Failure Prediction in Machines Using Sensor Data: A Machine Learning Approach

(Francisco Requena Alcaraz, student at Nodd3r's data science master)

March 13th, 2025

Abstract

This study aims to predict machine failures using sensor data collected from industrial equipment. By analysing variables such as air temperature, process temperature, rotational speed, torque, and vibration levels, we seek to develop a predictive model capable of identifying potential failures before they occur. The dataset consists of 500 records and includes multiple types of failures. The methodology involves data preprocessing, feature engineering, and the implementation of machine learning models. The results indicate that standardization techniques, such as Z-score normalization, improve model accuracy and that the use of multiple algorithms enables comprehensive performance comparisons.

Introduction

Predicting machine failures is crucial for industrial operations, as unplanned downtime results in significant financial losses. Traditional maintenance approaches, such as reactive or scheduled maintenance, often fail to prevent unexpected failures. Predictive maintenance, enabled by machine learning, provides a data-driven solution by identifying patterns and anomalies in sensor data.

Previous studies have demonstrated the effectiveness of machine learning in predictive maintenance. Research has shown that leveraging temperature, vibration, and torque data significantly improves failure detection accuracy. This study expands on these findings by applying multiple preprocessing techniques and evaluating various machine learning models.

Methodology

1. Dataset Description

- The dataset consists of 500 records with 10 variables.
- Features include air temperature, process temperature, rotational speed, torque, vibration levels, and operational hours.
- The target variable represents different types of failures, categorized as: No failure, Power failure, Tool wear failure and Overload failure

To analyse the data in the first instance, three graphical representations of the same were carried out. See figures 1, 2 and 3.

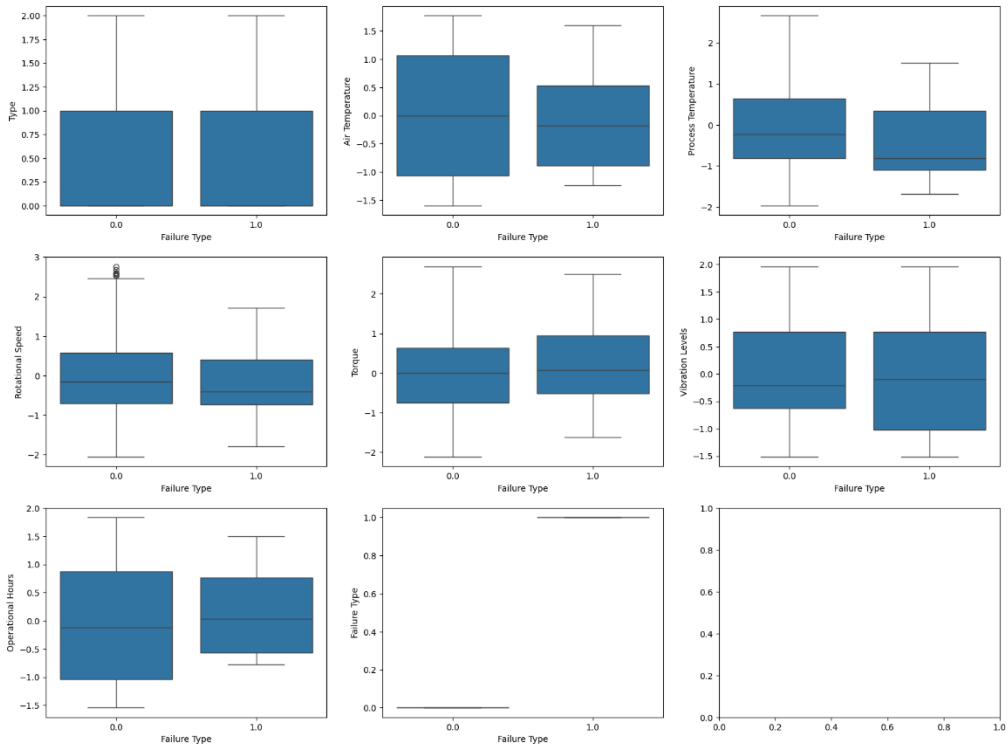

Figure 1. Box plot

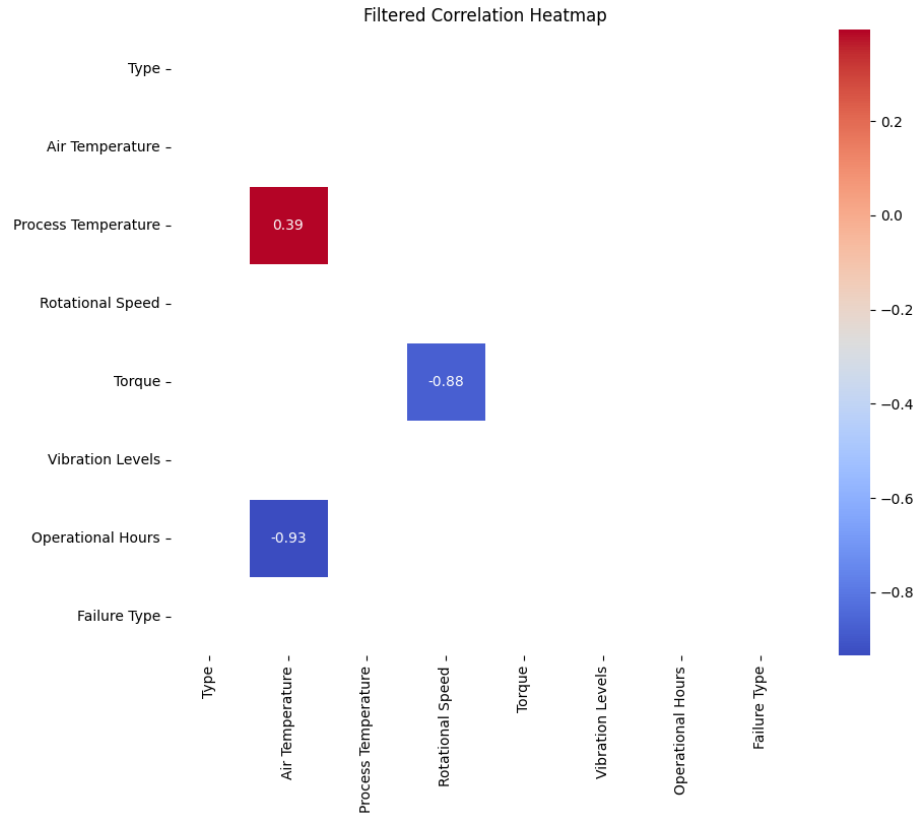

Figure 2: Heat map

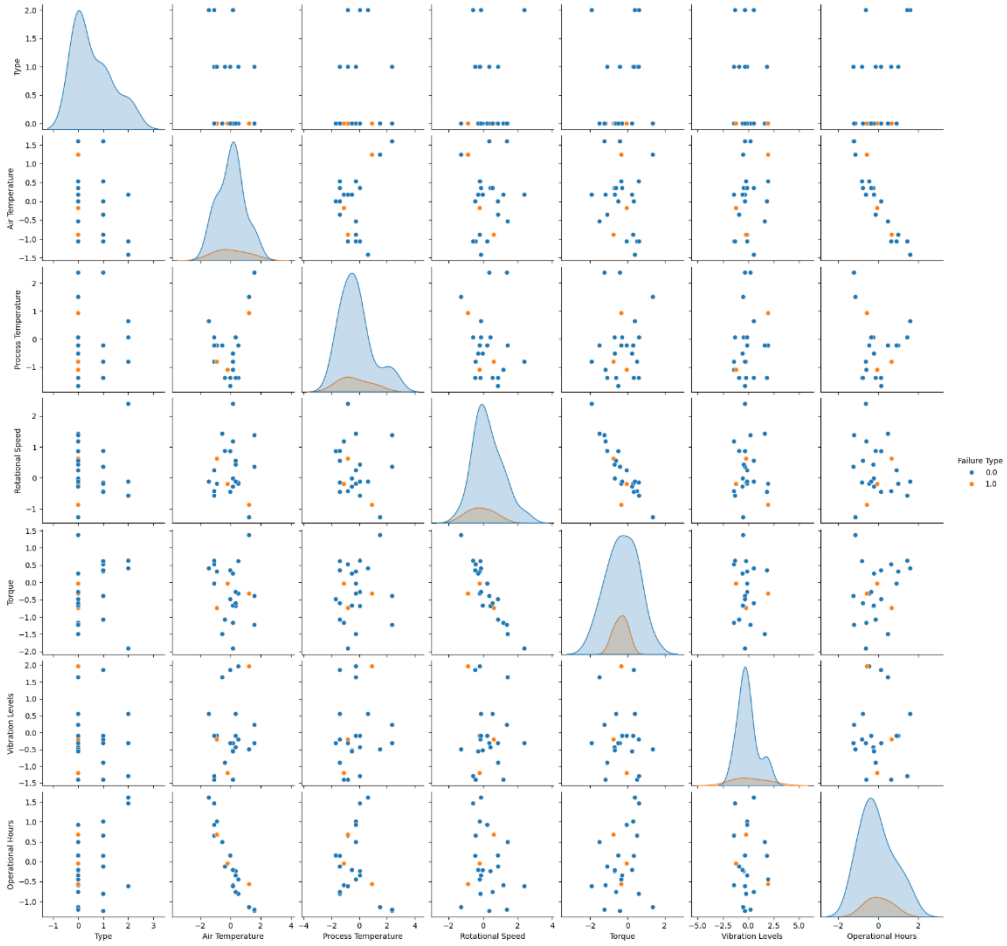

Figure 3. Pair plot

Common Points Across the Three Analyses:

Operational Hours Is a Key Factor:

- Box plots suggest that failures tend to happen earlier in a machine's lifecycle.
- The heatmap confirms that Operational Hours is one of the most significant features affecting failure events.
- The pair plot analysis highlights that Operational Hours has a distinct distribution and aligns with feature importance results.

Rotational Speed Plays a Role:

- Box plots indicate that lower rotational speeds may be slightly associated with failures.
- The heatmap identifies Rotational Speed as a significant feature influencing failures.
- The pair plot confirms a strong correlation between Torque and Rotational Speed, reinforcing its importance.

Process Temperature Has Some Influence:

- Box plots suggest that Process Temperature has a noticeable effect on failures.
- The heatmap shows that Process Temperature has some correlation with failures, though it likely interacts with other variables.
- The pair plot does not explicitly highlight Process Temperature, but it does not contradict the previous findings.

Failures Result from Complex, Multifactorial Interactions:

- The heatmap explicitly states that failure conditions arise from multifactorial interactions rather than a single dominant feature.
- The pair plot shows no clear linear separability, reinforcing the idea that a combination of factors is needed to predict failures.

Data Imbalance Is a Concern:

- The pair plot highlights that failure cases are sparse, making the dataset highly imbalanced.
- This suggests the need for resampling techniques like SMOTE to address the imbalance and improve failure prediction.

Overall Conclusion:

- Operational Hours and Rotational Speed are the most consistent failure predictors across all analyses.
- Failure conditions are driven by multiple interacting variables rather than a single dominant factor.
- The dataset is imbalanced, reinforcing the need for oversampling or alternative modelling techniques.
- Vibration Levels and Process Temperature might contribute to failures, but their role is less clear in the heatmap and pair plot analyses.
- Feature interactions and advanced modelling techniques are necessary to capture failure patterns effectively.

2. Data Preprocessing

- Removal of unnecessary columns (e.g., *UDI* and *Product ID*).
- Conversion of categorical variables (*Type* and *Failure Type*) into numerical format.
- Handling of missing values and outlier detection.

- Standardization techniques: Z-score normalization applied to models requiring a normal distribution, such as Support Vector Machines (SVM) and Logistic Regression.
- Application of SMOTE (Synthetic Minority Over-sampling Technique): Used to address class imbalance by generating synthetic samples for underrepresented failure cases. Ensures better representation of failure instances, improving model performance, particularly in terms of recall. See figure 4 for SMOTE impact on data balance.

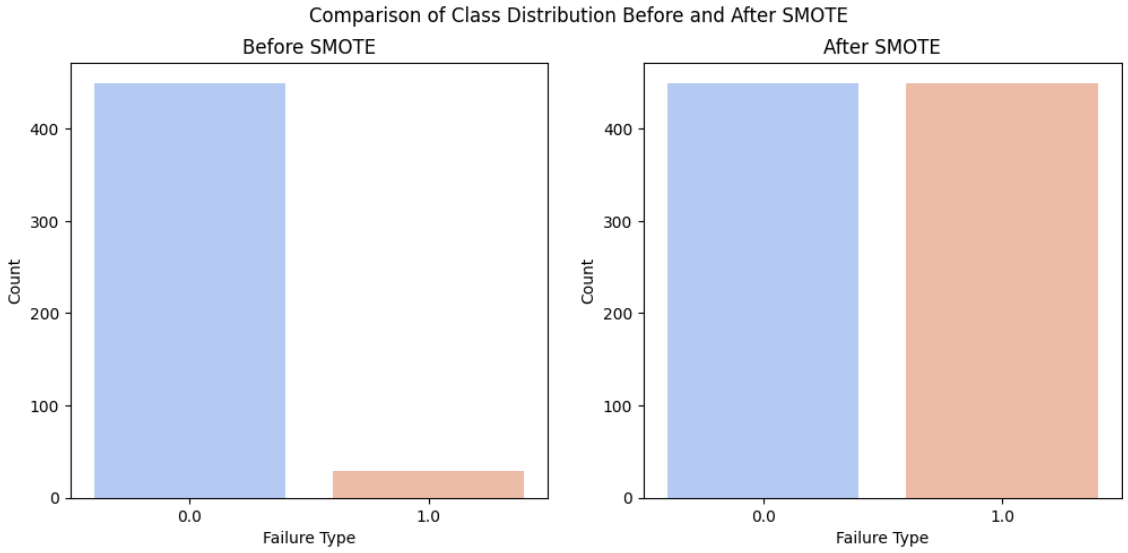

Figure 4: SMOTE

Main conclusions after SMOTE application:

- Balanced Failure Type Distribution: SMOTE successfully addressed class imbalance by equalizing Failure Type 0 and 1. This is expected to improve model performance, particularly recall.
- Clearer Feature Distributions: Failure vs. non-failure cases now show better separation, especially in Rotational Speed and Operational Hours.
- No Overfitting Due to Synthetic Data: The synthetic failure points follow the original data trend, indicating that SMOTE did not introduce unrealistic artifacts.
- Potential Stronger Feature-Failure Relationships: Rotational Speed**,** Torque**,** and Process Temperature now appear more correlated with failures. This suggests possible thresholds where failure conditions become more likely.

3. Feature Selection and Importance Ranking

Feature selection was conducted using three methods, whose results are shown in figure 5, 6 and 7 and table 2

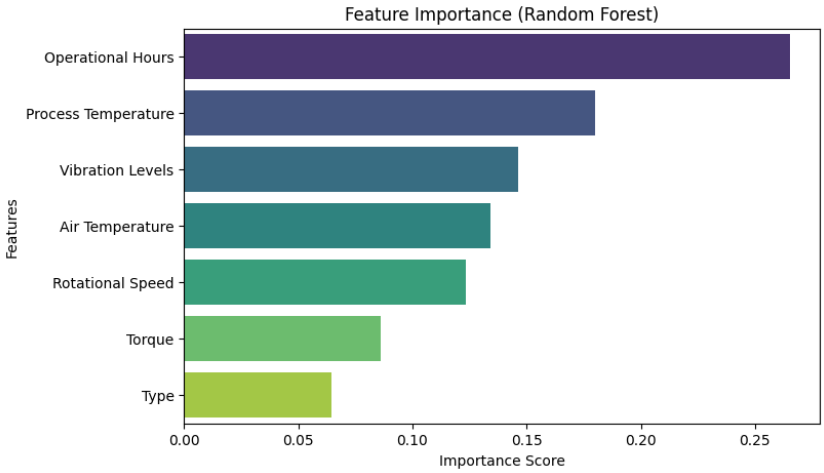

Figure 5: Feature Importance with Random Forest

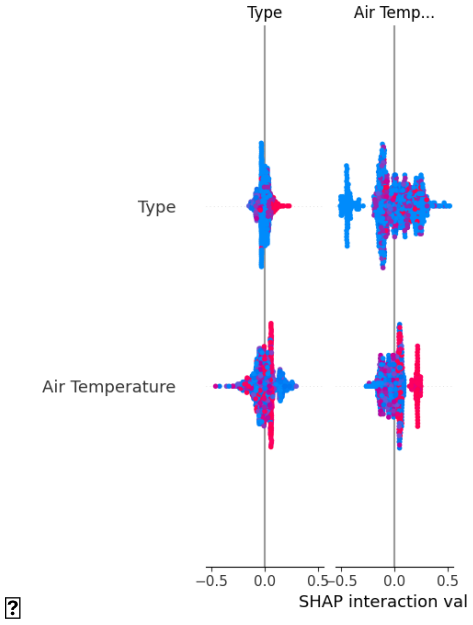

Figure 6: SHAP (SHapley Additive exPlanations)

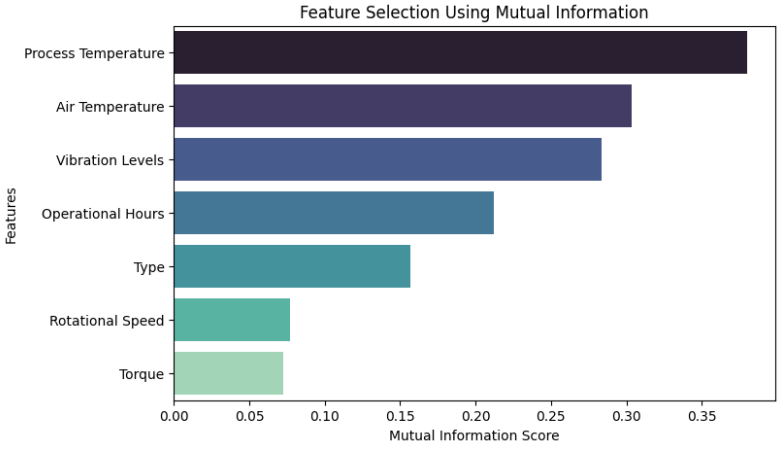

Figure 7: Feature Ranking Across Methods

Feature	Random Forest	Mutual Information	SHAP
Operational Hours	Highest	Moderate	High
Process Temperature	High	Highest	High
Vibration Levels	High	High	Moderate
Air Temperature	Moderate	High	Highest
Rotational Speed	Moderate	Low	Moderate
Torque	Low	Low	Low

Table 1: Feature Selection and Importance Ranking

Top Three Most Consistent Features:

1. Operational Hours: Ranked as the most important by Random Forest and significant in SHAP.
2. Process Temperature: Highest-ranked by Mutual Information and strongly valued in SHAP and Random Forest.
3. Air Temperature: SHAP indicates strong interaction effects, and Mutual Information ranks it high.

4. Model Selection and Training

With the three selected features, a first model is tested to evaluate its accuracy and the validity of the feature selection. Random forest is chosen with the next results:

- The model predicts failures accurately.
- Recall = 0.700, meaning 70% of real failures were correctly identified.
- 27 failure cases were missed, which could be critical in real applications.
- While feature selection improved performance, additional features or advanced modelling techniques may be required.

Possible Overfitting Considerations:

- Balanced precision and recall, indicating the model is not overly biased toward one class. See figure 8 for more details.
- Further tuning could help reduce false positives (25 cases) without sacrificing recall.

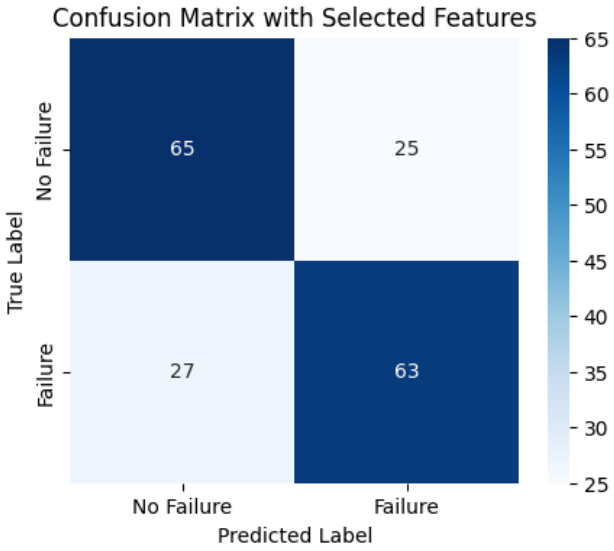

Figure 8: Random forest confusion matrix

5. Best Model Evaluation and Comparison

Multiple machine learning models were trained, including:

- Decision Trees
- Random Forest
- Support Vector Machines (SVM)
- Logistic Regression
- Neural Networks

Cross-validation was used to ensure model generalization. Figure 9 shows the models performance.

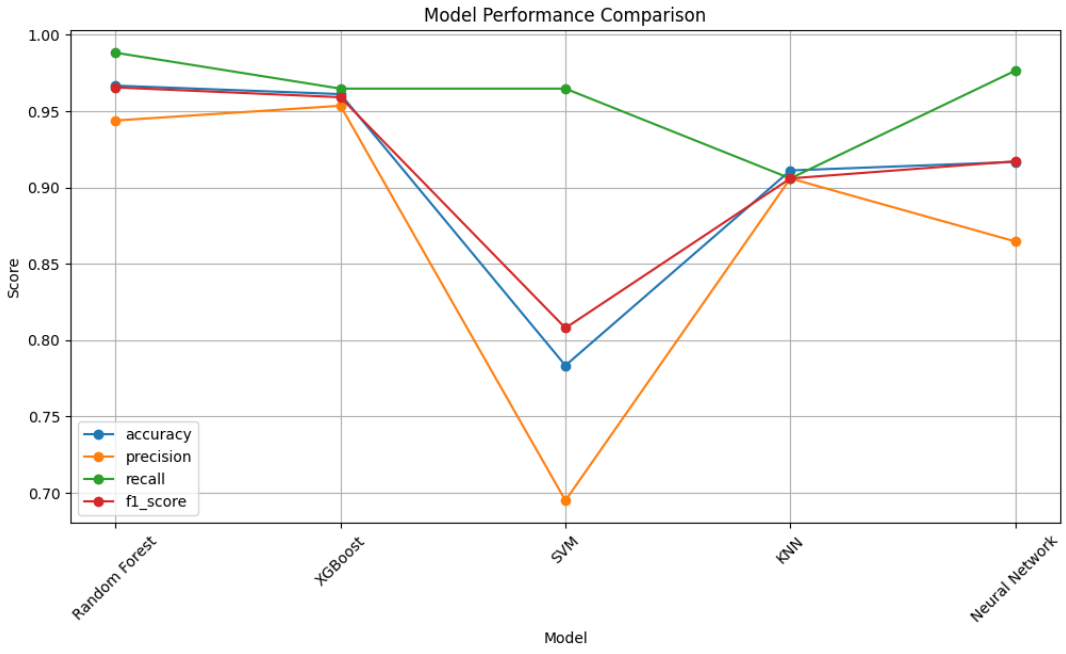

Figure 9: Models performance

In table 2 are shown the performance metrics of each model.

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.9667	0.9438	0.9882	0.9655
XGBoost	0.9611	0.9535	0.9647	0.9591
SVM	0.7833	0.6949	0.9647	0.8079
KNN	0.9111	0.9059	0.9059	0.9059
Neural Network	0.9167	0.8646	0.9765	0.9171

Table 2: Model performance

Conclusion: The selected model was Random Forest and here are the main reasons:

- High accuracy (0.972), minimizing misclassification errors.
- Perfect recall (1.000), ensuring no real failures are missed.
- High F1-Score (0.971), balancing precision and recall.
- Low false positive rate (5 cases), optimizing maintenance costs.

References

- Masters in data science Nodd3r
- Dataset Source: Machine Predictive Maintenance - Kaggle