



# Building a Dataset for Trip Style Assessment Based on Real Trip Data

Luís M. P. Loureiro<sup>1,3</sup>, Artur J. Ferreira<sup>1,2</sup><sup>a</sup> and André R. Lourenço<sup>1,3</sup><sup>b</sup>

<sup>1</sup>*ISEL, Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa, Portugal*

<sup>2</sup>*Instituto de Telecomunicações, Lisboa, Portugal*

<sup>3</sup>*CardioID Technologies, Portugal*

**Keywords:** Clustering, Driver Behavior, Feature Engineering, Pay-as-you-Drive, Trip Dataset, Trip Driver Style.

**Abstract:** In most countries, to have permission to drive vehicles on public roads one must have insurance against civil liability for vehicles. In many cases, the insurance fees depend on the age of the driver, the number of years one holds a driving license, and the driving history. The usual assumption taken by insurance companies that younger drivers are always more risky than others are not always correct, penalizing young good drivers. In this paper, we follow a pay-as-you-drive approach based on trip behavior data of different drivers. First, we build a dataset from real trip data. Then, we apply a two-stage clustering approach to the dataset to identify trip profiles. The experimental results show that we can cluster and identify distinct trip profiles in which many trips have a non-aggressive style, some have an aggressive style and only a few are risky style trips. Our solution finds application in fair insurance fee calculation or fleet management tasks, for instance.

## 1 INTRODUCTION

The European Road Safety Observatory (European Commission, 2021) reports that from 2010 to 2020 there were approximately 1.000.000 crashes and 1.250.000 injuries, per year, in the European Union. In most countries, car insurance is required for one to have permission to drive on public roads (NationMaster, 2014). The owner or driver of a vehicle is responsible for damages it may cause, in case of accident. The interests of injured parties must be protected, regardless of whether or not the person responsible for the accident is financially able to do so. The mandatory insurance against civil liability for motor vehicles assures this protection. The insurance companies typically define vehicle insurance fees as functions of static variables, such as the age of the driver, the number of years one has a driving license, and the driving history (Hutson, 2021).


Nowadays, the volume of data which can be acquired and processed has increased exponentially, yielding new opportunities. The acquisition of driving behavior data brings opportunities for insurance and transportation companies to provide solutions which are more fair than the existing ones.


### 1.1 Goals of This Work

The key idea of this work is that the more aggressive/risky the driver actually behaves, the more one should pay by the insurance. On the other hand, the more conservative, well-behaved drivers should pay less money, leading to an overall fair system. A company that needs to perform fleet management can also take better planning decisions and to identify the best drivers, based on the assessment of their real driving style. This work has the following specific goals:

- (1) from anonymous data records with events generated on trips from different drivers, perform feature engineering actions to build a dataset suited for driver style analysis;
- (2) identify and perform the necessary data preprocessing operations on the dataset;
- (3) analyze the output of clustering techniques on the dataset, to assign a style to each trip.

The remainder of this paper is organized as follows. In Section 2, we overview related work and the driver data acquisition setup. The developed approach and the dataset details are described in Section 3. The experimental results are reported in Section 4. Finally, Section 5 ends the paper with some concluding remarks and directions for future work.

<sup>a</sup> <https://orcid.org/0000-0002-6508-0932>

<sup>b</sup> <https://orcid.org/0000-0001-8935-9578>

## 2 RELATED WORK

In this section, we address related work and the state-of-the-art regarding driver style identification (Section 2.1), the key aspects and goals of the i-DREAMS project (Section 2.2) and the trip data acquisition setup, devices, and events (Section 2.3).

### 2.1 Driving Style Taxonomy

In the literature, driving styles are established at different levels of specification, ranging from simple and single indicators, such as speeding or hard acceleration, to general concepts, such as aggressive driving or risky driving. These general concepts are based on a combination of specific behavioral indicators. In Taubman-Ben-Ari et al. (2004), eight driving styles regarding as how one usually drives are identified. These styles are organized into two main categories named as *safe* and *unsafe*. Within the safe category, there are three styles, as follows. The *distress-reduction* driving style is related to muscle relaxation techniques while driving. The *patient* style refers to calm and safe behaviors based on the “*better safe than sorry*” motto. The *careful* style addresses cautious behaviors and a state of mind to always react well and quickly to unexpected actions of other drivers. The unsafe category has five styles. The *dissociative* driving style is based on unaware and misjudging behaviors. The *anxious* style is related to nervous and worrying behavior while driving. *Risky* applies to drivers that like to take risks. The *angry* style means that the driver will respond with aggressive reactions to other drivers actions. Finally, the *high-velocity* style corresponds to an impulsive attitude towards driving faster than allowed by the road conditions or feeling impatient when traffic slows down. Table 1 presents each style and its category.

A framework in which driving styles are seen in terms of driving habits established as a result of individual dispositions, as well as social norms and cultural values was proposed by Sagberg et al. (2015). In this framework, a *global driving style* is composed by a set of *specific driving styles*. A specific driving style

Table 1: Driving categories (safe/unsafe) and the corresponding driving styles (Taubman-Ben-Ari et al., 2004).

Driving Categories	
Safe	Unsafe
Distress-reduction	Dissociative
Patient	Anxious
Careful	Risky
-	Angry
-	High-velocity

is a commonly adopted behavior while driving.

Typically, the set of parameters and variables for profiling drivers cover the most important aspects of driving (Singh and Kathuria, 2021). These parameters refer to the dynamic driving environment, based on crash data gathered from self-reported surveys (Peck and Kuan, 1983; Singh and Kathuria, 2021) as follows: (i) speed; (ii) acceleration, braking, and jerks; (iii) annual mileage; (iv) lateral maneuver, such as swerving, lane changing, and sharp turns; (v) time and space factors, such as time of day, day of the week, and month; (vi) distraction.

Recently, a method to detect city potential hazard driving zones, using bus tracking data was proposed by Almeida et al. (2022). The approach consists in mapping geolocation coordinates into road segments, using bus speed, bus maximum allowed speed, and bus acceleration to classify the driving behavior into one of six categories to evaluate the driving behavior. They conclude that some roads exhibit problems that occur in the same period of the day while other roads present circulation issues regardless of the time period or day.

### 2.2 The i-DREAMS Project

This work arises as part of the European Horizon2020 i-DREAMS project (i-DREAMS Team, 2021). The project proposes a solution with a set of devices installed in vehicles to acquire data about the driver’s status and driving behavior. The i-DREAMS project aims to determine the *safety tolerance zone* (STZ) for driving and interventions for driver-vehicle-environment interactions.

Figure 1 depicts a global overview of the i-DREAMS project. The Gateway (GW, in Figure 1) is the central element of the system, by collecting the data originated from other components and handling data connectivity and transmission. It is an edge computing device that determines the STZ, allows the triggering of alarms, and real-time communication with an *application programming interface* (API) or storage for post-trip synchronization. The access to the data is carried out through this API, which defines that each data collection system is assigned to a vehicle and not with a specific driver. Data is acquired anonymously during trips from the moment the vehicle’s engine is activated until the moment it is deactivated. By interacting with this API, we retrieve trip event data to build a dataset.

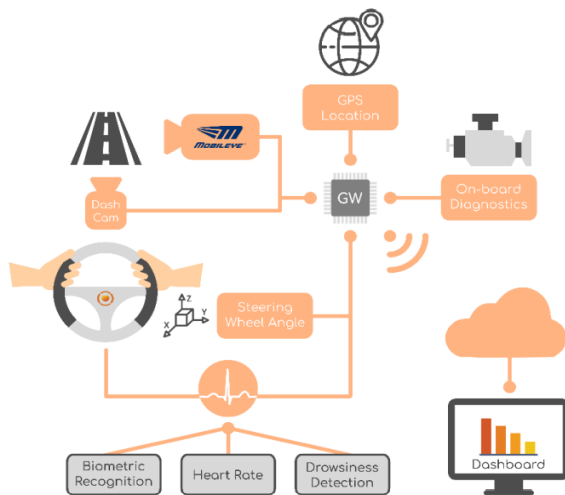


Figure 1: Global overview of the i-DREAMS on-vehicle systems (i-DREAMS Team, 2021).

### 2.3 Trip Data Acquisition

The devices in the i-DREAMS on-vehicle systems produce data events for the identification of the driving style. After data collection, the gateway makes it available through an API. The types of data provided by the main elements of the i-DREAMS architecture, are summarized in Table 2. The project also includes a smartphone application to detect driver distraction.

The CardioWheel device is installed on the steering wheel and the Mobileye (Mobileye, 2021) device is placed on the windshield. Figure 2 shows the sensors installed on a vehicle. CardioWheel can produce hands-on wheel detection events. The Drowsiness Detector device detects events of the driver’s drowsiness level according to the *Karolinska sleepiness scale* (KSS) (Shahid et al., 2011). KSS is a 10-point Likert scale (Joshi et al., 2015), to classify the states of human sleepiness in periods of five minutes, as described in Table 3.

Table 2: Data acquisition devices and data identifiers of the i-DREAMS project.

Device	Data Identifier
CardioWheel	LOD_Event
Driver App	Distraction
Drowsiness Detector	Drowsiness
Gateway	Ignition
Gateway Motion Sensor	DrivingEvents
Global Navigation Satellite System (GNSS)	Geolocation Coordinates
Mobileye	ME_AWS ME_TSR ME_CAR
Safety Tolerance Zone (STZ)	iDreams.Fatigue iDreams.Headway iDreams.Overtaking iDreams.Speeding

The Gateway generates data for ignition and driving events, corresponding to harsh driving detection events. The GNSS module provides satellite-based geolocation data, at a rate of one sample per second. The Mobileye device produces data about the safety and warning states in a certain timestamp (ME\_AWS). It also generates data about the traffic signs (ME\_TSR) and vehicle parameters (ME\_CAR). Finally, the STZ device produces fatigue, headway, overtaking, and speeding events.

## 3 THE DEVELOPED APPROACH

In this section, we describe the key aspects of the developed approach. We begin with the overview of the solution architecture, functionalities, and generated events (Section 3.1 and Section 3.2). Then, we describe the key steps and procedures to build the dataset from the events generated by the devices (Section 3.3). The dataset pre-processing steps are addressed in Section 3.4. Finally, Section 3.5 describes the clustering techniques we apply on the dataset.

### 3.1 Block Diagram of Our Approach

Figure 3 depicts the block diagram of the developed trip profiling system. It shows an example of some data types obtained by interacting with the i-DREAMS API. From the exported driving data events through the API, we perform a feature engineering step to build a dataset. From the set of events described in Table 2, we establish a set of features and we build an unlabeled dataset. Then, we apply unsupervised learning to cluster trips. The clustering output is analyzed by human experts to assess the trip style.

### 3.2 The i-DREAMS Generated Events

The CardioWheel device produces Hands-On Detection (LOD\_Event) events that signal the presence of

Table 3: The 10-point Karolinska Sleepiness Scale (KSS) (Shahid et al., 2011).

Level	Description
1	Extremely alert
2	Very alert
3	Alert
4	Rather alert
5	Neither alert nor sleepy
6	Some signs of sleepiness
7	Sleepy, but no effort to keep awake
8	Sleepy, but some effort to keep awake
9	Very sleepy, great effort to keep awake, fighting sleep
10	Extremely sleepy, can't keep awake



Figure 2: CardioWheel on the steering wheel (left). Mobileye and dashcam seen from the inside (middle) and from the outside of the vehicle (right).

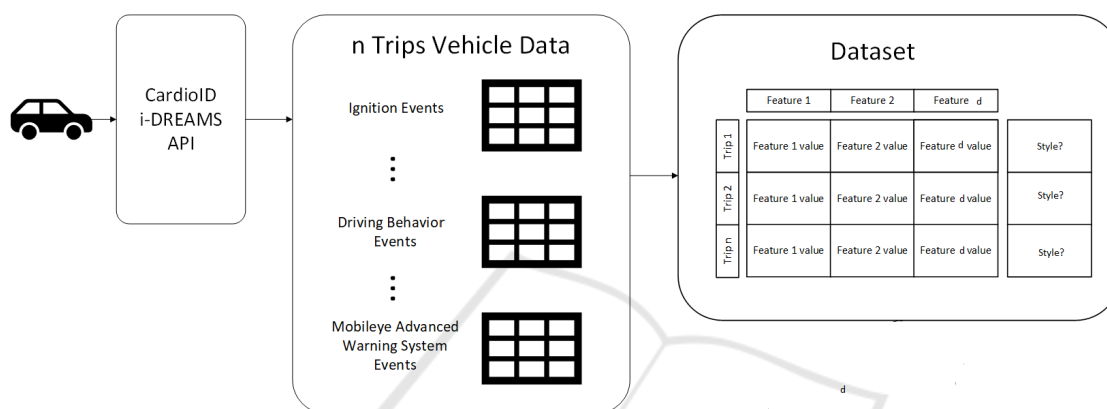


Figure 3: Proposed solution architecture to build and use a dataset with trip data for generic purposes. In this work, we also aim to identify the driving style for each trip, using the dataset.

the driver’s hands on the steering wheel. Each event consists of a timestamp and the Hands-On state (with four possible values). These events allow us to determine how much time the driver spent with no hands, left hand, right hand and both hands on the steering wheel. The Driver App provides real-time mobile phone use events, which are an indicator of distraction. Each sample is mapped by the type of event, mobile usage start, and mobile usage end. These events signal any distractions while driving. The Drowsiness Detector system is able to detect events of the driver drowsiness level based on the KSS scale, described in Table 3, at certain timestamps.

The Gateway Motion Sensor system generates driving behavior (DrivingEvents) data, which indicate the occurrence of harsh acceleration, harsh braking, and harsh cornering behaviors. Each event stores the timestamp and the type of event (harsh acceleration, braking, and cornering). Simultaneously, the maximum acceleration (in *g*-force) during an event is registered with the total duration of the event in seconds, and the event severity level is measured with one of three values: low, medium, or high. This indicator is useful to calculate the number of times each type of event occurs on a given trip. These events are widely used in the literature for driver profile classification

applications and are mostly associated with aggressive driving. The Gateway device is able to provide events regarding if the ignition is on or off.

The GNSS system offers satellite-based geolocation data, with a rate of about one sample per second. The Mobileye Advanced Warning System (ME\_AWS) produces data about the safety and warning state of the Mobileye system in a given timestamp. A significant part of this data refers to the system usage; for this work, the relevant events are:

- *fcw* - if true, forward collision warning is active;
- *hw\_level* - headway monitoring level (no car detected, green car, red car, warning);
- *ldw* - if true, lane departure warning is active (left or right);
- *ldw\_left* - if true, left lane departure warning is active;
- *ldw\_right* - if true, right lane departure warning is active;
- *pcw* - if true, pedestrian collision warning is active;
- *pedestrian\_dz* - if true, pedestrian detected in danger zone;
- *time\_indicator* - states the lighting conditions (day, dusk, or night);

- *tsr\_level* - traffic sign recognition level. Level of warning according to units over the speed limit (in km/h or mph);
- *zero\_speed* - if true, the vehicle is stopped.

The *fcw* parameter corresponds to the *forward collision warning* (FCW), which is activated when the system detects an imminent collision danger with cars ahead and triggers visible and audible warnings. The *hw* related parameters correspond to *headway monitoring & warning* (HMW), triggered when the vehicle is too close to the vehicle ahead and will raise visible and audible warnings. The *ldw* related parameters refer to the *lane departure warning* (LDW) event, activated when the vehicle departs from the current lane without turning signals and will also raise visible and audible warnings.

The Mobileye system also provides information about the vehicle parameters (ME\_CAR). Each data sample holds the current speed of the vehicle (expressed in km/h) and the following parameters:

- *brakes* - if brakes are on or off;
- *low\_beam* - if the low beam is on or off;
- *high\_beam* - if the high beam is on or off;
- *signal\_left* - if the left turn signal is on or off;
- *signal\_right* - if the right turn signal is on or off;
- *wipers* - if wipers are on or off.

In this work, we are not using the Mobileye Traffic Sign Recognition (ME\_TSR) events.

The STZ device generates real-time fatigue intervention (iDreams\_Fatigue) events with the following levels:

- 0 - no warning (normal driving);
- 1 - visual warning (dangerous driving);
- 2 - visual and auditory warning (dangerous driving);
- 3 - frequent warnings (avoidable accident).

STZ also generates the following real-time headway intervention (iDreams\_Headway) events:

- -1 - no vehicle detected (normal driving);
- 0 - vehicle detected, but headway  $\geq 2.5$  (normal driving);
- 1 - vehicle detected, headway  $< 2.5$ , but above warning threshold (normal driving);
- 2 - first warning stage (dangerous driving);
- 3 - second warning stage (avoidable accident).

STZ also provides real-time overtaking intervention (iDreams\_Overtaking) events with these levels:

- 0 - no warning (normal driving);
- 1 - visual warning (normal driving);

- 2 - visual and auditory warning (dangerous driving);
- 3 - frequent warnings (avoidable accident).

Finally, we have the real-time speeding intervention (iDreams\_Speeding) events:

- 0 - no warning (normal driving);
- 1 - visual indication (normal driving);
- 2 - visual speeding warning (dangerous driving);
- 3 - visual and auditory warning (avoidable accident).

### 3.3 Dataset Construction

After the analysis of the available events, the next step is to build a dataset through these events. To do this, for each event we devised several features, which we consider to be the most appropriate to determine the style of driving on a given trip. The events provided vary from trip to trip, since we are dealing with a naturalistic driving environment (et al, 2011). This includes aspects of vehicle movement, driver behavior, and the direct environment. As a result, there may be sensor failures on a given trip and as a consequence some features may have missing values, handled as described in Section 3.4. Although some events may not be available, all trips are represented by the following mandatory set of features:

- *trip\_start* - date and time trip started; this feature is also relevant to input missing values for other features such as the trip lighting condition;
- *trip\_end* - date and time trip ended; this feature is also relevant to input missing values for other features such as the trip lighting condition;
- *distance* - trip distance in kilometers;
- *duration* - trip duration in seconds.

The available data was collected and the features were designed. After a deeper analysis of the data, only the most relevant features were selected according to the literature. Through interaction with the i-DREAMS API, we gather trip data from April 1, 2021, to July 20, 2022, yielding a total of 17138 trips. The designed features are organized by the type of system that originated them. According to the four mandatory features, for each of the 17138 trips, 75 features were generated. However, for this collection of trips, the Hands-On and Drowsiness systems were not available, so only 67 features were kept. As a consequence, at this stage the dataset was composed of  $n = 17138$  instances (trips) and  $d = 67$  features.

### 3.4 Dataset Pre-Processing

After constructing the dataset, we performed data pre-processing by summarizing data and imputing missing values. We removed trips with less than one minute, trips with less than 1.5 km, and trips in which all features had missing values. To handle missing values, we locate features with such a problem. Subsequently, these values were imputed as follows:

- if the percentage of missing values for feature  $X_i$  is less than or equal to 50%, then these values are filled with the mean, median, or mode of the remaining values on that column;
- if this percentage is greater than or equal to 50%, the feature is removed.

We applied data normalization techniques because many clustering algorithms are based on distances, and the different orders of magnitude of the feature values influences the result (Witten et al., 2016). Thus, we normalize the data by trip distance or by trip duration, separately. Then, we evaluate both and choose the one with the best results. We have also removed the features unrelated with the objective of this work (to identify a trip style), yielding a dataset with  $n = 15002$  instances (trips) and  $d = 53$  features.

### 3.5 Clustering Algorithms

The clustering phase is organized into two clustering stages. In the first, we applied different clustering techniques. Initially, we use the K-Means algorithm (Hartigan and Wong, 1979). The strategy consisted in initially trying to identify the best value of  $K$  through the elbow and silhouette methods. Density-based spatial clustering (DBScan) (Ester et al., 1996) was also applied to check for outliers in the dataset. The parameters were set as in Sander et al. (1998). Another technique used was *gaussian mixture models* (GMM) (Reynolds, 2009) that states the probability that each instance of the dataset belongs to a Gaussian distribution. The advantage of using this type of clustering is that it can better deal with different shapes of the data distribution. The K-Means algorithm with Euclidean distance handles spherical shapes better. The number of components/gaussians used was the same as  $K$  in K-Means. Finally, *evidence accumulation clustering* (EAC) by Fred and Jain (2002) with K-Means (Okun and Valentini, 2008) was also evaluated.

The objective of the second stage of clustering was to use the best combination of the first stage, to check if the clusters found could be further decomposed into sub-clusters, given more details on the drivers pro-

files. After running the clustering algorithm, we assign a trip style to each cluster, by human analysis and inspection.

## 4 EXPERIMENTAL EVALUATION

In this section, we address the experimental evaluation of the developed solution, focusing on the goal to devise trip styles with clustering techniques, from the built dataset. The source code was written in Python. Section 4.1 describes the standard clustering evaluation metrics. In Section 4.2, we address the use of dimensionality reduction techniques. Finally, Section 4.3 reports the experimental results of clustering and the identification of trip styles.

### 4.1 Clustering Evaluation Metrics

The clustering algorithms results were evaluated with the Calinski-Harabasz (CH) (Caliński and Harabasz, 1974), Silhouette (Sil) (Rousseeuw, 1987), and Davies-Bouldin (DB) (Davies and Bouldin, 1979) scores. The CH metric, also known as the variance ratio criterion, is the ratio of the sum of between-clusters dispersion and inter-cluster dispersion for all clusters; the higher the score, the better the performance. The Silhouette score ranges from -1 (worst) to 1 (best); it is computed as functions of the average intra-cluster distance and the average inter-cluster distance. The DB score is a function of the ratio of the within cluster scatter to the between cluster separation and a lower value means better clustering.

### 4.2 Dimensionality Reduction

We have considered the use of dimensionality reduction techniques to reduce the number of features, and avoid the *curse of dimensionality* (Bishop and Nasrabadi, 2006). Since the dataset does not have pre-labeled data, we chose an unsupervised *feature reduction* (FR) technique, namely *principal component analysis* (PCA) (Jolliffe, 2002). The use of PCA yielded the following results: (i) with data normalization, the number of features was reduced from  $d = 53$  to  $m = 17$ ; (ii) without normalization, we got  $m = 1$  which is an extreme dimensionality reduction causing loss of information. Thus, we opted to consider the data with normalization by distance or by duration.

### 4.3 Clustering the Trip Dataset

For the first stage clustering, we use the normalization and dimensionality reduction techniques. The experi-

mental results of the best combination, for each algorithm, are reported in Table 4.

Table 4: First stage clustering results of the  $n = 15002$  trips, by K-Means, DBScan, GMM, and EAC algorithms evaluated with the CH, Sil, and DB scores. We aim to maximize the CH and Sil scores (denoted by the  $\uparrow$  symbol) and to minimize the DB score (the  $\downarrow$  symbol).

Scores	K-Means	DBScan	GMM	EAC
CH $\uparrow$	<b>7258.919</b>	2365.621	5602.887	1965.691
Sil $\uparrow$	0.488	0.650	0.382	<b>0.807</b>
DB $\downarrow$	1.164	1.372	1.377	<b>0.482</b>
# cluster 0	11916	14550	5287	54
# cluster 1	3086	452	9715	14948
Total	15002	15002	15002	15002

Although the results state that EAC has better DB and Silhouette scores, further cluster analysis has shown that its results yield large imbalance between clusters on different runs of the algorithm over the training data. We choose the K-Means algorithm because it was more consistent on successive runs and the scores were better overall when comparing to DB-Scan and GMM. In summary, normalization by distance, dimensionality reduction by PCA, and clustering with K-Means yield the best results. Moreover, using PCA evaluation we found that the most important features (for the first PCA component) were: speed, number of speeding events, number of times breaks are on, number of harsh cornering events, and the number of harsh acceleration events.

The second stage of clustering consisted in applying K-Means individually to cluster 0 and to cluster 1, in Table 4, to check if they can be further decomposed into sub-clusters. The experimental results showed that cluster 1 attained by K-Means was further decomposed in two clusters. Table 5 provides the scores for the second stage of clustering.

Table 5: Second stage clustering results with CH, Sil, and DB scores. The 3086 instances of # cluster 1 by K-Means, in Table 4, are further split into two clusters with 3033 and 53 instances, respectively.

CH $\uparrow$	Sil $\uparrow$	DB $\downarrow$	# cluster 1.1	# cluster 1.2
1144.656	0.735	0.548	3033	53

Table 6 and Figure 4 report the final results for unsupervised learning. After human inspection of the clusters and feature values, cluster 0 was assigned as *aggressive trips* while cluster 1 was defined as *non-aggressive trips*, and cluster 2 holds the *risky trips*. In detail, cluster 0 represents trips with low-speed values per km and low speeding events per km. Cluster 1 represents trips with low to medium speed per km and low to medium speeding events per km. In terms of the number of braking events per km, cluster 1 has

more events. For cluster 2, the number of events exceeding the speed limit was the most important feature to differentiate, aggressive trips from risky trips. Risky trips involve more speeding events than aggressive trips.

Table 6: Second stage final clustering results - distribution of the  $n = 15002$  trips per  $K = 3$  clusters, holding the trip styles, after human expert analysis.

Cluster ID	Description	Number of instances
0	Aggressive trips	3033
1	Non-Aggressive trips	11916
2	Risky trips	53
	Total	15002

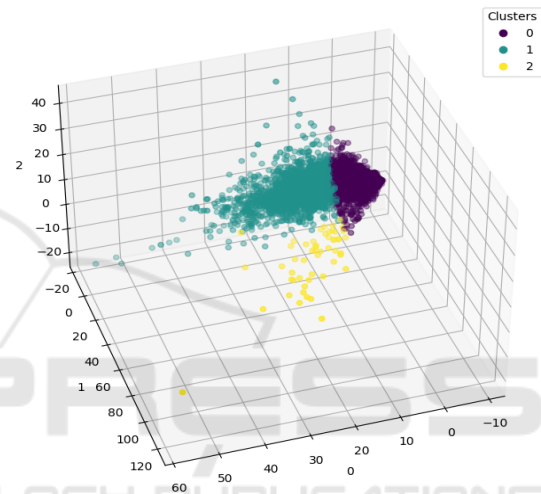


Figure 4: A 3D visualization of the final clustering results, with the *aggressive trips* (0), *non-aggressive trips* (1), and *risky* (2) trip styles.

## 5 CONCLUSIONS

The monitoring of driver behavior based on real trip data acquired from a vehicle yields many useful applications. In this paper, our key goal was to show that it is possible to devise a driver style identification solution, from a built dataset from the i-DREAMS project devices. Using anonymous trip data, our approach was to acquire data from a set of trip events and to perform feature extraction to build a dataset. Then, we apply some pre-processing techniques on the dataset. Afterwards, we perform clustering, evaluated with standard metrics. The dataset construction and pre-processing phases were the ones that most influenced the clustering results. By normalizing the trips by distance and performing dimensionality reduction by principal component analysis we achieved the best results. We applied a two-stage clustering

strategy, in which the K-Means algorithm has shown to be the best. The most challenging part of the clustering phase was to assign meaning (trip style) to the clusters, which implied further human analysis and inspection on the clustering results. Thus, the proposed solution was found to be able to identify the trip styles in a satisfactory way.

As future work, we may use different clustering algorithms and perform a deeper analysis of the evidence accumulation algorithm. We can use feature selection techniques instead of feature reduction, to identify the most relevant original features for the trip style identification task and the smallest subset of these features that yield robust classification.

## ACKNOWLEDGEMENTS

This work was co-funded by the EU H2020 i-DREAMS project (Project Number: 814761) funded by European Commission under the MG-2-1-2018 Research and Innovation Action (RIA), [www.idreamsproject.eu](http://www.idreamsproject.eu), and by the European Regional Development Fund (FEDER), through Portugal 2020, under the Operational Competitiveness and Internationalization (COMPETE 2020) and Lisboa 2020 programmes (grants no. 069918 “CardioLeather”).

## REFERENCES

- Almeida, A., Brás, S., Sargento, S., and Oliveira, I. (2022). Using bus tracking data to detect potential hazard driving zones. In *Pattern Recognition and Image Analysis*, pages 667–679, Cham. Springer International Publishing.
- Bishop, C. and Nasrabadi, N. (2006). *Pattern recognition and machine learning*, volume 4. Springer.
- Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27.
- Davies, D. and Bouldin, D. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227.
- Ester, M., Kriegel, H., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Knowledge Discovery in Databases (KDD)*, volume 96, pages 226–231.
- et al, I. S. (2011). Towards a large scale European naturalistic driving study: final report of PROLOGUE: deliverable D4.2. Technical report, SWOV Institute for Road Safety Research.
- European Commission (2021). Road safety, <https://ec.europa.eu/transport>.
- Fred, A. and Jain, A. (2002). Data clustering using evidence accumulation. In *International Conference on Pattern Recognition*, volume 4, pages 276–280 vol.4.
- Hartigan, J. and Wong, M. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. series c (applied statistics)*, 28(1):100–108.
- Hutson, D. (2021). How car insurance premiums are calculated <https://www.comparethemarket.com/car-insurance/content/what-impacts-upon-your-car-insurance/>.
- i-DREAMS Team (2021). A smart driver and road environment assessment and monitoring system.
- Jolliffe, I. (2002). *Principal component analysis*. Springer Series in Statistics.
- Joshi, A., Kale, S., Chandel, S., and Pal, D. (2015). Likert scale: Explored and explained. *British Journal of Applied Science & Technology*, 7(4):396–403.
- Mobileye (2021). Autonomous driving & ADAS (Advanced Driver Assistance Systems).
- NationMaster (2014). Motor vehicle ownership per 1000 inhabitants, <https://ourworldindata.org/>.
- Okun, O. and Valentini, G. (2008). *Supervised and Unsupervised ensemble methods and their applications*, volume SCI, 126. Springer.
- Peck, R. and Kuan, J. (1983). A statistical model of individual accident risk prediction using driver record, territory and other biographical factors. *Accident Analysis & Prevention*, 15(5):371–393.
- Reynolds, D. (2009). Gaussian mixture models. *Encyclopedia of biometrics*, 741(659–663).
- Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–65.
- Sagberg, F., Selpi, Piccinini, G., and Engström, J. (2015). A review of research on driving styles and road safety. *Human factors*, 57(7):1248–1275.
- Sander, J., Ester, M., Kriegel, H., and Xu, X. (1998). Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. *Data Mining and Knowledge Discovery*, 2(2):169–194.
- Shahid, A., Wilkinson, K., Marcu, S., and Shapiro, C. (2011). Karolinska sleepiness scale (KSS) - stop, that and one hundred other sleep scales. A. Shahid and K. Wilkinson and S. Marcu and C. Shapiro (eds), pages 209–210.
- Singh, H. and Kathuria, A. (2021). Profiling drivers to assess safe and eco-driving behavior—a systematic review of naturalistic driving studies. *Accident Analysis & Prevention*, 161:106349.
- Taubman-Ben-Ari, O., Mikulincer, M., and Gillath, O. (2004). The multidimensional driving style inventory—scale construct and validation. *Accident Analysis & Prevention*, 36(3):323–332.
- Witten, I., Frank, E., Hall, M., and Pal, C. (2016). *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann, 4th edition.