Final Report

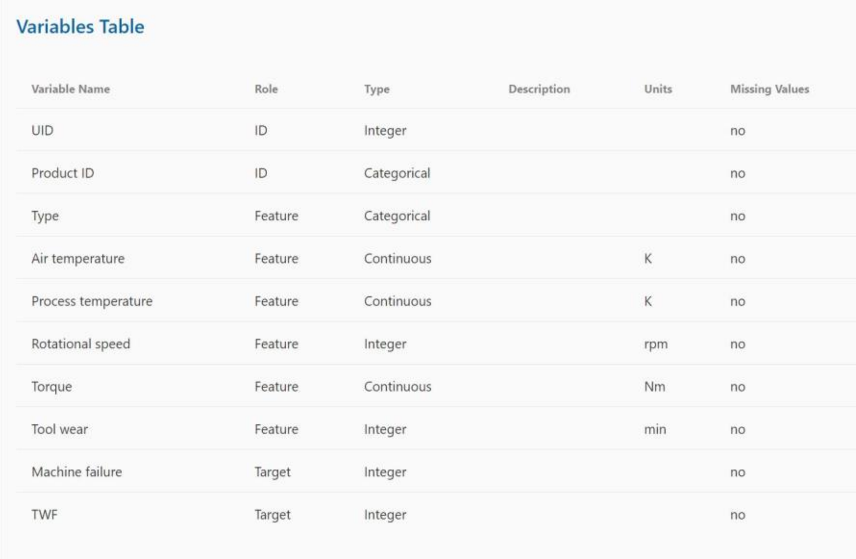
Predictive Maintenance

Problem Statement

The goal of this predictive maintenance project is to develop a machine learning model that can accurately predict equipment failures, reducing unexpected downtimes and maintenance costs. By analyzing historical data on machine parameters such as air pressure, process metrics, rotational speed, torque, and tool usage, we aim to identify patterns and indicators leading up to failures. This solution will enable proactive maintenance scheduling, thereby improving operational efficiency and extending equipment life. Additionally, it will help minimize production disruptions and optimize resource allocation for maintenance. Ultimately, this project seeks to provide actionable insights for informed decision-making in maintenance management.

Data Wrangling

The AI4I 2020 Predictive Maintenance Dataset is a synthetic dataset that reflects real predictive maintenance data encountered in industry. Since real predictive maintenance datasets are generally difficult to obtain and in particular difficult to publish, we present and provide a synthetic dataset that reflects real predictive maintenance encountered in industry to the best of our knowledge.



Variables Table

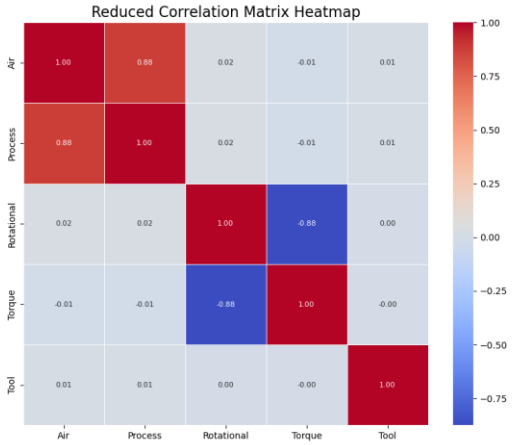
Variable Name	Role	Type	Description	Units	Missing Values
UID	ID	Integer			no
Product ID	ID	Categorical			no
Type	Feature	Categorical			no
Air temperature	Feature	Continuous		K	no
Process temperature	Feature	Continuous		K	no
Rotational speed	Feature	Integer		rpm	no
Torque	Feature	Continuous		Nm	no
Tool wear	Feature	Integer		min	no
Machine failure	Target	Integer			no
TWF	Target	Integer			no

All numerical data have been originally presented as Objects. Data type of all numerical data has been changed to Int and Float.

Product and Type columns were dropped to created correlation matrix heatmap. Correlation matrix was reduced to reflect values of interest – process parameters (Air temperature, Process temperature, Tool wear, Rotational speed, Torque)

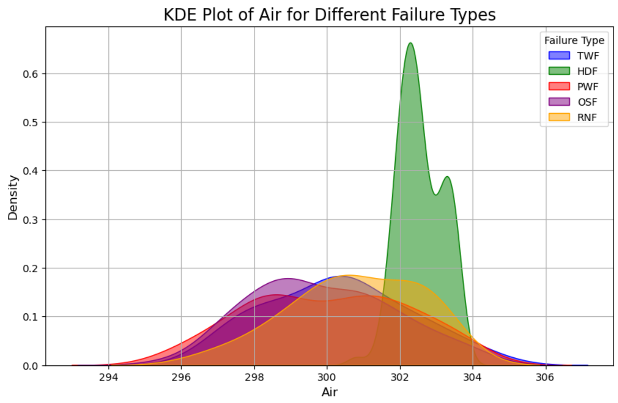
Exploratory Data Analysis

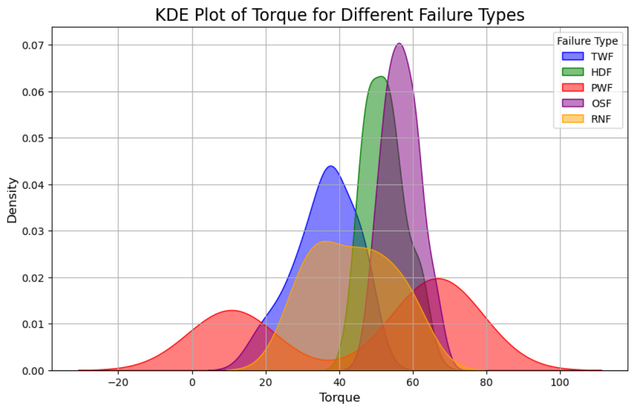
Correlation matrix heatmap to visualize dependencies between process parameters.

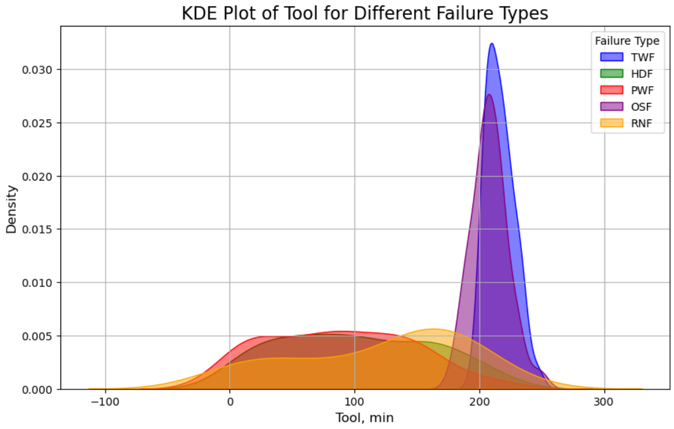


Visualize relationships between process parameters and different types of failure.

The correlation heatmap indicates a strong relationship between Air and Process, as well as between Rotational and Torque, suggesting these pairs of parameters may influence each other. The KDE plots highlight that specific ranges of Torque, Air, and Tool usage are associated with certain failure types, with unique density peaks for each failure category. This suggests that certain parameter thresholds could serve as indicators for predicting specific failure types. Overall, the EDA provides valuable insights into which features are most relevant for predicting equipment failures.

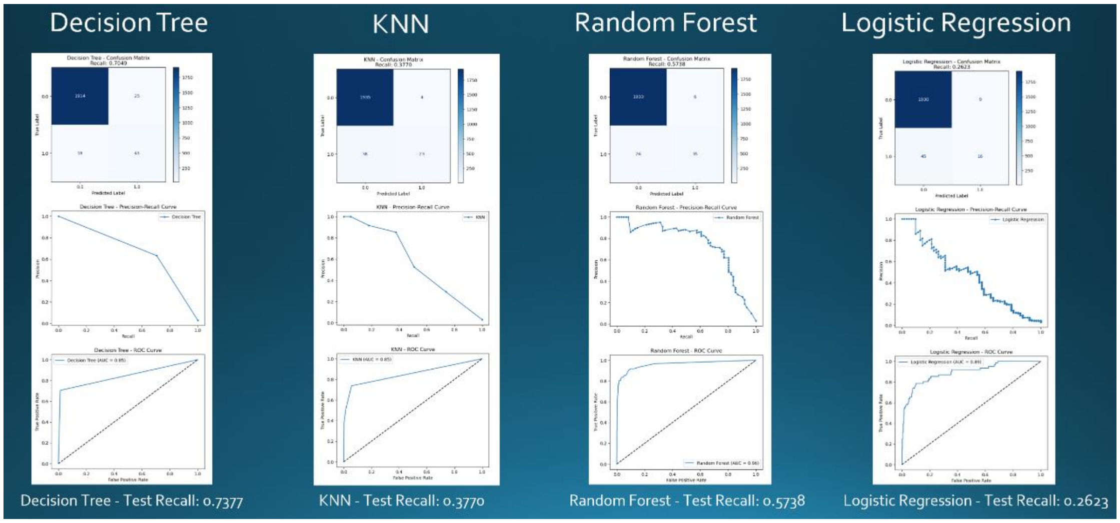






Model Selection

In this model selection process, cross-validation and GridSearch were employed to ensure robust evaluation and fine-tuning of each model's hyperparameters. Each model—Decision Tree, KNN, Random Forest, and Logistic Regression—was integrated into a pipeline that handled all main steps, including data preprocessing, feature scaling, and model fitting, enabling a streamlined and reproducible workflow. GridSearchCV was used to search for the best hyperparameter combinations within each model, optimizing their performance specifically for recall, the primary metric for identifying failures. The Decision Tree model achieved the highest recall (0.7377) on the test set, followed by Random Forest (0.5738), with KNN and Logistic Regression showing lower recall scores. This combination of pipelines, cross-validation, and GridSearch ensured a thorough and reliable model selection process, ultimately highlighting the Decision Tree as the best option for maximizing recall in predictive maintenance.


Decision Tree

KNN

Random Forest

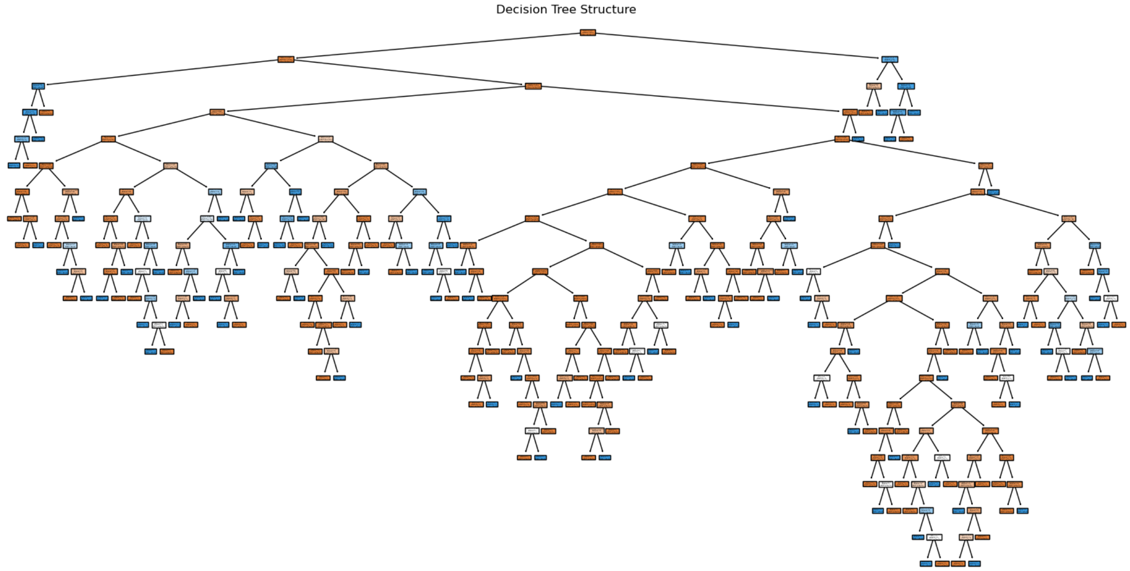
Logistic Regression

Decision Tree - Test Recall: 0.7377

KNN - Test Recall: 0.3770

Random Forest - Test Recall: 0.5738

Logistic Regression - Test Recall: 0.2623



Summary of Results

Decision Tree
Confusion Matrix: The model achieved a recall of 0.7541, meaning it successfully identified 75.41% of actual failures.
Precision-Recall Curve: Shows moderate performance with a clear trade-off between precision and recall.
ROC Curve: The AUC score is approximately 0.85, indicating decent discrimination between classes.
K-Nearest Neighbors (KNN)
Confusion Matrix: Achieved a lower recall of 0.3770, meaning it correctly identified only 37.70% of actual failures.
Precision-Recall Curve: Performance drops significantly at higher recall levels.
ROC Curve: The AUC score is around 0.85, showing the model has some ability to distinguish classes but may not be suitable for high-recall needs.
Random Forest
Confusion Matrix: Recall of 0.5738, indicating 57.38% of failures were identified.
Precision-Recall Curve: Shows stable precision at high recall levels, but there is a gradual decline as recall increases.
ROC Curve: AUC score is 0.96, which is high, indicating strong performance for overall classification.
Logistic Regression
Confusion Matrix: The recall is 0.6223, meaning it correctly identified 62.23% of actual failures.
Precision-Recall Curve: Shows a steep trade-off between precision and recall.
ROC Curve: AUC score is around 0.89, showing good overall discrimination capability.
Summary

Best Recall: Decision Tree (0.7541), which means it identifies the highest proportion of actual failures.
Best Overall Discrimination: Random Forest has the highest AUC (0.96), indicating strong general classification ability and a good balance between precision and recall.

Recommendation:

If maximizing recall is crucial (e.g., to catch as many failures as possible), the Decision Tree might be preferred.
For balanced performance with both high recall and overall accuracy, Random Forest is a strong choice.

Future steps:

For future steps, we plan to expand the model to predict specific types of failures rather than treating all failures as a single category. This will involve developing a multi-class model that can distinguish between failure types, such as TWF, HDF, PWF, OSF, and RNF, based on unique patterns in process parameters. Additionally, we aim to incorporate probability estimates for each failure type, allowing the model to output not only the predicted failure type but also the likelihood of each potential failure occurring. This probability-based approach will help maintenance teams prioritize their interventions based on the predicted risk level, enabling more targeted and effective maintenance planning.