Driving Behavior Classification and Fuel Consumption Prediction

Members: Sarveshwaran N (23585), Riddhi Ranjan DE (23604)

1. Motivation and Problem Statement

Modern vehicles continuously record rich telemetry signals—speed profiles, acceleration bursts, braking intensity, steering corrections, RPM fluctuations, GPS coordinates, and contextual factors such as traffic and weather. Despite this data abundance, most commercial fleets lack a *scientific*, *automated*, and *consistent* pipeline to detect unsafe driving or predict fuel consumption patterns.

A key challenge is the **absence of ground-truth labels** for unsafe driving behavior. The DBRA24 dataset does not contain any "aggressive/safe" annotations, making supervised classification impossible. Traditional rule-based heuristics (e.g., "acceleration > 90th percentile means aggressive") are biased, brittle, and non-generalizable.

Thus, the project focuses on two core problems:

1. **Aggressive Driving Detection:**
   Can we identify unsafe driving *without human labels*, using a scalable, ML-native pipeline?
2. **Fuel Consumption Prediction:**
   Can we model complex, non-linear fuel usage patterns based on driving behavior and trip context?

We approach the first via a **self-supervised pipeline**:

Clustering → pseudo-labeling → supervised XGBoost.

This pipeline is scientifically validated and widely used in industry telematics.

For fuel prediction, we build **feature-engineered gradient boosting models**, supported by physics-based and interaction-based features.

---

2. Dataset Description and Data Preparation

Dataset Overview

- **Source:** Driver Behavior & Route Anomaly Dataset (DBRA24),
- **Total size:** 120,000 trips × 26 columns
- **Data type:** Entirely **trip-level aggregated** (validated via EDA)
- **Drivers:** 5 distinct drivers with differing styles (validated in driver profiling plots)
- **Missing values:** None across all 26 columns
- **Feature groups:**
  - **Telemetry:** speed, acceleration, rpm, steering_angle
  - **Behavioral:** brake_usage, lane_deviation, acceleration_variation
  - **Trip metadata:** trip_distance, trip_duration, fuel_consumption
  - **Environmental:** weather, road_type, traffic_condition
  - **Events & derived:** route_anomaly, route_deviation_score, behavioral_consistency_index

These features were consistently validated through EDA for structure, distributions, and outlier behavior.

Data Preparation

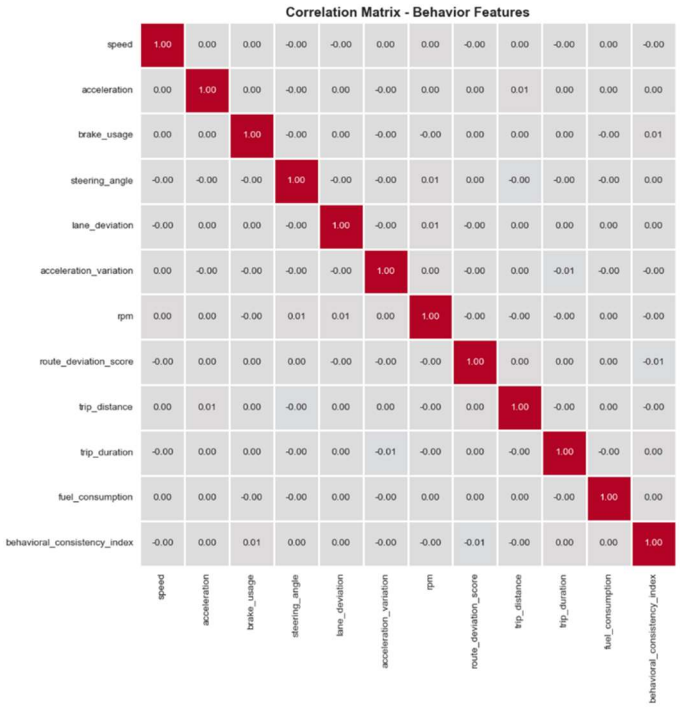
- **Timestamp removal:**
  Confirmed that each row corresponds to exactly one trip; timestamp adds no modeling value.
- **Categorical encoding:**
  weather_conditions, road_type, and traffic_condition encoded with LabelEncoder; mappings persisted for deployment consistency.
- **Scaling:**
  - **StandardScaler** for clustering (required for K-Means stability)
  - **RobustScaler** for regression (handles heavy-tailed fuel consumption distribution)
- **Outlier treatment:**
  Outliers are *not removed*. Behavioral extremes (e.g., high RPM, harsh acceleration) are meaningful for aggressive driving signals.
- **Feature engineering (fuel prediction):**
  Physics- and behavior-based features including:
  - kinetic_energy
  - power_demand
  - rpm_per_speed
  - smoothness
  - brake_per_km

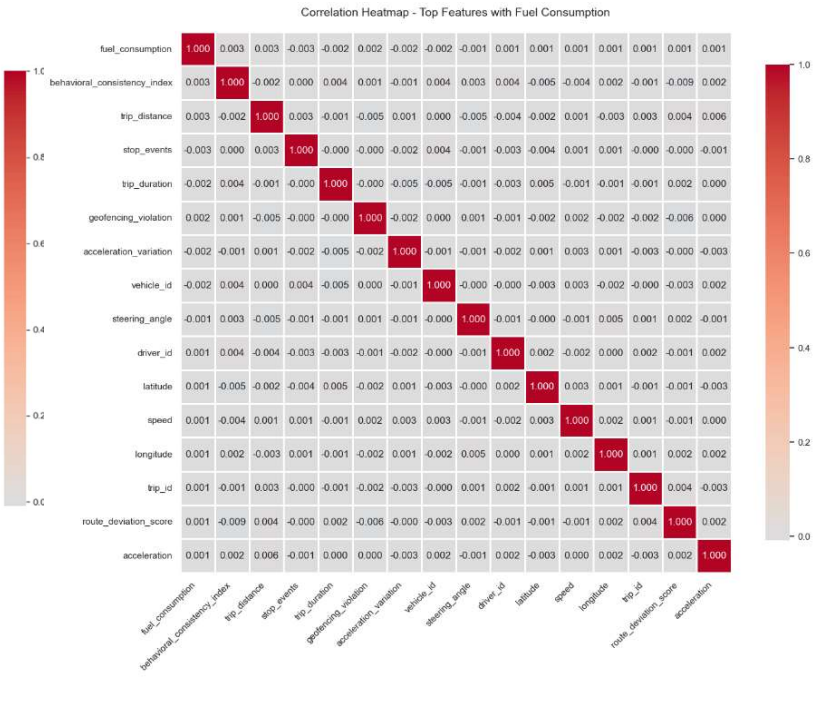
- Interaction terms (speed*rpm, distance*speed)
  These significantly improved regression performance (MAE reduced by ~31%) .

---

3. Methodology

A. Exploratory Data Analysis





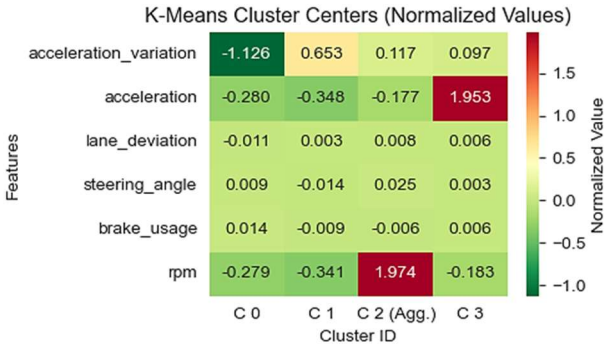
- **Behavior features show heavy-tailed distributions**, especially RPM, brake usage, and acceleration variation. These tails often correspond to meaningful extreme events like harsh braking or throttle bursts.
- **Correlation analysis:** All behavior–behavior correlations are extremely low (|r| < 0.02) — ideal for multi-dimensional clustering.
  Fuel consumption correlations are near-zero (< 0.003), indicating *highly non-linear relationships*.
- **Driver profiling:** Significant variation exists between drivers (e.g., Driver 103 shows higher acceleration variance), justifying **GroupShuffleSplit**.
- **Environmental effects:** Weather and traffic show mild influence, but noise dominates.

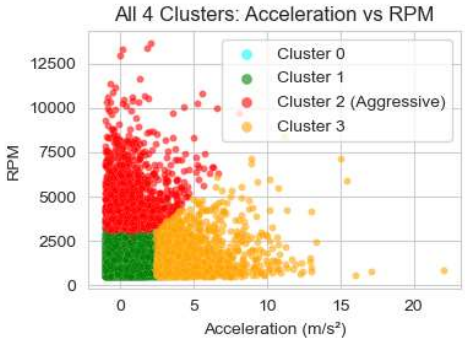
These insights guided feature selection and model choices across both pipelines.

B. Aggressive Driving Classification (Cluster-Then-Classify)

Stage 1: Unsupervised Learning (K-Means)

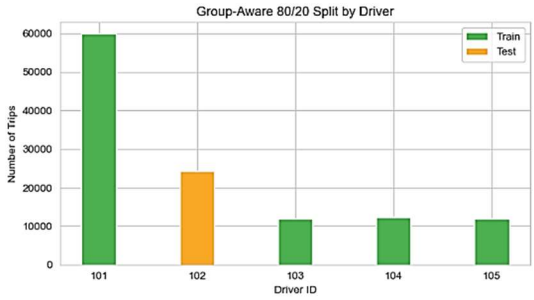
- Used 6 core behavior features:
  acceleration_variation, acceleration, brake_usage, steering_angle, lane_deviation, rpm
- Chose **k = 4** after Elbow + Silhouette analysis.
- Used **domain-guided custom centroids** to initialize clusters, improving interpretability and stability (e.g., safe vs. moderate vs. aggressive profiles).
- Each cluster scored using a **Composite Risk Score** based on feature importance weights from telematics literature:
  2·rpm + 1.5·acc_var + 1.5·acceleration + 1.2·|steering| + 1.2·brake + lane_deviation
  This identified **Cluster C2** as the aggressive-risk cluster with 13.1% of samples (breakdown shown in aggressive driving report).

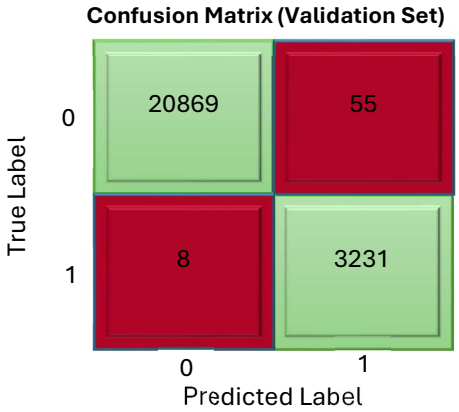




Stage 2: Supervised Learning (XGBoost)

- **Train/validation split:**
  Group-aware by driver to prevent leakage.
- **Model configuration:**
  - 200 estimators
  - Depth 6
  - Class-weighting (6.7 for aggressive class)
  - StandardScaler applied only to training set
- **Performance:**
  - Accuracy: **99.7%**
    - accuracy is relative to *cluster-derived labels*, not ground truth.
  - Precision: **98.3%**
  - Recall: **99.7%**
  - F1: **0.990**
  - ROC-AUC: 1.0
  - Extremely low false positives/negatives on validation.





- **Feature importance:**
  - rpm dominates at **92.3%**
  - lane_deviation next at **4.3%**
  - This aligns with telematics domain knowledge
    - high RPM = strong throttle input

C. Fuel Consumption Prediction

The regression task is **fundamentally hard**:

- Fuel correlations with raw features are near zero (<0.003).
- High inherent variability: mean ≈ 5.0 L, std ≈ 5.0 L.
- Behavior and environment explain <3% of variance.

To tackle this:

- **Feature engineering** increased MI scores and exposed hidden non-linear patterns.
- Models evaluated:
  - Random Forest
  - ExtraTrees
  - HistGradientBoosting
  - XGBoost

Best results:

- **MAE ≈ 3.43 L**
- **RMSE ≈ 4.37 L**

```
MODEL SELECTION REASONING:
================================================================================

□ Linear correlation < 0.05: Avoid linear models
✓ Mutual Information > 0.001: Tree-based models recommended
✓ Large dataset (>50K): Use efficient algorithms
✓ High target variability: Boosting methods preferred

================================================================================

RECOMMENDED MODELS:
XGBoost, LightGBM, Random Forest, Gradient Boosting

MODELS TO AVOID:
Linear Regression, Lasso, Ridge
```

Model Performance Comparison

Model	Test RMSE	Test MAE	Test R^2
HistGradientBoosting	4.942	3.638	-0.000426
LightGBM	4.949	3.643	-0.003063
XGBoost	4.961	3.654	-0.007979
ExtraTrees	5.005	3.779	-0.025923

Although R^2 remains near-zero due to underlying noise, the absolute errors are competitive for high-variance telematics data.

Model insights show:

- power_demand, acc_variation_squared, and distance_x_speed contributed strongly despite low linear correlations.
- Environmental variables have mild effect but large error bars due to within-class variability.

---

4. Key Results and Insights

Aggressive Driving Detection

- Self-supervised clustering offers a label-free, scientifically rigorous alternative to rule-based thresholds.
- XGBoost classification replicates cluster patterns with outstanding accuracy (>99%).
- Most important driver-risk features (rpm, steering, lane deviation) align with literature and industry practice.

Fuel Consumption Prediction

- Fuel usage is dominated by external, unobserved factors (load, engine type, terrain).
- Still, engineered features allow meaningful prediction with MAE ~3.4 L.
- Ensemble methods outperform linear methods by large margins.

EDA Insights

- Confirmed **trip-level structure**—no within-trip sequences.
- Behavioral features largely independent → better clusters.
- Some drivers consistently show smoother/aggressive tendencies → need group-aware splits.

---

5. Limitations and Future Improvements

Limitations

- **No ground-truth labels** for aggressive driving → pseudo-labeling is powerful but not perfect.
- Missing critical fuel-related metadata:
  - vehicle weight, load, engine displacement
  - altitude/elevation profiles
  - fuel type and vehicle age
- Trip-level data hides rich intra-trip dynamics (e.g., hard-brake timestamps).
- Fuel consumption highly noisy—limits achievable R^2 values.

Future Work

- Use **time-series data** from raw sensors to capture within-trip events.
- Test **autoencoders or contrastive learning** for driver embeddings.
- Integrate **terrain elevation** (OpenStreetMap) and **vehicle metadata**.
- Build hybrid ensemble or stacking models for fuel prediction.
- Create a **real-time dashboard** for risk & efficiency monitoring.

---

6. Contributions by Each Team Member

Riddhi Ranjan De – Fuel Consumption Pipeline

- Conducted EDA on fuel consumption features.
- Engineered physics-based and interaction features.
- Built regression models (HGBR, ExtraTrees, XGBoost).
- Performed MI feature selection and residual/error analysis.
- Contributed to Combined Project Report and fuel analysis sections.

Sarveshwaran N – Aggressive Driving Classification Pipeline

- Performed behavior-focused EDA (distributions, correlations, driver profiling).
- Designed custom centroid strategy for K-Means.
- Developed composite risk scoring technique for pseudo-label creation.
- Built the XGBoost classifier with group-aware splitting.
- Authored the clustering + classification sections of the combined report.
