# **Driving Behavior Classification at Highways Using Veichle Kinematics: Application of Unsupervized Machine Learning**

**Eleen Kmail1, Yamama Abu lail2, Rateba Sawalha3, Malak Mousa4, and Doaa Eid5**

1An-Najah National University, Faculty of Engineering, AI Department student, Nablus, Palestine .

(E-mail: [eleen.kmail@stu.najah.edu](mailto:eleen.kmail@stu.najah.edu))

2An-Najah National University, Faculty of Engineering, SoftwareEngineering Department student, Nablus, Palestine

(E-mail: [yamama.abuliel@stu.najah.edu](mailto:yamama.abuliel@stu.najah.edu))

3An-Najah National University, Faculty of Engineering, SoftwareEngineering Department student, Nablus, Palestine

(E-mail: [rateeba-sawalha@stu.najah.edu](mailto:rateeba-sawalha@stu.najah.edu))

4An-Najah National University, Faculty of Engineering, SoftwareEngineering Department student, Nablus, Palestine

(E-mail: [malak.mousa@stu.najah.edu](mailto:malak.mousa@stu.najah.edu))

5An-Najah National University, Faculty of Engineering, SoftwareEngineering Department student, Nablus, Palestine

(E-mail: [s11926144@stu.najah.edu](mailto:s11926144@stu.najah.edu))

## **ABSTRACT**

Drivers' behavior on the road is called into question as each driver's driving habit is unique and has a significant impact on road safety. In this paper we studied the behavior of drivers on the highway and the vehicles that were monitored were either a car or a truck, and relied on a highD dataset to monitor the data, K-means clustering algorithm was used as an unsupervised machine learning to cluster driving behavior into three classes (i.e., conservative, normal, and aggressive), and into two clusters and find that the behavior of the drivers was mostly normal driving. There are nine features that we used in classifying the driver. the classification is based on the standard deviation of speed and longitudinal deceleration or acceleration, coefficient of variation of speed, mean absolute deviation of speed and longitudinal acceleration, quantile coefficient of variation of normalized speed and longitudinal acceleration, percentage of time when the mean of longitudinal acceleration and deceleration exceeds the mean plus two standard deviations. This study is useful in knowing the possibilities that a car with automatic driving may face.

## 

## 

## 

## INTRODUCTION

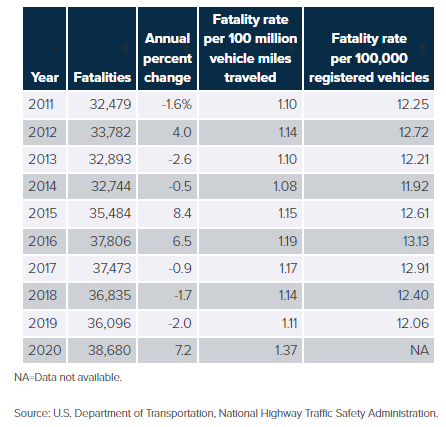
These days, the idea of using Automated viechles is widely spread. People are thinking about how to make automated vehicles which will be able to take them to any place they want. Several reasons led people to think about making Automated vehicles. The main reason is the impact of using it which the percentage of accidents will be reduced and the risk on people's lives will be reduced. There are also other causes resulting from manufacturing Automated vehicles, Reduced Congestion Several causes of traffic congestion could be addressed, Reduce the number of collisions and unexpected bend.

Maintaining a safe and consistent distance between vehicles, which helps reduce the number of stop-and-go waves that lead to road congestion. Environmental Gains Automated vehicles

have the potential to reduce fuel use and carbon emissions. Reduce greenhouse gases from needless idling. Automated vehicles depend on a large set of data from the real world Dataset. User paths are critical to many tasks as road user. Forecasting models or safety validation based on scenario.

## Traffic data is captured by drones, recently shown that drones are an effective way to obtain natural routes for road users. Compared to driving studies or ground-level infrastructure sensors, one of the main advantages of using a drone is the possibility of recording normal behavior, as road users do not notice that the measurements are being taken. Right.

## For the ideal viewing angle, a full crossover scenario can be measured with much less obstruction than with ground level sensors.



## RESEARCH QUESTION

On the highway, many people die in accidents,

so we have provided data analysis showing drivers' behavior on the highway to help us know the possibilities that an automated car will encounter. The table below shows statistics on the number of people killed in accidents.**[13]**

**Figure 1 :Table of traffic deaths, 2011-2020**

## 

## LITERATURE REVIEW

## The data-based approach uses real driving data and extracts scenarios from it. Hierarchical clustering is used to extract individual scenarios from the several minutes of recorded data. Uninteresting driving situations are sorted out to reduce the amount of data. The driving behavior pattern diagram can reflect the distribution of driving behavior and clearly present the driving style, habits, and driving risks of drivers. The evaluation results show that drivers with a high proportion of risky driving behavior scored higher on acceleration, deceleration, and overall driving behavior. Those who drove more consistently, with less risky driving, tended to have lower scores.

## The authors propose a novel method to measure data from an aerial perspective for scenario-based validation of highly automated vehicles. This dataset consists of 16.5 hours of measurements from six locations with 110 000 vehicles, a total driven distance of 45 000 km and 5600 recorded complete lane changes. The dataset used was the HighD, which contains post-processed data collected with a drone equipped with a high-resolution camera. It can also support smart cities technologies and can help authorities avoid accidents with immediate actions.

## The data include trajectories of 110,000 vehicles recorded for a duration of 16.5 h. They analyzed the acceleration time-series of the natural highD and artificial drivers using simulations of two car-following models. In contrast, artificial drivers followed the logical rules incorporated in the model, resulting in a smoother acceleration profile. The detection of abnormal driving is an important topic in the safety validation of autonomous vehicles. It can also support smart cities technologies and can help authorities avoid accidents with immediate actions.

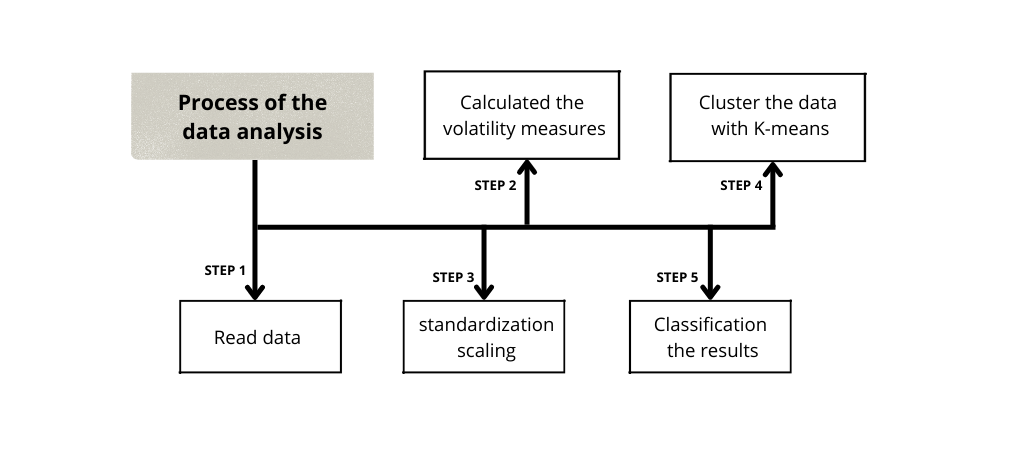
## A scenario was extracted from different traffic data collected in Germany using a drone, where five vehicle kinematics variables (lateral and longitudinal distances, speeds and acceleration) were used to know the correlation of the correlation, intersection and safety measures (collision time) and according to the data it was agreed that safety assessment methodologies need to be They are designed for different environments and regions to ensure safety and also the possibility of developing safety indicators that are applicable at the international level. We benchmark our approach against representative rule-based and black-box models as well as constant velocity models.

## 

## METHODS

Driver behavior can be described as aggressive, calm, and normal, with implications for traffic safety, emissions, and energy consumption that happened. In this study, the key objective is an analysis the data and classify individuals driving behavior that influence traffic safety on the highways (where aggressive, calm, and normal driver behavior has been studied), to reduce traffic safety issues and reduce fatal and seriously injured traffic accidents. We used highD dataset and this prepared dataset was taken as it was monitored by a drone. To study and classification the driver behavior in this study, we used volatility measures.

Figure 1. shows the operations we performed to obtain results about the driver’s behavior, where we obtained the data, read and analyzed it, calculated the volatility measures, then took the DVs columns because they are the ones that will help us to get the results, after that, we calculated standardization scaling for the data, cluster the data with K-means algorithm, and classification the results.



**Figure 2:** Process of Analysis Data to Cluster Driver Behavior

***Volatility Measures***

Table 1. shows the volatility measures equations that we used in the code to learn and analyze driver behavior on highDs [19].

[table—------------------------------------------------------------------]

***Standardization Scaling***

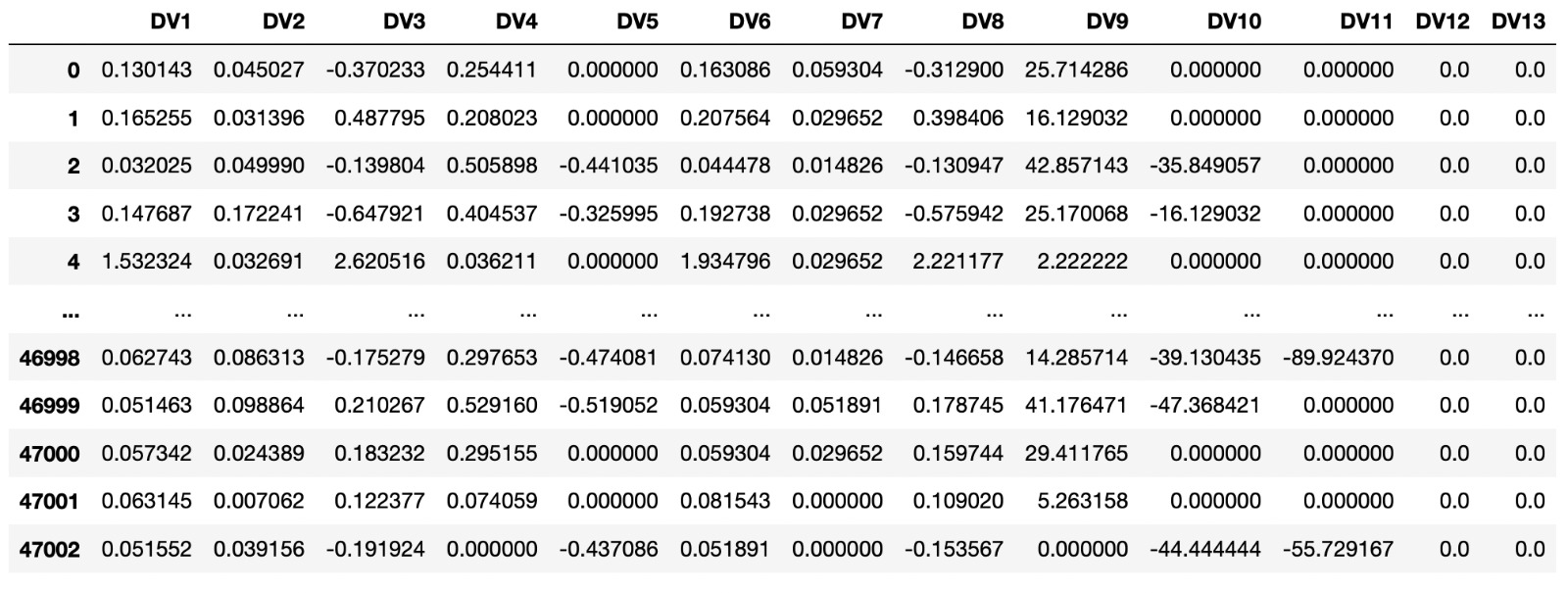
After calculating the volatility measures, we made a copy of the original dataset and took the DVs columns from them only because they are the ones that will help us in knowing and analyzing the driver’s behavior. After then we noticed that the values of the DVs columns are spaced apart, which slows down the running process, and when the results are represented on the graph, the data is separate. Therefore, we calculated the standardization scaling of the DVs column values until they are convergent and confined to a certain range.

Where x is the value of every cell on every DVs column, is distribution mean, and is standard deviation. The distribution mean and standard deviation are calculated for each

column, next subtract the mean from each value in the column (x), and then we divide the result values (mean is already subtracted) of each feature by its standard deviation:

[equation—-------------------------------------------]

Figure 2. shows the values confined to a certain range before do standardization scaling.



**Figure 2:** The Data before Calculate Standardization Scaling

Figure 3. shows the values confined to a certain range after do standardization scaling.

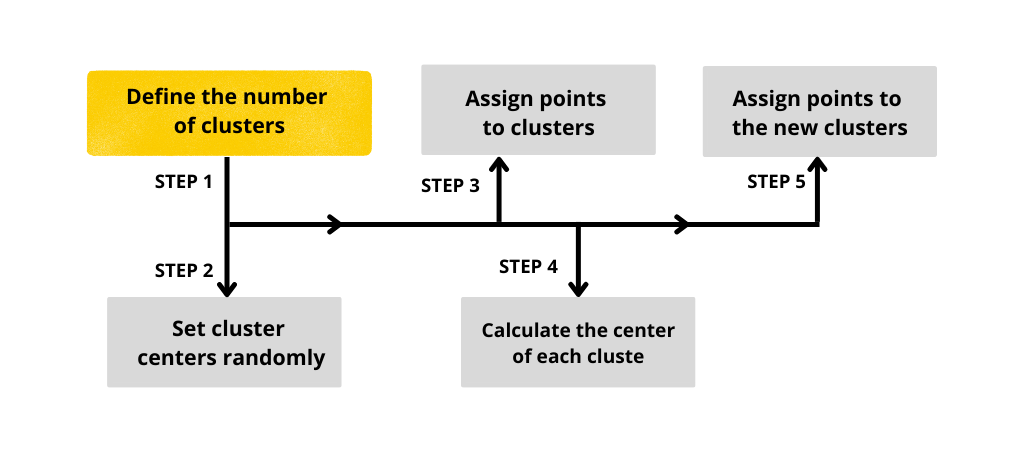


**Figure 3:** The Data after Calculate Standardization Scaling

***K-means Clustering Algorithm***

K-means is one of the simplest and most common methods of unsupervised machine learning algorithms for cluster analysis. The k-Means method clusters your data points on a given number of clusters, which means before anything should define the number of K (number of clusters).

Figure 4. shows how does the K-means cluster analysis work. Step one should define a number of clusters which is K to find the clusters of driver behavior. Step two sets cluster centers randomly each of the centers representing one cluster of driver behavior. In step three, we assign each element to one cluster, and the distance from every point to each of the cluster centers is measured and every point is then assigned to the cluster that is closest to it. In step four calculate the center of each cluster.



**Figure 4:** Steps of K-means Cluster Analysis

## **DATASET**

## ***How it was collected***

[16]For the highD dataset, traffic was recorded at German highways using unmanned aerial vehicles. In order to extract the trajectories from the recordings, state-of-the-art Computer Vision algorithms were used. Using neural networks the vehicles were detected and localized in every frame. To gain smooth trajectories from theses detections, the vehicles were tracked over time and their movement is smoothed using Bayesian smoothing

This dataset consists of 16.5 hours of measurements from six locations with **110 000 vehicles**, a total driven distance of **45 000 km** and **5600 recorded** complete lane changes

***The Features***

There are **9 features** that we used in classifying the driver . And the classification based on the **standard deviation of speed and longitudinal deceleration or acceleration, coefficient of variation of speed, mean absolute deviation of speed and longitudinal acceleration, quantile coefficient of variation of normalized speed and longitudinal acceleration, percentage of time when the mean of longitudinal acceleration and deceleration exceeds the mean plus two standard deviations.**

## ANALYSIS AND RESULTS

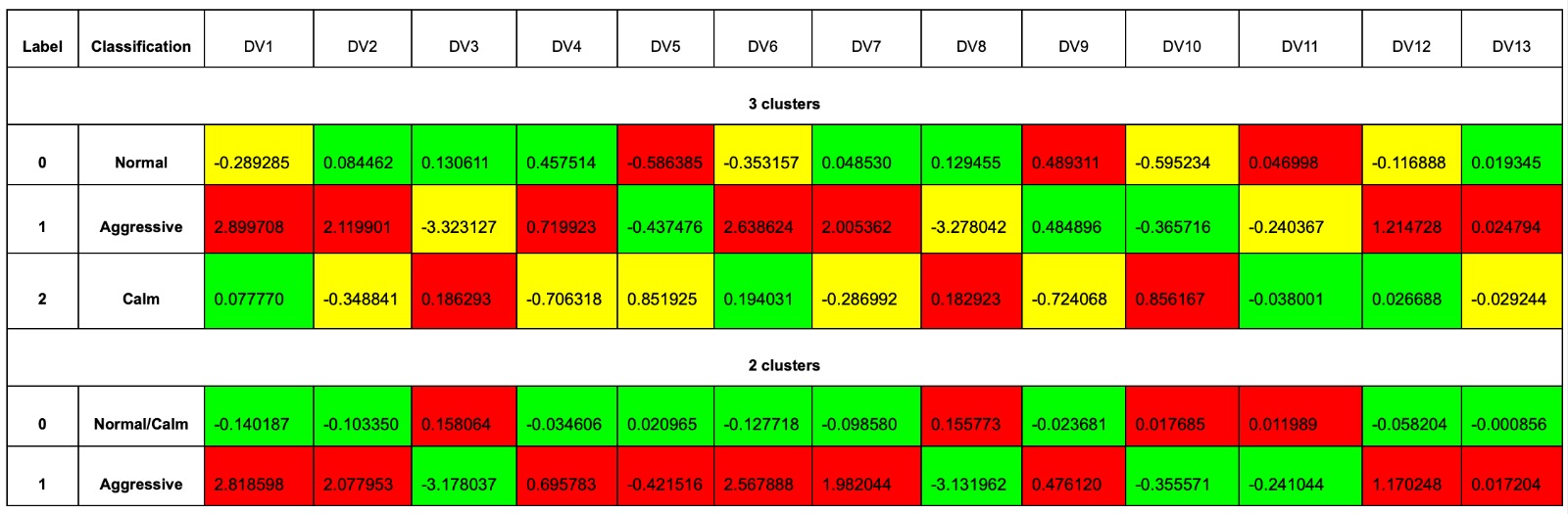
To prepare the data for clustering, we standardized data to make classification of each observation analogous across different features. Observations in the dataset denotes a driving behavior event across the two different conditions. Each observation will be clustered using K-means. In cluster analysis, we used elbow method, which is a heuristic used in determining the number of clusters in a dataset. The method consists of plotting the explained variation as a function of the number of clusters and choosing the elbow of the curve as the optimal number of clusters (k) to use [18.].

As a result, we found 2 clusters and 3 clusters. In 2 clusters there are 2 labels, label 0 is normal/clam, label 1 is aggressive.

In 3 clusters there are 3 labels, label 0 is normal, label 1 is aggressive and label 2 is calm. As the table below, the predominant behavior is normal driving.

## CONCLUSION

In this paper, we provided a classification of driving behavior for safety validation of automated driving vehicles and reduce traffic safety issues and reduce fatal and seriously injured traffic accidents. We used the highD dataset in which the data was captured by a drone. After analyzing the data and making a clustering by the K-means algorithm we find aggressive driving behavior is 2081, normal driving behavior is 25993, and calm driving behavior is 18929. That means the lowest number of driving behavior is aggressive, while the normal driving behavior and calm driving behavior are close to each other.



**Figure 3: Table of classfiction clusters**

**Contribution**

***Coding***

Eleen: All coding Part (Reading data, calculating DVs, and clustering).

***Classification:***

All members.

***Paper:***

Malak: Methods and Conclusion.

Rateeba: Abstract, Research question, and Analysis and Results.

Doaa: Intoduction and Referenes.

Yamama: Literature review and Dataset.

## REFERENCES

[15] Barry , Michael, and Riley Riley et al. “Facts + Statistics: Highway Safety.” *III Insurance Information Institute*, Inc, 2022, www.iii.org/fact-statistic/facts-statistics-highway-safety#:~:text=2019%3A%20According%20to%20the%20National,2018%20in%20vehicle%20miles%20traveled.

[13] D. Marutho, S. H. Handaka, and E. Wijaya, “The determination of cluster number at k-mean using elbow method and purity evaluation on headline news,” in 2018 International Seminar on Application for Technology of Information and Communication, 2018, pp. 533–538

[19] Arvin, Ramin, et al. “How Instantaneous Driving Behavior Contributes to Crashes at Intersections: Extracting Useful Information from Connected Vehicle Message Data.” *Accident Analysis & Prevention*, Pergamon, 7 Mar. 2019, www.sciencedirect.com/science/article/abs/pii/S0001457518304950?casa\_token=VjxhKF6jIF8AAAAA%3Ad20I4UdEVpP2MbLkAo9wE\_toJuVAmylY-rdnaW3M8ghWRgAcYcBOPZ4LivfeZO0Z8hyMVlTrRg.

[y20]

krajewski, Rober, and Julian Bock. “The Highway Drone Dataset.” *The HIGHD Dataset*, Zertifiziert Nach ISO 9001, 2015, www.highd-dataset.com/?fbclid=IwAR13515gAkVrZM80\_e3F\_ifKMcz5L4jFfk94-rwvfMfMWeVfu6Z9ecdbFHU.

[y21]

Kruber, Friedrich, et al. “Highway Traffic Data.” *The HIGHD Dataset*, ArXiv:1903.04249v1 [Eess.SP], 11 Mar. 2019.

[y22]

Krajewski, Robert, et al. “The HIGHD Dataset .” *: A Drone Dataset of Naturalistic Vehicle Trajectories on German Highways for Validation of Highly Automated Driving Systems*, USA, Nov. 2018.