Transportation Research Interdisciplinary Perspectives 41 (2027) 102344

Contents lists available at ScienceDirect

Transportation Research Interdisciplinary Perspectives

journal homepage: www.elsevier.com/locate/trip

Explainable machine learning for trip-level driving behaviour profiling from smartphone telematics in road freight operations☆

Pedro Entringer [a], Fátima Leal [a,*], Micaela Pinho [a,b]

[a] Research on Economics, Management and Information Technologies (REMIT), Portucalense University, Porto, Portugal
[b] Portucalense Legal Institute (IJP), Portucalense University, Porto, Portugal

ARTICLE INFO

Keywords:
Driver behaviour
Smartphone telematics
Road freight transport
Explainable machine learning
Fleet management
Eco-driving proxy

ABSTRACT

Low-cost smartphone telematics can support transparent trip-level profiling in road freight operations, but telemetry-derived labels require careful interpretation. We analyse an extract of 52 freight trips by 12 drivers (11,434.91 km) associated with a platform reporting 753,021 source location records. Distance-normalised acceleration, braking and route-alert indicators were used to construct four rule-based operational tiers (Excellent, Good, Moderate and Poor); K-means was used for exploratory description only. Decision tree (DT), random forest (RF) and k-nearest neighbours were evaluated using a driver-grouped hold-out (37 training and 15 test trips) and leave-one-driver-out cross-validation. In the hold-out, RF achieved accuracies of 0.867 for driving behaviour and 0.933 for the eco-driving proxy. Removing the score-generating indicators and their algebraically related features reduced RF accuracy to 0.467 and 0.533, respectively. The corresponding leave-one-driver-out RF accuracies were 0.942 and 0.962 using the full feature set, decreasing to 0.385 and 0.365, respectively, under strict feature ablation. Class-specific errors and local surrogate explanations further characterised model behaviour. The results support reproducible assignment of an operational telemetry taxonomy, not independently validated predictions of driver safety, energy consumption or emissions. The small driver cohort and missing vehicle, traffic and payload information limit generalisation and causal interpretation.

1. Introduction

Road freight transport is a data-rich operational environment in which driver behaviour affects safety, service reliability, vehicle wear, and energy use. From an operational economics perspective, these factors are relevant to transport costs, resource efficiency and the overall performance of freight operations. Fleet operators increasingly collect digital traces through smartphones and telematics platforms, creating opportunities to move from retrospective incident reporting towards continuous behavioural analysis. Smartphone-based systems are particularly attractive because they can be deployed without dedicated on-board hardware and can therefore support heterogeneous fleets at comparatively low cost (Castignani et al., 2013; da Silva et al., 2019; Bouhsissin et al., 2023).

The analytical opportunity is accompanied by two methodological problems. First, behavioural labels such as "excellent", "moderate", or "poor" are rarely observed directly. They are commonly constructed from thresholds, scores, or unsupervised groupings, which means that predictive performance may indicate consistency with an operational labelling scheme rather than validity against an external outcome. Second, high-performing models may be difficult for fleet managers and drivers to understand. In employment-related settings, opaque classifications can undermine trust and may encourage inappropriate punitive use. Explainable artificial intelligence (XAI) is therefore not an optional visual add-on; it is part of the governance architecture needed to connect model outputs with proportionate interventions.

Recent transport research illustrates the breadth of machine-learning applications to driver state, driver response, and freight decision-making (Gouribhatla and Pulugurtha, 2022; Meteier et al., 2024; Mascheroni et al., 2026). Nevertheless, relatively few studies combine real-world road-freight data, low-cost smartphone sensing, driver-grouped validation, and local explanations within a single operational pipeline. Moreover, sustainability claims are often inferred from harsh acceleration or braking without directly measuring fuel, electricity, or emissions. A more defensible approach is to treat such variables as *eco-driving proxies*: behavioural indicators that may be relevant to energy efficiency but do not constitute environmental outcomes by themselves.

☆ This article is part of a Special issue entitled: 'AI in Transportation' published in Transportation Research Interdisciplinary Perspectives.

* Corresponding author.
E-mail address: fatimal@upt.pt (F. Leal).

https://doi.org/10.1016/j.trip.2026.102344
Received 29 July 2026; Received in revised form 22 October 2026; Accepted 27 October 2026
Available online 9 November 2026
2590-1982/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

This study examines whether smartphone-derived event rates can support transparent behavioural profiling in road freight operations. The empirical setting is the Portuguese "Meu Rastreio" logistics platform, which records route locations through Android and iOS devices. The analysis is guided by three research questions:

RQ1. What operational trip-level behavioural profiles can be constructed from kilometre-normalised behavioural events derived from smartphone traces?

RQ2. How accurately can interpretable classifiers assign trip-level operational profiles to trips performed by drivers not represented in the training data?

RQ3. Which behavioural features contribute to individual model classifications, and what are the implications of these explanations for accountable fleet-management decision-making?

The study makes four contributions. First, it documents the construction and operational use of trip-level behavioural indicators derived from smartphone telematics, explicitly distinguishing the historical data-processing stage from the independently reproducible trip-level analysis. Second, it separates exploratory profile construction from supervised deployment and explicitly limits the interpretation of cluster- and rule-derived labels. Third, it evaluates classification performance using driver-grouped hold-out testing and leave-one-driver-out cross-validation, complemented by feature ablation to assess the dependence of predictive performance on the indicators used to construct the operational labels. Fourth, it combines model performance with decision-tree and LIME-style local explanations and discusses how the results can be used for coaching and operational review without treating the model as an independently validated safety or sustainability score.

2. Interdisciplinary background

2.1. Smartphone telematics and driver profiling

Driver profiling uses repeated behavioural observations to characterise how a person operates a vehicle. Signals may originate from dedicated vehicle sensors, on-board diagnostics, smartphones, cameras, or combinations of these sources. Smartphone approaches commonly use GPS, accelerometers, gyroscopes, and timing information to identify speeding, harsh manoeuvres, route deviations, and driver-state patterns (Castignani et al., 2013; Brahim et al., 2022). Their main advantages are scalability and low deployment cost, whereas their principal limitations include variable sensor quality, mounting position, missing connectivity, and lower temporal resolution than dedicated instrumentation.

The unit of analysis is consequential. Event counts are strongly affected by distance and exposure; a driver covering more kilometres has more opportunities to trigger alerts. Normalising events by distance therefore improves comparability. It does not, however, remove route, traffic, weather, vehicle, or scheduling effects. Behavioural profiles should consequently be interpreted as conditional operational indicators rather than immutable personal attributes.

Machine learning has been used to identify aggressive driving, fatigue, workload, and other driver states (Wu et al., 2024; Meteier et al., 2024). In road freight operations, the practical value of such models depends on whether they can be integrated into decisions such as training, route allocation, safety review, and maintenance planning. The recent review by Mascheroni et al. (2026) emphasises the need to connect technical models with decision-making processes and to validate them using real operational data.

2.2. From profile discovery to operational classification

Unsupervised learning can reveal structure where no accepted labels exist. K-means is often used because it is computationally efficient and easy to operationalise, although its results depend on scaling, initialisation, the selected number of clusters, and the assumption of approximately spherical groups (Timmerman et al., 2013; Miraftabzadeh et al., 2023). Cluster membership is not ground truth. It is a data-dependent taxonomy whose substantive meaning must be established through external outcomes or domain evidence.

A supervised model trained to reproduce cluster- or rule-derived labels serves a different purpose from a conventional predictive model. It can provide a deployable approximation of an operational taxonomy, but its accuracy does not establish that the taxonomy corresponds to crash risk, fuel efficiency, or environmental performance. This distinction is central to the present study. The supervised stage is evaluated as a profile-assignment mechanism for unseen drivers, not as external validation of the profile labels.

2.3. Explainability and human-centred fleet decision support

Decision trees provide global and local explanations through explicit paths, while random forests generally require post-hoc methods. LIME approximates a complex model locally around one observation and identifies the features supporting or opposing a specific prediction (Ribeiro et al., 2016). Such explanations can help fleet managers move from a generic score to a review of concrete behaviours, for example recurrent speeding alerts or a high rate of adverse curve events.

Explanations must nevertheless be used with caution. Local explanations are model- and instance-specific, may vary with the perturbation process, and do not establish causality. In an employment context, appropriate use should include human review, access to contextual information, opportunities to contest erroneous records, and a preference for supportive interventions over automated sanctions.

2.4. Eco-driving proxies and the limits of sustainability claims

Harsh acceleration, abrupt braking, excessive speed, and poor anticipation are associated in the literature with fuel use, emissions, and vehicle wear (Ferreira et al., 2017; van Nes et al., 2019; Al Hla et al., 2019). However, a behavioural event is not equivalent to a measured environmental outcome. Vehicle mass, engine or powertrain, road gradient, load, weather, congestion, and regenerative braking can substantially change the energy consequences of the same manoeuvre.

Accordingly, this study uses the term *eco-driving proxy profile* for a label based on smoothness- and alert-related events. It does not claim to estimate fuel consumption, electricity consumption, carbon emissions, or life-cycle sustainability. This narrower terminology improves construct validity and clarifies what must be added in future external validation.

3. Materials and methods

3.1. Research design and sample structure

The study uses route records retained from the Portuguese "Meu Rastreio" logistics platform, which collected smartphone-based location and speed information over approximately three months. The original system reported 753,021 location observations recorded at nominal spatial intervals of approximately 100 m. The retained analytical extract contains 52 distinct trips (`shipment_id`) from 12 pseudonymised drivers (`driver_id`), totalling 11,434.91 km. The mean and median trip distances were 219.90 and 105.67 km, respectively; each driver contributed between one and eleven trips. Trips, rather than raw location records or intermediate windows, are the statistical observations used in all classification results. Vehicles could not

Table 1
Source telemetry and retained trip-level analytical sample.

Characteristic	Value
Location records reported by the original platform	753,021
Source collection period	Approximately three months
Trips in the retained analytical dataset	52
Unique drivers	12
Total trip distance	11,434.91 km
Mean (median) trip distance	219.90 (105.67) km
Trips per driver (minimum–maximum)	1–11
Unique vehicles and window count	Not recoverable

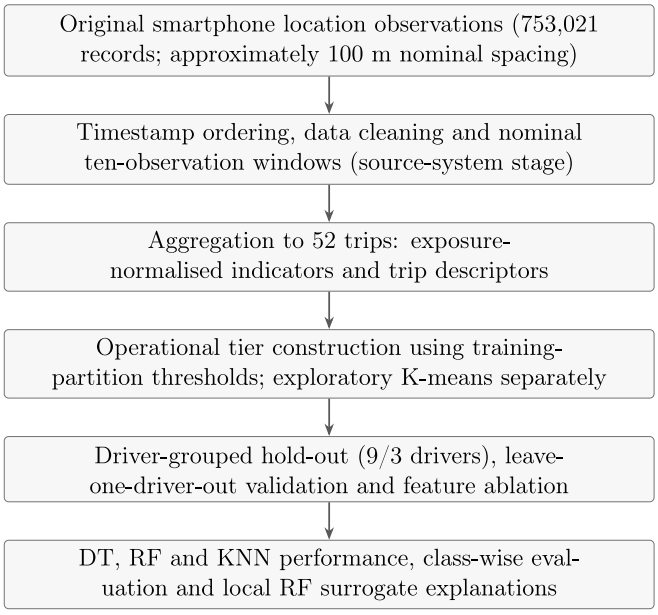


Fig. 1. Analytical hierarchy and evaluation pipeline. The point-level processing stage is described from archived documentation; the trip-level experiments are independently reproduced from the retained extract.

be counted reliably because no persistent trip-linked vehicle identifier is available in the analytical extract.

The hierarchy is location observations → approximately kilometre-scale groups of observations → accumulated trip-level counts and descriptors → driver-grouped evaluation. A driver identifier defines the train–test grouping and is never provided to a classifier. Because the original point-level observations were not retained in the available extract, the direct correspondence between all 753,021 source records and the 52 selected trip summaries, including the number of intermediate windows, cannot be independently re-established. Table 1 summarises the effective sample (see Fig. 1).

3.2. Source preprocessing and trip-level feature construction

According to the available platform documentation, location observations were ordered by acquisition time, including buffered observations transmitted after a period offline, and duplicates were removed before feature extraction. Ten successive observations were organised into intermediate, approximately kilometre-scale groups. This is a nominal spatial approximation, not a guarantee of exactly 1 km per group. The original executable rules for duplicate detection, incomplete groups, temporal gaps and de-duplication of adjacent flagged events were not available for independent verification. Event counts in the analytical extract were therefore retained as delivered and were not reconstructed from synthetic point records.

The 18 non-identifier columns in the trip-level extract consist of distance (`distance`, km), maximum and average speed (`max_speed` and `avg_speed`, m/s), offline-record percentage (`offline_records`), a platform-defined residual category (`default`), counts of sudden braking and acceleration, general speeding alerts, five route-geometry event counts (`line`, `smooth_curve`, `moderate_curve`, `sharp_curve`, `tight_curve`), and their five `_alert` counterparts. Geometry counts describe platform-tagged events, not kilometres driven on each road type. The platform's reported angle breakpoints and operating speed caps are given in Table 2; exact decision boundaries were not recoverable from source code.

Table 2
Archived route-geometry categories and associated operational speed caps. These are platform settings, not independently verified road-design measurements.

Category	Nominal direction-change breakpoint	Speed cap (km/h)
Straight segment	0°	120
Smooth curve	30°	100
Moderate curve	60°	80
Sharp curve	90°	60
Tight curve	120°	40

Sudden-acceleration and sudden-braking events were based on speed changes and the source-system threshold magnitude of 2.74 m s^{-2} (Hwang et al., 2018). For speed v_i at timestamp t_i, the conventional discrete longitudinal-acceleration proxy is

$$\hat{a}_i = \frac{v_i - v_{i-1}}{t_i - t_{i-1}}, \qquad t_i > t_{i-1}. \tag{1}$$

Under this conventional signed definition, values above +2.74 m s^{-2} indicate sudden-acceleration candidates and values below −2.74 m s^{-2} sudden-braking candidates. Eq. (1) explains the intended physical interpretation; it is not a verified reconstruction of the inaccessible original timestamp-gap and event-detection code. The relatively sparse sampling limits manoeuvre precision. Fig. 2 illustrates the source-system event display.

For trip t, let $L_t > 0$ denote travelled distance in kilometres and $N_{t,k}$ the count for event type k. The exposure-normalised rate is

$$r_{t,k} = \frac{N_{t,k}}{L_t} \quad \text{(events/km)}. \tag{2}$$

No factor of ten is applied: ten-record source groups and normalisation per kilometre are separate operations. All 52 analytical trips had non-missing positive distance values; no trip-level imputation was needed. The platform's offline-record percentage is treated as a data-quality measure, not evidence of driver compliance.

3.3. Operational scores, thresholds and profile labels

Two separate trip-level scores were defined from adverse-event counts:

$$E_t = \frac{N_{\text{acc},t} + N_{\text{brake},t}}{L_t}, \tag{3}$$

$$B_t = \frac{N_{\text{lineAlert},t} + N_{\text{smoothAlert},t} + N_{\text{moderateAlert},t} + N_{\text{sharpAlert},t} + N_{\text{tightAlert},t}}{L_t}. \tag{4}$$

Here E_t is an *eco-driving proxy* (not measured energy or emissions) and B_t a route-specific alert score for *driving behaviour*. Each event has unit weight, without geometry- or severity-dependent weighting. In the retained data, the sum of all five geometry-specific alert counts equals the sum of sudden-acceleration, sudden-braking and general speeding-alert counts in every trip. The two families therefore contain algebraically related information and should not be treated as independent measurements.

For either score x, three data-driven thresholds were estimated *from the training trips only*:

$$T_2 = \text{mean}(x_t : t \in \mathcal{T}_{\text{train}}), \tag{5}$$

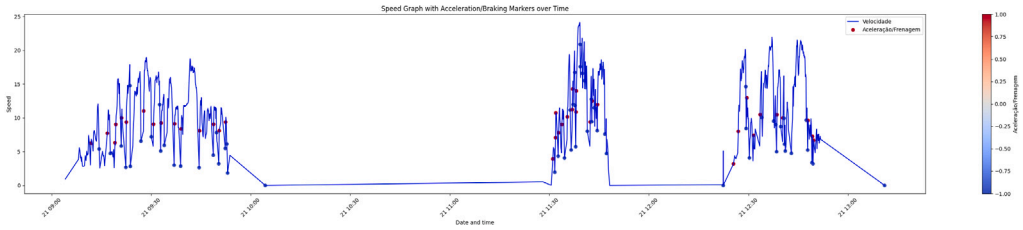


Fig. 2. Illustrative speed series and event flags from the source platform; not a direct reconstruction from the retained aggregate dataset.

Table 3
Operational-score thresholds in adverse events/km. Descriptive thresholds use all 52 trips; the independent hold-out thresholds use its 37 training trips only.

Profile task	Estimation sample	T_1	T_2	T_3
Driving behaviour	All trips (descriptive)	0.7319	1.2308	1.7297
	Hold-out training trips	0.7662	1.2292	1.7179
Eco-driving proxy	All trips (descriptive)	0.6300	1.1301	1.6301
	Hold-out training trips	0.6584	1.1256	1.6187

Table 4
Prespecified model settings of the reproducible trip-level re-analysis.

Model	Settings
DT	Gini; `max_depth=3`; `max_features=sqrt`; `random_state=42`; `min_samples_leaf=1`; `min_samples_split=2`.
RF	100 trees; Gini; `max_depth=3`; `max_features=sqrt`; `random_state=42`; `oob_score=True`; `n_jobs=1`.
KNN	$k = 3$; inverse-distance weighting; automatic neighbour search; training-only `StandardScaler`.

$$T_1 = \text{mean}(x_t : t \in \mathcal{T}_{\text{train}},\ x_t < T_2), \tag{6}$$

$$T_3 = \text{mean}(x_t : t \in \mathcal{T}_{\text{train}},\ x_t \geq T_2). \tag{7}$$

A held-out trip's label is assigned using these unchanged training-derived thresholds:

$$\text{Profile}(x_t) = \begin{cases} \text{Excellent}, & x_t \leq T_1, \\ \text{Good}, & T_1 < x_t \leq T_2, \\ \text{Moderate}, & T_2 < x_t \leq T_3, \\ \text{Poor}, & x_t > T_3. \end{cases} \tag{8}$$

Thresholds are re-estimated in each independent cross-validation fold. Table 3 reports both full-sample descriptive cutoffs (for characterising the 52 trips) and the exact training-derived hold-out cutoffs (for evaluating the models). These should not be interchanged.

The labels are rule-defined, context-dependent descriptions of trips. They are not independent ground truth for crash risk, driver quality, fuel efficiency or environmental impact.

3.4. Exploratory K-means clustering

K-means was used *separately* on the standardised one-dimensional E_t and B_t score distributions (52 trips per task), not to generate the supervised labels. Scores were z-standardised and clustered with Euclidean distance, `n_init=50`, `max_iter=300` and `random_state=42`. Values $k = 2, \ldots, 6$ were assessed with silhouette coefficients and within-cluster inertia. The historical three-group representation was retained for descriptive comparison, although it does not maximise silhouette. Because clustering is performed on one scalar score at a time, z-standardisation does not change membership for a fixed k. Stability of the $k = 3$ solutions was assessed using 300 driver-level bootstrap resamples (sampling drivers with replacement), refitting the clustering and assigning the original trips to bootstrap centroids; the adjusted Rand index (ARI) was computed against the reference partition. This is an exploratory cluster-stability analysis on the available cohort, not an external validation of the four tiers.

3.5. Classifiers, validation and strict ablation

An independent trip-level re-analysis was carried out to provide a reproducible alternative to the historical student models, whose fitting code and original split could not be recovered. The full predictor matrix contains 20 variables: the 18 non-identifier trip-level columns and the two derived scores in Eqs. (3)–(4). Neither driver nor shipment identifier is included. A single `GroupShuffleSplit` (`test_size=0.25`, `random_state=0`) assigned nine drivers (37 trips) to training and three previously unseen drivers (15 trips) to testing. The same division was used for all models and tasks. Operational thresholds were estimated solely on the training trips; test labels were assigned afterwards using those thresholds. The analysis therefore evaluates reproduction of training-defined operational rules, not independently measured outcomes.

For a broader within-cohort evaluation, *leave-one-driver-out cross-validation* (LODO-CV) comprised 12 folds, each retaining one complete driver's trips for testing and the other 11 drivers for training. Thresholds and preprocessing parameters were re-estimated in the training portion of each fold, and test predictions were pooled across folds before computing accuracy, macro F1 and weighted F1. Per-class precision, recall, F1 and support were computed on the same predictions; a zero-division convention of zero was used when a class was never predicted. With only 12 drivers and unequal trip counts, the estimates remain imprecise and dependent on the observed cohort.

The classifier settings in Table 4 were fixed before the re-analysis, using one configuration described in the archived source manuscript. This was *not* a new data-driven hyperparameter search; the earlier student's model-selection procedure cannot be verified. KNN predictors were standardised using training-only means and population standard deviations (`StandardScaler`); trees were trained on the original numerical scales. No preprocessing statistics, class thresholds or model settings were selected using test trips. RF out-of-bag scoring was not used for model selection.

To quantify target–feature dependence, both derived scores, trip distance, sudden-acceleration and sudden-braking counts, general speed alerts, and all five geometry-specific alert counts were *jointly* removed before refitting each classifier. This strict ablation is necessary because geometry-alert totals equal acceleration, braking and speeding totals in the extract: ablating only the obvious numerator for one task would allow reconstruction from correlated event columns. Nine predictors remained (maximum and mean speed, offline-record percentage, the `default` count, and five geometry-specific non-alert counts). Both full and ablated models used identical train/test drivers, parameter settings and evaluation rules.

The trip-level experiments used Python 3.13.5, pandas 2.2.3, NumPy 2.3.5 and scikit-learn 1.8.0. The reproducibility archive includes the fitting script, feature definitions, driver partitions, fold-specific thresholds, per-trip predictions, confusion matrices and summary metrics. Its release is subject to permission for the privately sourced trip-level records. It does not reproduce the inaccessible original GPS-to-event-processing code.

Table 5
Observed trip counts by operational tier: descriptive full dataset and independent hold-out test set.

Tier	Driving behaviour		Eco-driving proxy	
	All 52	Test 15	All 52	Test 15
Excellent	14	5	13	5
Good	12	2	13	2
Moderate	12	4	14	4
Poor	14	4	12	4
Total	52	15	52	15

3.6. Contextual sensitivity analysis

Because adverse-event rates may be influenced by exposure conditions, Spearman rank correlations were calculated between both operational scores and trip distance, average speed, offline-record percentage and a route-event composition indicator. The latter is $q_{\text{tight},t} = N_{\text{tight},t} / \sum_{g \in G} N_{g,t}$, with G comprising the five geometry categories. It measures the share of *geometry-tagged events*, not the proportion of route length driven through tight curves. Associations were accompanied by percentile 95% confidence intervals from 1200 driver-level bootstrap samples, preserving within-driver trip grouping. All analyses are descriptive: neither traffic nor vehicle type, payload, weather, road gradient or independent geometric exposure is available per trip, precluding their independent adjustment.

3.7. Local model explanations

Decision-tree paths provide direct descriptions of learned splitting rules. To examine RF predictions, an *independently implemented LIME-style* local surrogate procedure was applied to a held-out trip (`shipment_id=341`) for each task, following the local model-approximation principle of Ribeiro et al. (2016). The approach is not an exact execution of the official `lime` Python package. Training-trip distributions defined quartile bins for each of the 20 numeric predictors. For each explained instance, 5000 observations were generated by independently sampling these bins and sampling values within bins; the held-out instance itself was included. The trained RF returned class probabilities for perturbed records. Binary indicators denoting whether each feature belonged to the observed instance's quartile interval were weighted by an exponential proximity kernel based on their Euclidean distance, with kernel width $0.75\sqrt{p}$ ($p = 20$). A weighted ridge surrogate was fitted for the predicted class; the eight selected locally influential intervals were displayed. The random seed was 42, and local fidelity was quantified by weighted R^2. Code and settings are supplied for reproduction.

A positive local coefficient supports the predicted-class probability within the fitted *surrogate*; a negative coefficient opposes it. Coefficients do not quantify causal effects or real-world outcomes, and the independent perturbations can violate the algebraic relationships among event features. The explanations must be understood as local approximations of this classifier, not as counterfactual recommendations or evidence of individual fault.

4. Results

4.1. Operational profiles and sample distribution

The scores and four rule-derived tiers were computed at trip level. Table 5 describes class counts under the descriptive full-sample thresholds and under the distinct training-derived thresholds used on the 15 held-out trips. As operational thresholds vary with their estimation sample, these two distributions have different roles.

Table 6
Silhouette scores for candidate one-dimensional K-means partitions of trip-level scores.

k	Driving behaviour	Eco-driving proxy
2	**0.639**	**0.645**
3	0.611	0.570
4	0.552	0.603
5	0.634	0.620
6	0.621	0.598

Table 7
Spearman associations with trip-level event-rate scores; driver-bootstrap 95% percentile intervals in brackets ($n = 52$ trips, 12 drivers).

Trip descriptor	Driving behaviour	Eco-driving proxy
Distance (km)	−0.718 [−0.852, −0.526]	−0.721 [−0.844, −0.555]
Mean speed (m/s)	−0.477 [−0.625, −0.263]	−0.542 [−0.678, −0.328]
Offline records (%)	0.009 [−0.181, 0.242]	0.060 [−0.160, 0.272]
Tight-curve event share	−0.338 [−0.568, −0.177]	−0.353 [−0.556, −0.163]

Table 8
Reproducible trip-level hold-out performance (37 training trips from nine drivers; 15 test trips from three other drivers).

Profile task	Model	Accuracy	Macro F1	Weighted F1
Driving behaviour	DT	0.733	0.712	0.752
	RF	**0.867**	0.867	0.858
	KNN	0.533	0.463	0.543
Eco-driving proxy	DT	0.667	0.646	0.681
	RF	**0.933**	0.937	0.932
	KNN	0.400	0.313	0.383

4.2. Exploratory cluster selection and stability

Table 6 reports silhouette coefficients for candidate k. The largest coefficient was obtained with $k = 2$ for both scores (0.639 and 0.645); the historical $k = 3$ exploratory solution yielded 0.611 and 0.570. For $k = 3$, median driver-bootstrap ARI was 1.000 (interquartile range 0.897–1.000) for driving behaviour and 0.753 (0.608–0.857) for eco-driving. The latter was less stable under changes in sample composition. The three clusters do not map one-to-one onto four operational categories; all supervised targets derive from the explicit threshold rule, not cluster membership.

4.3. Contextual associations

The exploratory correlations in Table 7 demonstrate remaining associations after distance normalisation. Distance was negatively associated with both scores (Spearman $\rho = -0.718$ and -0.721), and average speed showed moderately negative associations. Offline-record percentages were not clearly associated with either score. The tight-curve event share reflects the composition of platform-tagged events and cannot be interpreted as measured road-curvature exposure. Neither correlation magnitudes nor their bootstrap intervals identify independent causal effects of travel distance, speed or geometry.

4.4. Driver-grouped hold-out performance

Table 8 presents the 15-trip test results from three entirely unseen drivers. RF achieved the highest accuracy in both tasks (0.867 for driving behaviour and 0.933 for eco-driving). DT was lower (0.733 and 0.667), and KNN was markedly weaker (0.533 and 0.400). These metrics are derived from the independent re-analysis, *not* the untraceable historical student hold-out. The 15-trip test sample is too small to support precise population-level estimates.

Table 9
Class-specific results for driving-behaviour profiles on the 15-trip hold-out.

Model	Tier	Precision	Recall	F1	Support
DT	Excellent	1.000	0.800	0.889	5
	Good	0.500	0.500	0.500	2
	Moderate	1.000	0.750	0.857	4
	Poor	0.500	0.750	0.600	4
RF	Excellent	1.000	1.000	1.000	5
	Good	1.000	1.000	1.000	2
	Moderate	1.000	0.500	0.667	4
	Poor	0.667	1.000	0.800	4
KNN	Excellent	1.000	0.600	0.750	5
	Good	0.000	0.000	0.000	2
	Moderate	0.500	0.750	0.600	4
	Poor	0.500	0.500	0.500	4

Table 10
Class-specific results for eco-driving proxy profiles on the 15-trip hold-out.

Model	Tier	Precision	Recall	F1	Support
DT	Excellent	1.000	0.800	0.889	5
	Good	0.500	0.500	0.500	2
	Moderate	0.750	0.750	0.750	4
	Poor	0.400	0.500	0.444	4
RF	Excellent	1.000	1.000	1.000	5
	Good	1.000	1.000	1.000	2
	Moderate	0.800	1.000	0.889	4
	Poor	1.000	0.750	0.857	4
KNN	Excellent	1.000	0.600	0.750	5
	Good	0.000	0.000	0.000	2
	Moderate	0.375	0.750	0.500	4
	Poor	0.000	0.000	0.000	4

4.5. Class-specific precision, recall and F1

Tables 9 and 10 describe every tier of the same 15 held-out trips used in Table 8; the corresponding confusion matrices are presented in Figs. 3 and 4. For driving behaviour, RF correctly classified 13/15 trips but recognised only two of four Moderate trips (recall 0.500), assigning the remaining two to Poor. For eco-driving, RF classified 14/15 trips correctly, with one Poor trip assigned to Moderate. KNN failed to correctly identify any of the two Good trips in either task and any of the four Poor trips in the eco-driving task. These differences explain why accuracy should not be interpreted without class-specific assessment, especially with only two Good observations in the test partition.

4.6. Strict feature ablation and target–feature dependence

Removing target-generating and algebraically related features reduced accuracy in five of the six model–task comparisons (Table 11). The exception was KNN on the eco-driving proxy, for which accuracy increased from 0.400 to 0.467; this small-sample fluctuation should not be interpreted as evidence that ablation improves predictive validity. The RF accuracy decreased from 0.867 to 0.467 for driving behaviour and from 0.933 to 0.533 for eco-driving, i.e., by 40 percentage points in both tasks. The corresponding macro F1 decreased from 0.867 to 0.462 and from 0.937 to 0.527. These results identify dependence on scoring-related features as a major contributor to the full-model performance. Because the ablated models use different predictors and do not validate independent outcomes, their residual accuracy should not be interpreted as evidence of true safety or environmental predictability.

4.7. Leave-one-driver-out validation

In the 12-fold LODO-CV (52 pooled held-out trip predictions per task), RF achieved accuracy of 0.942 (macro F1 0.943) for driving behaviour and 0.962 (macro F1 0.962) for eco-driving (Table 12). The corresponding RF accuracies after strict feature ablation were 0.385 and 0.365. LODO-CV mitigates dependence on one arbitrary test-driver selection; it does not remove target–feature dependence or expand the sample beyond these 12 drivers. Although DT and KNN were less accurate, the RF's near-rule-reproduction performance was consistent across the observed cohort and should not be interpreted as proof of external transferability.

Table 11
Trip-level driver-held-out ablation results; full and ablated classifiers use identical partitions and model settings.

Task	Model	Predictors	Accuracy	Macro F1	Weighted F1
Driving	DT	Full	0.733	0.712	0.752
		Ablated	0.267	0.296	0.350
	RF	Full	0.867	0.867	0.858
		Ablated	0.467	0.462	0.489
	KNN	Full	0.533	0.463	0.543
		Ablated	0.400	0.338	0.405
Eco proxy	DT	Full	0.667	0.646	0.681
		Ablated	0.333	0.368	0.392
	RF	Full	0.933	0.937	0.932
		Ablated	0.533	0.527	0.562
	KNN	Full	0.400	0.313	0.383
		Ablated	0.467	0.396	0.467

Table 12
LODO-CV performance from 52 pooled out-of-fold trip predictions per task (12 held-out-driver iterations).

Task	Model	Accuracy	Macro F1	Weighted F1
Driving behaviour	DT	0.673	0.681	0.680
	RF	**0.942**	0.943	0.944
	KNN	0.538	0.523	0.524
Eco-driving proxy	DT	0.731	0.730	0.732
	RF	**0.962**	0.962	0.962
	KNN	0.538	0.520	0.520

Note: Results refer to the full predictor set. Under strict feature ablation, RF accuracy decreased to 0.385 for driving behaviour and 0.365 for the eco-driving proxy.

4.8. Decision rules and local surrogate explanations

The retrained DT provides interpretable split conditions, but those splits must not be confused with T_1–T_3, which construct the targets. Fig. 5 presents the DT fitted to the hold-out training data for driving behaviour; it replaces the historical tree rather than reproducing its original fit.

Fig. 6 displays the local ridge-surrogate explanations for held-out trip 341, predicted Excellent in both tasks. RF predicted-class probabilities were 0.414 for driving behaviour and 0.466 for eco-driving. Weighted local surrogate R^2 was 0.721 and 0.702, respectively, indicating imperfect local fidelity. These are local *model* contributions, not causal effects of adverse driving on measured risk or energy use. Because inputs include mathematically related event counts and composite rates, some independently perturbed combinations are not physically feasible.

5. Discussion

5.1. What the evaluation establishes

The results demonstrate that conventional classifiers, especially RF, can reconstruct an explicitly defined operational taxonomy on previously unseen drivers within this small dataset. The driver-grouped hold-out RF accuracies (0.867 and 0.933) and LODO-CV accuracies (0.942 and 0.962) are reproducible from the retained trip-level extract and fixed model settings. These figures do *not* estimate the accuracy of predicting independently observed driving safety, crash risk, fuel

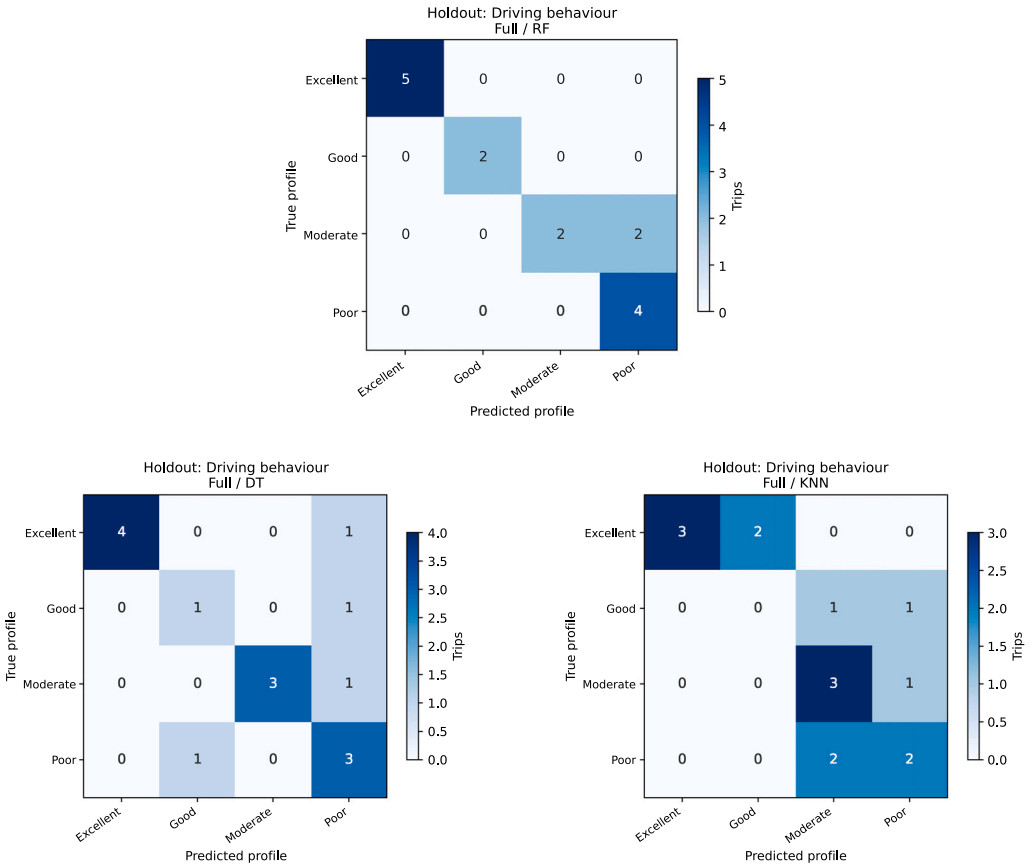


Fig. 3. Driving-behaviour confusion matrices for the same 15 held-out trips: RF (top), DT (bottom left) and KNN (bottom right). Rows denote true tiers and columns predicted tiers.

consumption or emissions. The almost direct relationship between predictor event counts and rule-derived targets is a central limitation of the interpretation.

The strict ablation gives direct empirical support for this qualification: RF hold-out accuracy decreased to 0.467 and 0.533 when both operational scores and their algebraically related components were removed. Residual prediction may reflect indirect associations rather than independent behavioural validity. The varying class-wise errors also matter: in the hold-out, RF missed two Moderate driving-behaviour trips, and KNN failed to recognise the Good category in both tasks. With only two Good trips in the test set, per-class metrics remain highly unstable.

LODO-CV extends the analysis beyond one selected three-driver test partition but remains confined to 12 drivers and unequal numbers of trips. The differences between DT, RF and KNN, and between full and ablated predictors, also emphasise that excellent taxonomy reproduction with one model does not establish transportability of the profiling rule itself.

5.2. Clustering, context and construct validity

Clustering supports a descriptive summary of variation in adverse-event rates but does not establish four naturally occurring driving types. The silhouette criterion favoured two clusters, while the three-group solution was retained only to compare with exploratory material from the earlier study. Bootstrap stability was weaker for the eco-driving proxy. The operational tiers remain the product of explicit mean-based thresholds, not validated latent driver identities.

Distance-normalisation reduces the most obvious exposure difference but does not remove it. In particular, the strong association of both event rates with trip distance may reflect denominator effects, journey composition and operating conditions. Average speed is confounded by congestion, route type and vehicle capabilities; curve-tagged event fractions do not measure actual road curvature exposure. Without trip-linked vehicle type, payload, traffic, weather, gradient or independently measured road geometry, the analysis cannot isolate driver effects from context. The appropriate interpretation is a *trip-level context-sensitive flag for human examination*, not a permanent behavioural assessment of a worker.

5.3. Explainability and human-centred governance

Decision-tree paths and local RF surrogates can expose which telemetry features and numerical intervals contribute to model outputs, but no explanation is a causal estimate or a validated counterfactual. Here the independently sampled LIME-style perturbations may violate natural feature constraints, and local fidelity was moderate rather than perfect. Feature contributions must therefore be used diagnostically and should never be treated as evidence that a given manoeuvre caused a real crash risk or emissions outcome.

A responsible deployment would require an explicit governance protocol rather than relying on explainability plots alone. Before data collection and profiling, the operator should communicate the monitoring purpose, categories of data, retention period, access rights and foreseeable uses in plain language; the lawful basis and any requirement for consent must be determined under applicable privacy and employment rules, rather than assuming that employee consent is inherently valid. Drivers should be able to inspect the trip observations and alerts underlying their scores and request correction of inaccurate records, with a documented review and rectification channel. Supervisors should examine route, vehicle and payload context, document

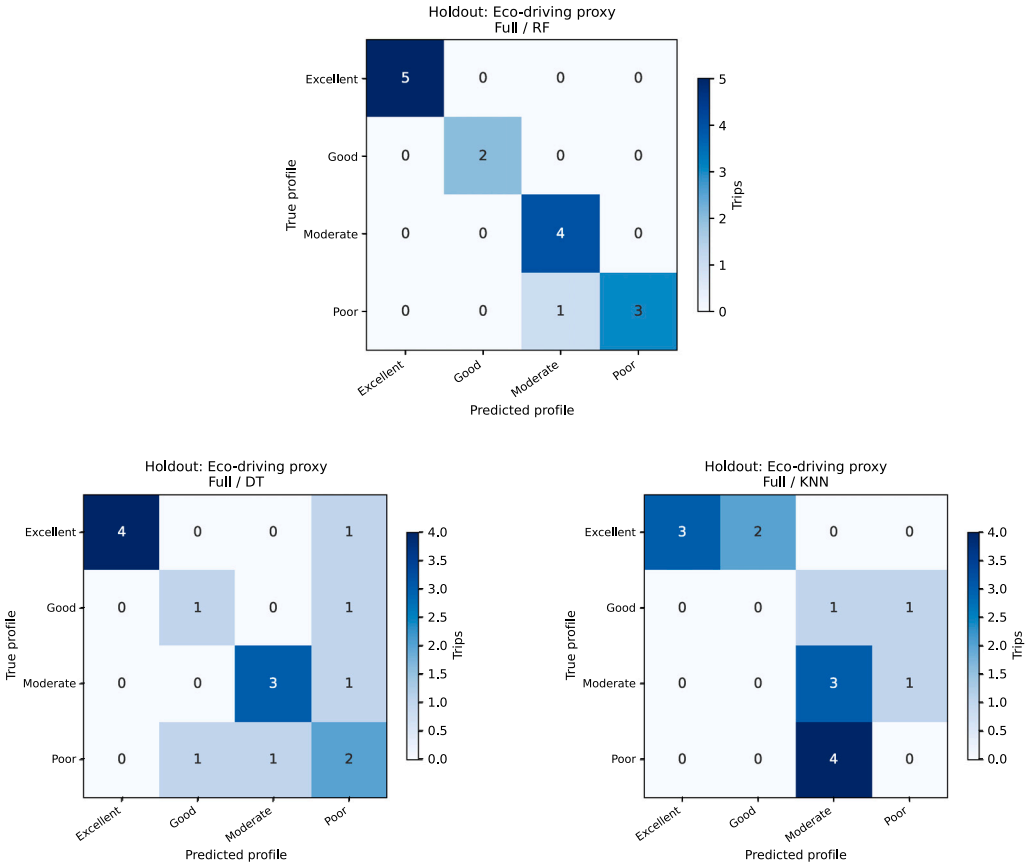


Fig. 4. Eco-driving proxy confusion matrices for the same 15 held-out trips: RF (top), DT (bottom left) and KNN (bottom right). Rows denote true tiers and columns predicted tiers.

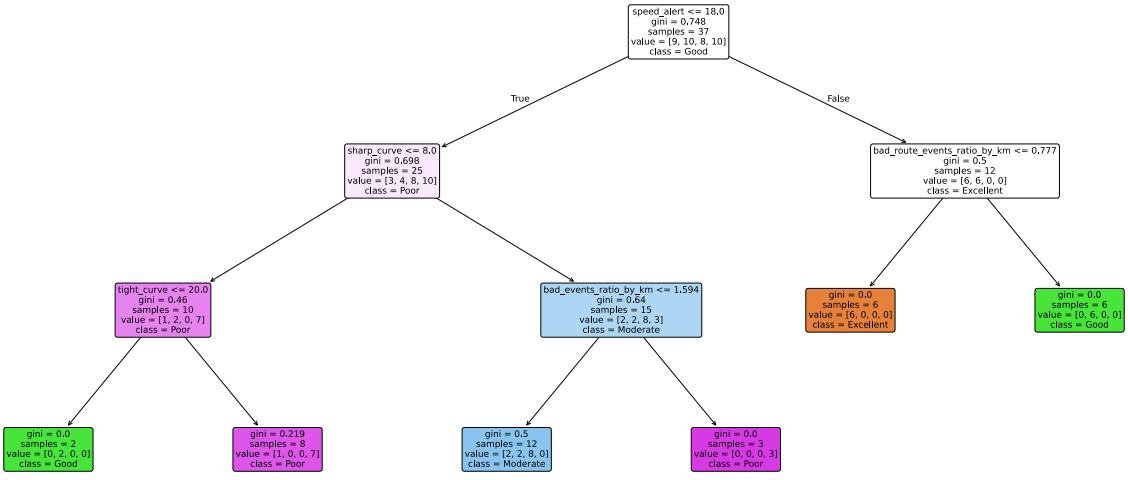


Fig. 5. Decision tree reconstructed from the training trips for the driving-behaviour task. Displayed decision thresholds are classifier splits, not operational-profile cutoffs.

reasons for any intervention and provide a meaningful opportunity to contest the interpretation. From an organisational perspective, context-sensitive performance indicators are also important for avoiding distorted incentives and potentially inequitable assessments of drivers operating under different conditions. The system should be restricted to proportionate coaching or safety review unless a separate, properly governed evidence base supports consequential uses. No such operational governance procedure was evaluated empirically in this study; these are safeguards recommended for deployment.

5.4. Limits and external validation priorities

The first priority for outcome validation should be independently verified safety measures (e.g., incident and near-miss reports, ideally exposure-adjusted and independently adjudicated) together with objectively logged energy or fuel consumption per kilometre, accounting for vehicle and payload. Second priorities include maintenance and wear records and blinded expert assessments of trip behaviour. These outcomes address distinct constructs and should not be collapsed into

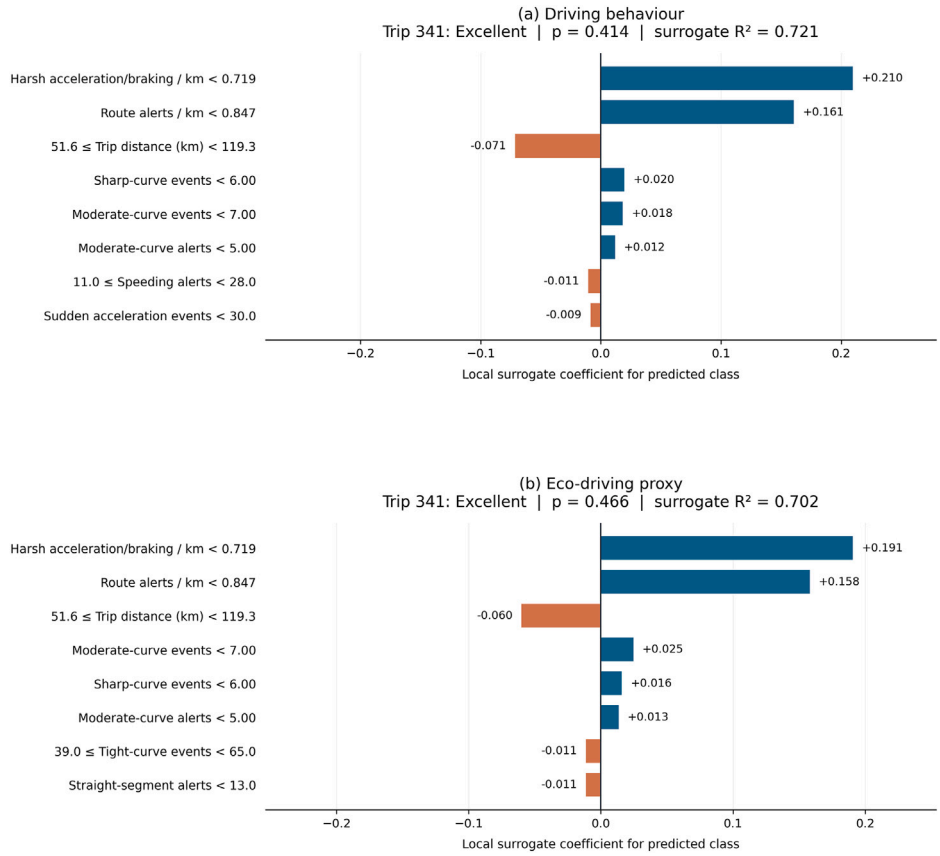


Fig. 6. Two independent LIME-style local surrogate explanations of the RF-predicted Excellent class for held-out trip 341: (a) driving behaviour; (b) eco-driving proxy. Positive and negative weights support or oppose the local surrogate estimate of the predicted probability. These weights are not causal effects.

one unvalidated score. Validation should be repeated across operators, vehicles, weather conditions and geographic settings, and assessed with larger numbers of drivers and trips.

The present data have several constraints: one commercial platform, three months of source collection, only 52 retained trips/12 drivers, no vehicle identifiers, and no trip-linked traffic or payload. The approximately 100 m location spacing and absent original GPS processing code prevent verification of instantaneous manoeuvre thresholds and window handling. The independent re-analysis reproduces aggregate-trip modelling, not the original student's executable telemetry pipeline or historical evaluation partition. Hyperparameters were fixed a priori rather than tuned in a nested search; performance estimates may be sensitive to model choices and split composition. Independent perturbations in local explanation may produce physically implausible feature combinations. These limitations preclude claims of definitive driver-level or causal validity.

6. Conclusions

The manuscript presents a transparent re-analysis of smartphone-telematics-derived trip profiles from 52 journeys and 12 drivers, together with exploratory clustering, driver-grouped validation, strict feature ablation, class-wise performance and local model explanations. RF reproduced operational targets with hold-out accuracies of 0.867 (driving behaviour) and 0.933 (eco-driving proxy), but strict removal of score-related predictors reduced those values to 0.467 and 0.533. Thus, the principal finding is reproducible *operational taxonomy assignment*, not independent prediction of safe, efficient or environmentally sustainable driving. Context-adjusted modelling and validation against independently measured safety and energy outcomes are essential before stronger inferences or consequential applications. More broadly, the study highlights the importance of transparent label construction, rigorous validation and human oversight in the development of fleet-management decision-support tools, where operational efficiency and fair performance assessment must be considered alongside technological capabilities.

CRediT authorship contribution statement

Pedro Entringer: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Fátima Leal:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Micaela Pinho:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is funded by national funds through FCT - Fundação para a Ciência e Tecnologia, I.P., under the Programme Contract UID/05105/2025 https://doi.org/10.54499/UID/05105/2025.

Data availability

A reproducibility package containing the independently developed trip-level fitting scripts, feature definitions, model settings and aggregate evaluation outputs was prepared. Access to the original platform records and the trip-level extract is subject to approval from the data controller; no public release is implied. The original point-level event-detection code and complete raw traces were not available for this re-analysis.

References

Al Hla, Y.A., Othman, M., Saleh, Y., 2019. Optimising an eco-friendly vehicle routing problem model using regular and occasional drivers integrated with driver behaviour control. J. Clean. Prod. 234, 984–1001.

Bouhsissin, S., Sael, N., Benabbou, F., 2023. Driver behavior classification: a systematic literature review. IEEE Access 11, 14128–14153.

Brahim, S.B., Ghazzai, H., Besbes, H., Massoud, Y., 2022. A machine learning smartphone-based sensing for driver behavior classification. In: 2022 IEEE International Symposium on Circuits and Systems. ISCAS, IEEE, pp. 610–614.

Castignani, G., Frank, R., Engel, T., 2013. Driver behavior profiling using smartphones. In: 16th International IEEE Conference on Intelligent Transportation Systems. ITSC 2013, IEEE, pp. 552–557.

Ferreira, J., Carvalho, E., Ferreira, B.V., de Souza, C., Suhara, Y., Pentland, A., Pessin, G., 2017. Driver behavior profiling: An investigation with different smartphone sensors and machine learning. Plos One 12, e0174959.

Gouribhatla, R., Pulugurtha, S.S., 2022. Drivers' behavior when driving vehicles with or without advanced driver assistance systems: A driver simulator-based study. Transp. Res. Interdiscip. Perspect. 13, 100545. http://dx.doi.org/10.1016/j.trip.2022.100545.

Hwang, C.p., Chen, M.S., Shih, C.M., Chen, H.Y., Liu, W.K., 2018. Apply scikit-learn in python to analyze driver behavior based on obd data. In: 2018 32nd International Conference on Advanced Information Networking and Applications Workshops. WAINA, IEEE, pp. 636–639.

Mascheroni, M., Moretti, E., Tappia, E., Melacini, M., 2026. Machine learning in road freight transportation: A decision-making perspective. Transp. Res. Interdiscip. Perspect. 37, 101936. http://dx.doi.org/10.1016/j.trip.2026.101936.

Meteier, Q., Favre, R., Viola, S., Capallera, M., Angelini, L., Mugellini, E., Sonderegger, A., 2024. Classification of driver fatigue in conditionally automated driving using physiological signals and machine learning. Transp. Res. Interdiscip. Perspect. 26, 101148. http://dx.doi.org/10.1016/j.trip.2024.101148.

Miraftabzadeh, S.M., Colombo, C.G., Longo, M., Foiadelli, F., 2023. K-means and alternative clustering methods in modern power systems. IEEE Access.

van Nes, N., Bärgman, J., Christoph, M., van Schagen, I., 2019. The potential of naturalistic driving for in-depth understanding of driver behavior: Udrive results and beyond. Saf. Sci. 119, 11–20.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144.

da Silva, D.A., Torres, J.A.S., Pinheiro, A., de Caldas Filho, F.L., Mendonça, F.L., Praciano, B.J., de Oliveira Kfouri, G., de Sousa, R.T., 2019. Inference of driver behavior using correlated iot data from the vehicle telemetry and the driver mobile phone. In: 2019 Federated Conference on Computer Science and Information Systems. FedCSIS, IEEE, pp. 487–491.

Timmerman, M.E., Ceulemans, E., De Roover, K., Van Leeuwen, K., 2013. Subspace k-means clustering. Behav. Res. Methods 45, 1011–1023.

Wu, Y., Chen, S., Ma, Y., Cheng, W., Zhang, F., Xing, G., 2024. Identification of aggressive driving behavior of online car-hailing drivers based on association classification. Hum. Factors Ergon. Manuf. Serv. Ind. 34, 118–131.