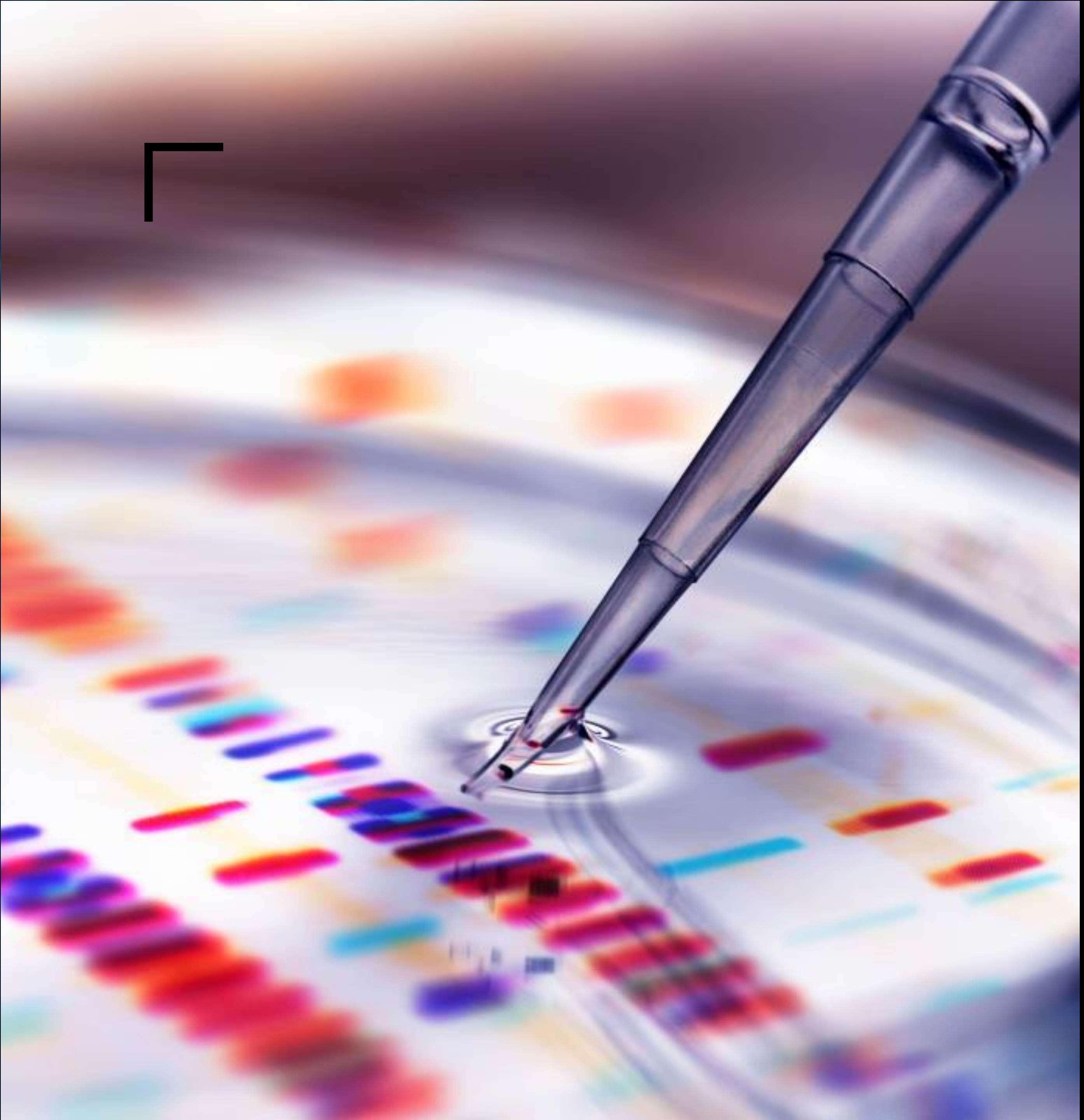
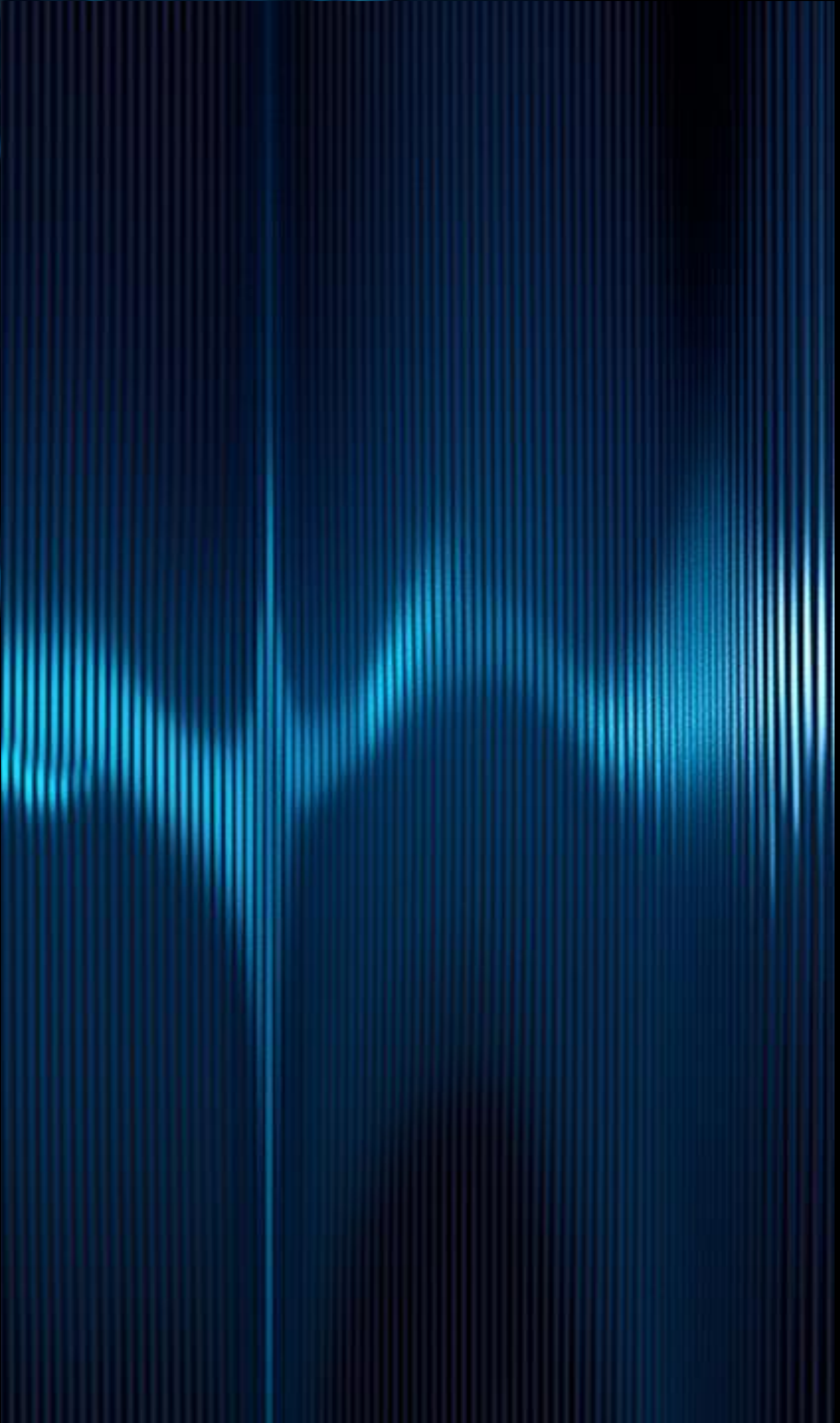
WEEK 13 PROGRESS UPDATE

The end of Sprint 3



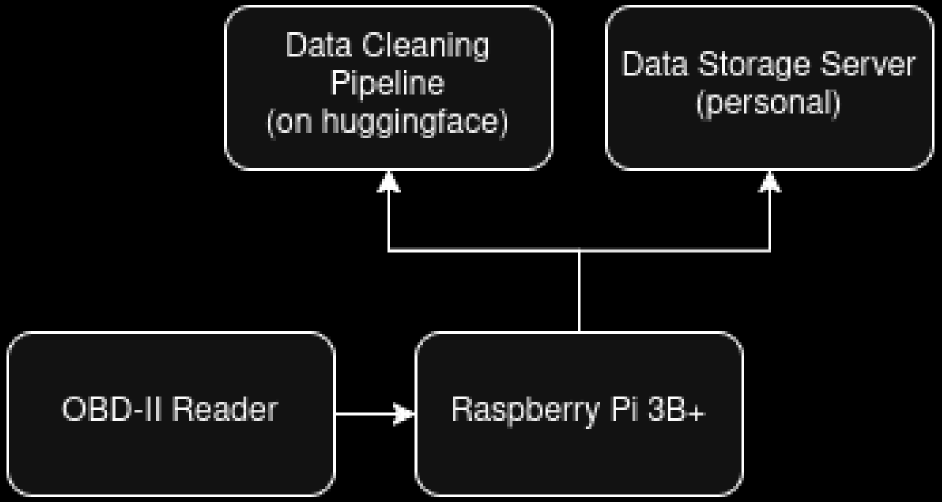
INTRODUCTION

- This week conclude week 13, sprint 3, as well as the first semester – phase of this project.
- We want to demonstrate our progress up-to-date, and discuss on future planning of the project.



AUTOMATED LOGGING PIPELINE

- Collected logs are uploaded to the cleaning API hosted on huggingface
- Raw data is also saved on a storage server for future reference and ease of access
- We will continue to collect logs over the break



Data Cleaning Pipeline (on huggingface)
Data Storage Server (personal)
OBD-II Reader
Raspberry Pi 3B+

UNSUPERVISED LEARNING

How we tackle manual labelling

DATA SET A

'RPM',

'THROTTLE_POS',

'SPEED',

'ENGINE_LOAD',

'MAF',

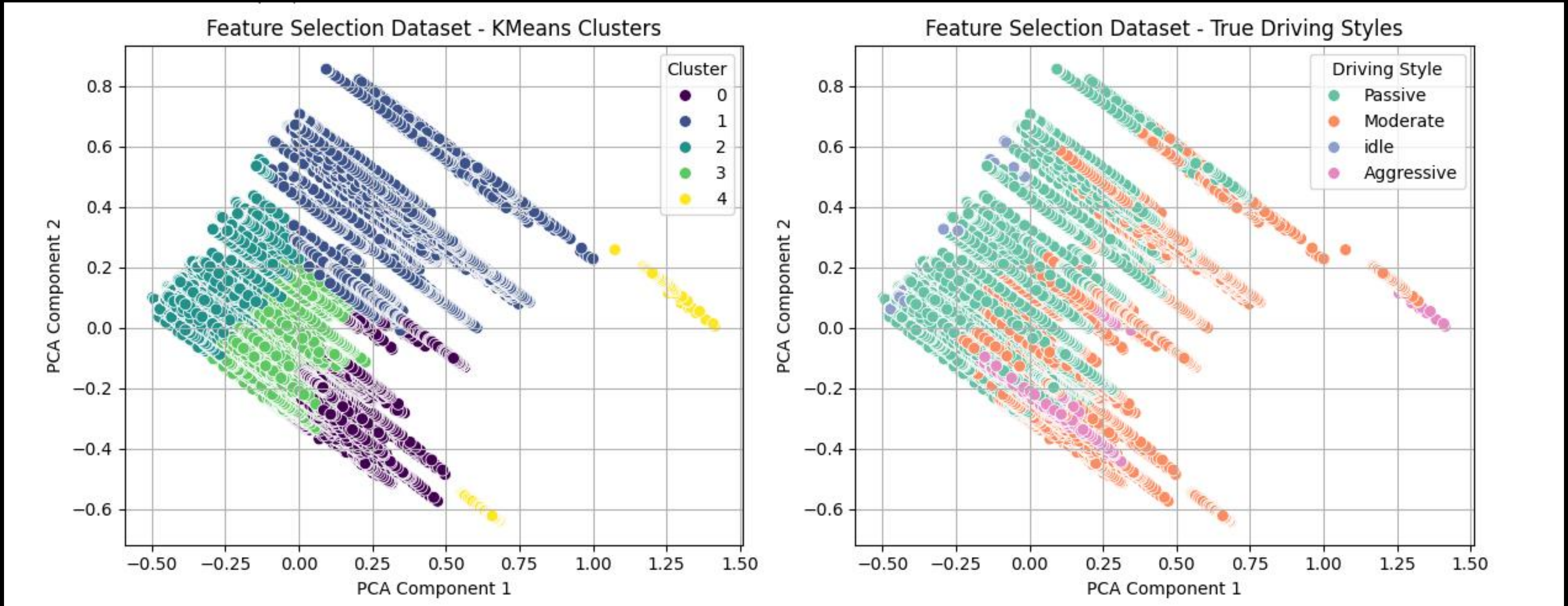
'FUEL_PRESSURE',

'INTAKE_PRESSURE'

FEATURED ENGINEERING DATASET

- INTAKE_PRESSURE *(kept as-is)*
- RPM_per_speed *(ratio of RPM to SPEED)*
- throttle_per_rpm *(ratio of throttle position to RPM)*
- speed_rolling_std *(rolling standard deviation of speed over 3 samples)*
- delta_throttle *(change in throttle position from previous sample)*
- engine_activity *(average of ENGINE_LOAD and MAF before dropping them)*
- speed_rpm_index *(normalized and averaged SPEED and RPM)*

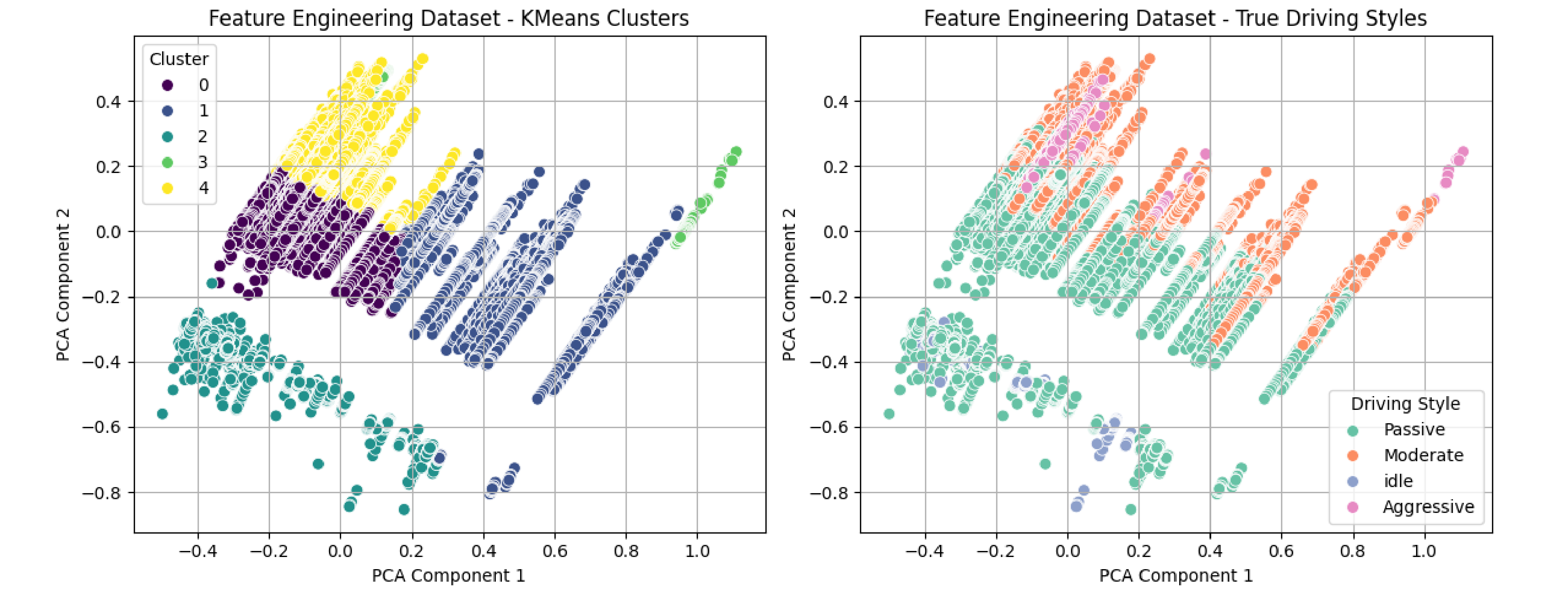
5 CLUSTER



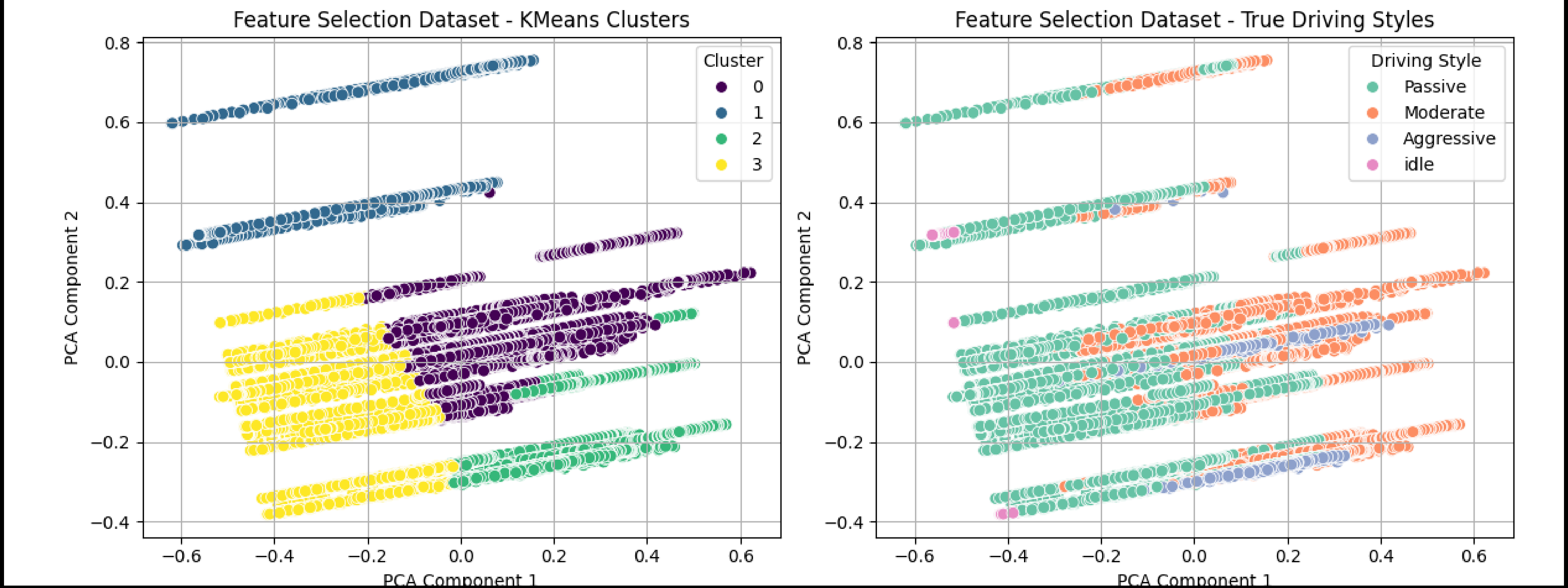
Feature Selection Dataset - KMeans Clusters

Feature Selection Dataset - True Driving Styles

FEATURED ENGINEERED DATASET



HIGH ACCURACY DATASET



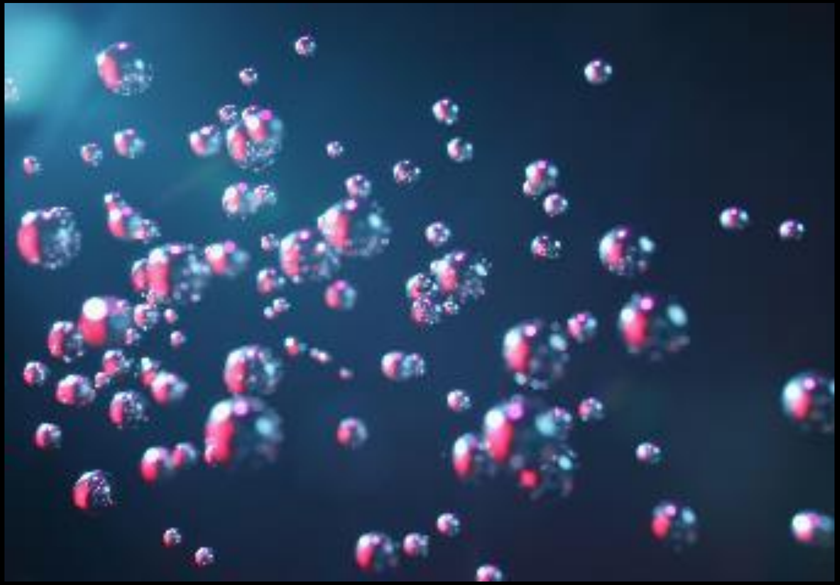
6/6/2025

PATHWAY AND INITIAL PROGRESS TO PREDICTIVE MAINTENANCE MODELLING

COLAB NOTEBOOK ACCESSIBLE URL

HTTPS://COLAB.RESEARCH.GOOGLE.COM/DRIVE/1WHF9CHSG__VUZCEVWH0G1FRIQ69HJOZS?USP=SHARING

COLLECTION



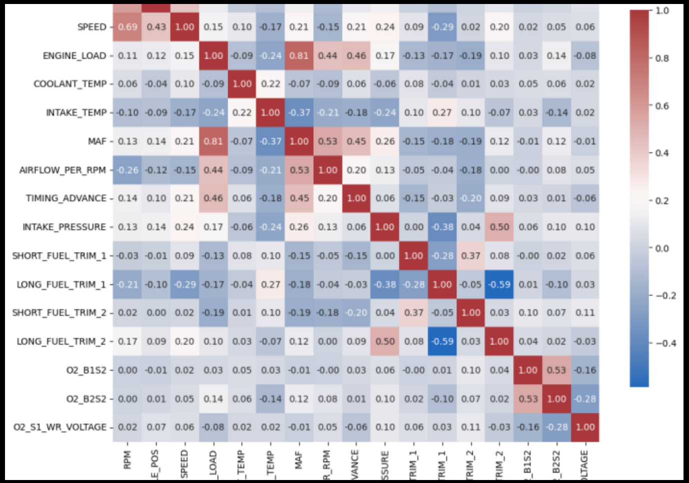
We collect and merge all dataset from the beginning until now and that to prepare for the initial modelling prototype.

EDAs run through over details and key findings from this merged data.

Although data has been through the automation cleaning pipeline (https://binkhoale1812-obd-logger.hf.space), minimal cleaning procedures also have to undergone to kept everything consistently in-place.

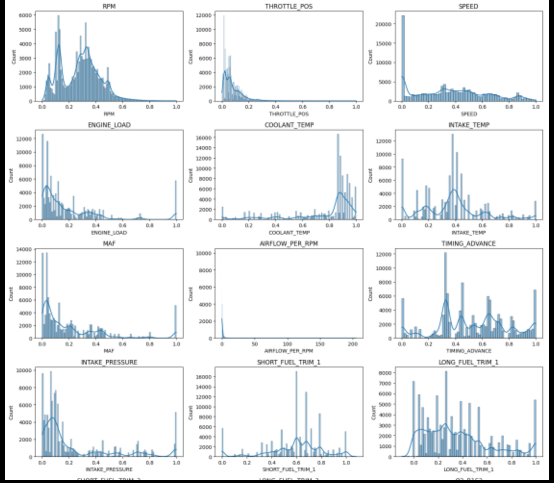
EDA

Correlation Heatmap



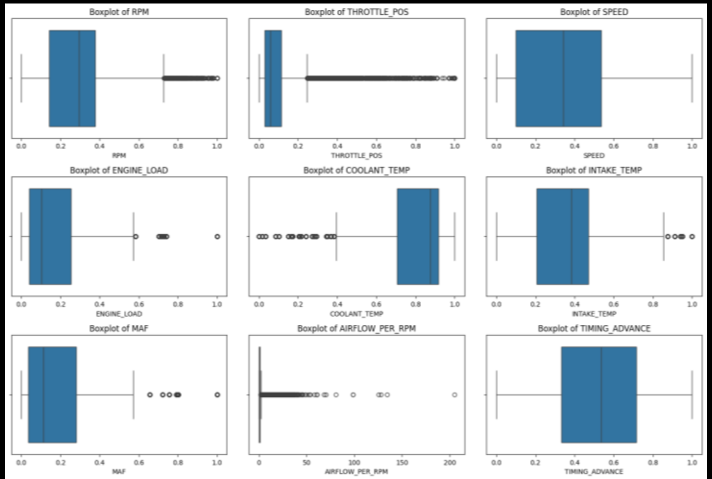
Action: Note which feature pairs exceed |0.8| correlation. This may inform feature removal or engineered interactions.

Distribution Plot

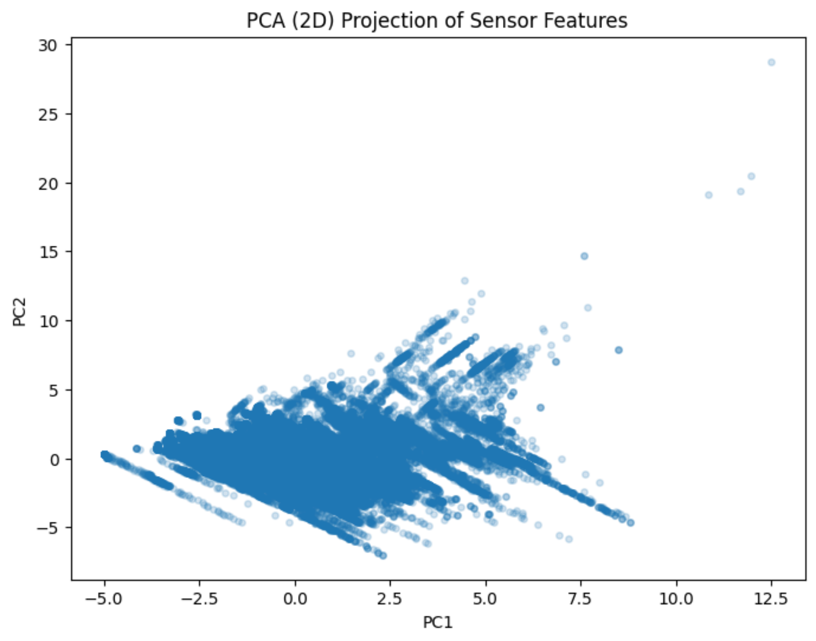


Action: Check for skewness, multimodality, heavy tails

Box Plot

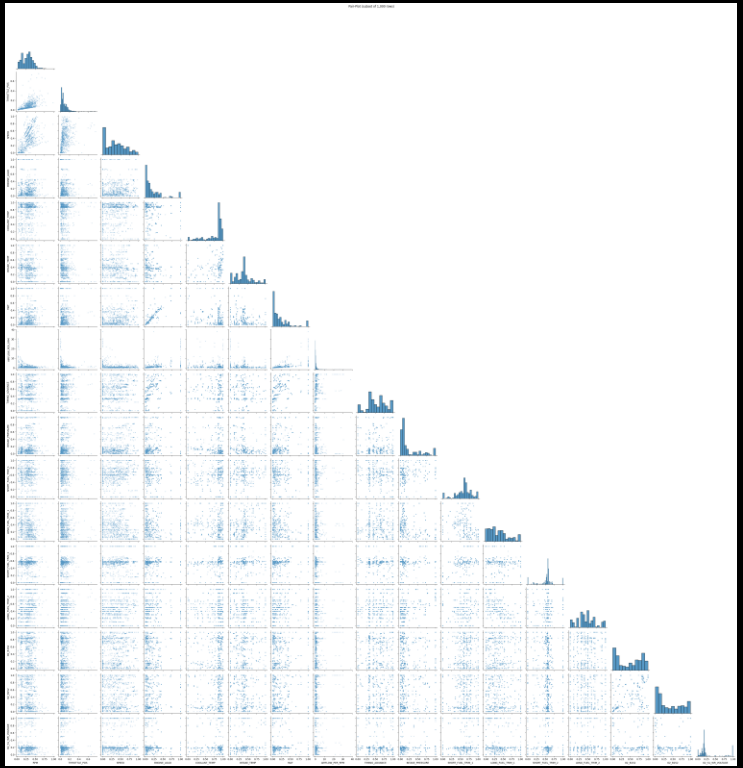


Action: Identify whether data has high-stress or outliers



PCA

Action: If outliers are too extreme, drop or investigate them



Scatter-Matrix

Action: Confirm any non-linear relationships, clusters, or obvious anomalies.

EDA

Data Distribution & Outliers

•Most variables are normalized to **[0, 1]**, except:

- •**AIRFLOW_PER_RPM**: Extreme outliers (max ≈ 205, 75% ≤ 0.982).
- •**Fuel trims (SHORT_FUEL_TRIM_1, etc.)**: Slightly exceed 1 (e.g., 1.08).

•**ENGINE_LOAD**: Mostly < 0.3, occasional spikes near 1.0.

•**COOLANT_TEMP**: Bimodal—cold starts (~0.0) vs. hot engine (~1.0).

Strong Correlations (Potential Redundancies)

•**ENGINE_LOAD ↔ MAF (0.81)**: Consider combining/dropping one.

•**Fuel trims & O_2 sensors**: Moderate correlations (0.5–0.53).

•**Others**: Weak correlations (|r| < 0.4).

Distribution Patterns

- **Bimodal/Multimodal**:
  - **RPM**: Peaks at low (~0.2–0.4) and high (~0.6+) ranges.
  - **INTAKE_TEMP**: Multiple modes (ambient vs. warmed air).
- **Heavy Tails**:
  - **AIRFLOW_PER_RPM**: Highly right-skewed (most < 5, rare > 200).
  - **SPEED**: Mostly idle (0.0–0.1), with occasional high-speed events.

PCA & Anomalies

- **Outliers (PC1 > 10 or PC2 > 20)**: Likely linked to extreme AIRFLOW_PER_RPM or sensor errors.
- **Rare events (|PC1| > 5 or |PC2| > 10)**: Potential malfunctions/noise.

Actionable Takeaways

- **Drop/combine highly correlated features** (e.g., ENGINE_LOAD & MAF).
- **Investigate outliers** in AIRFLOW_PER_RPM and fuel trims.
- **Segment analysis by engine state** (cold vs. hot) for clearer patterns.

ACTION:

Cap AIRFLOW_PER_RPM at 99th percentile

Combine O2 sensors into O2_COMBO

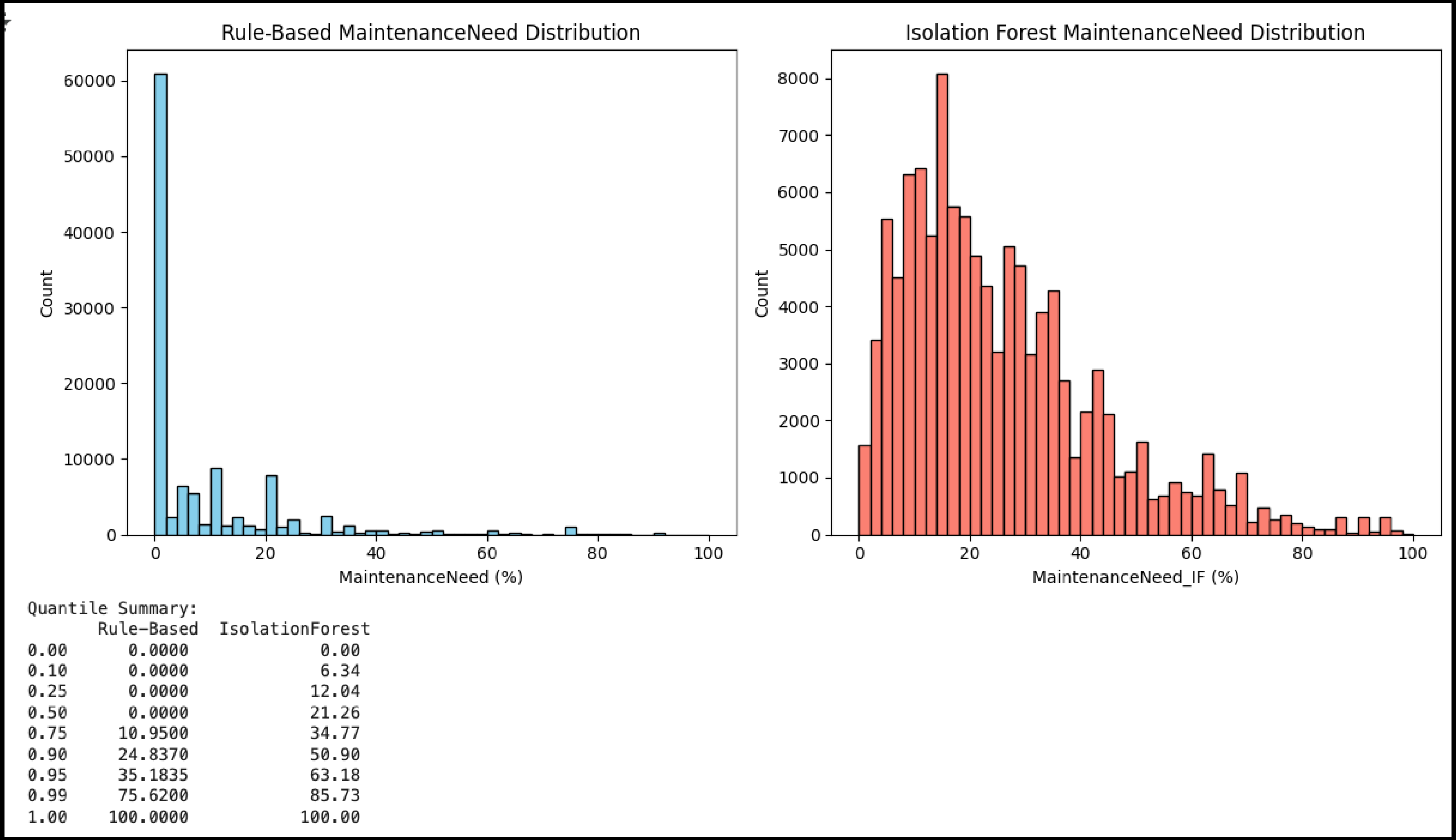
Decide to drop MAF because it's 0.81 correlated with ENGINE_LOAD

EDA

DECISIVE VARIABLES

'ENGINE_LOAD'
'COOLANT_TEMP'
'AIRFLOW_CAPPED_NORM'
'SHORT_FUEL_TRIM_1'
'TIMING_ADVANCE'
'O2_COMBO'

HEURISTIC TARGET LABELLING

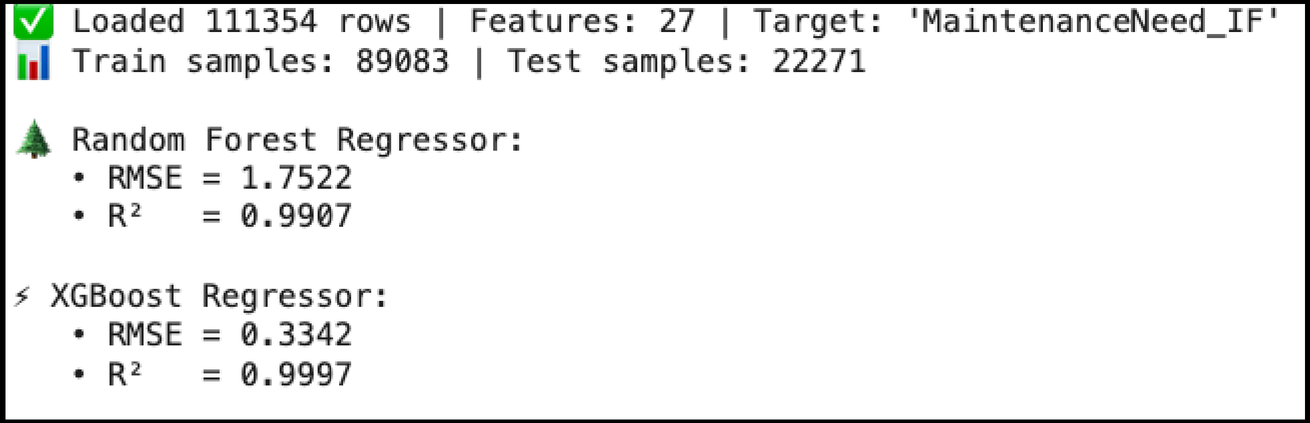


```
Quantile Summary:
      Rule-Based  IsolationForest
0.00      0.0000             0.00
0.10      0.0000             6.34
0.25      0.0000            12.04
0.50      0.0000            21.26
0.75     10.9500            34.77
0.90     24.8370            50.90
0.95     35.1835            63.18
0.99     75.6200            85.73
1.00    100.0000           100.00
```

INITIAL MODELLING

The current maintenance need score is based on inferred stress from engine sensor patterns. This assumes that high-stress readings indicate mechanical risk, which is a reasonable proxy in the absence of actual failure data. However, in real-world applications, labels should ideally reflect confirmed maintenance events — such as breakdowns or part replacements — rather than estimated risk.

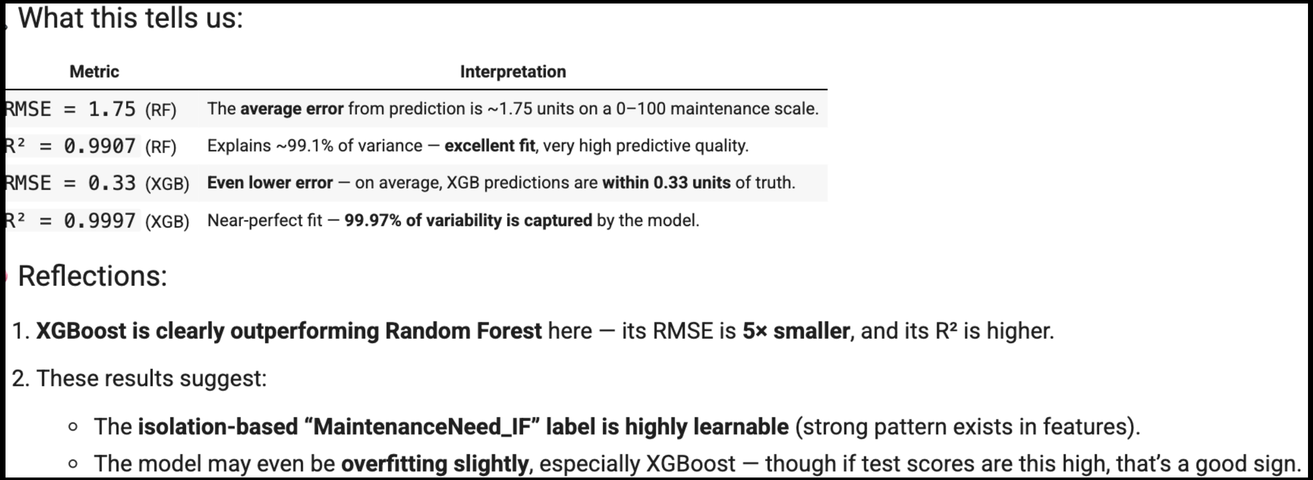
To improve future models, real maintenance records should be collected and used for supervised learning. Additional context like driving behavior, time-series patterns, or fleet-wide data could also help refine predictions and make the system more reliable in practice.



```
✅ Loaded 111354 rows | Features: 27 | Target: 'MaintenanceNeed_IF'
📊 Train samples: 89083 | Test samples: 22271

🌲 Random Forest Regressor:
  • RMSE = 1.7522
  • R²   = 0.9907

⚡ XGBoost Regressor:
  • RMSE = 0.3342
  • R²   = 0.9997
```



What this tells us:

Metric	Interpretation
RMSE = 1.75 (RF)	The **average error** from prediction is ~1.75 units on a 0–100 maintenance scale.
$R^2 = 0.9907$ (RF)	Explains ~99.1% of variance — **excellent fit**, very high predictive quality.
RMSE = 0.33 (XGB)	**Even lower error** — on average, XGB predictions are **within 0.33 units** of truth.
$R^2 = 0.9997$ (XGB)	Near-perfect fit — **99.97% of variability is captured** by the model.

Reflections:

1. **XGBoost is clearly outperforming Random Forest** here — its RMSE is **5× smaller**, and its R^2 is higher.
2. These results suggest:
   - The **isolation-based "MaintenanceNeed_IF" label is highly learnable** (strong pattern exists in features).
   - The model may even be **overfitting slightly**, especially XGBoost — though if test scores are this high, that's a good sign.

Model save at:
https://drive.google.com/drive/folders/1PnICqfScUvd37CgXDCmC4pxeeYfqWnQW?usp=sharing

Q & A

Questions and instructions for future phases.