Road Accident Analysis: Clustering and Classification of High-Risk Areas*

Hela Bouhlel
Ecole Polytechnique Sousse
Hela.Bouhlel@polytechnicien.tn

Ines Jellali
Ecole Polytechnique Sousse
Ines.Jellali@polytechnicien.tn

***Abstract*—This article uses the Traffic Accident Database, which provides detailed information on road accidents in the United States, including geographic (GPS coordinates) and temporal (accident dates and times) data. The main objective is to apply clustering techniques to identify geographic areas at high risk of accidents and to develop a classification model to predict accident severity based on various factors, such as accident time, weather conditions, and road characteristics. To predict accident severity, we employ K-Nearest Neighbors (KNN), Random Forest (RF), LightGBM, and XGBoost, achieving an accuracy of 91% and f1-score of 82%. Additionally, we apply clustering algorithms, including K-Means, DBSCAN, and Agglomerative Clustering, to uncover geographic areas with distinct patterns related to location, time, severity, and weather conditions.**

***Index Terms*—road accident, machine learning, classification, clustering, imbalenced data.**

I. Introduction

Road safety is a major public health concern worldwide, particularly in the United States. Every year, road traffic accidents result in numerous fatalities, serious injuries, and significant economic losses.

According to the World Health Organization, approximately 1.19 million people lose their lives annually in road traffic accidents, while an additional 20 to 50 million sustain non-fatal injuries—many of which lead to long-term disabilities [3].

In this context, transportation systems must be designed to meet user needs and ensure safety for all road users. A key element of this approach is the systematic collection, sharing, and analysis of crash data. These practices are crucial for understanding the root causes of accidents, identifying high-risk areas, and developing systems that account for human error, hazardous road structures, and adverse weather conditions. Addressing these factors is essential to reducing traffic-related fatalities and injuries and to implementing more effective preventive measures.

The U.S. Roads Accidents dataset, covering the years 2016 to 2023, offers a comprehensive resource, providing detailed information on millions of accidents, including GPS coordinates, dates and times, weather conditions, and road characteristics [4]. This dataset enables the application of advanced predictive models to analyze road safety issues at a large scale.

Although numerous studies have examined traffic accidents, few have combined unsupervised clustering methods to identify high-risk zones with supervised classification techniques to predict accident severity—while simultaneously incorporating temporal, geographical, and environmental variables [1], [2].

In this paper, we propose a dual approach:

- We employ machine learning classification algorithms, including K-Nearest Neighbors (KNN), Random Forest (RF), LightGBM, and XGBoost, to predict accident severity based on various features such as time, weather conditions, and road characteristics.
- We apply clustering techniques, particularly K-Means, DBSCAN and the Agglomerative clustering algorithm, to detect geographic zones with high concentrations of incidents.

The proposed methodology involves several stages: cleaning and imputing missing data, converting GPS coordinates for geospatial analysis, encoding categorical variables, and applying temporal clustering on multiple time scales (e.g., day/night, seasons).

The results of this study aim to enhance the understanding of accident dynamics and provide actionable insights for local authorities and road safety stakeholders.

The remainder of this article is organized as follows: Section 2 describes the dataset used, Section 3 outlines the methodology and presents the experimental results, and Section 5 offers a conclusion.

II. Exploratory Data Analysis (EDA)

The dataset used in this study is an open-source CSV (Comma-Separated Values) file [4], containing information on road traffic accidents across 49 U.S. states, comprising approximately 7 million records spanning the period from 2016 to 2023. Each record includes 46 attributes related to accident characteristics, environmental conditions, and geographic information.

Given that a minimum of three years of data is typically necessary to conduct meaningful incident prediction, this study focuses on a subset of the dataset, covering the period from January 2020 to December 2022. A summary of the main attributes is presented in Table I.

Due to its large size, the dataset was hosted on Google Drive and loaded using the `Dask` library. Dask enables parallel and lazy data loading, which is particularly useful for handling large datasets without exceeding memory limitations.

TABLE I
SUMMARY OF VARIABLES FROM THE U.S. ACCIDENTS DATASET

Variable	Type
ID	Object
Source	Object
TMC	Float64
Severity	Int64
Start_Time	Object
End_Time	Object
Start_Lat	Float64
Start_Lng	Float64
End_Lat	Float64
End_Lng	Float64
Distance (mi)	Float64
Description	Object
Street	Object
City	Object
County	Object
State	Object
Zipcode	Object
Country	Object
Timezone	Object
Airport_Code	Object
Weather_Timestamp	Object
Temperature (F)	Float64
Humidity (%)	Float64
Pressure (in)	Float64
Visibility (mi)	Float64
Wind_Direction	Object
Wind_Speed (mph)	Float64
Precipitation (in)	Float64
Weather_Condition	Object
Amenity	Boolean
Bump	Boolean
Crossing	Boolean
Traffic_Signal	Boolean
Sunrise_Sunset	Object

To gain initial insights into the geographic distribution of accidents, the number of incidents was analyzed by state and by city. The results reveal a significant imbalance, with states such as California (CA), Florida (FL), and Texas (TX) accounting for a large proportion of recorded accidents.

These states are characterized by high population density, extensive road networks, and diverse weather patterns, which may explain their higher incident rates. California, in particular, contributes a substantial share of the dataset and was therefore selected for deeper analysis in subsequent stages of the project.

Figure 1 illustrates the states with the highest accident frequencies, while Figure 2 shows those with the lowest. Figure 3 highlights the cities most affected by road accidents, notably:

- Miami, Florida
- Los Angeles, California
- Houston, Texas
- Dallas, Texas

Urban areas are generally more prone to accidents due to heavier traffic volumes, more complex infrastructure, and increased exposure to risk factors such as congestion and pedestrian activity.

Accidents are not evenly distributed throughout the year. The data reveal a higher frequency of incidents during the

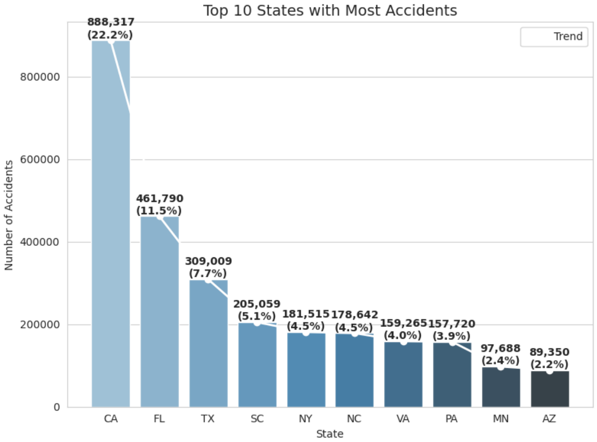

Fig. 1. Top states with the highest number of accidents in the U.S.

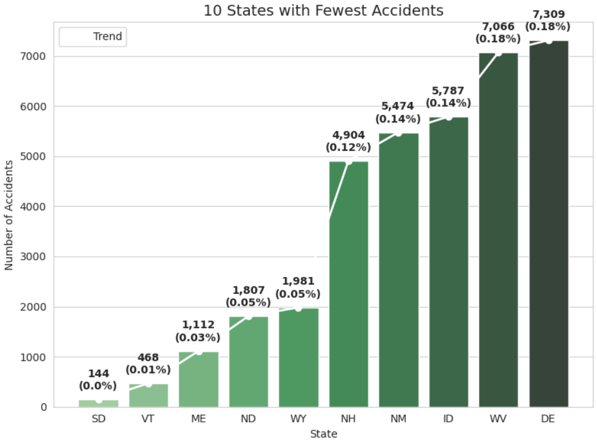

Fig. 2. States with the lowest number of accidents in the U.S.

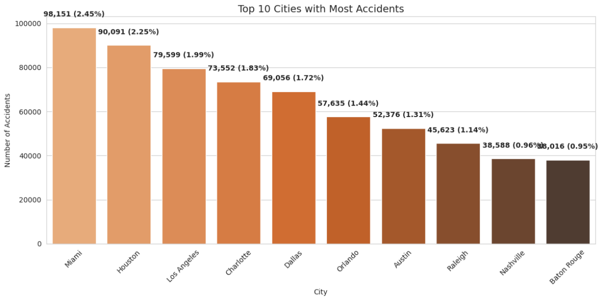

Fig. 3. Top cities with the highest number of accidents in the U.S.

winter months (see Figure 4), likely due to adverse weather conditions such as rain, snow, and reduced visibility. Conversely, the number of accidents tends to decrease slightly during the summer, possibly due to improved weather and road conditions.

An analysis of weekly distribution shows that accidents are more frequent on weekdays, particularly from Monday to Friday, which may reflect increased traffic density and work-related travel, as well as potential fatigue associated with commuting.

Time of day is also a critical factor influencing accident frequency. The dataset highlights two distinct peaks:

- Between 7:00 a.m. and 9:00 a.m., coinciding with morning rush hours.
- Between 4:00 p.m. and 6:00 p.m., during evening commutes.

These peaks suggest that traffic congestion significantly contributes to accident occurrence (see Figure 6). A notable decline is observed during late-night hours, particularly between 1:00 a.m. and 5:00 a.m., likely due to reduced traffic volume.

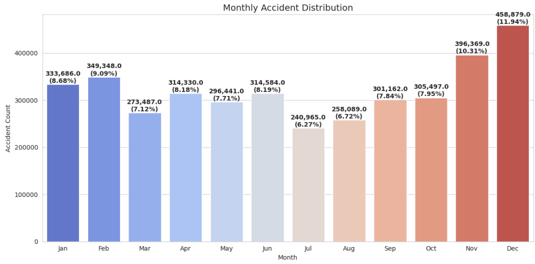

Fig. 4. Monthly distribution of accidents.

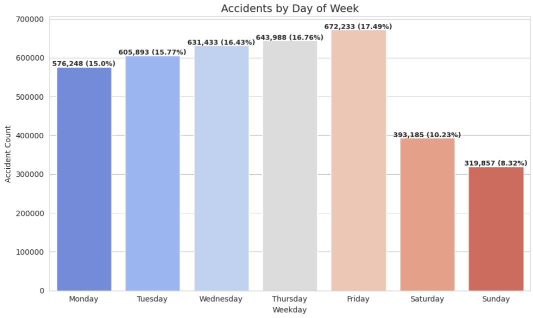

Fig. 5. Distribution of accidents by day of the week.

The primary weather conditions observed during accidents include cloudy, very cloudy, partly cloudy, and overcast skies. Although not extreme, these conditions can negatively impact visibility and driver perception, thus increasing the likelihood of accidents.

The `Severity` variable serves as the target variable for the classification task and is a four-level categorical variable:

- **Severity 1**: Minor impact
- **Severity 2**: Moderate impact

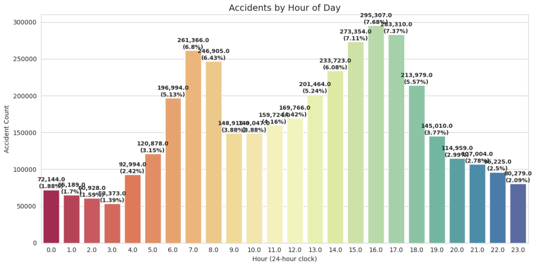

Fig. 6. Distribution of accidents by hour of the day.

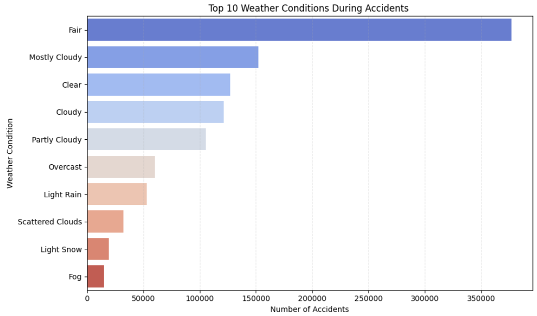

Fig. 7. Weather conditions associated with the highest number of accidents.

- **Severity 3**: Significant impact
- **Severity 4**: Severe impact

A preliminary analysis of severity levels reveals a marked imbalance (see Figure 8):

- Severity 2 is the most frequent class.
- Severity 3 is moderately represented.
- Severities 1 and 4 are rare.

This is typical in real-world traffic data, where more serious accidents occur less frequently but are critical to predict accurately.

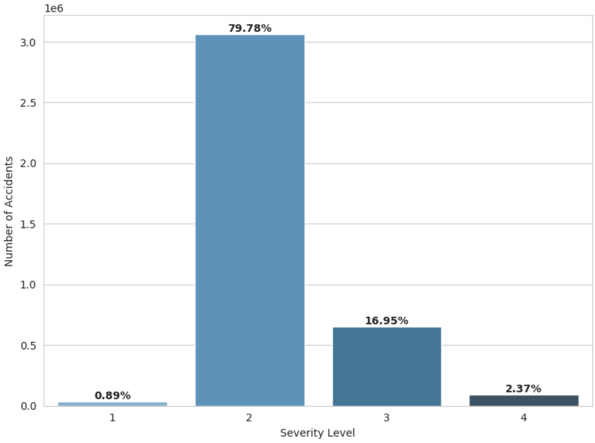

Fig. 8. Distribution of accidents by severity level.

III. Methodology

A. Data Processing

Due to the large volume of the dataset, the analysis was restricted to the year 2022. This subset still contains a significant number of records, ensuring the reliability and representativeness of the study while reducing computational complexity.

TABLE II
MISSING VALUES IN THE DATASET (YEAR 2022)

Variable	Missing Count	Missing Percentage (%)
Precipitation (in)	90,456	33.15
Duration	29,205	10.70
Wind Speed (mph)	26,142	9.58
Humidity (%)	7,375	2.70
Wind Direction	7,037	2.58
Temperature (F)	7,009	2.57
Visibility (mi)	6,218	2.28
Weather Condition	6,136	2.25
Pressure (in)	5,716	2.10
Street	416	0.15
Airport Code	343	0.13
Nautical Twilight	223	0.08
Astronomical Twilight	223	0.08
Civil Twilight	223	0.08
Sunrise/Sunset	223	0.08
Timezone	83	0.03
Zipcode	83	0.03
City	2	0.00

Although the dataset is large, it is not advisable to delete rows with missing values, due to the unbalanced distribution of the target variable, especially for minority classes. Deleting missing data would disproportionately reduce the already limited number of samples in these classes, potentially harming model performance and fairness.

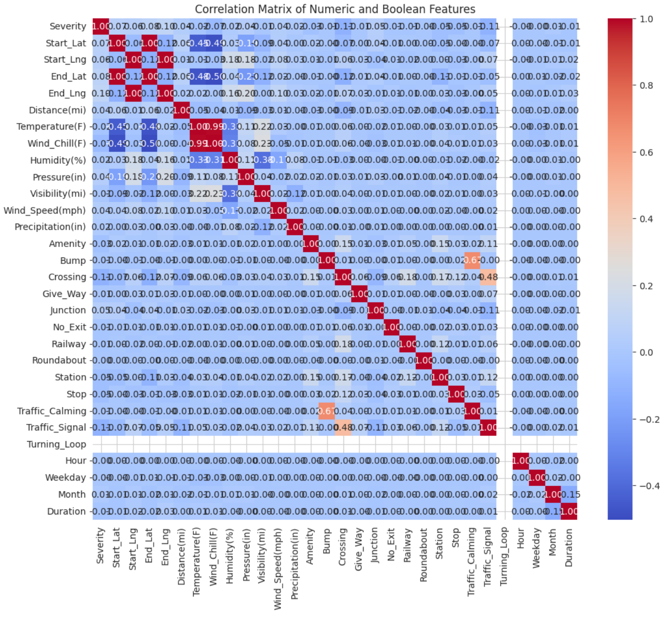

Fig. 9. Correlation matrix of key variables.

The following strategy was applied:

- For Duration, Temperature (F), and Pressure (in), missing values were filled with the **mean**.
- For Precipitation (in), Wind Speed (mph), Visibility (mi), and Humidity (%), missing values were filled with the **median**.

Variables such as Source, ID, and Description were removed as they provided no relevant information. End_Lat and End_Lng were highly correlated with Start_Lat and Start_Lng, and thus omitted.

A strong linear relationship between Temperature (F) and Wind_Chill (F) led to the removal of Wind_Chill (F).

Variables such as Country (always 'US') and Turning_Loop (always 'False') were discarded because they contained only a single class.

To improve model performance, some categorical variables were grouped:

- **Weather Condition**: Initially 121 categories, merged into 13 based on keyword grouping (e.g., clear, cloudy, rain, snow, etc.).
- **Wind Direction**: Reduced from 24 to 9 major directions (e.g., N, NE, E, SE, etc.).

Categorical variables such as Weather_Condition were transformed using *one-hot encoding*. High-cardinality categorical variables (e.g., Timezone, Wind Direction, Street, City, County, Zipcode, Airport Code, State) were encoded based on their relative frequency.

Finally, variables whose absolute correlation with the target variable Severity was lower than 0.005 were eliminated to reduce noise.

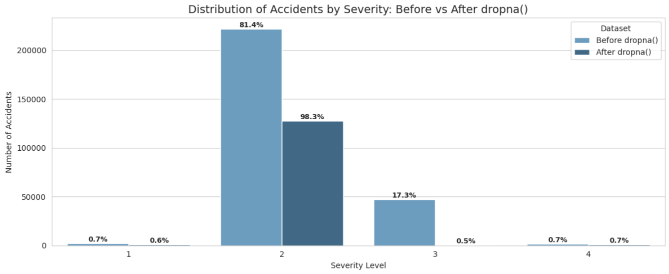

Fig. 10. Distribution of severity bafor and after dropna().

B. Classification

Due the importance of extracting information from accident data, classification can be a helpful tool in road safety research. The data were stratified into two sets: 80% for training and 20% for testing. To address the multi-class classification of accident severity, we employed several supervised learning algorithms: **K-Nearest Neighbors (KNN)**, **Random Forest (RF)**, **LightGBM**, and **XGBoost**.

1) Knearest neighbor (KNN): The k-nearest neighbors (KNN) algorithm is a non-parametric, supervised learning classifier that uses proximity to make classifications or predictions about the clustering of a data point. It is the one the simplest classification and regression methods in machine learning today. Implemented with $k = 5$, using distance-based weighting and Euclidean distance ($p = 2$) under the Minkowski metric.

TABLE III
CLASSIFICATION REPORT FOR THE KNN MODEL

Class	Precision	Recall	F1-score
Severity 1	0.00	0.00	0.00
Severity 2	0.88	0.92	0.90
Severity 3	0.47	0.38	0.42
Severity 4	0.23	0.04	0.06
Accuracy		0.82	
Macro Avg	0.39	0.33	0.35
Weighted Avg	0.80	0.82	0.81

2) Random Forest(RF): is an ensemble of decision trees in which each decision tree is trained with a specific random noise. Random forests are the most widely form of decision tree ensemble .

The Random Forest model was trained with the following hyperparameters:

- `n_estimators` = 100
- `max_depth` = 50
- `min_samples_split` = 6
- `min_samples_leaf` = 2
- `max_features` = `'sqrt'`
- `bootstrap` = `False`
- `random_state` = 42

TABLE IV
CLASSIFICATION REPORT FOR THE RF MODEL

Class	Precision	Recall	F1-score
Severity 1	0.00	0.00	0.00
Severity 2	0.91	0.98	0.94
Severity 3	0.81	0.58	0.67
Severity 4	1.00	0.03	0.06
Accuracy		0.90	
Macro Avg	0.68	0.40	0.42
Weighted Avg	0.89	0.90	0.89

while the Random Forest (RF) model shows a good overall accuracy of 0.90, the low F1-score of 0.42 (macro average) suggests that the model has an imbalance in predicting the severity classes, especially for the less frequent classes like Severity 1 and Severity 4. The F1-score is a more informative metric in such imbalanced datasets as it balances both precision and recall.

3) LightGBM: (Light Gradient Boosting Machine) is an open-source, distributed, high-performance gradient boosting framework based on decision tree algorithms. It is specifically designed for efficiency and speed, particularly with large-scale datasets and high-dimensional data. LightGBM uses a histogram-based algorithm, resulting in faster training speed and lower memory usage compared to traditional gradient boosting methods.

The complete set of hyperparameters is summarized below:

  - `objective` = `'multiclass'`
  - `num_class` = 4
  - `boosting_type` = `'gbdt'`
  - `num_leaves` = 31
  - `max_depth` = -1
  - `learning_rate` = 0.05
  - `n_estimators` = 100
  - `subsample` = 0.9
  - `colsample_bytree` = 0.9
  - `random_state` = 42

TABLE V
CLASSIFICATION REPORT FOR THE LIGHTGBM MODEL

Class	Precision	Recall	F1-score
Severity 1	0.43	0.03	0.08
Severity 2	0.91	0.97	0.94
Severity 3	0.81	0.59	0.64
Severity 4	0.50	0.03	0.05
Accuracy		0.90	
Macro Avg	0.66	0.40	0.43
Weighted Avg	0.89	0.90	0.88

4) XGBoost: (eXtreme Gradient Boosting) is an open-source, distributed machine learning library that uses decision trees to which gradient boosting is applied, the latter being a supervised learning boosting algorithm that uses gradient descent. It is known for its speed, efficiency, and ability to scale to large datasets. The complete set of hyperparameters is summarized below:

- `objective` = `'binary:logistic`
- `eval_metric` = `'logloss'`
- `random_state` = 42
- `reg_alpha` = 1
- `reg_lambda` = 1.5
- `min_child_weight` = 1

TABLE VI
CLASSIFICATION REPORT FOR THE XGBOOST MODEL

Class	Precision	Recall	F1-score
Severity 1	0.20	0.02	0.03
Severity 2	0.92	0.97	0.94
Severity 3	0.78	0.61	0.68
Severity 4	0.50	0.06	0.11
Accuracy		0.90	
Macro Avg	0.60	0.41	0.44
Weighted Avg	0.89	0.90	0.89

Since our dataset is highly imbalanced, the model's performance, as reflected by an F1-score of 0.44, remains insufficient for reliable prediction, as detailed in TableVI. So, several strategies were implemented to improve the quality of the data, especially for minority classes:

As shown in Table VII, although resampling methods improved the class balance, the F1-scores remain relatively low, indicating that the problem persists and suggesting the need to explore alternative solutions.

C. Binary classification

To better understand accident severities, we performed a new clustering based only on the **Severity** variable and its most

TABLE VII
MODEL PERFORMANCE WITH SMOTE AND ADASYN AUGMENTATION

Model	Resampling Method	Accuracy	F1-Score
XGBoost	SMOTE	0.88	0.50
XGBoost	ADASYN	0.88	0.50
Random Forest	SMOTE	0.88	0.48
Random Forest	ADASYN	0.88	0.47

correlated environmental factors: *Temperature (F)*, *Humidity (%)*, *Wind Speed (mph)*, and *Precipitation (in)*.

The objective was to investigate whether natural groupings could emerge, particularly a separation between **mild** and **severe** accidents.

We applied the KMeans clustering algorithm to this subset of features and evaluated different cluster numbers using the silhouette score. The results indicated that **two clusters** ($k = 2$) offered the best separation.

- **Cluster 0** predominantly contained accidents originally labeled as Severity 1 and Severity 2.
- **Cluster 1** mostly included accidents of Severity 3 and Severity 4.

This observation validated our hypothesis: accident severities could indeed be divided into two major groups.

TABLE VIII
CLUSTER ANALYSIS BASED ON SEVERITY AND CORRELATED FEATURES

Cluster	Sev_1 (%)	Sev_2 (%)	Sev_3 (%)	Sev_4 (%)
0	0.7	99.3	0.0	0.0
1	0.0	0.0	95.6	4.4

Based on this result, we redefined the severity classes as follows:

- **Mild Class:** Includes original Severity 1 and Severity 2.
- **Severe Class:** Includes original Severity 3 and Severity 4.

This new classification simplifies further analysis and enables a more robust modeling approach for predicting accident risks.

By transforming the problem into a binary classification task, we aim to simplify the modeling process and potentially improve the overall performance of the classifiers. We will therefore reapply our models using the new binary target.

TABLE IX
PERFORMANCE COMPARISON OF DIFFERENT MODELS.

Model	Accuracy	Macro F1-Score
KNN	0.83	0.64
Random Forest	0.91	0.82
XGBoost	0.91	0.82
XGBoost (Optuna)	0.91	0.82

The KNN model shows moderate performance with an overall accuracy of 83%. It performs well on Severity 1 but struggles with Severity 2.

Random Forest This model performs significantly better, with an accuracy of 90%. It achieves high precision and recall for Severity 1, with moderate performance for Severity 2.

XGBoost: The XGBoost model also achieves an accuracy of 91%, showing similar performance to Random Forest. However, its recall for Severity 2 can still be improved.

The XGBoost model optimized with Optuna also reaches an accuracy of 91%. This model improves the balance of performance between both severity levels, especially for Severity 2.

The results demonstrate that Random Forest and XGBoost models, along with their Optuna-optimized versions, perform similarly, achieving the highest accuracy of 91%, while KNN lags behind at 83%.

D. Clustering

Cluster analysis is often used as a preliminary step to organize heterogeneous data into homogeneous groups or behavioral patterns. In the context of traffic accident data, choosing an appropriate clustering algorithm and determining the optimal number of clusters can be complex and challenging, especially when dealing with a large and high-dimensional dataset.

For this study, we focused on 10000 traffic accidents that occurred in California during the year 2022.

1) K-Means: To begin the analysis, we applied the K-Means clustering algorithm, a commonly used method due to its simplicity and efficiency. To determine the optimal number of clusters, we relied on the Silhouette Index, which measures how similar an object is to its own cluster compared to other clusters. Based on this metric, we identified 14 clusters as the most appropriate configuration for our dataset.

Table X summarizes the main characteristics of each cluster.

These clusters reveal diverse patterns in terms of location, time, severity, and environmental conditions. A detailed interpretation is provided below:

- **Geographical Trends:**
  - Clusters such as 3 and 8 have higher average latitudes ($\sim$38), suggesting northern California regions.
  - Clusters 0, 5, and 10 have latitudes around 34, indicating southern California areas like Los Angeles.
- **Peak Accident Hours:**
  - Most clusters report peak accidents during afternoon hours (15:00–17:00), corresponding to typical rush hours.
  - Clusters 0 and 7 peak around 7:00 AM, indicating risks during morning commutes.
- **Severity Distribution:**
  - Clusters 0, 5, 10, and 11 report nearly **100% Severity 2** accidents, reflecting frequent but less severe incidents in urban zones.
  - Cluster 6 is notable for its **61% Severity 4**, marking it as a high-risk area for severe accidents.
- **Visibility and Weather Conditions:**
  - Most clusters have good visibility (9–10 miles), suggesting weather is not a primary factor.

TABLE X

Summary of Clusters in California Traffic Accident Dataset (2022)

Cluster	#Accidents	Lat	Lng	Peak Hr	Sev 1 (%)	Sev 2 (%)	Sev 3 (%)	Sev 4 (%)	Visib.	Night (%)	Top Weather
0	1196	34.11	-118.04	7.0	0.0	100.0	0.0	0.0	9.55	6.9	Clear
1	687	36.05	-119.62	15.0	0.3	99.7	0.0	0.0	2.86	40.8	Fog/Haze
2	193	34.72	-118.98	16.0	0.0	99.5	0.5	0.0	9.27	37.8	Clear
3	1472	38.23	-121.71	16.0	1.0	99.0	0.0	0.0	9.52	64.3	Clear
4	8	36.53	-120.84	13.0	12.5	87.5	0.0	0.0	10.00	50.0	Clear
5	1749	34.03	-118.01	16.0	0.0	100.0	0.0	0.0	9.58	3.8	Clear
6	123	35.69	-119.61	15.0	0.0	0.0	39.0	61.0	9.39	25.2	Clear
7	796	35.43	-119.42	7.0	1.4	98.6	0.0	0.0	9.33	39.4	Clear
8	1408	38.07	-121.50	17.0	1.5	98.5	0.0	0.0	10.16	9.2	Clear
9	303	36.18	-119.95	17.0	0.7	99.3	0.0	0.0	9.27	38.0	Clear
10	1521	34.08	-118.02	16.0	0.1	99.9	0.0	0.0	9.50	97.2	Clear
11	73	36.02	-119.88	17.0	0.0	100.0	0.0	0.0	2.79	47.9	Rain
12	386	36.21	-119.98	16.0	0.8	99.0	0.3	0.0	9.19	37.0	Clear
13	85	35.50	-119.65	17.0	0.0	97.6	0.0	2.4	9.55	28.2	Clear

  - Clusters 1 and 11 are exceptions, showing visibility under 3 miles and being dominated by *Fog/Haze* or *Rain*, highlighting weather-related risks.
- **Night-Time Accidents:**
  - Cluster 10 shows a **97.2% night accident rate**, suggesting poor lighting or risky night-time driving conditions.
  - Clusters 3 and 1 also show high night accident percentages (64% and 40.8%, respectively).

Figure 11 illustrates the spatial distribution of these clusters across California.

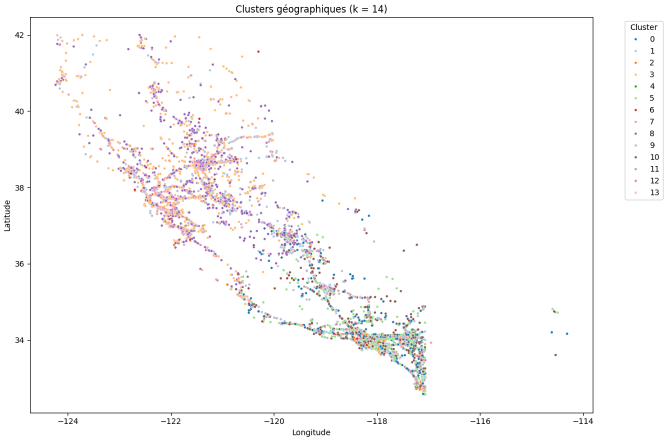

Fig. 11. Cluster Visualization for K-Means Clustering

2) DBSCAN: We also applied the DBSCAN algorithm (Density-Based Spatial Clustering of Applications with Noise). DBSCAN is particularly suitable for detecting dense areas and identifying noise points (isolated or anomalous accidents), which is highly relevant when analyzing traffic accidents.

Table XI presents a summary of the DBSCAN clustering results:

The following interpretation is based on the cluster summary and the severity levels:

- **Cluster -1 (Noise Points):** This group includes accidents that DBSCAN could not assign to any dense region. It is characterized by a relatively high percentage of nighttime accidents (40.3%) and predominantly rainy weather. Severity is mixed but largely moderate (Severity 2), indicating diverse accident profiles.
- **Cluster 0 (Daytime, Clear Weather, Moderate Severity):** The largest cluster (5962 accidents) is located around (35.58, -119.41). All accidents have Severity 2 (moderate impact), occur during the daytime under clear weather with good visibility. This cluster represents typical daytime accidents under normal driving conditions.
- **Cluster 1 (Nighttime, Clear Weather, Moderate Severity):** Similar in location to Cluster 0, but accidents occur entirely at night (100%). Despite nighttime conditions, the accidents remain of moderate severity, indicating that nighttime driving did not significantly worsen the outcomes.

TABLE XI
DBSCAN Cluster Summary

Cluster	#Acc.	Lat	Lng	Peak Hr	S1 (%)	S2 (%)	S3 (%)	S4 (%)	Vis.	Night (%)	Top Weather
-1	419	36.88	-120.26	17.0	12.4	64.2	9.8	13.6	7.33	40.3	Rain
0	5962	35.58	-119.41	16.0	0.0	100.0	0.0	0.0	9.22	0.0	Clear
1	3574	35.77	-119.51	17.0	0.0	100.0	0.0	0.0	9.04	100.0	Clear
2	16	37.71	-121.80	6.0	100.0	0.0	0.0	0.0	10.00	100.0	Clear
3	10	35.90	-119.58	13.0	0.0	0.0	0.0	100.0	10.00	0.0	Cloudy
4	9	34.46	-118.46	9.0	0.0	100.0	0.0	0.0	2.67	100.0	Rain
5	10	37.69	-120.98	14.0	0.0	100.0	0.0	0.0	9.80	100.0	Windy

- **Cluster 2 (Morning, Clear Weather, Minor Impact):** A small cluster (16 accidents) characterized by accidents occurring around 6 AM. All accidents are of Severity 1 (minor impact) under clear weather with excellent visibility (10 miles), suggesting minor incidents during early morning low-traffic hours.
- **Cluster 3 (Daytime, Cloudy Weather, Severe Impact):** This cluster contains 10 accidents, all classified as Severity 4 (severe impact). Accidents occur around 1 PM under cloudy conditions, implying serious accidents possibly linked to high speeds or complex situations despite daylight.
- **Cluster 4 (Nighttime, Rainy Weather, Moderate Impact):** Consists of 9 accidents near Los Angeles, under rainy weather with low visibility (2.67 miles). All accidents are Severity 2 (moderate impact) and occur at night, highlighting the effect of poor weather and visibility on accident risks.
- **Cluster 5 (Nighttime, Windy Conditions, Moderate Impact):** Composed of 10 accidents happening at night under windy conditions. All accidents are of Severity 2 (moderate impact), suggesting that strong winds may influence nighttime driving stability.

Weather and visibility conditions have a noticeable impact on accidents, especially at night. Most accidents in the dataset are of moderate severity (Severity 2), with severe accidents (Severity 4) identified distinctly in Cluster 3 under cloudy weather conditions during the day.

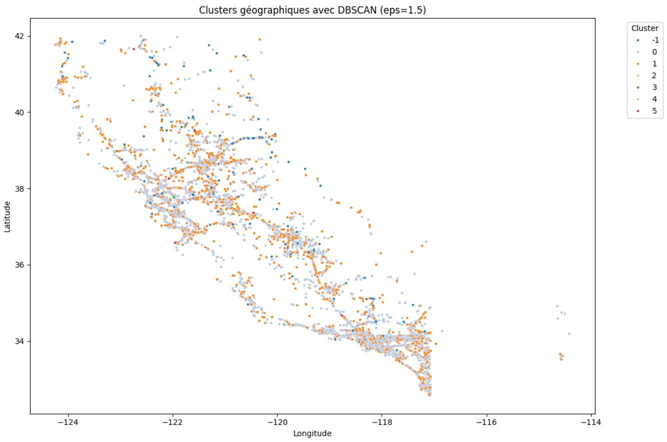

Fig. 12. Cluster Visualization for DBSCAN Clustering

3) Agglomerative Clustering: Agglomerative clustering is a hierarchical method that groups data points into clusters based on their similarity. Below is the table summarizing the results of the Agglomerative clustering for accident data, including key features such as accident count, location, peak hour, severity, visibility, and weather conditions.

The following interpretation is based on the cluster characteristics from the Agglomerative clustering algorithm. The severity levels are as defined: Severity 1 for minor impact,

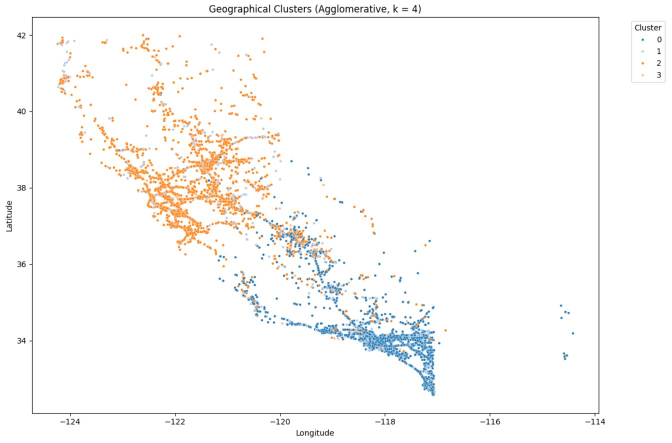

Fig. 13. Cluster Visualization for Agglomerative Clustering

Severity 2 for moderate impact, Severity 3 for significant impact, and Severity 4 for severe impact:

- **Cluster 0 (Daytime, Clear Weather, Moderate Severity):** Cluster 0, containing 5200 accidents, is characterized by accidents occurring during the daytime (around 5 PM) under clear weather conditions. All of the accidents in this cluster are of Severity 2 (moderate impact), indicating that accidents in clear weather during the day tend to be of moderate severity. The visibility is good (9.64 miles), and 37.6% of the accidents occur at night, suggesting a relatively balanced time distribution. This is the largest cluster, indicating that daytime accidents under clear conditions are most common.
- **Cluster 1 (Daytime, Fog/Haze, Moderate Severity):** This cluster contains 1093 accidents and is characterized by accidents occurring in foggy or hazy conditions. The accidents occur around 4 PM with moderate severity (Severity 2). The visibility is lower (3.72 miles), and 43.5% of accidents happen at night. The weather conditions, including fog or haze, may be contributing to an increased accident rate under these specific conditions. The higher proportion of accidents at night suggests that reduced visibility in foggy conditions is particularly hazardous during evening hours.
- **Cluster 2 (Daytime, Clear Weather, Moderate Severity):** Similar to Cluster 0, Cluster 2 contains 3599 accidents under clear weather conditions during the daytime, with accidents occurring around 5 PM. The accidents are mostly Severity 2 (moderate impact), indicating typical accident conditions in clear weather. The visibility is high (9.88 miles), and 36.6% of accidents occur at night, showing that even under clear weather, night accidents are still a significant factor. This cluster suggests that clear weather accidents are not limited to the daytime but also occur at night.
- **Cluster 3 (Daytime, Clear Weather, Severe Impact):** Cluster 3 contains only 108 accidents, but they are notably severe (Severity 4). These accidents occur during the day around 5 PM under clear weather conditions. The visibility is 9.25 miles, and 29.6% of the accidents happen at night. The higher severity in this cluster might indicate specific high-risk situations that warrant further investigation, such as complex driving conditions or high-speed collisions. Despite the small number of accidents, the high severity of these cases makes them an important focus for analysis.

The Agglomerative clustering highlights that most accidents occur under clear weather conditions during the daytime, with moderate severity (Severity 2) being the most common outcome. Fog and haze conditions are associated with a slightly higher frequency of accidents at night. Cluster 3, while small, is marked by a higher severity of accidents, which may indicate dangerous driving conditions in particular areas.

These insights can help in further refining safety measures, identifying hazardous conditions, and improving accident prevention strategies, particularly in areas where certain weather conditions, such as fog or haze, contribute significantly to accidents.

IV. Conclusion

Through our analysis, we have demonstrated that classifying accidents into two severity levels (Severity 1 and Severity 2) provides an effective solution. Given the available features, such as weather conditions, time of day, and location, the distinction between these two classes appears sufficient for accurate classification. We observed that a four-class classification might not be as meaningful or beneficial with the current dataset.

However, it is important to note that if additional information were included, such as the number of deaths, injuries, or damaged vehicles, a four-class classification could provide more valuable insights. Such enriched data could capture more granular levels of severity, improving the robustness and applicability of the model in safety-critical applications.

By applying the clustering algorithms, we were able to identify distinct patterns related to accident severity, time of occurrence, geographical location, weather conditions, and visibility. These insights provide a clearer understanding of the factors contributing to accidents, which can be leveraged to inform targeted safety measures.

Key findings suggest that accidents are more frequent during rush hours, particularly in urban areas, with certain regions exhibiting higher risks due to adverse weather conditions, low visibility, and nighttime driving. By identifying high-risk zones and times, stakeholders such as transportation authorities, urban planners, and law enforcement can implement more focused interventions, such as improving road lighting, enhancing weather-related signage, or adjusting traffic management strategies during peak hours.

Overall, the findings of this study contribute to the ongoing effort to reduce traffic-related accidents, ultimately aiming to save lives and minimize the economic and social costs of road incidents. Moving forward, further research can explore the integration of additional factors such as driver behavior, road infrastructure, and vehicle types, to further refine accident prediction and prevention strategies.

In conclusion, this project lays a strong foundation for improving road safety by providing valuable insights that can guide proactive measures to prevent accidents and protect individuals on the road.

References

[1] Khosravi, Y., Hosseinali, F., & Adresi, M. (2024). Identifying accident prone areas and factors influencing the severity of crashes using machine learning and spatial analyses. Scientific Reports, 14(1), 29836.

[2] Rezashoar, S., Kashi, E., & Saeidi, S. (2024). A hybrid algorithm based on machine learning (LightGBM-Optuna) for road accident severity classification (case study: United States from 2016 to 2020). Innovative Infrastructure Solutions, 9(8), 319.

[3] World Health Organization. Road traffic injuries. *WHO Fact Sheets*. Retrieved from https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

[4] U.S. Accidents (2016 - 2020) dataset. Available at: https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

[5] Sinclair, C., & Das, S. (2021, January). Traffic accidents analytics in UK urban areas using k-means clustering for geospatial mapping. In 2021 International conference on sustainable energy and future electric transportation (SEFET) (pp. 1-7). IEEE.

[6] Yan, R., Hu, L., Li, J., & Lin, N. (2024). Accident severity analysis of traffic accident hot spot areas in Changsha City considering built environment. Sustainability, 16(7), 3054.

TABLE XII
Cluster Characteristics (New Results)

Cluster	#Acc.	Lat	Lng	Peak Hr	S1 (%)	S2 (%)	S3 (%)	S4 (%)	Vis.	Night (%)	Top Weather
0	5200	34.11	-118.07	17.0	0.0	100.0	0.0	0.0	9.64	37.6	Clear
1	1093	35.93	-119.56	16.0	0.1	99.9	0.0	0.0	3.72	43.5	Fog/Haze
2	3599	37.93	-121.50	17.0	1.9	98.1	0.0	0.0	9.88	36.6	Clear
3	108	36.11	-119.84	17.0	0.0	0.0	38.0	62.0	9.25	29.6	Clear