1,3(✉)], Miguel Realpe[1], Angel D. Sappa[1,2], and Alice Gomez[1]

[1] ESPOL Polytechnic University, Escuela Superior Politécnica del Litoral, ESPOL, Campus Gustavo Galindo Km. 30.5 Vía Perimetral, P.O. Box 09-01-5863, Guayaquil, Ecuador
{hvelesac,mrealpe,asappa,alivagom}@espol.edu.ec

[2] Computer Vision Center, 08193 Bellaterra, Barcelona, Spain
asappa@cvc.uab.es

[3] Software Engineering Department, University of Granada, 18014 Granada, Spain
hvelesaca@correo.ugr.es, jholgado@ugr.es

Abstract. This study presents the application of various clustering techniques for the analysis of a traffic accident dataset. Methods such as Gaussian Mixture Model, Agglomerative Clustering, MiniBatchK-Means and K-Means are used to identify hidden patterns in the data, including features such as geographic coordinates, accident severity, cause and type of accident. The dataset is preprocessed by removing null values, encoding categorical variables, and robust scaling of features. Furthermore, PCA is applied to reduce the dimensionality of the dataset. The performance of the clustering techniques is evaluated using metrics such as the Silhouette Score, Davies-Bouldin Score and Calinski-Harabasz Score. The results indicate that K-Means, with 5 clusters, provides the best overall performance, according to the Elbow method and the evaluated metrics. Visualizations of 2D is included for a better interpretation of the clusters, highlighting the distribution and features of the groups formed. The code and dataset are available at Kaggle: https://www.kaggle.com/code/hvelesaca/smarttech-paper-id-42, facilitating further research.

Keywords: Machine learning · Clustering · MiniBatch K-Means · Traffic accidents · Unsupervised technique

1 Introduction

Traffic accidents are one of the main causes of mortality and injuries worldwide, and Ecuador is no exception to this tragic reality [2,9,11]. The growing concern about road safety and the urgency to reduce the number of accidents have led to the search for innovative and effective solutions. In this sense, data analysis and machine learning techniques are presented as powerful tools to understand and mitigate the factors that contribute to traffic accidents, allowing for more precise and timely intervention [1,7,10].

F. Narváez et al. (Eds.): SmartTech-IC 2024, CCIS 2393, pp. 96–111, 2026.
https://doi.org/10.1007/978-3-031-98290-3_8

This study focuses on the analysis of traffic accidents in Ecuador through the application of machine learning techniques. The dataset used comes from the accident statistics of the National Traffic Agency (ANT) of Ecuador[1], a reliable source that provides a detailed overview of road incidents in the country [8]. The main objective of this analysis is to identify underlying patterns and trends that can inform the design of policies and strategies aimed at improving road safety. Through the collection and exhaustive analysis of historical accident data, the aim is to build predictive models capable of anticipating high-risk areas and conditions that increase the probability of accidents [5,14,19].

This analysis not only offers a comprehensive view of traffic accidents in Ecuador, but also highlights the transformative potential of machine learning in improving road safety. The results of this research seek to contribute significantly to the creation of a safer and more efficient road environment, thus minimizing the devastating impact that traffic accidents have on Ecuadorian society.

To address this work, the manuscript is organized as follows. Section 2 presents works related to clustering problem. Section 3 presents the proposed methodology to carry out the clustering task, including the clustering algorithms and ending with the choice of the best technique for the problem presented. Then, Sect. 4 shows the analysis on the cluster and the interpretation of the results obtained for the best technique. Finally, conclusions are presented in Sect. 5.

2 Background

The analysis of traffic accidents and the identification of patterns that allow reducing their incidence is a crucial field of study in road safety. Various investigations have addressed this topic from multiple perspectives, including public health approaches, accident characterization, and the use of advanced machine learning techniques for prediction and data analysis.

The first works to be reviewed, Nandurge and Dharwadkar [15] analyze accident data using machine learning paradigms such as regression and clustering methods. The strength of this approach is its ability to handle large volumes of data and improve prediction accuracy. On the other hand, and similar to previous work, Taamneh et al. [17] use clustering techniques and artificial neural networks to classify traffic accidents. The main characteristic of this approach is its ability to group similar data, and obtain good accuracy in the classification task. Similar to the previous approaches, Islam et al. [12] explores the application of clustering algorithms, such as DBSCAN and OPTICS, to identify traffic accident-prone areas. These algorithms stand out for their ability to handle spatial data and detect anomalies in areas of high accident density, using metrics such as Calinski-Harabasz Index, Davies-Bouldin Index, and Silhouette Coefficient.

Another of the studies to be reviewed, Kumeda et al. [13] use machine learning classification algorithms, including decision trees, support vector machines

[1] https://www.ant.gob.ec/visor-de-siniestralidad-estadisticas/.

(SVM), and neural networks. The strength of this approach is its ability to handle large volumes of data and identify complex patterns. Similar to the approach presented previously, Bokaba et al. [4] carry out a comparative study of machine learning classifiers, evaluating techniques such as decision trees, random forests, SVM and neural networks, the main characteristic of this study is its comparative approach that identifies the best techniques for different scenarios.

On the other hand, Santos et al. [16] explore machine learning approaches for accident analysis and hotspot prediction, using techniques such as logistic regression, decision trees, and ensemble methods. The main contribution of this work is its accuracy in predicting accident hotspots. Similar to the previous approach, Yadav et al. [20] propose a framework for the analysis of traffic accidents using machine learning paradigms such as logistic regression and neural networks. The methodological structure of the approach is clear and applicable to different types of accident data. On the other hand, Banerjee et al. [3] focus on traffic accident risk prediction using machine learning techniques such as regression analysis and neural networks. The characteristic of this approach is its correct predictive ability, while the weakness may lie in the complexity of the models and the need for high-quality data.

Finally, reviewed work such as the presented by Comi et al. [5] focus on cluster analysis to identify patterns and causes of accidents in the 15 districts of the city (2016–2019) for urban planning purposes, on the other hand, the article presented by Esenturk et al. [7] use the ROCK algorithm and Market Basket to analyze the UK STATS19 database to generate test scenarios for autonomous vehicles. Both studies demonstrate the benefits of data mining to identify meaningful patterns and improve road safety, whether through planning preventive interventions [5] or developing tests for autonomous vehicles [7]. However, they share similar limitations related to the dependence on the quality and completeness of historical data, the geographical specificity of their respective databases, and the challenges in generalizing results, either to other cities or to ensure that the scenarios The test results generated are truly representative of all potential risk situations.

Machine learning techniques for the analysis of traffic accidents present certain common weaknesses, among which the complexity in the interpretation of the clusters formed, the need for specific, detailed and high quality data, as well as a robust computational infrastructure (e.g., use of GPUs). Additionally, these techniques typically require large amounts of labeled data, which can result in variability in model performance depending on the quality and quantity of available data. Finally, the processing required to implement these methodologies can be another significant challenge to consider [6,18].

3 Methodology

The dataset[2] to be used in this work contains detailed information on traffic accidents in Ecuador from January 2017 to April 2024. The dataset has 55 columns

[2] https://www.ant.gob.ec/visor-de-siniestralidad-estadisticas/.

and 166684 records. Among the data contained in the dataset, geolocation information of the accident site and data on its location are present, as well as the type of vehicles involved, possible causes of the accident, type of accident, number of deaths and injuries involved in the accident. The objective of this study is to apply cluster techniques to find relationships between the different variables of the dataset. On the other hand, Fig. 1 presents the correlation matrix between the variables of the dataset. This analysis allows identifying linear relationships between the variables, which is useful for building predictive models.

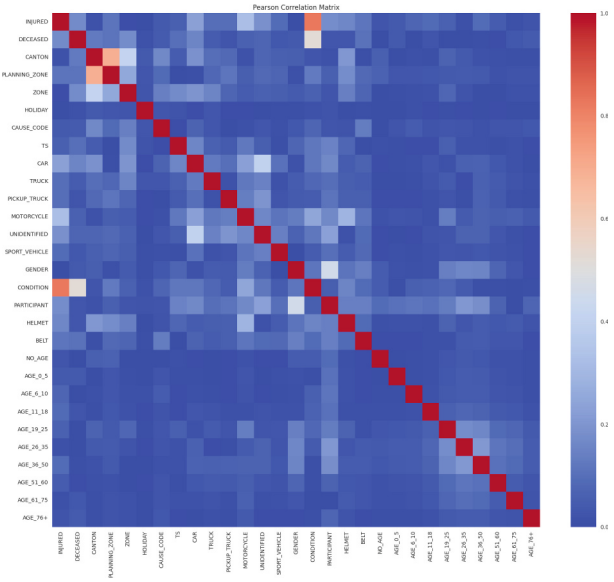

Fig. 1. Correlation matrix for all features.

3.1 Base Models

The objective of this section is to apply various clustering techniques to the traffic accident dataset, and evaluate their performance using cuantivativo y cualitativa evaluaciones. Methods such as Gaussian Mixture Model, Agglomerative Clustering, MiniBatch K-Means and K-Means are implemented. These clustering techniques are applied to identify hidden patterns in the data, such as groups of accidents with similar features (for example, based on accident type, severity, or location). The justification of the techniques used, the preparation of the dataset, and the interpretation of the results obtained are included below.

The original dataset contained accident data, including geographic coordinates, severity, cause and type of accident, and other related features. Below are the cleaning criteria applied to the dataset.