10th International Conference on

RSS2026

Road Safety & Simulation

23-26 June 2026, Naples, Italy

Machine Learning-Based Analysis of Driving Behavior and Road Safety Using Connected Vehicle Data

Andreas Georgios Englezos[a*], Armira Kontaxi[a*], Eleonora Papadimitriou[a], George Yannis[a]

[a]National Technical University of Athens (NTUA), 5 Heroon Polytechniou Str., GR-15773 Athens, Greece

*Corresponding author: andreas_englezos@mail.ntua.gr

INTRODUCTION AND OBJECTIVES

Road safety remains a critical global challenge, strongly influenced by driver behavior and traffic conditions. Connected vehicle technologies offer new opportunities to monitor and improve driving behavior through real-time data collection and advanced analytics.

A Connected Vehicle (Uhlemann, 2015) is a vehicle equipped with technology that enables data exchange with other vehicles and with transport infrastructure (C-ITS).

Research on connected vehicle data generally falls into three directions (Zhou & Bridgelall, 2020): driving behavior analysis, Surrogate Safety Measures (SSMs), and traffic signal/intersection management.

Machine learning methods (K-Means, DBSCAN, SVM, Decision Trees) are widely applied to connected vehicle data for clustering, classification, and pattern recognition.

Objectives of this study:

1. Investigate driving behavior through connected vehicle data (speed, engine temperature, ABS, and other driving characteristics).
2. Identify distinct driving profiles using K-Means clustering.
3. Assess which variables influence ABS activation using Random Forest and Binary Logistic Regression.

LITERATURE REVIEW

Connected Vehicles & C-ITS

First defined by Uhlemann (2015) as vehicles equipped with technology enabling data exchange with other vehicles and infrastructure, within the Cooperative Intelligent Transport Systems (C-ITS) framework.

Machine Learning in Driving Behavior Research

ML algorithms such as K-Means, DBSCAN, and SVM have been widely applied to connected vehicle data for driving behavior clustering and classification (Kontaxi et al., 2025; Petraki et al., 2025; Theofilatos et al., 2018).

Several studies have used such methods to classify driving profiles and identify high-risk road segments (Mohammadnazar et al., 2021; Theodoraki et al., 2025; Koliou et al., 2025; Ziakopoulos et al., 2025).

METHODOLOGY AND DATA COLLECTION

This study combines clustering-based driving profile analysis with predictive modelling (Random Forest + Binary Logistic Regression) of ABS activation, using a naturalistic connected vehicle dataset of 420,855 observations across 262 trips.

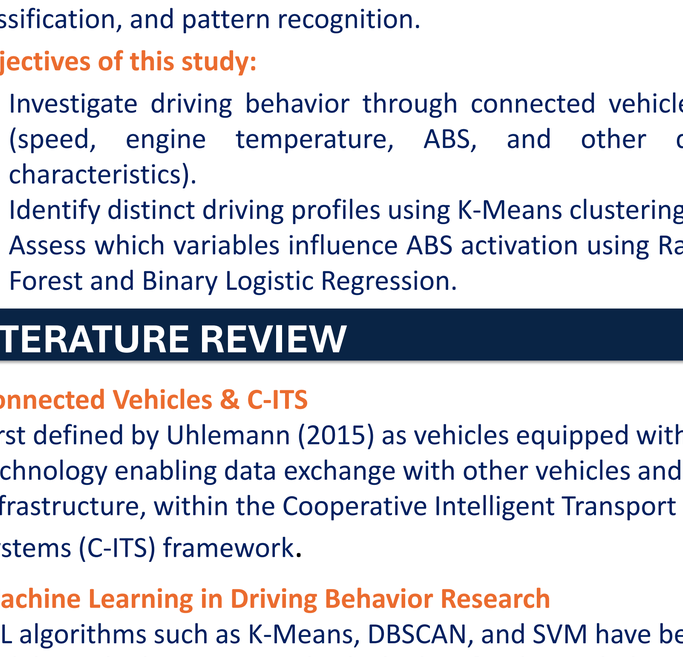

Fig. 1. The OSeven data flow system

Data Source

Data were generated by sensors in two vehicles rented for three months and driven by three employees, within the OSeven Telematics R&D project "07Connected." A total of 262 trips were recorded (152 from Vehicle 1, 110 from Vehicle 2), comprising 420,855 observations. Variables include geolocation, vehicle speed, odometer, engine status, fuel data, collision indicators, ABS, Lane Keep Assist, parking assist, cruise control, and engine oil temperature.

RESULTS

K-Means Clustering

Trips were classified per vehicle using mean/max speed, speed variability, ABS activations, and harsh braking events (z-score standardized). Elbow and Silhouette methods identified K = 3 as the optimal number of clusters; 25 random initializations (nstart = 25) ensured stable results.

Random Forest

ABS activation (active/inactive) was used as the dependent variable. The dataset (420,855 observations) was split 80/20 at the observation level. The model used 300 trees, permutation-based variable importance, and inverse class-frequency weighting to address class imbalance. Accuracy: 98%, sensitivity: 98.5%, specificity: 65.4%.

Binary Logistic Regression

Models the probability of ABS activation as a function of the independent variables, using a 95% significance level and Odds Ratios. VIF values (1–2) indicate no multicollinearity. Pseudo-R^2: McFadden = 0.154, Cox & Snell = 0.163.

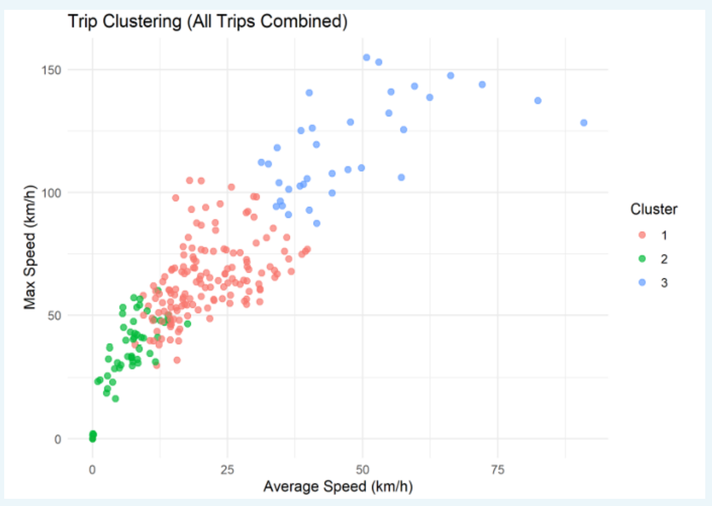

Fig. 2. Trip classifications based on average and maximum speed

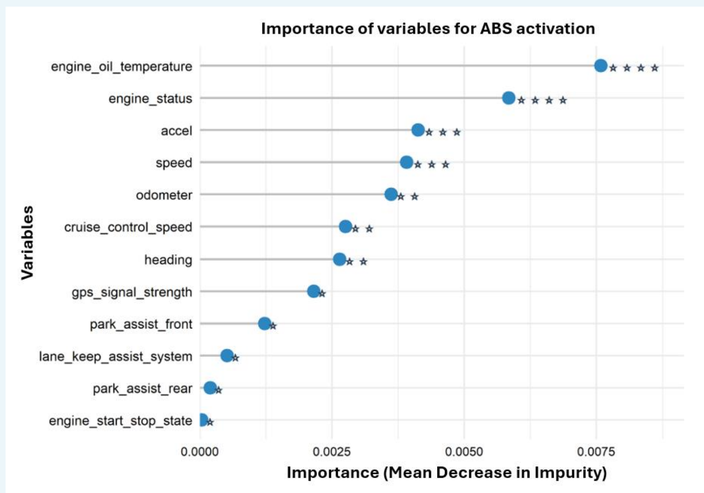

Fig. 3. Variable importance diagram – Random Forest

DISCUSSION

Speed & Driving Patterns

Vehicle 1: 20–65 km/h (occasional >100 km/h, SD≈18 km/h). Vehicle 2: 20–100 km/h (peaks 120–150 km/h, SD≈15 km/h). Speeds drop during peak hours (07:00–09:00, 15:00–19:00) and rise off-peak (10:00–13:00, 21:00–24:00). Harsh braking more frequent for Vehicle 2 (peak ~06:00, also elevated at 11:00 & 16:00).

Cluster Analysis

- Cluster 1: Typical driving - intermediate speed, moderate variability, highest trip frequency.
- Cluster 2: Cautious driving - low speed, minimal fluctuation, highest ABS activation.
- Cluster 3: Aggressive driving - high average and maximum speeds.

Random Forest & Variable Importance

Top predictors: engine oil temperature, engine status, acceleration, speed. Moderate: odometer, cruise control speed, heading. Minimal: parking assist, lane-keeping system.

Model Performance

Accuracy 98%, Sensitivity 98.5%, Specificity 65.4%.

Table 1. Overall results of the Binary Logistic Regression model

Independent Variables	estimate	p.value	Odds Ratio	Importance	VIF
(Intercept)	2,526	<0,001	12,499	<0,001	-
engine_oil_temperature	0,001	0,204	1,001	-	1,960
engine_status (on)	1,300	<0,001	3,671	<0,001	2,033
accel	-0,049	0,199	0,952	-	1,005
speed	0,056	<0,001	1,058	<0,001	1,183
odometer	0,000	0,078	1,000	-	1,006
cruise_control_speed	0,002	0,007	1,002	0,007	1,008
heading_sin	-0,076	<0,001	0,927	<0,001	1,003
heading_cos	-0,101	<0,001	0,904	<0,001	1,009
Pseudo-R^2	McFadden=0,154		Cox & Snell = 0,163		

CONCLUSIONS

- Connected vehicle data combined with machine learning effectively supports understanding of driving behavior and road safety. K-Means clustering identified three distinct profiles: cautious, typical, and aggressive driving. Random Forest and Binary Logistic Regression showed that engine-related variables and speed are the most critical predictors of ABS activation.
- Recommendations: improve road infrastructure (safety + data quality); raise road-safety awareness early through school programs; use aggressive-driving findings to design driver interventions; strengthen traffic enforcement; conduct driver-behavior research more regularly to inform policy.
- Future work: expand the dataset to more vehicles/drivers and diverse road conditions; test alternative variables and ML algorithms; incorporate geographic/slope data to clarify the road-morphology effects observed in this study.