ANALYZING MACHINE USAGE PATTERNS IN INDUSTRIAL IOT DATA:

A HYBRID CLUSTERING-CLASSIFICATION APPROACH

Presented by : **Farideh Tavakoli**

Open in Colab Clustering

Open in Colab Classification

BUSINESS MOTIVATION

- Industrial machines generate massive IoT data from sensors tracking daily activity.
- They capture how long machines spend in Execution, Ready, Failure, and Power-Off.
- With thousands of machines and millions of records, manual analysis is impossible.

➡ ***How can we automatically identify underperforming machines based on real usage data to enable proactive maintanance?***

OBJECTIVE

IoT Data → Pattern Analysis → Detect Underperformance → Notify Customer → Offer Maintenance

Goal: *Transform IoT data into proactive service insights that enhance customer reliability.*

Pattern Analysis → Clustering
Detect Underperformance → Classification

DATASET

- IoT data collected from industrial machines equipped with embedded sensors.
- Data was integrated and aggregated in Power BI, producing ~*36,000* daily records.
- Each representing a machine's operational summary for a single day.
- Machine IDs and subtechnologies anonymized (GDPR compliant).

Features:

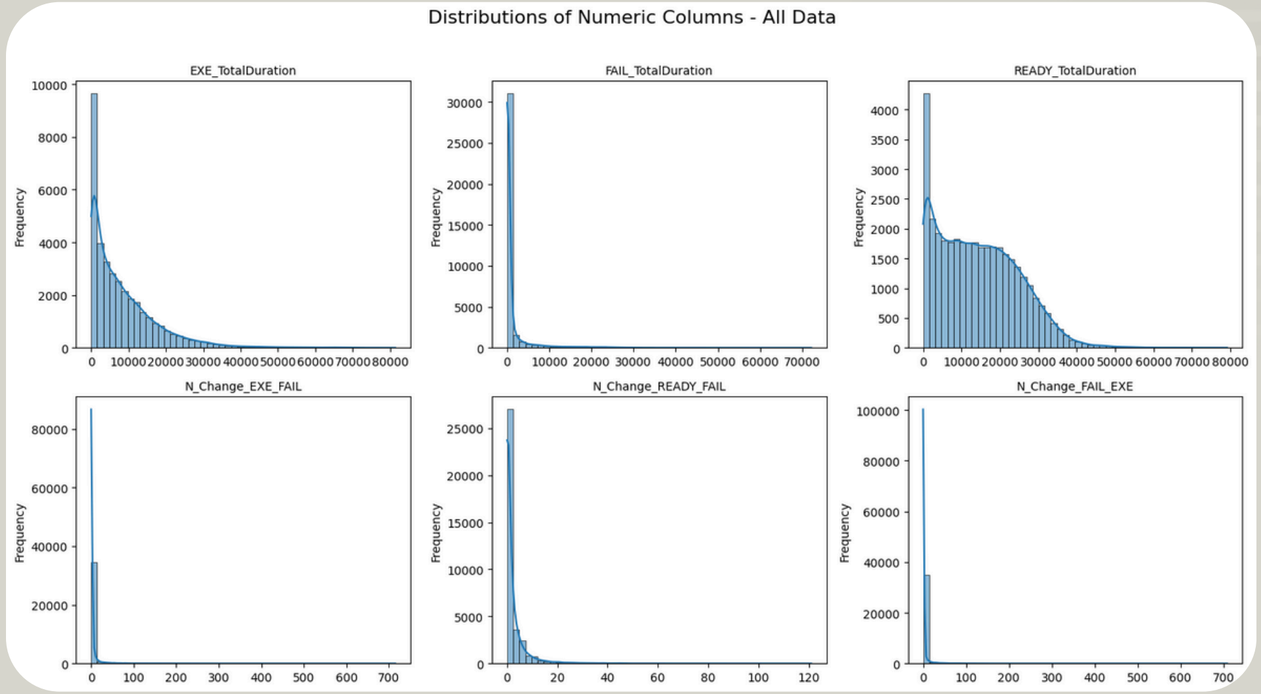
- Operational Duration in **EXE**, **READY**, **FAIL**, **POWER_OFF**
- State Transition between them, e.g. **EXE→FAIL, FAIL→READY**
- Contextual Identifiers like **Serial Number**, **Date**, **Subtechnology Name**, **Shipment** and **Manufacturing dates**.

DATA UNDERSTANDING

Exploratory Data Analysis

- *Right-skewed distribution*
- *Outliers →*

  real abnormal behavior, not noise
- To balance feature ranges and reduce distortion → Min–Max scaling



DATA PREPARATION

To ensure the dataset is clean, consistent, and ready for clustering and classification.

Missing Values:

- Imputed missing operational durations with 0.
- Removed rows where all durations = 0 (inactive machines).
- Dropped records missing subtechnology name or shipment date.

Outlier Handling:

- Right-skewed numeric features; outliers retained as meaningful rare events.
- Log transformation tested → reduced cluster quality → discarded.
- Outlier impact mitigated via scaling.

DATA PREPARATION

Data Type Verification:

- *Ensured correct formats: numeric → int / float; dates → datetime.*

Relevant Data Selection:

- Removed non-informative identifiers (*Serial Number*, *Date*) and metadata (*Shipment* and *Manufacturing Dates*) from modeling.
- Focused on *sub_3* and *sub_4* for relevance and data volume.

Feature Engineering & Scaling:

- Converted durations to minutes for interpretability.
- *RobustScaler* vs *MinMaxScaler* → chose MinMax for better cluster separation.

FEATURE SELECTION

Variance Analysis

- Tested after MinMax scaling
- Many features showed low variance (normalization effect)
- Log / robust scaling → increased variance but reduced cluster quality.
- *Decision:* Keep all features since they capture meaningful behaviors.

Correlation Analysis

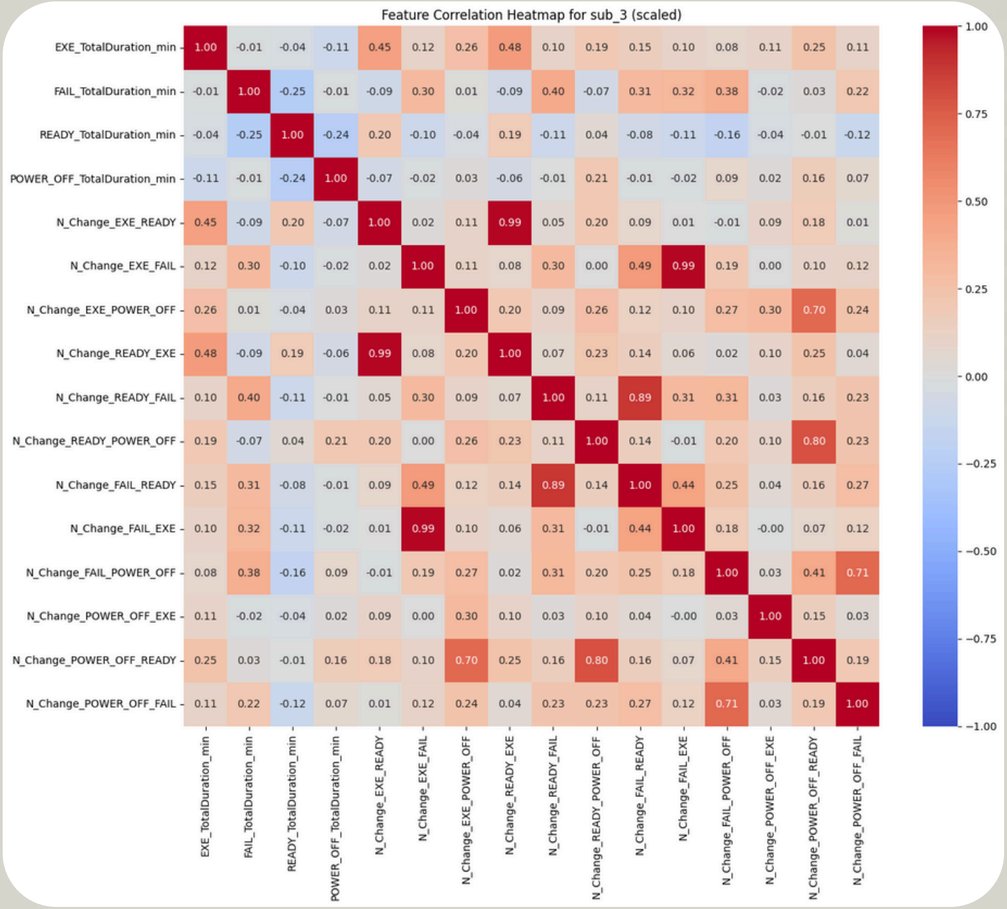
- Computed Pearson correlation matrix.
- Removed one feature if correlation ***> 0.7***
- *Result:*
  - Reduced redundancy
  - improved stability & interpretability.

FEATURE SELECTION

Correlation Analysis

Strong positive correlations among transition-related features

- e.g. N_Change_EXE_READY ↔ N_Change_READY_EXE
- This reflect bidirectional machine state transitions



MODELING

Goal: Combine clustering (to discover machine behavior patterns) with classification (to predict those patterns in new data).

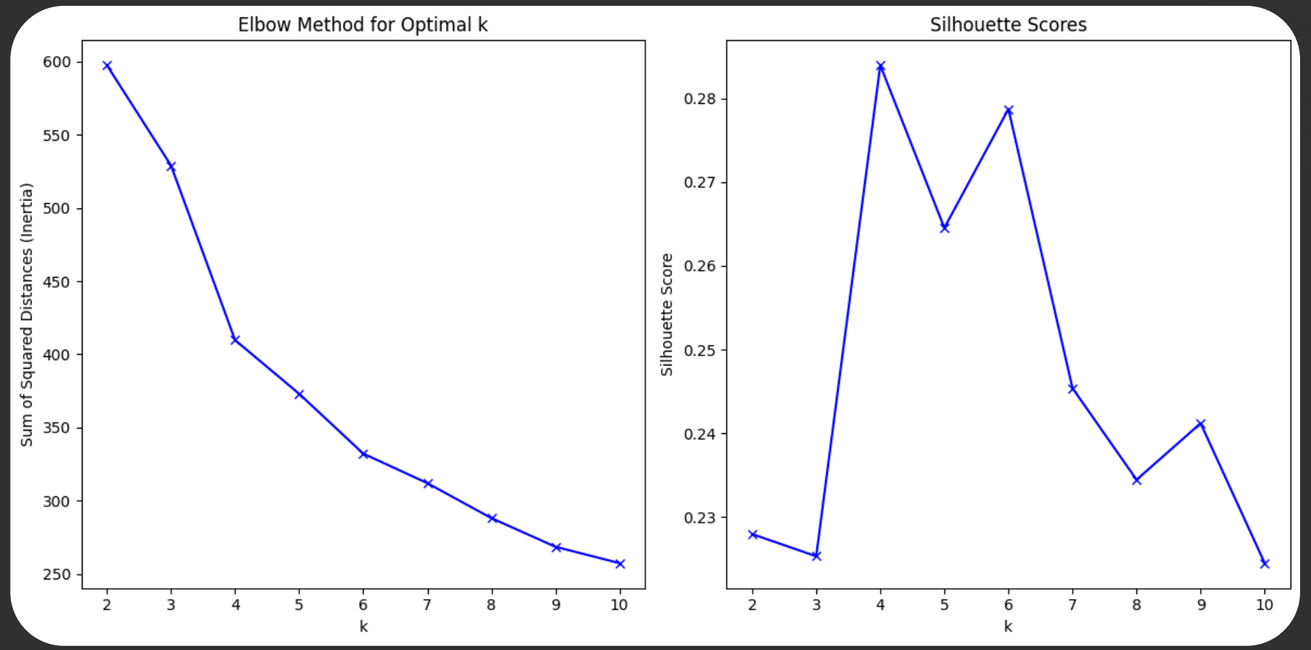
Raw Data → K-Means → Cluster Labels → Classifiers (SVM, D Tree, KNN)

- ***Step 1:*** Unsupervised K-Means clustering → discovers natural machine groups
- ***Step 2:*** Clusters become labels → used to train supervised models
- ***Outcome:*** A hybrid system that both explores and predicts machine behavior

CLUSTERING & RESULTS

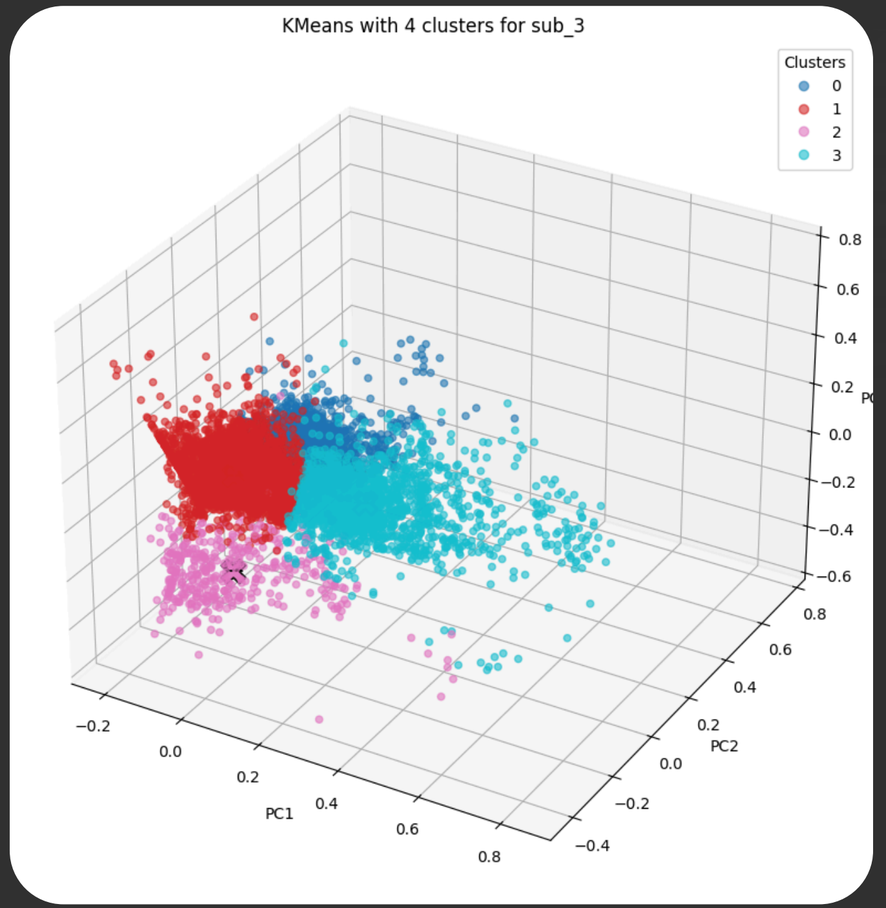
- Algorithm: ***K-Means*** (chosen for simplicity, scalability, interpretability) →
- Optimal clusters: *k = 4* (Elbow + Silhouette methods)

- ***DBSCAN*** → detected noise but unsuitable (incomplete labeling)



CLUSTERING & RESULTS

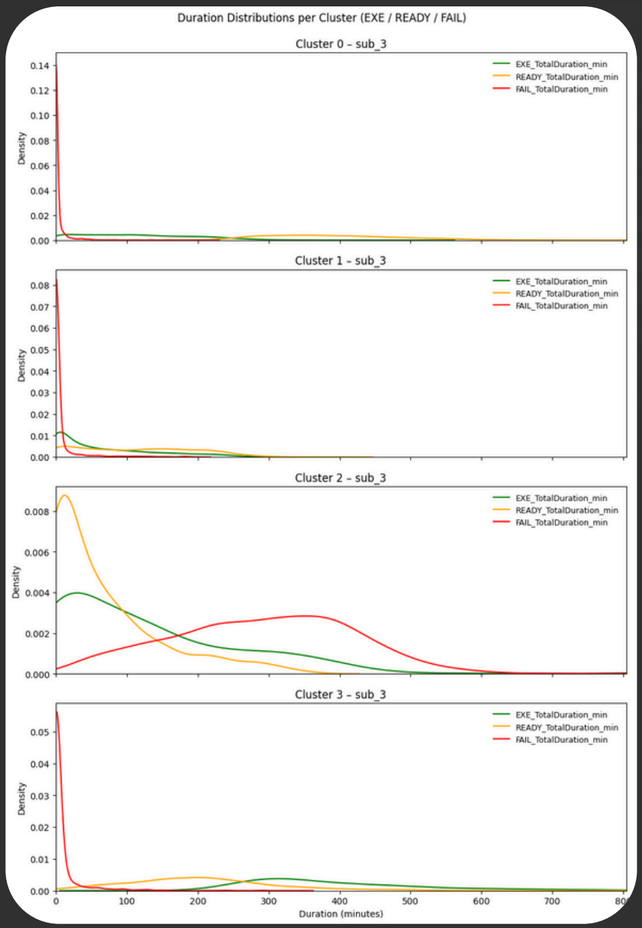
- PCA used for 3D visualization of behavioral separation
- Separately applied on each subtechnology, to capture design-specific behaviors
- Clusters serve as interpretable machine behavior categories → foundation for supervised classification.



CLUSTERING & RESULTS

Cluster Interpretation (sub_3)

(k)	Total_N	Distinct_N	EXE_D	FAIL_D	READY_D	POWER_OFF_D	Label
0	4809	514	1:55:43	0:03:36	6:35:24	0:27:05	Balanced Use
1	3594	600	1:06:56	0:06:30	1:55:30	1:54:36	Idle
2	419	62	2:16:57	4:51:55	4:51:55	0:49:10	Faulty
3	2181	332	7:03:29	0:09:17	0:09:17	0:39:47	Productive



Clustering outputs serve as input labels for classification stage.

CLASSIFICATION & RESULTS

Goal: Predict machine behavior labels (from clustering) for new unseen data.

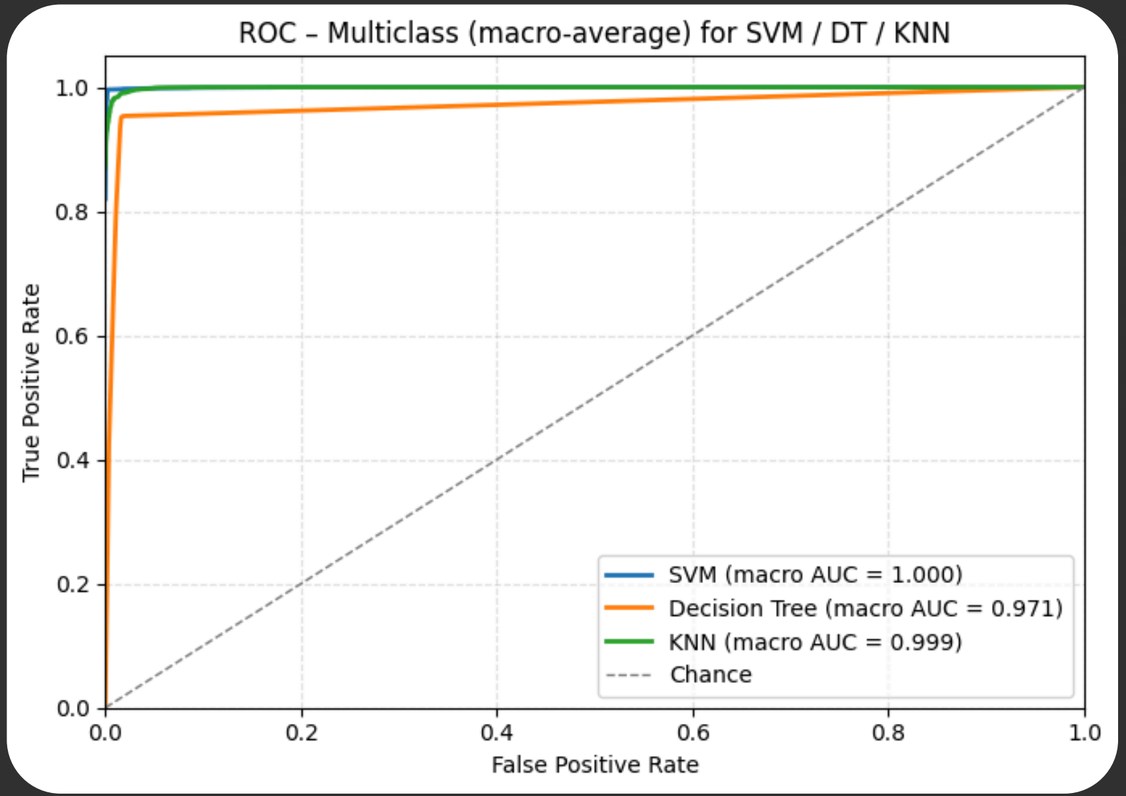
Approach:

- Models: ***SVM, Decision Tree, KNN***
- Implemented via Pipelines (MinMax Scaler + Model) → prevents data leakage
- GridSearchCV + Stratified 5-Fold CV for tuning
- Evaluation metric: Macro F1-score (balances precision & recall, minimizing false alarms and missed failures)

CLASSIFICATION & RESULTS

Performance Evaluation

Model	Accuracy	W_F1	AUC
SVM	0.994	0.994	~1.0
Decision Tree	0.965	0.965	0.97
KNN	0.977	0.977	0.999



All models achieved high predictive accuracy

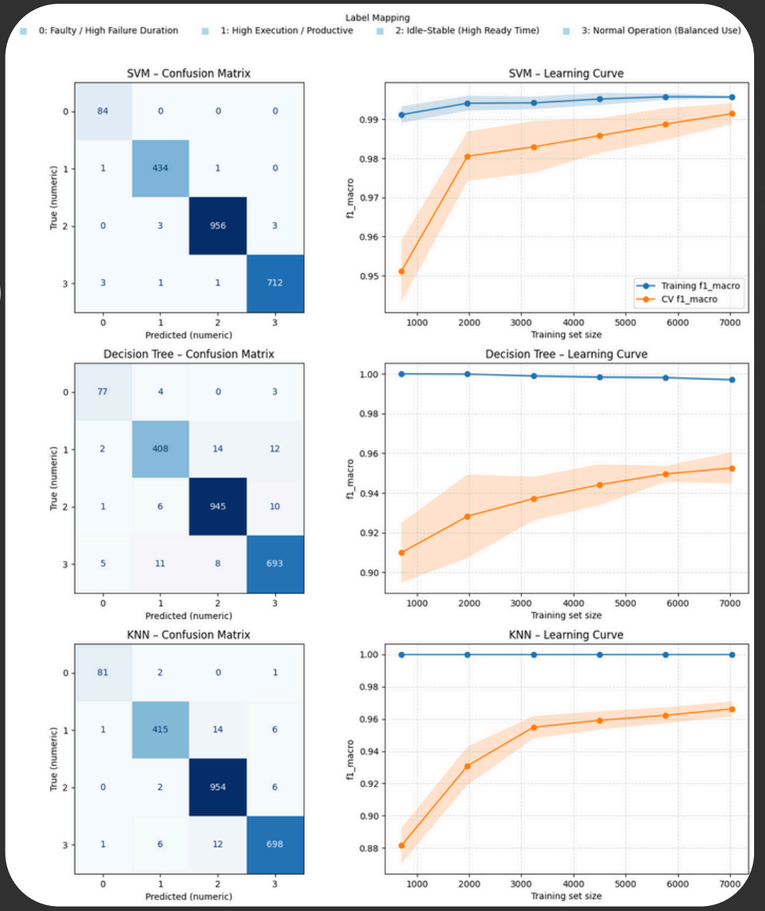
SVM achieved best generalization (AUC ≈ 1.0, minimal overfitting)

CLASSIFICATION & RESULTS

Performance Evaluation

- ***SVM*** → stable performance, minimal rain–val gap
- ***Decision Tree*** → mild overfitting (train ≈ 1.0, val ≈ 0.95)
- ***KNN*** → good generalization, moderate overfitting typical of instance-based models

SVM selected as final model for deployment (saved as MinMaxScaler + SVM pipeline).



CONCLUSION

Data-driven framework for machine behavior understanding

- K-Means clustering (behavioral segmentation) + classification (pattern prediction)
- Clustering revealed clear usage profiles → used as behavioral labels
- All classifiers showed strong accuracy and consistency, validating cluster quality
- SVM achieved the best F1-score, generalization, and stability
- Final SVM pipeline was saved with joblib for real-time prediction on IoT data

Impact:

Transforms raw machine data → actionable insights for performance monitoring and predictive maintenance

CONCLUSION

Limitations

- Only 2 subtechnologies analyzed
- Semantic labeling bias risk
- Rsk of concept drift over time

Future Work

- Expand the dataset
- Periodically retrain clustering
- Explore more algorithms
- A monitoring dashboard

THANK YOU

Presented by : **Farideh Tavakoli**