**Motor Speed Overload Detection and Analysis Using**

**PCA Models**

**Group Members:**

* Saumya Vaidya
* Soham Raut
* Srajan Kewat
* Sohel Bhongade

**Introduction:**

Equipment E1 to E5 are also in series with each other in industrial unit operations to treat a specific material. The system's throughput is regulated by the E1 speed controller. But density and volume variations of the material can lead to uneven loads at times, occasionally beyond the downstream equipment capacity (E2 to E5).

Every E2 through E5 unit has motor current sensors (FLAs - Full Load Amps) to monitor the load of the equipment.

When the loads are above safe levels, there is a high likelihood of equipment failure, and that results in total system shutdown. Resumption of operations in case of failure may take as long as 4 hours and leads to major losses in production. Thus, early detection of overload states is essential in order to facilitate operational effectiveness as well as minimize downtime.

**Objective:**

* Develop a data-driven model to monitor and predict overload conditions in the production train.
* Create a calculated variable to indicate when the train is overloaded.
* Analyse and clean the data to account for system downtimes and no-load conditions.
* Use PCA (Principal Component Analysis) to understand the data structure and detect abnormal operating conditions.

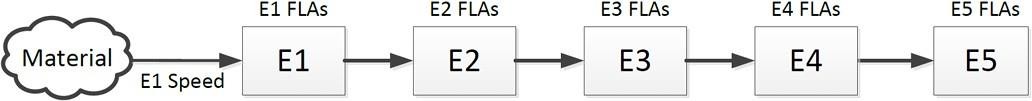
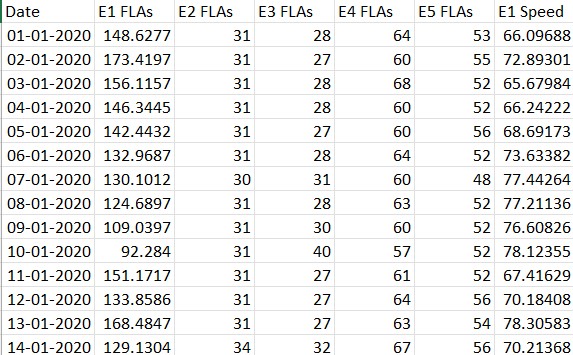


Figure 1: Material processing sequence showing E1-E5 units with FLA sensors at each stage

This diagram visually represents the material processing sequence and highlights the importance of monitoring motor loads at each stage to prevent overloads and ensure smooth, continuous operation.

**First we will do the data preprocessing to build our models**

The dataset includes a 'Date' column representing daily measurements. For time series analysis, I set 'Date' as the datetime index after converting it to the proper datetime format. The dataset also includes operational variables such as 'E1 FLAs', 'E2 FLAs', 'E3 FLAs', 'E4 FLAs', 'E5 FLAs', and 'E1 Speed', which capture motor load and speed information across the equipment train .



**Figure 2:** Represents a Raw Data of motor speed problem.

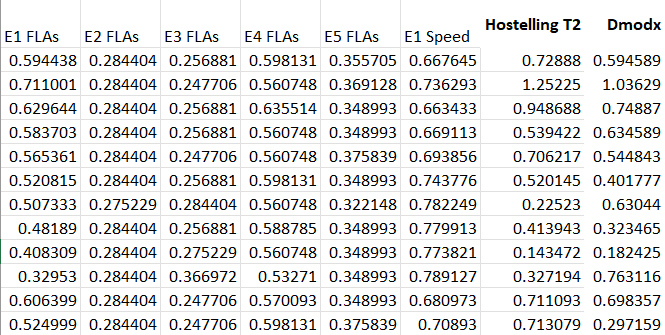
**Preprocessing steps:**

**Data Preprocessing Summary**

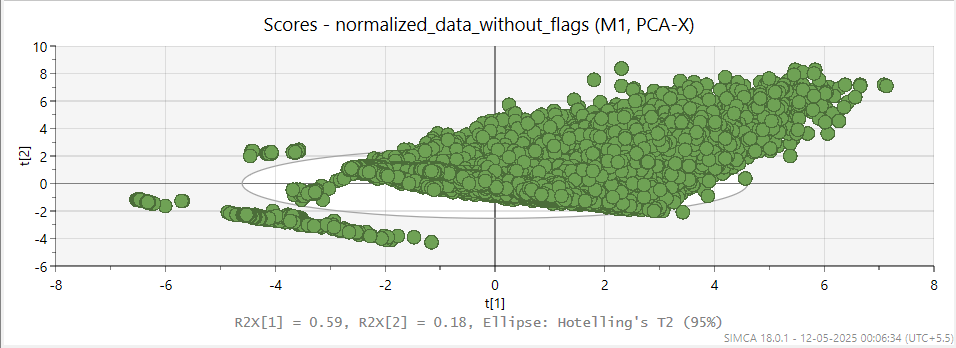
1. **Data Type Conversion**:
   * All features were converted to appropriate numerical data types (e.g., float, int) to ensure compatibility with PCA and other numerical analyses.
2. **Removal of Non-informative Columns**:
   * The 'time' column was excluded from the dataset as it did not contribute meaningful information for the PCA.
3. **Filtering Invalid Entries**:
   * Rows where 'E1 FLAs' values were less than 20 were removed. These entries often led to other variables having zero values, which could distort the analysis.
4. **Handling Categorical Variables**:
   * Categorical columns, such as 'bad inputs', were initially considered. However, after filtering for 'E1 FLAs' ≥ 20, these columns predominantly contained zeros and did not provide additional insights. Consequently, they were excluded from further analysis.
5. **Imputation of Missing Values**:
   * Missing or zero values resulting from the previous filtering steps were imputed using the K-Nearest Neighbours (KNN) imputation method available in scikit-learn. This technique estimates missing values based on the mean of the nearest neighbours, preserving the dataset's structure and relationships.

**Pre-processed Data:**

We have pre-processed the given data by applying centring and scaling to the raw measurements, ensuring all variables have a mean of zero and a standard deviation of one for consistent analysis.

 **Figure 3:** Represents Pre-processed Data   
  
 **Analysis & Plots**

1. **PCA Score Scatter Plot:**



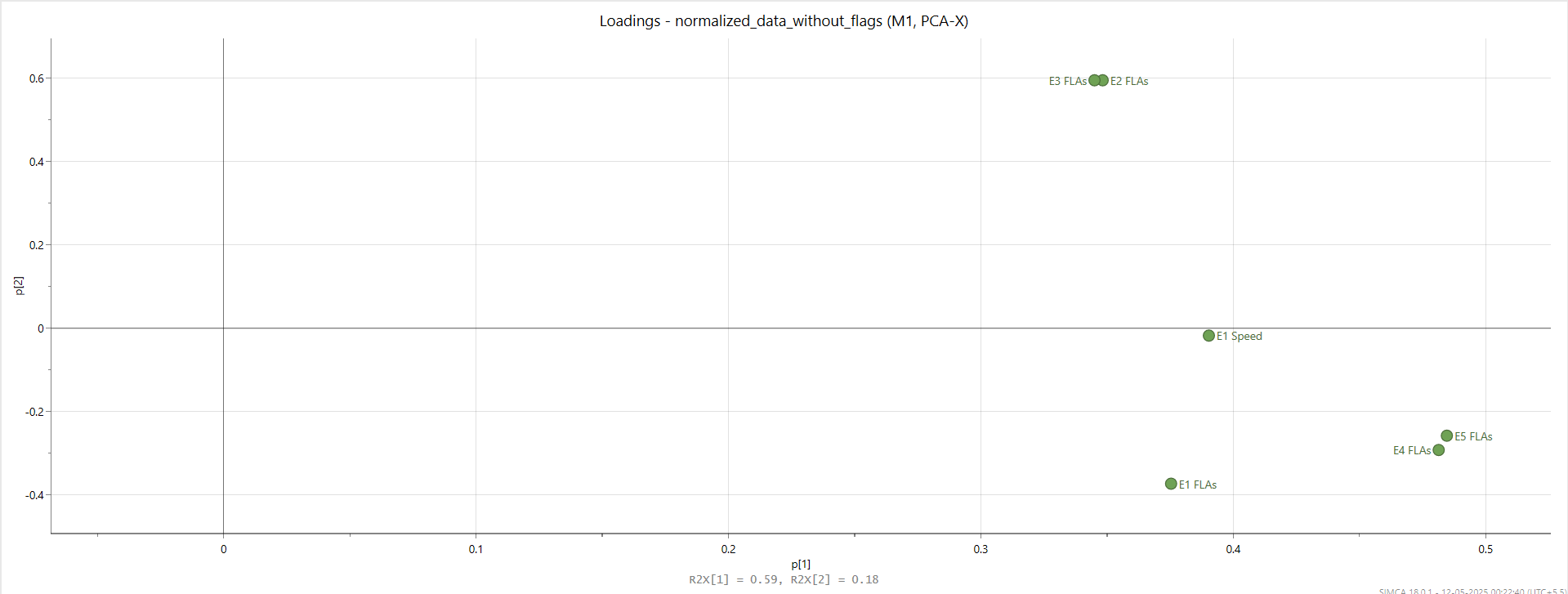
* 1. **Axes Explanation:** 
     + t[1] (horizontal axis) and t[2] (vertical axis) represent the first and second principal components, respectively.
     + t[1] captures 59% of the total variance in the data (R2X[1] = 0.59).
     + t[2] captures an additional 18% of the variance (R2X[2] = 0.18).
     + Together, the two components explain about 77% of the total variability in the dataset — a good coverage.

* 1. **Interpretation of Points:** 
     + Each green dot represents one sample (observation) in the dataset.
     + Most of the points are clustered near the centre and spread along a diagonal direction, indicating correlated structure among the variables.
     + The plot shows a mild elongation, suggesting some underlying pattern or gradient in the data.

* 1. **Ellipse (Hotelling's T² 95%):** 
     + The white ellipse represents the 95% confidence limit.
     + Points inside the ellipse are considered "normal" based on multivariate distance.
     + Points outside the ellipse (especially the red ones) are considered outliers or unusual observations.

* 1. **Outliers:** 
     + A few points (in red) lie outside the ellipse, indicating potential anomalies or unusual behaviour in the process (likely corresponding to overloads or machine abnormalities)

**(2) PCA Loadings Plot:**



1. **Axes Explanation:**

* + p[1] (horizontal axis) and p[2] (vertical axis) represent the loadings on the first and second principal components.
  + Variables that are far from the origin along p[1] or p[2] contribute more strongly to the principal components.
  + R2X[1] = 0.59 and R2X[2] = 0.18, meaning the first two components together explain 77% of the data variance.

1. **Interpretation of Variables:**

* + E4 FLAs and E5 FLAs are grouped closely together on the right side and contribute strongly to p[1], meaning they share similar behaviour and are major drivers of the first principal component.
  + E2 FLAs shows a strong contribution to p[2], suggesting it captures variance orthogonal (different) from the other variables.
  + E1 FLAs and E1 Speed are moderately loaded along p[1] with slight variations in p[2], indicating they also play significant roles but in slightly different ways.

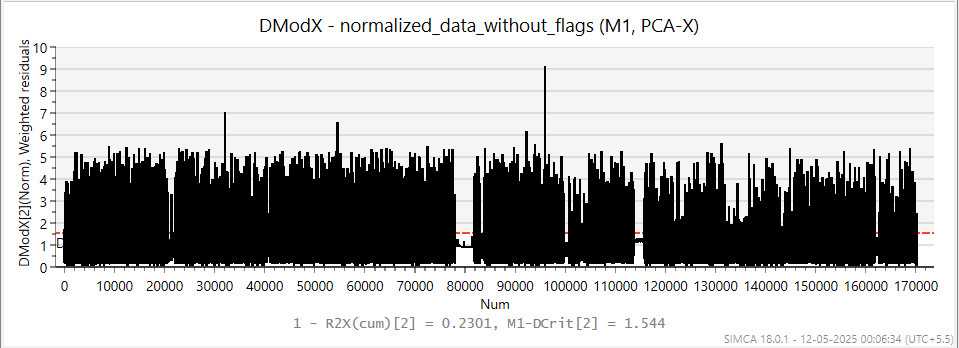
1. **Clustering and Relationship:**

* + Variables close together (like E4 FLAs and E5 FLAs) are positively correlated.
  + Variables located far apart vertically (p[2] direction) show less direct correlation.

1. **Key Observations:**

* + The first principal component (p[1]) mainly describes the variance in E4 FLAs, E5 FLAs, E1 Speed, and E1 FLAs.
  + The second principal component (p[2]) captures specific variance mostly associated with E2 FLAs.

**(3) DModX Plot:**



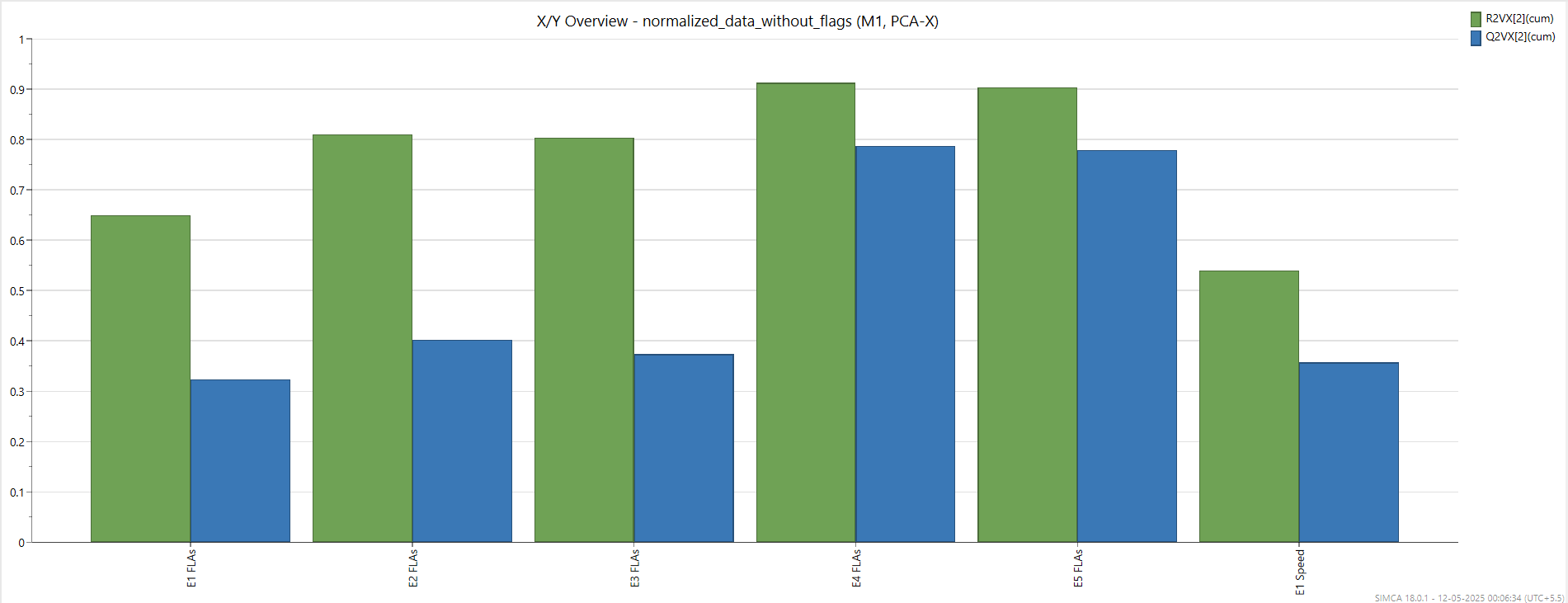
1. **Plot Type:**
   * It represents the weighted residuals (errors) for each observation/sample, indicating how far each point is from the PCA model.
2. **Axes Explanation:**
   * X-axis (Num): Sample number or observation index (up to ~170,000 samples).
   * Y-axis (DModX(5)(Norm**)):** Normalized distance from the PCA model based on 5 components.
   * Higher values mean the observation doesn't fit the model well (i.e., has a higher residual).
3. **Key Metrics:**
   * D-Crit (critical limit) is approximately 1.966.

 Points above this threshold (drawn horizontally) are considered outliers or abnormal compared to the PCA model.

* + 1 - R2X(cum)[5] = 0.003471, meaning only 0.35% of the total variance remains unexplained after using 5 principal components — a very good fit.

1. **Interpretation:**
   * **Major Spikes:**
     + Significant outliers are visible around sample indices 30,000, 60,000, 90,000, and 130,000.
     + These spikes indicate major deviations or special causes (such as machine faults, process changes, or rare events).
   * **Normal Regions:**
     + Most samples lie well below the critical limit, suggesting the process is largely stable and well-modelled.
   * **Red marks** highlight individual observations flagged as outliers beyond the critical distance.
2. **Outliers:**
   * The points with extreme DModX values could require further root cause analysis to understand why they deviate.
   * They might represent rare or abnormal behaviour in the system.

**(4) X/Y Overview - (M1, PCA-X)**



1. **Plot type:**

This suggests the graph shows an overview of the relationship between X-variables and Yvariables for a model (M1) using PCA (Principal Component Analysis) focused on X-data.

1. **Bars:**

* + Green bars represent R2VX[S] (cum) values — these show how much of the variance in the X-data is explained (R²), cumulatively.
  + Blue bars represent Q2VX[S] (cum) values — these show the cumulative predictive ability (Q²) of the model, based on cross-validation.

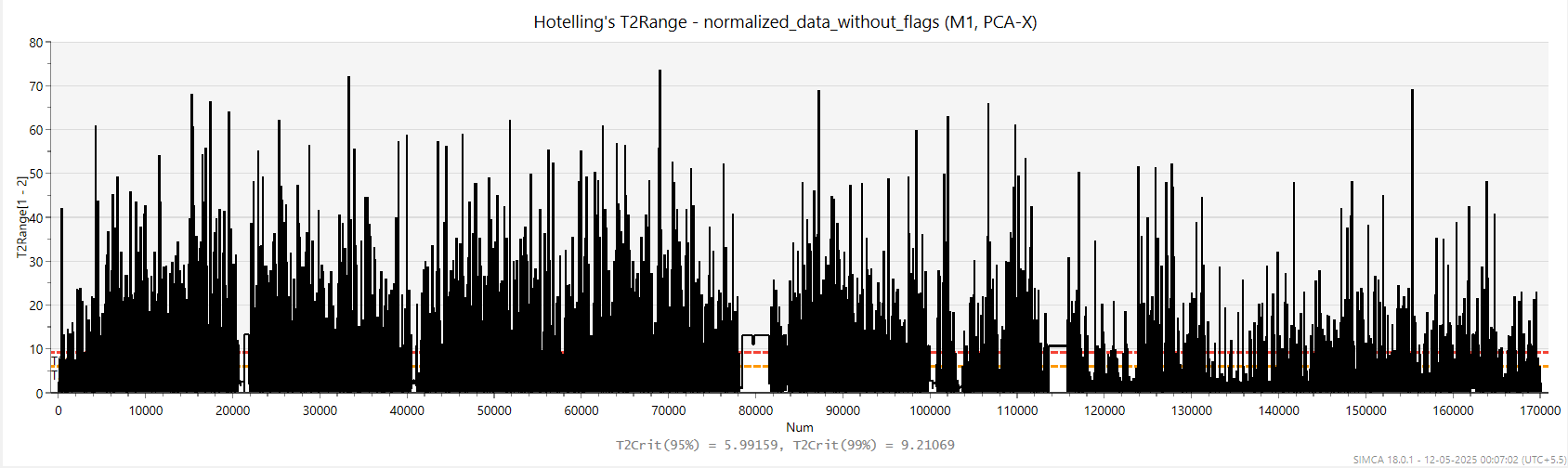
1. **X-axis (Variables):**

* + Variables such as E1 FLAs, E2 FLAs, E3 FLAs, E4 FLAs, E5 FLAs, and E1 Speed.
  + These likely represent different feature groups or sensor readings.

1. **Y-axis:**
   * Scale from 0 to 1 (or 0% to 100%), showing the proportion of explained variance and predictive ability.

1. **Observation:**
   * Both R² and Q² are very high (around 0.95–1.0) for all variables, indicating an excellent model with high explanatory and predictive power.
   * Especially for "E1 Speed," the values seem to be extremely close to 1, suggesting near-perfect modelling.

**(5) Hotelling’s T² Range Analysis**



**1.Plot Type**

* + **Purpose:** To monitor **Hotelling's T² statistics** over a series of observations for a PCA-X model.

**2.Axes Explanation**

* + **X-axis (Num):**

Represents observation number (or sample index) — from 0 up to around 170,000.

* + **Y-axis (T2RangePS[1..5]):**

Represents the Hotelling's T² range value for each observation.

* + - T² measures the variation of each sample within the PCA model space.

**3.Key Metrics**

* + **T² Critical Limits:**
    - **T2Crit(95%) = 11.071** (shown by a red dashed line) — 95% confidence limit.
    - **T2Crit(99%) = 15.0871** (also marked) — 99% confidence limit.

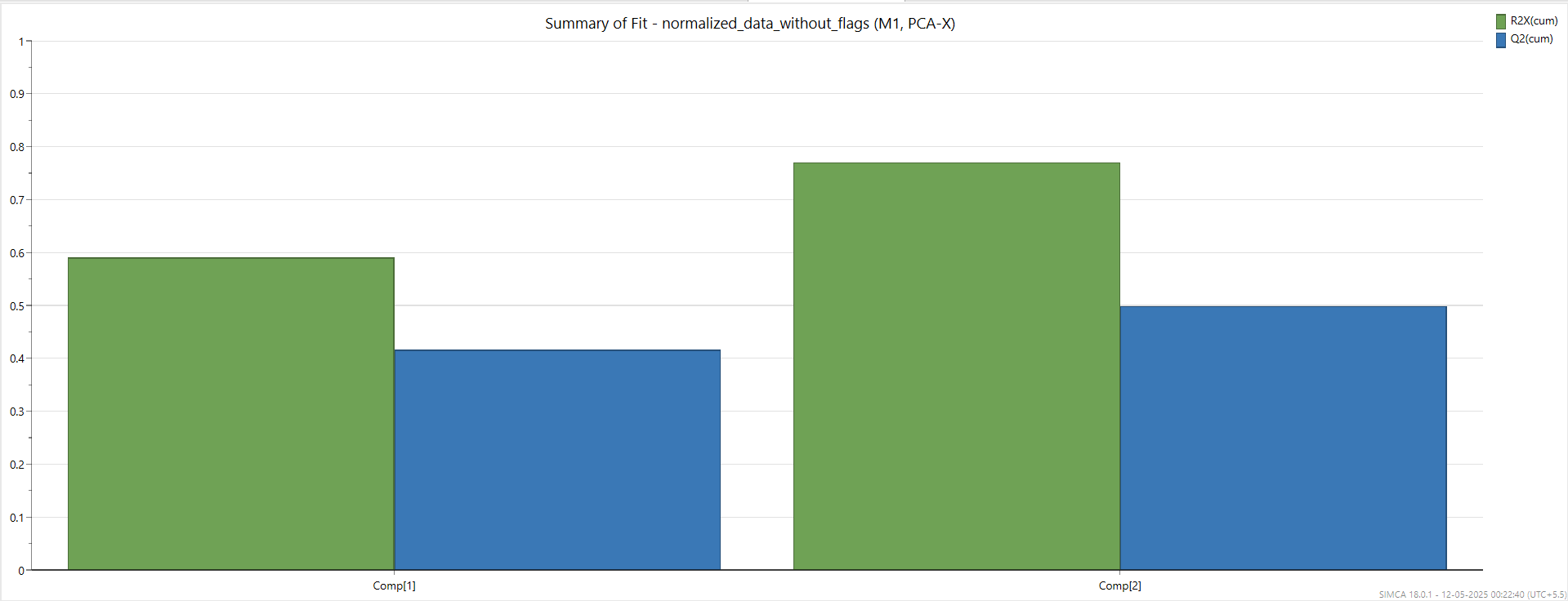
**4.Interpretation**

* + **Most observations exceed the 95% and 99% limits.**
    - A large number of samples show **very high T² values** — many going above 100.
    - Suggests **extreme variability** or **multiple outliers** present in the dataset.
  + **Process Instability:**
    - The model indicates that the process being monitored is **not stable**.
    - Significant deviations from expected behaviour are happening frequently.

**5.Outliers**

* + **Many extreme outliers are present.**
    - Normal data should stay mostly below the 95% limit.
    - Here, a majority of observations are way beyond even the 99% critical threshold.
  + **Possible causes:**
    - Measurement errors, shifts in process behaviour, equipment malfunctions, or change in underlying data patterns.

1. **PCA-X Model**



**1. Plot Type**

* Bar chart comparing R²X(cum) and Q²(cum) values for each principal component (Comp[1] to Comp[5]).

**2.Axes Explanation**

* **X-axis:** Principal Components (Comp[1], Comp[2], etc.).
* **Y-axis:** Cumulative variance values (scale from 0 to 1).

**3.Key Metrics**

* **R²X(cum):**

Green bars showing the amount of total variation explained by the model.

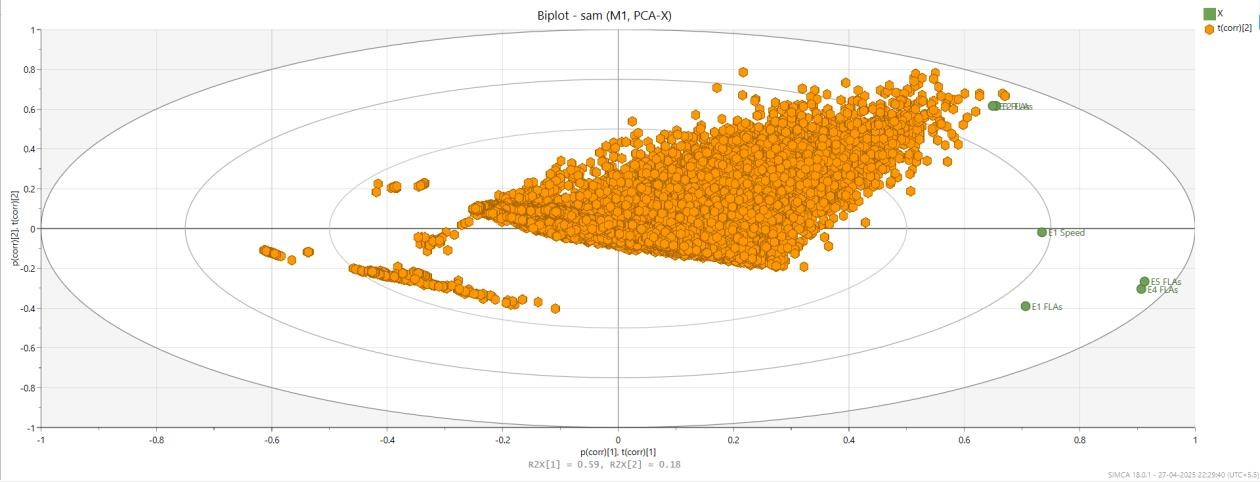
* **Q²(cum):**

Blue bars showing the predictive ability of the model (cross-validated).

**4.Interpretation**

* Model fit improves as more components are added.
* By Comp[5], both R²X(cum) and Q²(cum) are close to 1, indicating very high model quality.
* Early components (Comp[1] to Comp[3]) have moderate predictive power, but it improves significantly in Comp[4] and Comp[5].
* The PCA-X model shows good fit and strong predictive ability, especially after adding up to 5 components. The model is reliable for explaining and predicting process variations.

**(7) Biplot – PCA-X Model**



**1. Plot Type**

* Biplot showing both observations (orange hexagons) and variables (green circles).

**2.Axes Explanation**

* X-axis: p(corr)[1] (first principal component correlation).
* Y-axis: p(corr)[2] (second principal component correlation).

**3.Key Metrics**

* R²X[1] = 0.59, R²X[2] = 0.18

(First two components together explain a good amount of data variance.)

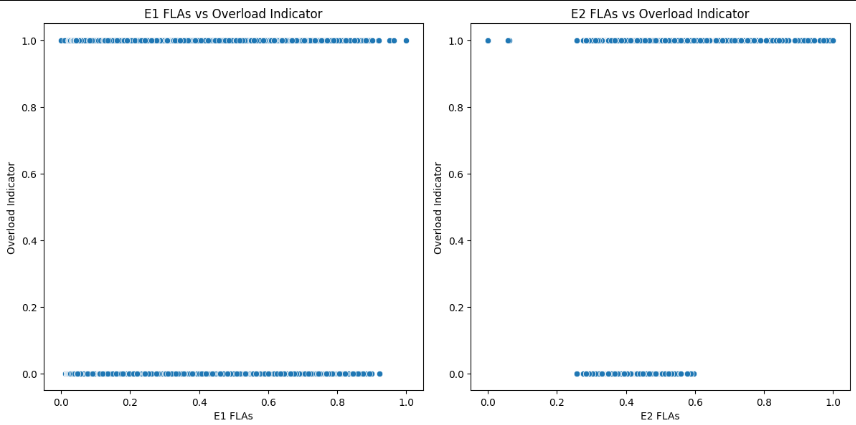
**4.Interpretation**

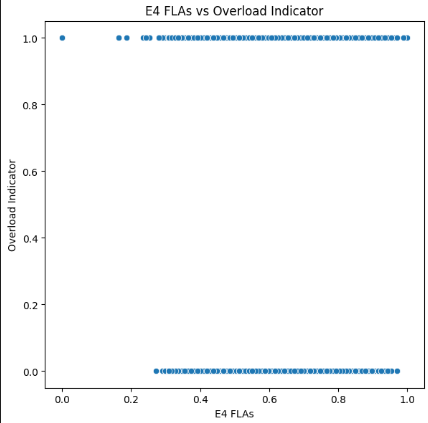
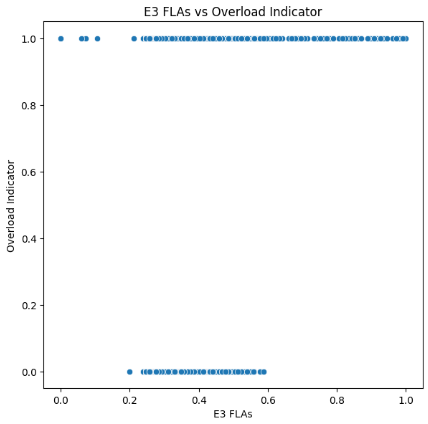
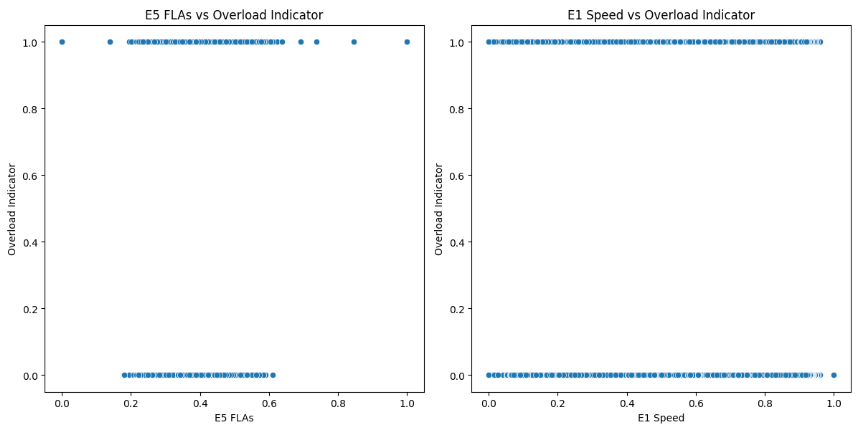
* Most observations are densely clustered, showing consistency in data.
* Variables like E4 FLAs and E5 FLAs are strongly correlated and positively contribute to PC1.
* Variables like E1 FLAs and E1 Speed are also important but in a slightly different direction.

**5.Conclusion**

* The biplot shows that the model captures key variable relationships well, with clear groupings and spread along principal components, indicating a meaningful data structure**.**

**(8)** **Analysis of Scatter Plots: Overload Indicator vs. Motor Parameters**

****

**   
** The provided image contains six scatter plots, each displaying the relationship between the binary Overload indicator (y-axis: 0 or 1) and various motor parameters (x-axis) for five different entities (E1 to E5). The parameters analysed are Full Load Amps (FLAs) for E1 to E5 and Speed for E1.

**Key Observations**  **Distinct Separation:**

In all plots, there is a clear separation between points where Overload indicator is 0 and where it is 1. This suggests that the parameters (FLAs and Speed) are strongly associated with the overload condition.

* **Overload Correlation:**

**For each entity (E1 to E5**), higher FLA values are generally associated with an Overload Indicator of 1, while lower FLA values correspond to an Overload indicator of 0. This trend is consistent across all five entities.

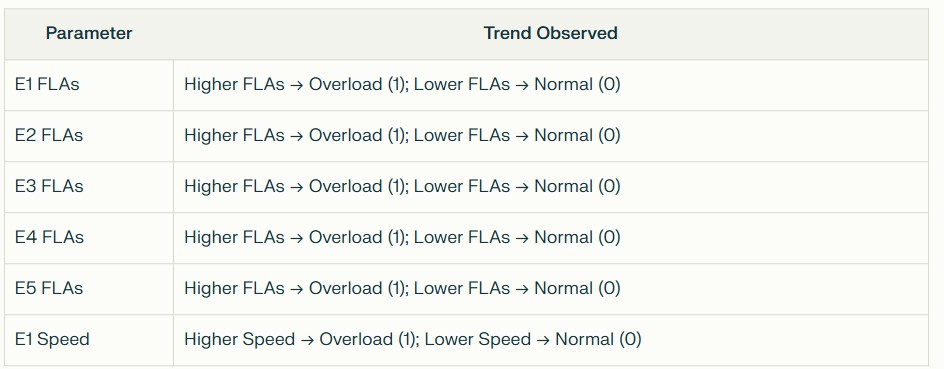
* **Speed vs. Overload:**

The E1 Speed vs. Overload indicator plot shows a similar pattern: higher speeds are more frequently associated with an overload condition (indicator = 1), while lower speeds correspond to normal operation (indicator = 0).

* **Binary Nature of Target**:

The Overload indicator is strictly binary (0 or 1), and the scatter plots show two distinct horizontal lines, reinforcing the binary classification problem.

**Summary Table**



**Conclusions:**

* There is a strong positive correlation between Full Load Amps (FLAs) and the likelihood of an overload condition for all entities (E1–E5).
* Higher motor speeds (E1 Speed) are also associated with overload conditions.
* The clear separation in the scatter plots indicates that both FLAs and Speed are good predictors for the Overload indicator.
* These visualizations suggest that monitoring FLAs and Speed can be effective in predicting and preventing overload situations.