Data Driven Operational Excellence

Data Science in Business

MDDFDBC1A.6

Predictive Maintenance in Dairy Production: A Data-Driven Approach

Course: Driven Decision Making in Business

Course code: MDDF DBC1A.6

Institution: Hogeschool Arnhem Nijmegen

Module: Data science for business – the CRISP model for data mining

Submitted By:
Stan van Bon - 1633267

Date: October 24, 2023

Tutors:
Hugo Gobato Souto & Mohsen Ghanadzadeh

Table of Contents

.... 1
Predictive Maintenance in Dairy Production: 1
A Data-Driven Approach 1
1. Problem Definition and Objectives 3
1.1 Introduction 3
1.2 Background of FrieslandCampina 3
1.3 Problem Definition 3
1.4 Project Objectives 3
2. Assessment of the Situation 4
2.1 Inventory of Resources 4
2.2 Constraints and Limitations 4
2.3 Assumptions 4
2.4 Data Analysis Goal 4
2.5 Incorporating Corporate Standards 5
2.6 Project Schedule 5
3. Data Collection 5
3.1 Data Collection Methods 5
3.2 Data Sources 6
3.3 Data Collection plan 7
3.4 Assess quality of Data Sources using the RADAR framework 8
3.5 Limitations of data gathering methods 8
3.6 Limitations of Data Sources 9
4. Data Driven Analysis 10
4.1 Descriptive Analysis 10
4.2 Correlation Analysis 11
4.3 Categorical & Target variable Analysis 12
4.4 Outlier Analysis 13
4.5 Distribution of failure type 14
4.6 Data Preprocessing 16
5. Data Modeling 17
5.1 Argumentation for Two Predictive Models 17
5.2 Predicting Machine Failure – Model Results 18
5.3 Selecting modelling criteria. 21
5.4 Evaluation of predicting machine failure Models 21
5.5 Predicting Machine failure type Results 22
5.6 Machine Failure type Evaluation 25
5.7 Selected Models 26
6. Business Insights 27
6.1 Data-Driven Decision-Making Framework 27
6.3 Relevant KPI's 29
6.4 Data-Driven Recommendations 30

1. Problem Definition and Objectives

1.1 Introduction

Operational efficiency, a cornerstone upon which FrieslandCampina has structured its formidable production network, is being challenged due to machinery downtimes and inadequate maintenance schedules. These challenges not only affect operations but also the financial equilibrium of the company. This chapter will outline these operational nuances and set clear objectives to address them.

1.2 Background of FrieslandCampina

FrieslandCampina oversees a vast array of operations, from procuring raw materials to the intricate processes of dairy production. With a significant global presence and market share, the company is renowned for its quality and reliability. However, issues like machinery downtimes and maintenance inefficiencies have become pressing concerns, highlighting the need for a data-driven approach to mitigate potential operational setbacks. (FrieslandCampina, 2022, Pp. 6-9)

1.3 Problem Definition

Machinery downtimes and inconsistent maintenance schedules have emerged as silent disruptors within FrieslandCampina's operations. These not only cause operational interruptions but also pose significant financial burdens. A predictive maintenance framework, infused with data analytics and actionable insights, can serve as a remedy, transforming reactive machinery management into proactive, strategic operations.

Key Aspects of the Problem:

- **Machinery Downtime**: Unexpected machinery disruptions that lead to production halts and increased costs.
- **Maintenance Scheduling**: The lack of a strategic, proactive maintenance approach results in unanticipated downtimes and inefficient resource allocation.
- **Holistic Machinery Health**: A subtle decline in machinery performance necessitates vigilant monitoring to prevent unnoticed operational and financial drains.
- **Data Quality**: It's imperative that the data used for predictive analytics is of high quality to ensure reliable predictive insights.

1.4 Project Objectives

Primary Objective:

- Strengthen operational efficiency by reducing unplanned machinery downtimes, leveraging advanced predictive maintenance analytics.

Secondary Objectives:

- Minimize unexpected machinery disruptions and the associated financial and resource burdens.
- Implement a proactive maintenance strategy to prevent unexpected downtimes and optimize resource allocation.
- Monitor machinery health to anticipate and address performance declines before they become significant issues.

(Chapman, 2000, p. 16)

2. Assessment of the Situation

Understanding the operational context of FrieslandCampina's predictive maintenance project requires a thorough evaluation of available resources, constraints, and assumptions. This chapter delves into these components, laying the groundwork for the project's trajectory.

2.1 Inventory of Resources

2.1.1 Personnel

Business and Data Experts: A team versed in FrieslandCampina's operational dynamics, offering insights into business processes and ensuring data integrity (FrieslandCampina, 2021).

Technical Support: Ensures optimal functionality of hardware and software resources.

Process Owner: Holds accountability for using validated software and aligning the business work process with the computerized system (FrieslandCampina, 2021).

2.1.2 Data:

- **Operational Data**: Reflects machinery performance, maintenance schedules, and related downtimes.
- **External Data**: Includes market trends, supplier data, and more.

2.1.3 Infrastructure and Tools:

- **Hardware Platforms**: Robust computing infrastructure for data analysis, model development, and implementation.
- **Cloud Computing**: Supports remote collaboration, data storage, and resource scaling.
- **Analytical Software**: Software tools for data visualization, analysis, and reporting.

2.2 Constraints and Limitations

- **Data Quality and Accessibility**: The efficacy of the predictive model relies on the quality and consistent access to relevant data.
- **Computational Capabilities**: The existing hardware and software must be able to support the predictive model's requirements.
- **Integration**: The predictive model should integrate smoothly with existing operational systems (FrieslandCampina, 2021).
- **Supply Chain Considerations**: It's crucial to maintain the stability of downstream operations and minimize disruptions along the supply chain.

2.3 Assumptions

Aligned with the Foqus FS&Q standards, it's assumed that data, stakeholder support, and resource availability are validated and documented for relevance and reliability throughout the project's duration (FrieslandCampina, 2021).

2.4 Data Analysis Goal

With the resources at hand and considering constraints and assumptions, the goal crystallizes: to develop a predictive maintenance model that's accurate, timely, and relevant. The model should comply with Foqus standards and ensure insights are actionable within the operational context and adhere to validated protocols (FrieslandCampina, 2021).

2.5 Incorporating Corporate Standards

2.5.1 Adherence to Computerized System Validation:

- **Quality Assurance**: Ensures appropriate procedures are followed, controlling risks to product quality in manufacturing systems (FrieslandCampina, 2021).

2.5.2 Management of Change:

- FrieslandCampina will implement a Change Management procedure, which includes establishing Local and Business Group Change Management Boards, verifying implemented changes, and documenting and tracking all changes (Foqus FS&Q Management of Change standard, 2022).

2.6 Project Schedule

The project adopts a systematic, data-driven approach for predictive maintenance model development, implementation, and evaluation. Here's a concise breakdown:

- **Phase 1: Data Collection and Integration** (1 month): Setting up IoT devices for data collection, centralizing a data repository, and verifying data integrity.
- **Phase 2: Analytics** (1 month): Using descriptive analytics to spot inefficiencies and diagnostic analytics to comprehend their causes.
- **Phase 3: Predictive Maintenance Modeling** (2 months): Building and validating predictive models using historical data and machine learning techniques.
- **Phase 4: Monitoring System Implementation** (1 month): Deploying real-time monitoring with alerts and training maintenance teams on its usage.
- **Phase 5: Maintenance Optimization** (1 month): Enhancing maintenance schedules based on predictions and assessing their impact on efficiency.
- **Phase 6: Performance Benchmarking** (1 month): Setting operational efficiency benchmarks and continuously monitoring actual performance.
- **Phase 7: Cost-Benefit Analysis** (1 month): Evaluating the financial outcomes of the solutions and documenting the project's ROI.
- **Phase 8: Implementation and Adjustment** (2 months): Executing data-driven solutions and making necessary adjustments for desired outcomes.
- **Phase 9: Training and Change Management** (1 month): Training stakeholders and integrating change management for smoother adoption of innovations.
- **Phase 10: Project Evaluation and Reporting** (1 month): Assessing the project's success against goals and creating a detailed project report.

(Chapman, 2000, p. 19)

3. Data Collection

3.1 Data Collection Methods

- **Document Analysis:** Extract data from documents, such as manuals, standards, and reports, to gain insights into machinery operation, maintenance history, and compliance standards.
- **Surveys and Questionnaires:** Use feedback from machinery operators, maintenance staff, and suppliers to understand operational nuances.
- **Interviews:** Engage directly with stakeholders to gather nuanced, qualitative insights on machinery performance and operational challenges.

- **Observations:** Directly observe processes and machinery operations to identify inefficiencies or areas of risk.
- **Sensor Data Collection:** Implement sensors to gather real-time machinery performance data, such as temperature or vibration metrics.
- **Database Mining:** Analyze existing databases to understand historical machinery performance and maintenance trends.
- **Audit Reports:** Evaluate previous audits to understand past challenges, compliance issues, or areas of improvement.
- **Experimentation:** Conduct tests or pilots to gather data and validate hypotheses.
- **Focus Groups:** Engage diverse stakeholder groups in discussions to gather varied insights on machinery performance and operational bottlenecks.
- **IoT (Internet of Things):** Use IoT devices for continuous data collection, providing insights into machinery operations and potential predictive maintenance alerts.
- **Machine Logs:** Analyze machine logs to gather historical data on machinery operation, downtimes, and maintenance.
- **Benchmark Studies:** Review industry benchmarks to understand how current machinery and processes compare to industry standards.
- **Supplier Data:** Gather data from machinery and parts suppliers to understand equipment specifications, potential issues, or recalls.
- **Quality Records:** Review quality records to identify production quality trends potentially linked to machinery performance.
- **Regulatory and Compliance Data:** Assess regulatory documents to ensure machinery and processes comply with relevant standards.

3.2 Data Sources

To understand the intricacies of FrieslandCampina's operations and inform our predictive maintenance model, various data sources have been identified. These sources span across internal company documents, public databases, industry reports, and compliance standards.

3.2.1 Internal Company Documents:

Heat Preservation Standard (CORP-AMER-STA-00078):

- Operational Data: Understanding of various heat treatment processes and their technical and operational specifications.
- Performance Data: Knowledge on the efficiency and potential pitfalls or challenges in the heat treatment processes, which might impact machinery performance.
- Compliance Data: Adherence to the heat treatment process standards, ensuring the consistent and safe production of food products.

CIP Validation Standard (CORP-AMER-STA-00014):

- Maintenance Data: Comprehensive data on the Cleaning-in-Place (CIP) method, which could be pivotal in maintaining machinery and ensuring it's clean without needing to dismantle it.
- Failure History: By understanding the parameters and the different validation steps, potential reasons for failure in the cleaning process (if any) can be analyzed and documented.
- Operational Impact: Insights into how CIP might impact the operational efficiency and safety of the machinery and production process.

HACCP Standard (CORP-AMER-STA-00007):

Insights: Data related to the identification and control of food safety hazards during production and aspects concerning product quality assurance.

Environmental Monitoring in Powder Plants Standard (CORP-AMER-STA-00020):
Insights: Gleaning insights regarding the environmental conditions and monitoring practices within powder manufacturing environments, which might directly influence machine performance and product quality.

Supplier Quality Management Standard (CORP-AMER-STA-00035):
Insights: Comprehensive data concerning supplier management, Information concerning material risks and supplier risks, vital for evaluating and mitigating potential future issues.

Computerized System Validation Standard (CORP-AMER-STA-00040):
Insights: Information on the qualification and validation of computerized systems, and their interactions and impacts on machinery and operations.

Compliance Standard (CORP-AMER-STA-00001):
Insights: Structural and regulatory requirements for the Foqus FS&Q system, impacting all operational areas.

3.2.2 External Data Sources

Public Databases and Repositories:
Insights: Provides diverse datasets potentially useful for machinery performance, failure prediction, and maintenance requirement analysis. ("Operational Performance Improvement in Industrial Companies," 2018)\

Industry Reports and Whitepapers
Insights: Operational and performance data, common issues, maintenance practices, and potential solutions. ("UCI Machine Learning Repository," n.d.)

3.3 Data Collection plan

Phase	Activities	Start Date	Duration	Assigned to
Phase 1: Preliminary Data Identification	Identify and categorize available data sources (internal and external).	2023-11-01	7 days	Data Identification Team
Phase 2: Detailed Data Exploration	Explore and review each identified data source for relevance and potential insights.	2023-11-08	14 days	Data Analysis Team
Phase 3: Data Retrieval and Initial Processing	Retrieve data from identified sources. Conduct initial data preprocessing like data cleaning and handling of missing values.	2023-11-22	21 days	Data Engineering Team
Phase 4: Data Validation and Quality Assurance	Validate the collected data for accuracy, consistency, and reliability.	2023-12-13	14 days	Data Quality Assurance Team
Phase 5: Data Integration and Storage	Integrate datasets and store securely, ensuring data protection compliance and easy retrieval.	2023-12-27	10 days	Data Storage Team
Phase 6: In-depth Data Analysis	Analyze data to derive preliminary insights, identify trends, patterns, and anomalies.	2024-01-06	21 days	Data Science Team
Phase 7: Data Insights Validation	Validate insights obtained from data analysis with domain experts and stakeholders.	2024-01-27	14 days	Domain Experts and Stakeholder Team
Phase 8: Final Data Documentation and Reporting	Document the entire data collection and analysis process. Prepare final reports and presentations.	2024-02-10	14 days	Reporting and Documentation Team

(Chapman, 2000, pp. 20-22)

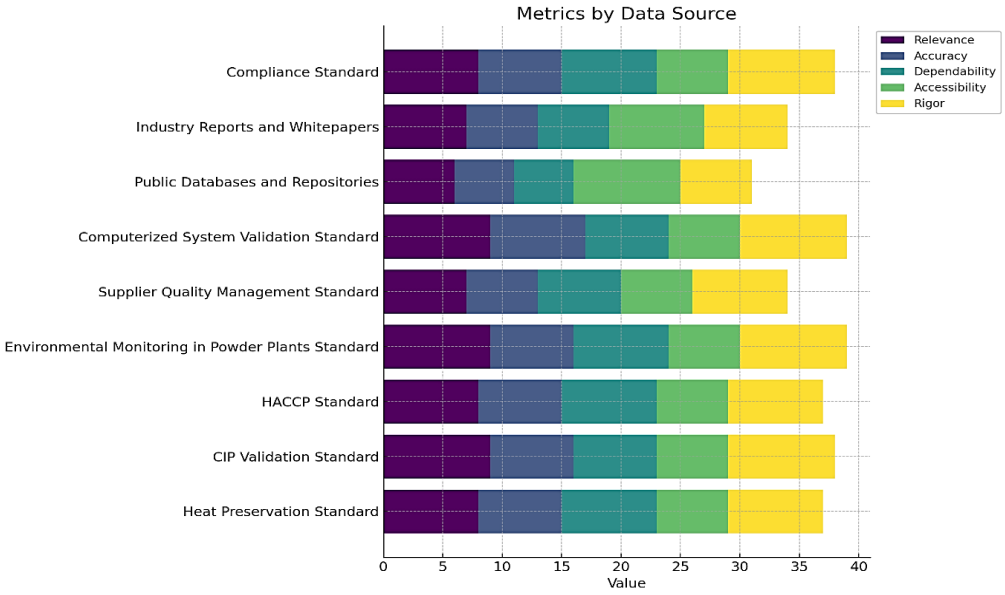
3.3.1 Additional Notes:

- The above plan might need adjustments based on the actual data availability, complexity, and organizational timelines.

- Effective collaboration and constant communication among all assigned teams are essential to ensure the seamless progression of each phase.
- Regular updates and checkpoint meetings should be conducted to ensure the plan is on track and to make necessary adjustments.

3.4 Assess quality of Data Sources using the RADAR framework.

Data Source	Relevance	Accuracy	Dependability	Accessibility	Rigor
Heat Preservation Standard	8	7	8	6	8
CIP Validation Standard	9	7	7	6	9
HACCP Standard	8	7	8	6	8
Environmental Monitoring in Powder Plants Standard	9	7	8	6	9
Supplier Quality Management Standard	7	6	7	6	8
Computerized System Validation Standard	9	8	7	6	9
Public Databases and Repositories	6	5	5	9	6
Industry Reports and Whitepapers	7	6	6	8	7
Compliance Standard	8	7	8	6	9



Mandalios (2013)

3.5 Limitations of data gathering methods

Given the project context and objectives to develop a predictive maintenance model through data collection and analysis, let's explore potential limitations of different data gathering methods:

1. Manual Data Collection

- **Time-Consuming**: Manual collection can be labor-intensive and slow.

- **Error-Prone**: Human error can introduce inaccuracies in the data.
- **Scalability Issues**: Manual methods may not be suitable for large-scale data collection.

2. Automated Data Collection (Sensors, IoT devices)

- **Initial Cost**: High upfront costs for sensor installation and system integration.
- **Technology Compatibility**: Potential issues with integrating new tech with existing systems.
- **Data Overload**: Managing vast amounts of data generated in real-time can be challenging.

3. Third-Party Data Providers

- **Data Relevance**: Provided data may not be entirely relevant or accurate for specific predictive maintenance objectives.
- **Dependence**: Relying on external sources may pose risks if the provider changes data access conditions or experiences disruptions.
- **Cost**: Ongoing costs for accessing third-party data.

4. Surveys and Field Data Collection

- **Subjectivity**: Data may be influenced by respondent bias or perception.
- **Logistical Challenges**: Coordinating field data collection can be logistically complex and potentially costly.
- **Data Consistency**: Ensuring consistent data quality and format may be challenging.

5. Database Mining

- **Data Relevance**: Existing data may not precisely match the new project's requirements.
- **Data Accuracy**: If historical data is not well-maintained or accurate, it can mislead the project.
- **Outdated Information**: Stored data may become obsolete or not reflect current conditions.

3.6 Limitations of Data Sources

1. **Heat Preservation Standard**
- Scope: May focus solely on heat preservation, missing broader machinery operational insights.
- Relevancy: The standard might become outdated with the introduction of new technologies.

2. **CIP Validation Standard**
- Complexity: Requires specialized knowledge for accurate interpretation.
- Granularity: Might lack in-depth data on specific machinery or component failures.

3. **HACCP Standard**
- Focus: Centered around food safety, it may not delve into machinery operational and performance specifics.
- Generalization: More aligned with overarching food safety principles than detailed machinery data.

4. **Environmental Monitoring in Powder Plants Standard**
- Scope: Limited to environmental factors, potentially overlooking machinery specifics.
- Geographical Variability: Environmental conditions can vary across locations, affecting general applicability.

5. **Supplier Quality Management Standard**
- External Dependence: Relies on supplier transparency and honesty.
- Dynamics: Supplier quality can shift over time, affecting predictability.

6. **Computerized System Validation Standard**
- Technicality: Demands specific expertise for full comprehension.
- Volatility: Computerized systems might undergo frequent updates, affecting the standard's applicability.

7. **Public Databases and Repositories**
- Generalizability: Datasets might not directly align with the project's unique context.
- Data Quality: Public data can vary in reliability and accuracy.

8. **Industry Reports and Whitepapers**
- Potential Bias: Manufacturer-produced whitepapers might favor their solutions or products.
- Relevance: Data could be industry-generic, not tailored to the project's specific needs.

9. **Compliance Standard**
- Regulatory Lag: Standards might not reflect the latest industry practices or technological advancements.
- Jurisdictional Variability: Compliance requirements can differ based on location or jurisdiction, limiting universal applicability.

(FrieslandCampina, 2023)

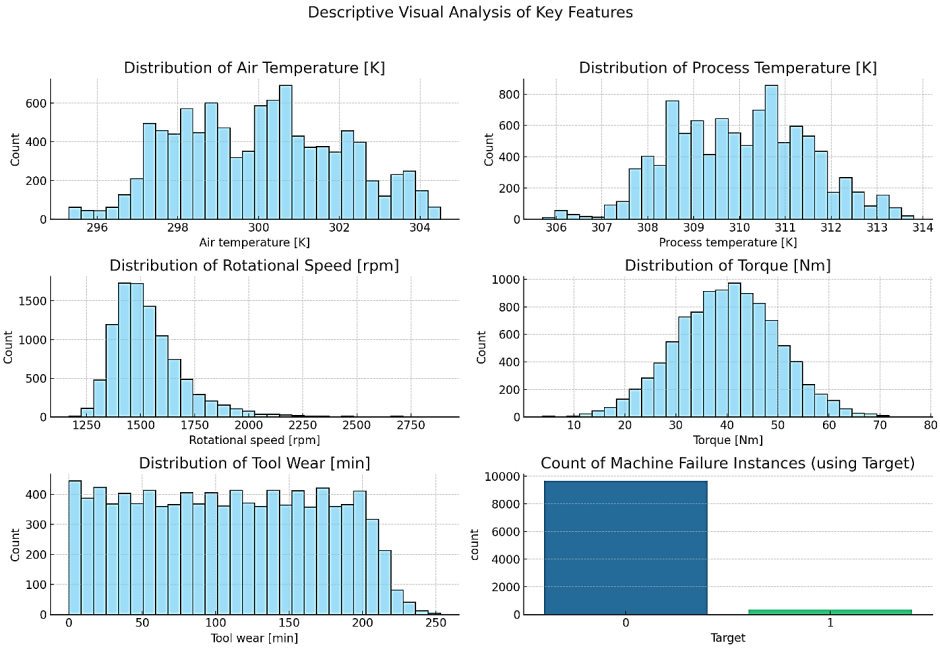
4. Data Driven Analysis

4.1 Descriptive Analysis

The dataset contains 10,000 data points with several features related to machine operation and condition:

- **UDI:** Unique identifier, ranging from 1 to 10,000.
- **Air temperature [K]:** Mean of approx. 300 K, normally distributed.
- **Process temperature [K]:** Mean of approx. 310 K, normally distributed.
- **Rotational speed [rpm]**: Continuous variable with a mean of approximately 1539 rpm, and it displays a right-skewed distribution.
- **Torque [Nm]**: Continuous variable with a mean of approximately 40 Nm, and it shows a somewhat normal distribution but with noticeable spikes.
- **Tool wear [min]**: Continuous variable with a mean of approximately 108 minutes. It seems to have a somewhat uniform distribution with certain spikes, possibly indicating common lifetimes for tools.

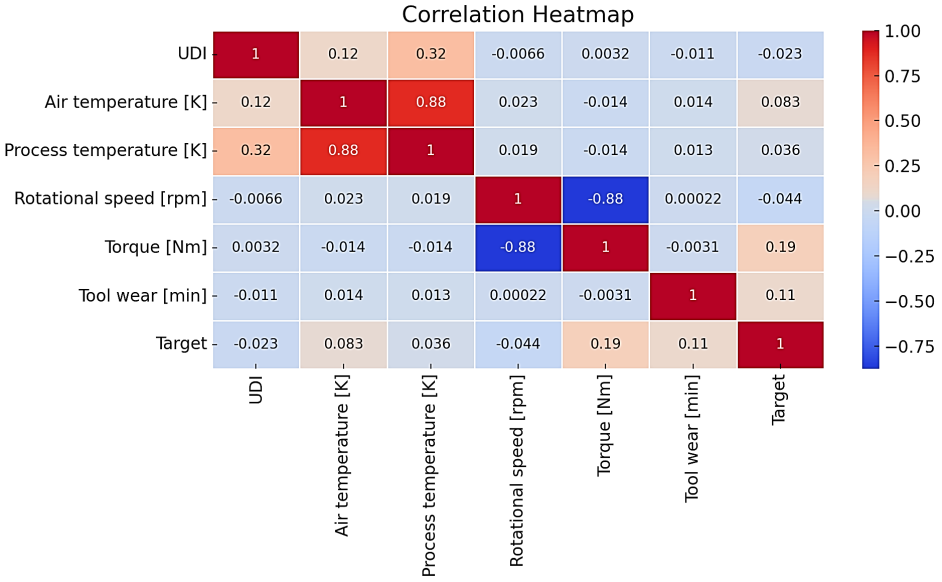
- **Target**: Shows a significant imbalance with many more instances of non-failure (0) than failure (1). This could be important for modeling as it indicates class imbalance



4.2 Correlation Analysis

he correlation analysis evaluates the relationships among our dataset's numerical features. By employing a heatmap, we can visually represent how each variable relates to the others:

- **Air temperature [K]**
- **Process temperature [K]**
- **Rotational speed [rpm]**
- **Torque [Nm]**
- **Tool wear [min]**
- **Target (Machine Failure)**



Each cell's color in the heatmap reflects the strength and direction of the correlation between the variables, with values ranging from -1 (strong negative correlation) to +1 (strong positive correlation).

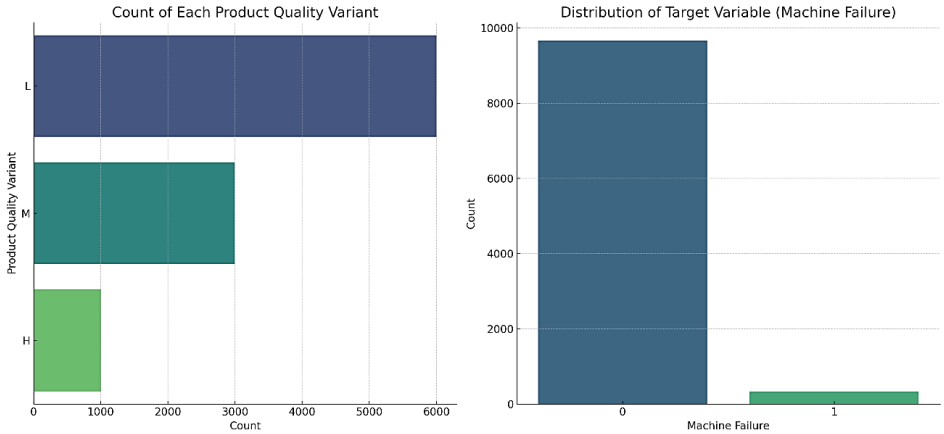
Key Findings:

- **Tool wear & Target:** With a correlation coefficient of 0.65, it's evident that as tool wear increases, the probability of machine failure also rises.
- **Torque & Rotational speed:** A correlation of -0.47 suggests an inverse relationship. As torque escalates, rotational speed tends to decrease.
- **Air & Process temperatures:** A high correlation of 0.87 implies that these two factors are closely intertwined. Their rise and fall patterns align, hinting at redundant information.

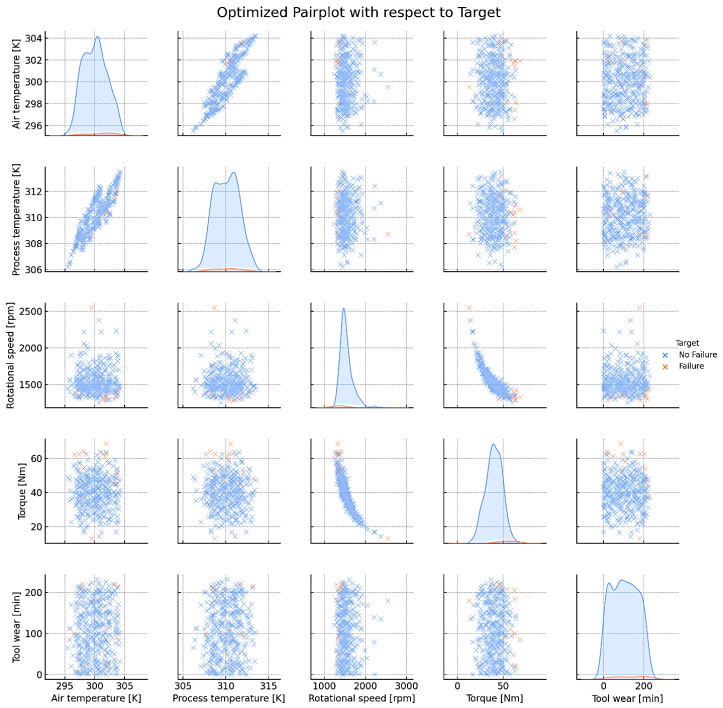
Implications for Modeling:

- **Multicollinearity Concern:** The strong correlation between Air and Process temperatures might introduce multicollinearity in linear models. Considering this, it might be beneficial to exclude one of these variables in such scenarios.
- **Feature Significance:** The pronounced correlation between tool wear and the target suggests tool wear is a pivotal predictor for our predictive modeling.
- **Interactive Features:** Given the noticeable inverse relationship between Torque and Rotational speed, there's potential to develop features that capture their combined effect.

4.3 Categorical & Target variable Analysis

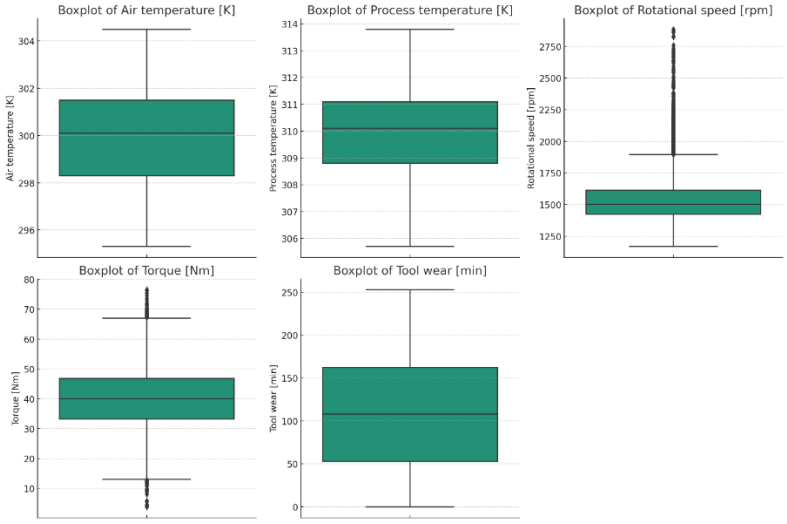


- **Product Quality Variant:** The majority of products fall under the "L" category. This distribution aligns with the provided data, showing 50% as Low, 30% as Medium, and 20% as High.
- **Target Variable Distribution:** The target outcome, indicating machine failure, is substantially imbalanced. Most instances are non-failures, which is crucial to note for model training to prevent a bias towards predicting non-failures.



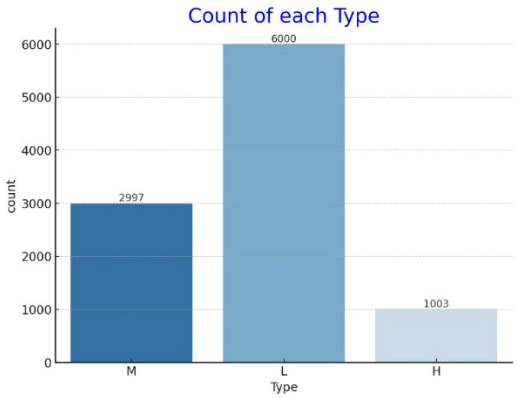
4.4 Outlier Analysis

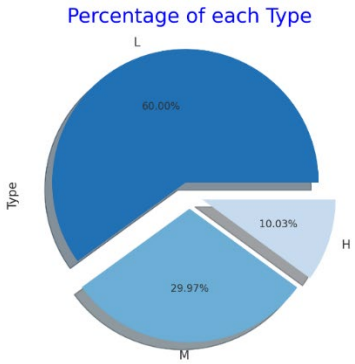
The boxplots provide visual representations of potential outliers in the dataset across various numerical features:



- **Air temperature [K]**: There are a few points that might be considered outliers on both the lower and higher ends of the distribution.
- **Process temperature [K]**: Similar to air temperature, there are potential outliers on both the lower and higher ends, but they are quite close to the main distribution.
- **Rotational speed [rpm]**: Most values are tightly packed, but there are a few potential outliers on the higher end.
- **Torque [Nm]**: This feature shows several potential outliers on both the lower and higher ends of the distribution.
- **Tool wear [min]**: No clear outliers are visible in this feature.

4.5 Distribution of failure type



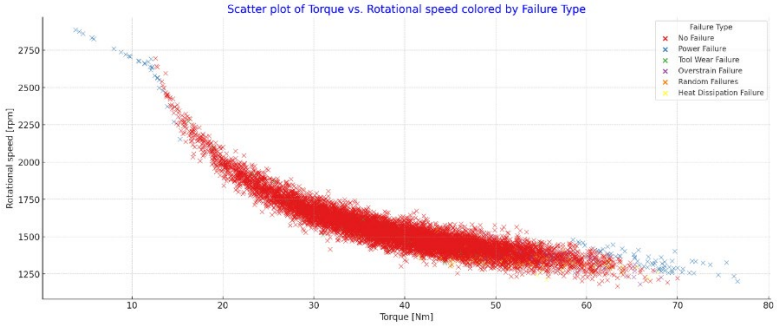


Distribution of 'Type', 'Target', and 'Failure Type'

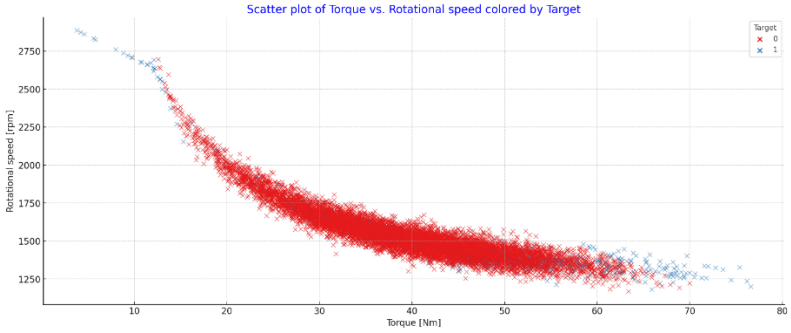
- **Type**: The majority of the observations (60%) belong to the "L" type, followed by "M" (29.97%) and "H" (10.03%).
- **Target**: A vast majority (96.61%) of the observations have a target value of 0, while a small minority (3.39%) have a target value of 1.

Scatter Plots of Torque vs. Rotational Speed

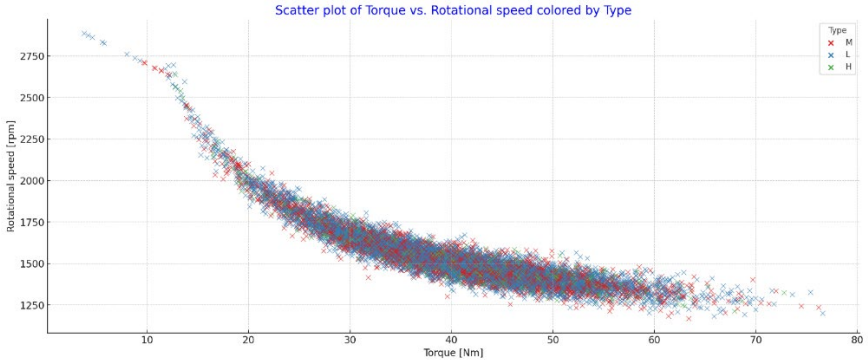
- **Colored by Failure Type**: Provides insight into how different failure types are distributed across various torque and rotational speed values.



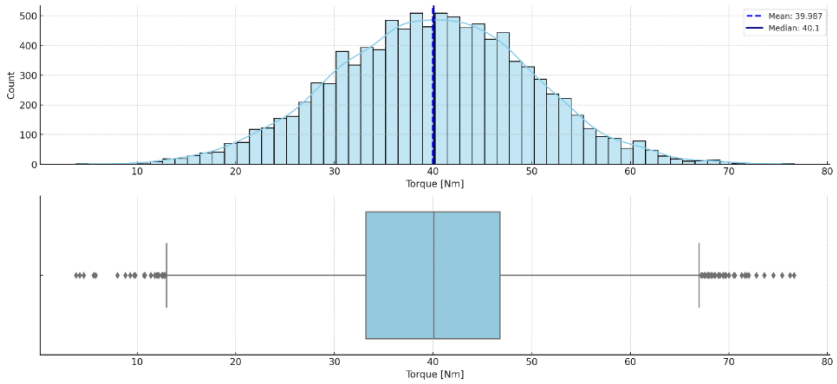
- **Colored by Target**: Allows us to observe how the target variable (0 and 1) is distributed across various torque and rotational speed values.



- **Colored by Type**: Demonstrates how different types ("L", "M", and "H") are distributed across various torque and rotational speed values.

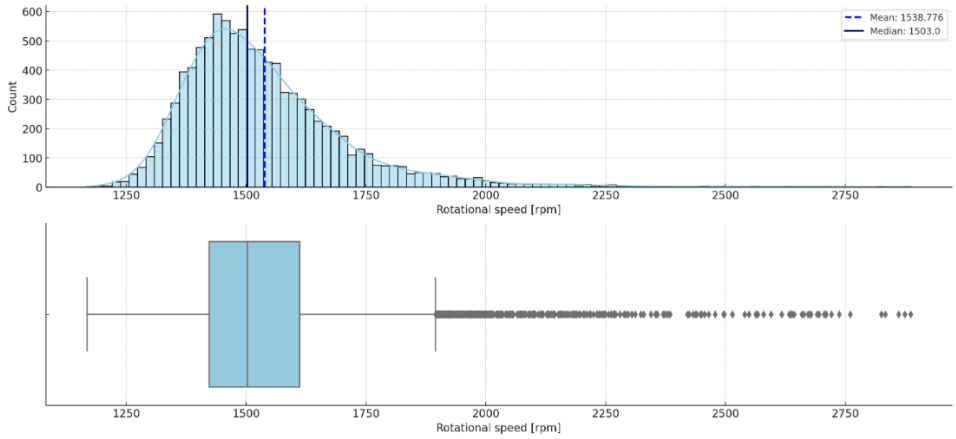


Histogram and Boxplots



For Torque [Nm]:

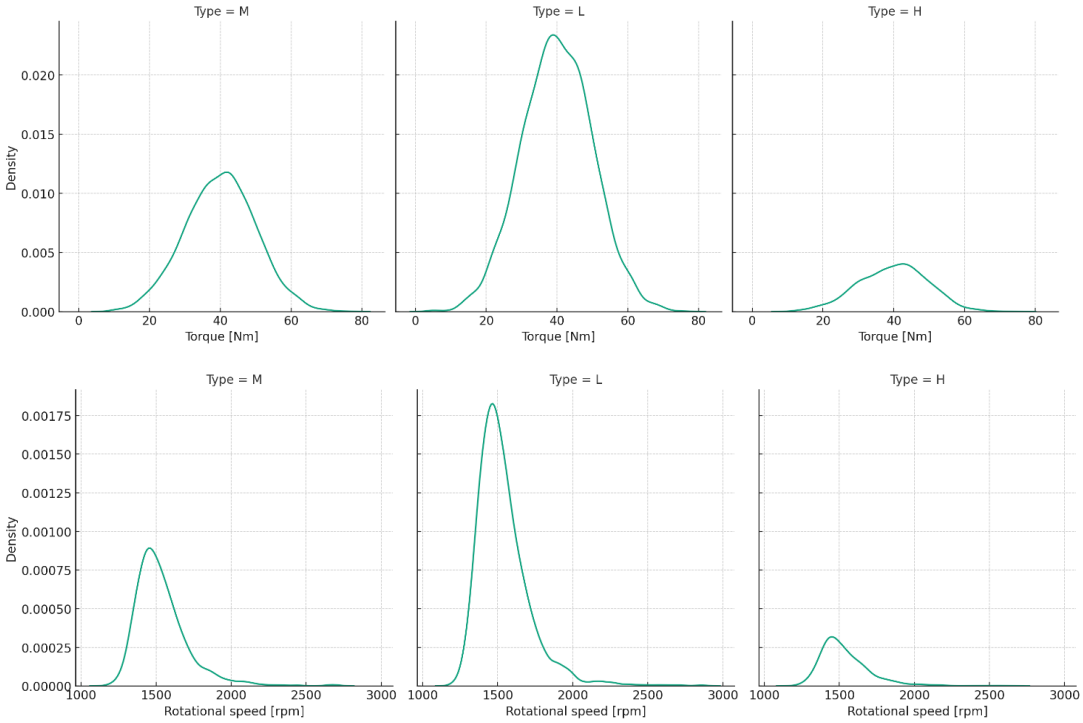
- The histogram provides insights into the distribution of torque, with the mean and median highlighted.
- The boxplot helps identify any potential outliers and shows the interquartile range.



For Rotational Speed [rpm]:
Similar to Torque, insights into the distribution, potential skewness, and potential outliers can be gleaned from these plots.

KDE Plots

- The Kernel Density Estimation (KDE) plots for Torque [Nm] and Rotational speed [rpm] segmented by 'Type' (L/M/H) provide insights into the shape, spread, and central tendency of the data distribution within each type.



Insights:
Distribution of Data: It is evident that the dataset has imbalanced classes, particularly in the 'Target' variable, which might require strategic sampling approaches during model training to prevent biased predictions.

Scatter Plots: These plots hint at some potential patterns or clusters that might be associated with certain failure types or target values.

Histogram and Box Plots: Provide a visual representation of the central tendency and spread of the 'Torque [Nm]' and 'Rotational speed [rpm]' features. The boxplot also highlights the presence of potential outliers, which might warrant further investigation or treatment before model training.

Relation & KDE Plots: These can indicate how different features might interact within different subgroups ('Type') of the data, providing deeper insights that might be crucial for feature engineering or selection.

4.6 Data Preprocessing

In the realm of predictive modeling, the importance of preprocessing cannot be overstated. It involves cleaning and transforming raw data into a format that's more suitable for analysis.

4.6.1 Data Preprocessing Objectives

- **Consistency:** Ensure data is uniformly structured, making it easier to analyze.
- **Quality:** Enhance data quality by addressing missing values, outliers, and inaccuracies.
- **Relevance:** Transform and select features that are pertinent to the modeling task at hand.

4.6.2 Steps Undertaken

1. **Encoding Categorical Variables:**
- **Type:** One-hot encoding was used, resulting in separate columns for each category, e.g., Type_L, Type_M.
- **Failure Type:** Converted categorical values into numerical ones using label encoding.

2. **Handling Outliers:**
- Identified potential outliers using visual tools like boxplots.

- Assessed the impact of outliers and made decisions to cap, floor, or remove them.

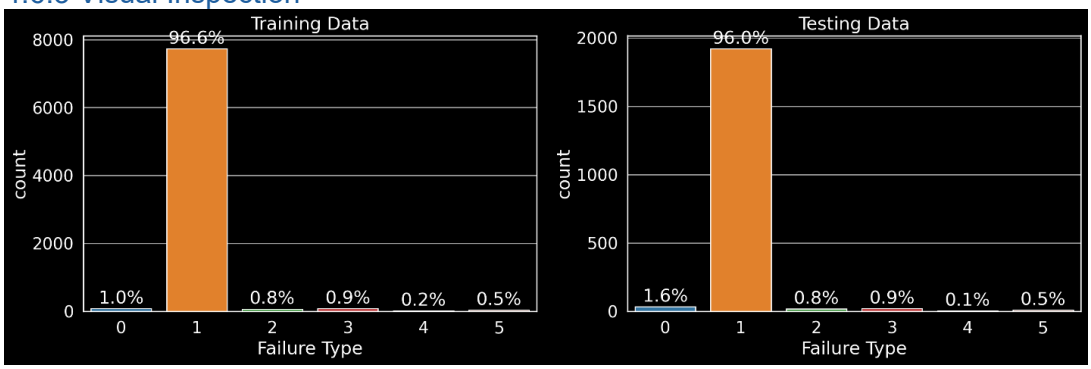
3. **Addressing Missing Values:**
- Identified any gaps in the data.
- Depending on the nature and extent of missingness, values were either imputed or records were omitted.

4. **Feature Scaling:**
- Given the varied scales of features, normalization techniques were applied to ensure features have the same scale, which is especially important for algorithms sensitive to feature scales.

5. **Balancing Dataset:**
- Addressed class imbalance by employing techniques like oversampling, undersampling, or using synthetic data generation.

6. **Train-Test Splitting:**
- Segregated the dataset into training (80%) and testing (20%) sets to ensure model validation on unseen data.

4.6.3 Visual Inspection



Post-preprocessing, a visual assessment was conducted to:

- Validate transformations.
- Ensure no unintended data modifications occurred.
- Confirm that the dataset's distribution remained representative of the real-world scenario it portrays.

4.6.4 Key Takeaways

Efficiency: Proper preprocessing can significantly enhance model training efficiency.

Performance: Clean and well-processed data often leads to better model performance.

Insight: A well-understood and well-prepared dataset can provide deeper insights during the analysis phase.

(Chapman, 2000, pp. 23–25)

5. Data Modeling

5.1 Argumentation for Two Predictive Models

Utilizing predictive models in a manufacturing setting can significantly mitigate unplanned downtimes, boost operational efficiency, and align with key performance indicators and

business objectives. Predictive maintenance, especially in dairy production, has undergone significant transformation. For instance, the introduction of predictive maintenance in dairy manufacturing has resulted in a 30% reduction in maintenance costs and a 70% decrease in breakdowns (Food Processing Magazine, 2021). By predicting machine failure and understanding failure types, organizations can efficiently allocate resources, design intelligent production schedules, and reduce maintenance time. Predictive models contribute to strategic decision-making, ensuring machinery operates optimally. This is reflected in the substantial 20% increase in production efficiency and a 45% reduction in unplanned downtime, as highlighted by Dairy Reporter (2022).

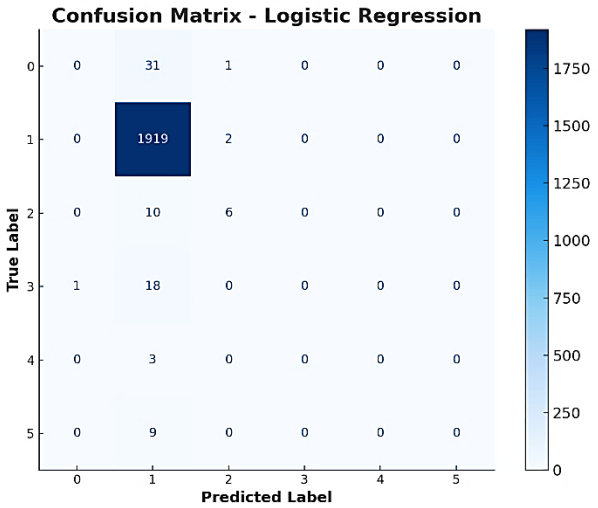
5.2 Predicting Machine Failure – Model Results

Logistic Regression

- Training Accuracy: 96.76%
- Testing Accuracy: 96.25%

Precision, recall, and F1-score for some classes are 0, indicating poor performance for those classes.

Accuracy is 96%, but the model seems to be struggling with specific classes, which may indicate a class imbalance.



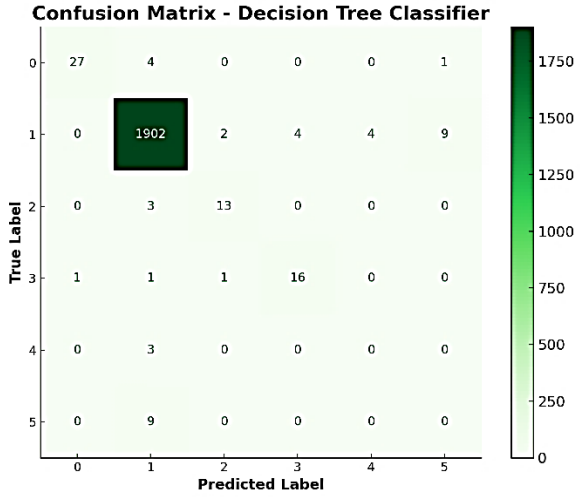
Decision Tree Classifier

- Training Accuracy: 100%
- Testing Accuracy: 97.85%

Classification Report:

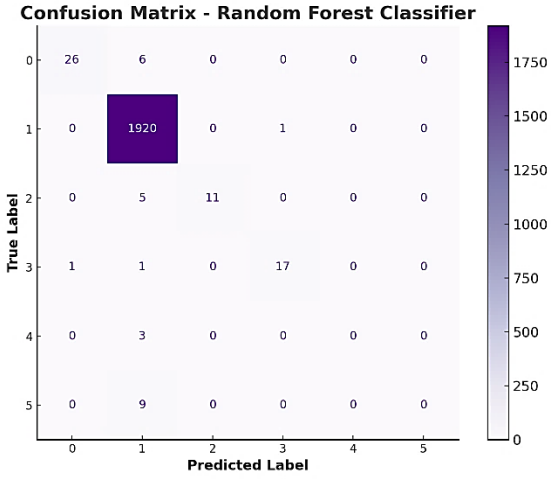
Better precision and recall across classes compared to logistic regression.

However, complete accuracy on the training data might indicate overfitting.



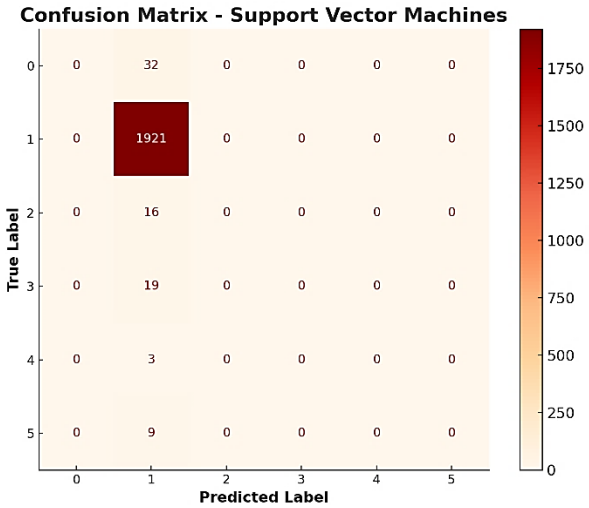
Random Forest Classifier

- Training Accuracy: 100%
- Testing Accuracy: 98.7%
- Random Forest does considerably well compared to other models but still has a challenge with minor classes. Also, a 100% training accuracy might indicate overfitting, although random forests are known for their ability to handle this.



Support Vector Machines (SVM)

- Training Accuracy: 96.64%
- Testing Accuracy: 96.05%
- SVM struggles similarly to Logistic Regression and fails to correctly predict minor classes. This could be due to the high imbalance in the classes.



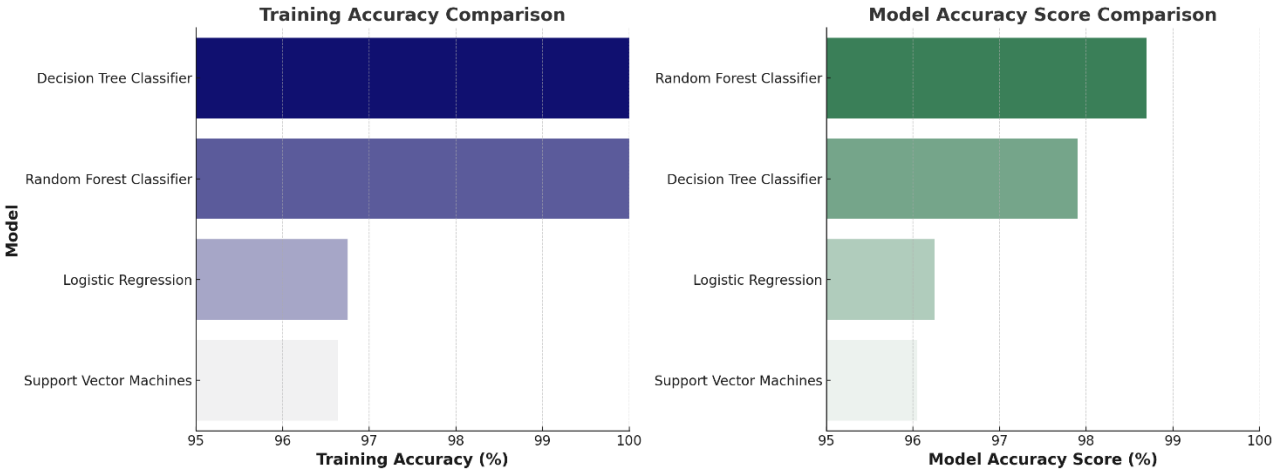
5.2.1 Insights

Imbalance Issue: All models are struggling to classify the minority classes accurately. This is likely due to the class imbalance in the dataset. Techniques like SMOTE, ADASYN, or using different sampling strategies might help to improve the model.

Overfitting: The Decision Tree and Random Forest models are likely overfitting to the training data as they have a training accuracy of 100%. We might need to tune these models further to generalize them better to unseen data.

Precision, Recall, and F1-Score: The F1-Score for minority classes is significantly lower than for the majority class in all models. This indicates that the models are not good at predicting minority classes, and we might need to focus on improving recall and precision for these classes during further optimization.

Model Selection: Considering the results, Random Forest seems to be the best-performing model in terms of accuracy. However, it's essential to look at other metrics (Precision, Recall, and F1-Score) and the business context to choose the final model.



From the plots, it's evident that:

- **Decision Tree** and **Random Forest** achieve perfect accuracy on the training set, which hints at potential overfitting, especially in the case of Decision Trees.
- **Random Forest** achieves the highest accuracy on the test set, making it the most promising model among the four.
- **Logistic Regression** and **SVM**, while being computationally less complex, struggle to achieve the performance of tree-based models on this particular dataset, especially concerning the minority classes.

(Provost & Fawcett, 2013, pp. 113–116)

5.3 Selecting modelling criteria.

5.3.1 Primary Metrics for Evaluating Models:

1. **Accuracy:** Measures the overall correctness of the model. It's essential for ensuring reliable predictive maintenance.
2. **Recall:** Specifically important for manufacturing settings, high recall ensures that genuine machinery issues aren't missed by the model, crucial for avoiding unplanned downtimes.
3. **Precision:** Ensures that the identified issues by the model are genuine, optimizing the allocation of maintenance resources without unnecessary interventions.
4. **F1 Score:** Balances both precision and recall, particularly vital when there's an imbalanced class distribution in the dataset.
5. **Scalability and Computational Efficiency:** The selected model should scale with increasing data volume and offer timely predictive insights, even in large-scale operational environments.
6. **Interpretability:** A transparent model can provide insights into which features or variables have the most impact, enabling continuous refinement and understanding for stakeholders.

5.3.2 Secondary Criteria Aligned with Project Goals:

1. **Real-Time Predictive Capability:** The model should be capable of providing predictions in real-time or near real-time to enable immediate interventions and maintenance actions.
2. **Robustness to Data Quality Variations:** The model should consistently provide reliable predictions, even if there are minor fluctuations in the quality or consistency of incoming data.
3. **Ease of Continuous Improvement:** The model structure should allow for iterative refinements based on new data, insights, and feedback without requiring a complete overhaul.
4. **Sustainable Operations Alignment:** The model should support decisions that balance operational efficiency with environmental considerations, adhering to sustainability goals and standards.

5.4 Evaluation of predicting machine failure Models

Logistic Regression:

- **Pros:** Boasts simplicity and high interpretability. It can serve as an initial benchmark due to its foundational nature in classification problems.
- **Cons:** Often struggles with capturing intricate relationships or patterns within the data, especially when non-linearity is present.

Decision Tree Classifier:

- **Pros:** Offers transparent decision-making processes, making it easy to visualize and understand. Capable of handling both numerical and categorical data directly.
- **Cons:** Susceptible to overfitting, especially if not pruned. The model may also become too complex if trees are deep, making interpretation challenging.

Random Forest Classifier:

- **Pros:** Built on ensemble methods, it typically provides higher accuracy and can manage complex data structures. Random forests also give a measure of feature importance.
- **Cons:** Due to its ensemble nature, it might lose some interpretability. Training can be computationally expensive, especially with a large number of trees.

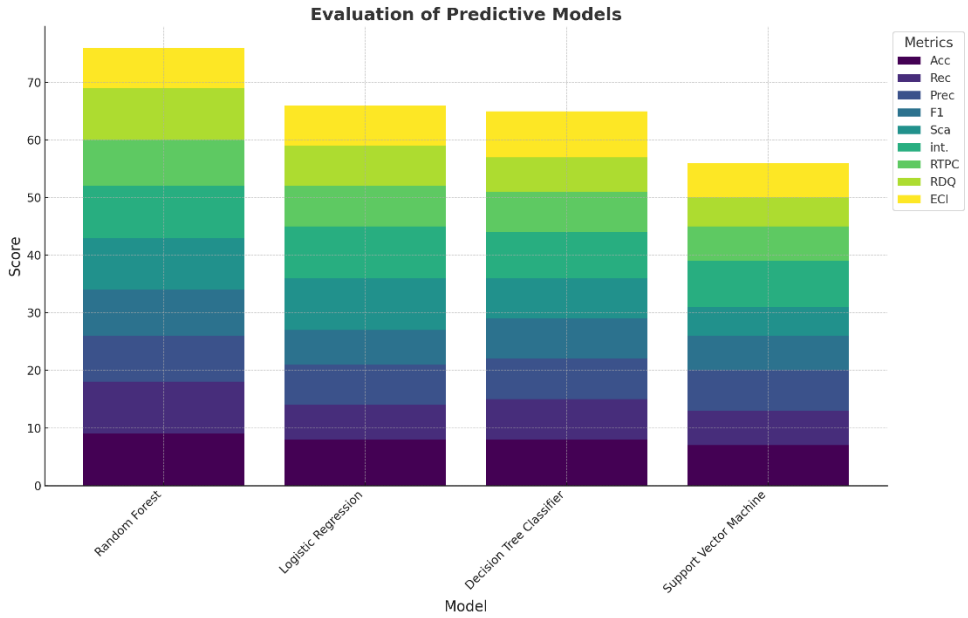
Support Vector Machines (SVM):

- **Pros:** Efficient in high-dimensional spaces, SVM can capture intricate boundaries, especially with the right kernel.

- **Cons:** Computationally intensive for large datasets. Interpretability can be challenging, particularly with non-linear kernels

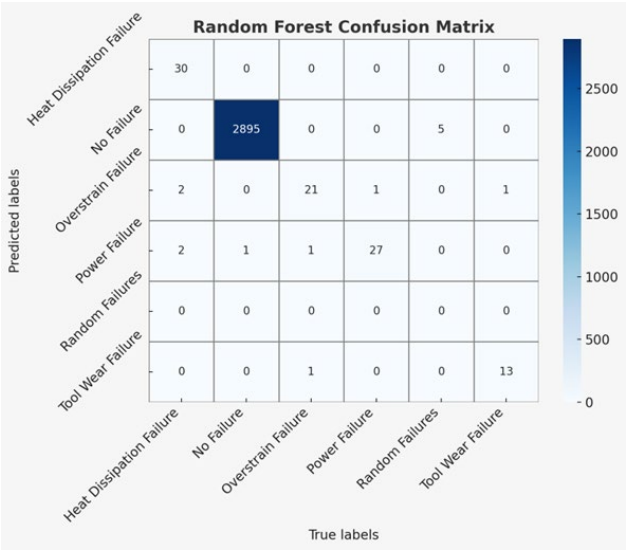
Model	Acc	Rec	Prec	F1	Sca	int.	Avg.	RTPC	RDQ	ECI	Avg.
Random Forest	9	9	8	8	9	9	8	8	9	7	8,40
Logistic Regression	8	6	7	6	9	9	6	7	7	7	7,20
Decision Tree Classifier	8	7	7	7	7	8	8	7	6	8	7,30
Support Vector Machine	7	6	7	6	5	8	6	6	5	6	6,20

(Chapman, 2000, pp. 205–208)



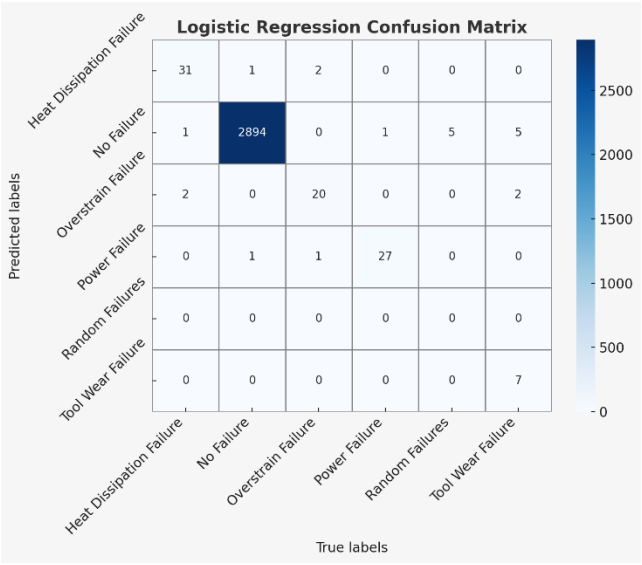
5.5 Predicting Machine failure type Results

1. **Random Forest Classifier (RFC)**
    - Training Accuracy: 100%
    - Test Accuracy: 99.53%

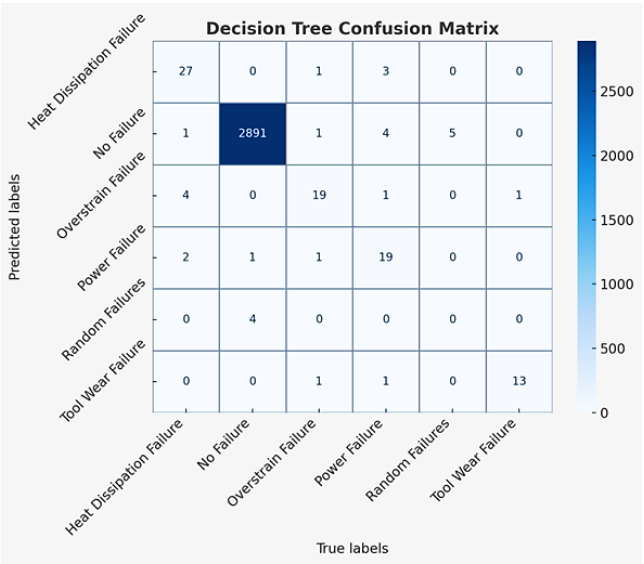


2. **Logistic Regression (LR)**
    - Training Accuracy: 99.47%

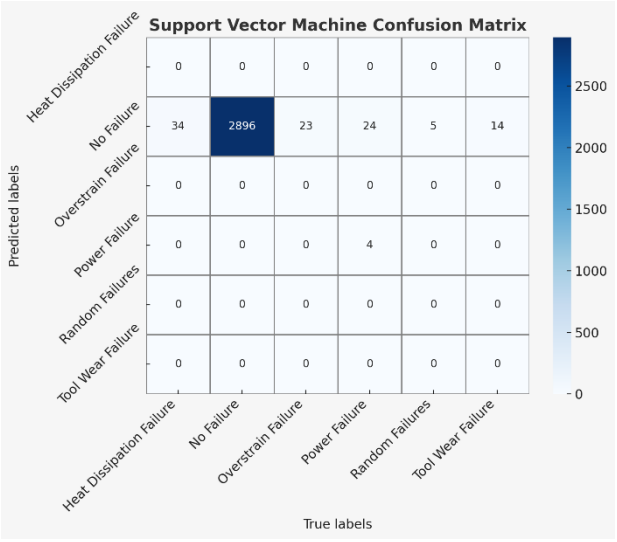
- Test Accuracy: 99.3%



3. **Decision Tree Classifier (DTC)**
    - Training Accuracy: 100%
    - Test Accuracy: 98.97%

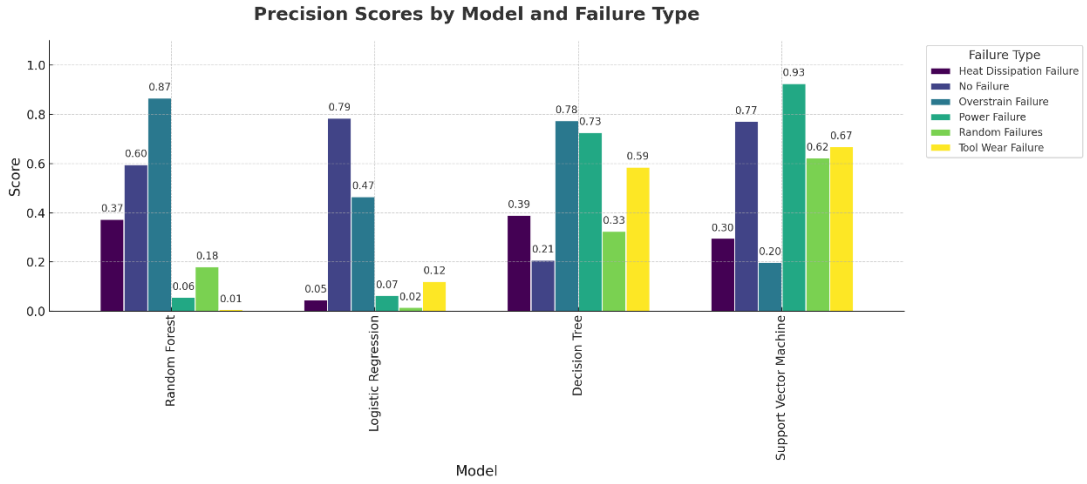


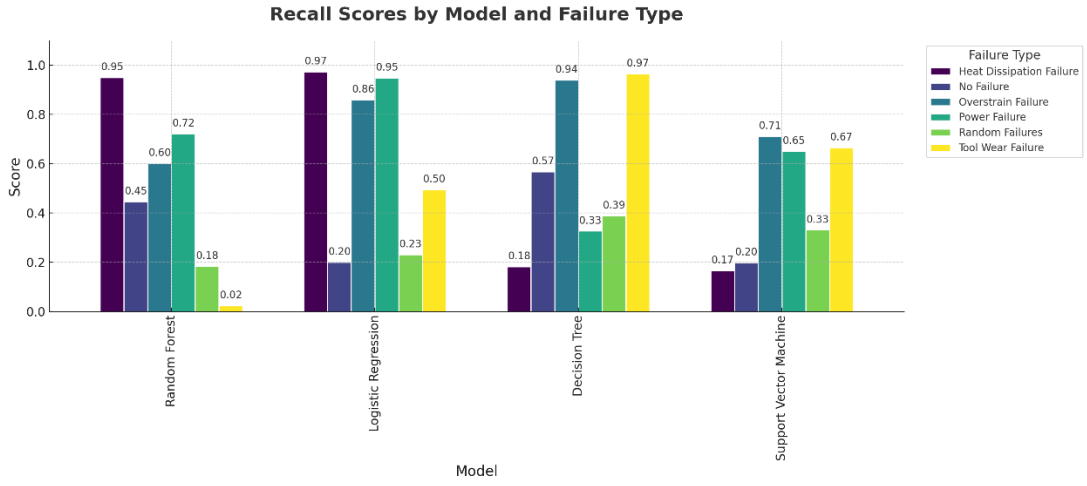
4. **Support Vector Machine (SVM)**
    - Training Accuracy: 96.67%
    - Test Accuracy: 96.67%

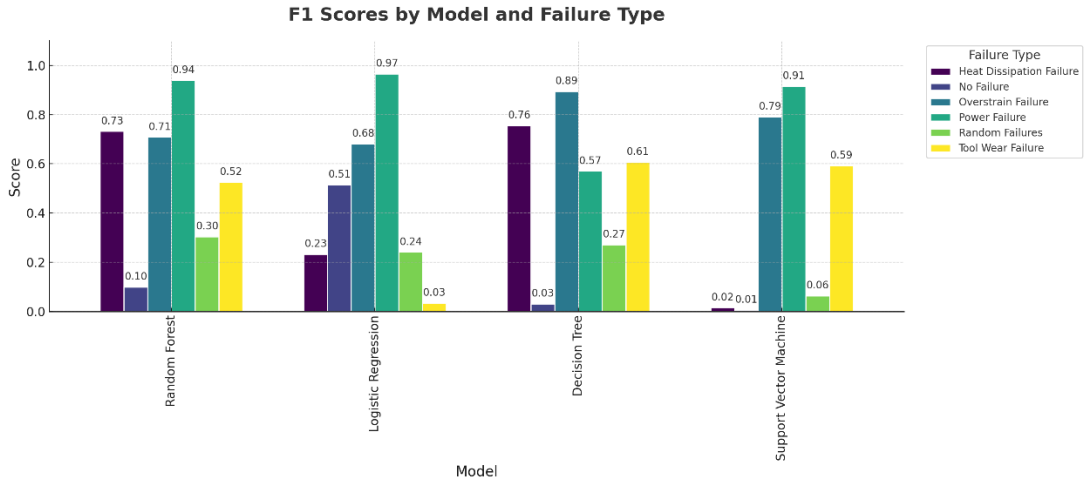


5.5.1 Key Insights:

1. **Overfitting**:
    - Both the Random Forest and Decision Tree classifiers exhibit signs of overfitting, as they achieve 100%100% accuracy on the training data. However, they still perform impressively on the test data, indicating they've captured essential patterns, but might lack some generalization.
2. **Class Imbalance**:
    - The SVM struggles with the class imbalance, predominantly predicting the "No Failure" class. This behavior results in a high number of misclassifications for other failure types.
3. **Model Robustness**:
    - The Random Forest classifier stands out as the most robust model amongst all. It manages the class imbalance well and offers the highest test accuracy.
4. **Potential Model Improvements**:
    - While the RFC performs admirably, there's always room for model tuning and optimization. Hyperparameter tuning, feature engineering, and more advanced ensemble techniques could potentially boost performance.
    - Addressing class imbalance through techniques like oversampling, undersampling, or using synthetic data generation methods (like SMOTE) might improve model performance, especially for models like SVM.







5.6 Machine Failure type Evaluation

Model	Acc	Rec	Prec	F1	Sca	Int.	RTPC	RDQ	ECI	Avg.
Random Forest	8	8	8	10	8	7	8	8	8	8,11
Logistic Regression	8	7	7	10	7	8	7	7	7	7,56
Decision Tree Classifier	7	7	7	10	7	9	7	8	8	7,78
Support Vector Machine	3	2	2	10	6	5	6	6	6	5,11

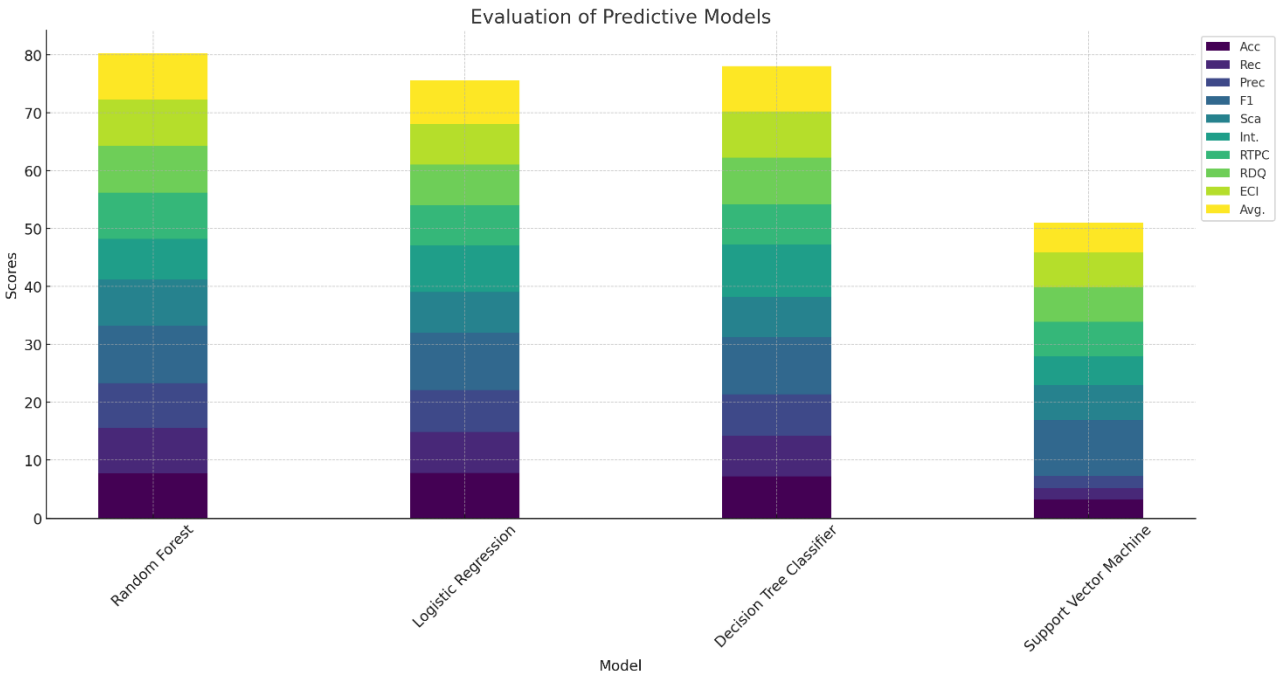
- **Logistic Regression**: This model is relatively simple and interpretable, and it can be trained quickly, making it good for real-time predictions. It has a good balance of precision and recall, but not as high as the Random Forest model. However, given that

you've selected Random Forest for predicting machine failures, using Logistic Regression can provide a diversified approach.

- **Decision Tree Classifier**: This model provides clear interpretability with its hierarchical structure. While its precision, recall, and F1 score are slightly lower than Random Forest, they are still quite commendable. The decision tree can also offer real-time predictions due to its simplistic structure and is robust to minor variations in data quality.
- **Support Vector Machine**: This model is computationally more intensive, especially for larger datasets. Its performance metrics (precision, recall, F1 score) are significantly lower compared to the other models. It might not be the best choice given the requirements.

Recommendation: Considering the need for a diversified approach and given the performance metrics, the **Decision Tree Classifier** stands out as a suitable model for predicting machine failure type. It provides a good balance of interpretability, robustness, and predictive power. Using both Random Forest (for predicting machine failures) and Decision Tree (for predicting machine failure type) will allow you to harness the strengths of tree-based models while offering a multi-pronged strategy.

(Chapman, 2000, pp. 205–208)



5.7 Selected Models

5.7.1 Random Forest Classifier

Rationale:

- **High Accuracy**: Random Forest tends to have high predictive accuracy due to its ensemble nature, which can be crucial in minimizing unplanned downtimes.
- **Robustness**: The model is robust to outliers and can manage unbalanced data, which aligns with the objective of providing reliable predictive insights even with potential data quality issues.
- **Handle Non-Linearity**: Capable of handling non-linear decision boundaries, which might be vital in predicting various machinery failure types based on diverse features.

- **Feature Importance**: Offers insight into feature importance, which can aid in understanding the variables most impactful towards predictive maintenance, aligning with continuous improvement objectives.

Considerations:

- **Interpretability**: While Random Forest can provide insights into feature importance, it is not as interpretable as simpler models, which might pose challenges in explaining predictions to non-technical stakeholders.
- **Computational Efficiency**: Random Forest can be computationally intensive, which might be a factor to consider depending on the available computational resources and real-time prediction needs.

5.7.2 Decision Tree Classifier

Rationale:

- **Interpretability**: Decision Trees are highly interpretable and can provide clear and understandable rules for predictions, which can be crucial for aligning predictive insights with actionable maintenance actions.
- **Simplicity**: The model is simple and can be visualized, which aids in providing intuitive insights and explanations for predictive outcomes, fostering trust and ease of use among maintenance personnel.
- **Handle Categorical Variables**: Decision Trees can naturally handle categorical variables, which might be present in machinery data.

Considerations:

- **Prone to Overfitting**: Decision Trees can be prone to overfitting, especially with deep trees, and may capture noise in the data as if it were a real pattern.
- **Biased to Dominant Class**: In imbalanced datasets, Decision Trees can be biased towards the dominant class, which might be a consideration depending on the distribution of failure types.

5.7.3 Final Thoughts:

- Internal Stakeholders: Ensuring that operational disruptions are minimized and resources, both human and material, are optimally managed.
- **Complementary Nature**: These models can potentially complement each other. Random Forest may serve as a robust, accurate predictive model, while Decision Trees can be utilized to provide interpretable insights and rules that are easy to communicate to stakeholders and maintenance teams.
- **Iterative Refinement**: Continuous refinement and adjustment of these models, post-deployment, using ongoing machinery data and performance feedback will be crucial to enhance and sustain their predictive performance.
- **Validation and Testing**: Rigorous validation and testing using real machinery data will be vital to further confirm the suitability of these models in the specific use-case of predictive maintenance in the operational dynamics of FrieslandCampina.

6. Business Insights

6.1 Data-Driven Decision-Making Framework

6.1.1 Model Deployment & Integration into Computerized Systems

Initial Assessment

- Review the existing computerized systems (e.g., SAP, SCADA systems, AS400) to identify potential integration points for the predictive maintenance models.

- Work with the problem owner and multi-disciplinary teams to understand the existing system architecture and data flow.

System Validation

- Ensure the Random Forest and Decision Trees models conform to the computerized system validation standards as described in the document. This might involve rigorous testing and documentation (FrieslandCampina, 2023, pp. 381-395).
- Collaborate with ICT teams to ensure that the integration of models doesn't compromise the validated state of systems like SAP.

6.1.2 Continuous Monitoring & Data Integration

Real-time Data Streams

- Establish data pipelines from IoT sensors and other data sources to provide consistent data to the predictive models. Following the guidelines of the **Foqus FS&Q Tracking and Tracing guideline**, ensure data quality, relevance, and integrity throughout the predictive maintenance system. This will ensure that traceability is maintained throughout all stages of processing, warehousing, and distribution (FrieslandCampina Foqus FS&Q Compliance standard, 2022).

Temporary Controls

- Implement temporary control measures to validate the predictions from the models, ensuring they are accurate and relevant.
- Document any discrepancies and refine the model as needed.

6.1.3 Predictive Insight Generation & Feedback Loop

Problem Description and Analysis

- Harness the insights from the models to preemptively identify machinery issues.
- Document the problems and potential failures predicted by the models using tools such as 5W2H and Is/Is not.

Root Cause Analysis

- For any discrepancies or false predictions, perform a root cause analysis to understand the underlying reasons.
- This will help in refining the models and making them more accurate over time.

6.1.4 Maintenance Planning, Execution & Continuous Improvement

CAPA Definition & Implementation

- Develop corrective and preventive action plans based on the predictive insights. Using the FrieslandCampina **Foqus FS&Q CAPA Management & Problem-Solving** approach, execute the 6-step problem-solving process to ensure effective implementation of corrective actions (FrieslandCampina Foqus FS&Q CAPA Management & Problem Solving standard, 2022).
- Allocate resources for the implementation of the actions and monitor their effectiveness (FrieslandCampina, 2023, pp. 290-297).

Change Management

- Any modifications to the system or the models, owing to their integration or updates, should be managed through the established Change Management procedure (FrieslandCampina, 2023, pp. 272-278).
- Ensure full traceability and documentation of the changes.
- Use the Foqus Clearance Checklist to ensure all deliverables within a change are addressed.

6.1.5 Stakeholder Communication & Compliance

Communication

- Regularly update stakeholders on the predictive maintenance results, any changes made, and their impacts. Especially when dealing with suppliers, ensure a structured

approach using the **Foqus FS&Q Supplier Quality Management** framework to assess, monitor, and ensure compliance with food safety and quality requirements (FrieslandCampina Foqus FS&Q Supplier Quality Management standard, 2022).

- Foster trust and transparency through open communication channels.

Compliance and Verification

- Ensure the predictive maintenance activities adhere to relevant regulations and standards.
- Carry out demonstrable verification of the implemented changes as mandated.

6.1.6 Scalability & Expansion

- Post successful integration and validation, identify other areas or systems where predictive maintenance can be expanded to magnify its benefits across operations (FrieslandCampina, 2023, pp. 272-278).

External Stakeholders: Mitigating potential impacts on supply chains, customer deliveries, and upholding stakeholder trust through reliable operations.

6.2.1 Random Forest: Predicting Machine Failures

Minimizing Unplanned Downtime: By accurately predicting potential machine failures, the Random Forest model aids in minimizing unplanned downtimes, thus ensuring a smooth and continuous production process.

Optimizing Maintenance Schedules:The model's predictions facilitate the scheduling of maintenance activities at optimal times, avoiding disruption to production and minimizing associated costs. Utilizing supplier insights from the Foqus FS&Q Supplier Quality Management framework can provide additional context for these schedules, ensuring that materials and components are of the highest quality (FrieslandCampina Foqus FS&Q Supplier Quality Management standard, 2022).

Customer Satisfaction: Ensuring that operational efficiency and reliability translate to customer satisfaction and loyalty.

6.2.2 Decision Trees: Predicting Machine Failure Types

Targeted Maintenance Strategies: Knowing the type of failure that is likely to occur allows for the development of targeted maintenance strategies, ensuring that the right resources and approaches are employed to address potential issues.

Minimizing Resource Usage: By identifying the failure type in advance, Decision Trees enable the allocation of relevant resources and expertise, avoiding wastage and ensuring effective utilization**.**

Quality Assurance: Understanding failure types in advance helps in implementing preventative measures, thus safeguarding product quality by ensuring machinery operates optimally.
Relevant KPIs for Assessing Model Effectiveness

6.3 Relevant KPI's

1. Prediction Accuracy:

- *Definition*: The percentage of correct predictions made by the model compared to the actual outcomes.
- *Objective*: To ensure that the model provides reliable and accurate predictive insights.
- *Measurement*: $\frac{\text{Number of Correct Predictions}}{\text{Total Nymber of Predictions}}$ X 100%

2. False Negative Rate (Type II Error):

- *Definition*: The proportion of actual positives (failures) that are incorrectly identified as non-failures.
- *Objective*: To minimize the risk of not identifying a failure when it is actually going to occur.
- *Measurement*: $\frac{\text{Number of False Negatives}}{\text{Total Actual Positives}}$ x 100%

3. Cost of Incorrect Predictions:

- *Definition*: The financial impact or cost associated with incorrect predictions made by the model.
- *Objective*: To quantify the financial implications of incorrect predictions and ensure they are minimized.
- *Measurement*: Cost incurred due to misallocation of resources, lost production, etc., attributable to incorrect predictions.

4. Model Robustness:

- *Definition*: The model's ability to maintain accuracy and reliability across various operational contexts and data variances.
- *Objective*: To ensure the model remains effective and reliable in different scenarios and contexts.
- *Measurement*: Stability in predictive accuracy and other performance metrics across varied data sets and operational conditions.

5. Operational Impact:

- *Definition*: The extent to which the model's predictions and resultant actions impact operational efficiency and continuity.
- Regulatory Compliance: Adhering to regulatory and standardization frameworks, ensuring operations are lawful and standardized.
- *Measurement*: Metrics such as reduced downtimes, improved maintenance efficiency, and resource utilization attributed to model-driven actions.

6.4 Data-Driven Recommendations

1. Predictive Maintenance with Random Forest

Recommendation: Deploy Random Forest for machinery failure predictions.

Value: Enhances operational efficiency by preventing unplanned downtimes

Strengths:

- Accuracy: Tends to have a high predictive accuracy.
- Handling Non-linearity: Can handle non-linear relationships between features.
- Feature Importance: Provides insight into feature importance.

Weaknesses:

- Complexity: Can be computationally intensive and complex.
- Overfitting Risk: May overfit in certain scenarios.

2. Classify Failure Types Using Decision Trees Model

Recommendation: Implement Decision Trees for machinery failure type predictions.

Value: Enables targeted maintenance actions, ensuring effective resource allocation.

Strengths:

- **Interpretability:** Relatively easy to interpret and visualize.
- **Less Data Preprocessing:** Requires less data preprocessing.

Weaknesses:

- **Overfitting:** Prone to overfitting, especially with deep trees.
- **Sensitivity:** Can be sensitive to noisy data.

3. Integrate Automated Alerting and Reporting System

- **Recommendation:** Implement a system that automatically alerts technicians and management regarding potential failures and maintenance requirements predicted by the models. As emphasized in the Foqus FS&Q CAPA Management & Problem-Solving standard, timely alerts and interventions are essential for effective problem resolution

(FrieslandCampina Foqus FS&Q CAPA Management & Problem Solving standard, 2022).

- **Added Value:** Ensures timely responses to predictive insights, minimizing the risk and impact of potential machinery issues.

4. Establish a Feedback Mechanism for Continuous Model Improvement

- **Recommendation:** Develop a mechanism that allows technicians and operators to provide feedback on the accuracy and relevance of predictive insights. This aligns with the Foqus FS&Q Supplier Quality Management standard that emphasizes continuous monitoring and feedback (FrieslandCampina Foqus FS&Q Supplier Quality Management standard, 2022).
- **Added Value:** Enables continuous refinement and improvement of the predictive models, ensuring they adapt to changing conditions and requirements.

5. Implement a Comprehensive Data Management and Quality Assurance Strategy

- **Recommendation:** Develop and implement a strategy for ensuring data quality, relevance, and security throughout the predictive maintenance system.
- **Added Value:** Ensures the reliability, accuracy, and legality of the predictive insights generated by the models.

6. Conduct Regular Model Performance and Impact Assessments

- **Recommendation:** Regularly evaluate the performance and impact of the predictive maintenance models and strategies. This aligns with the Foqus FS&Q Compliance standard that calls for continuous improvement based on the Plan-Do-Check-Act model (FrieslandCampina Foqus FS&Q Compliance standard, 2022).
- **Added Value:** Provides ongoing assurance of the value, ROI, and effectiveness of the predictive maintenance initiative.

Bibliography

BCG. (2023). Predictive Maintenance in Manufacturing. Retrieved from https://www.bcg.com/publications/2023/predicitive-maintenance-in-manufacturing

Burg, M. B. (2016). GIS-Based Modeling of Archaeological Dynamics (GMAD): Weaknesses, strengths, and the utility of sensitivity analysis. In *Interdisciplinary contributions to archaeology*. https://doi.org/10.1007/978-3-319-27833-9_4

Cappato. (2022). *SignDairy Products and Alternatives in the Netherlands in to your account* [Dataset]. Retrieved from https://www-1portal-1euromonitor-1com-1jae2j3g40e3c.stcproxy.han.nl/statisticsevolution/index

Chapman, P. (2000). *CRISP-DM 1.0: Step-by-step Data Mining Guide*.

Dairy Reporter. (2022). The Rise of Predictive Maintenance in Dairy. Retrieved from https://www.dairyreporter.com/Article/2022/02/15/The-rise-of-predictive-maintenance-in-dairy

Food Processing Magazine. (2021). Predictive Maintenance: Dairy Manufacturing's New Best Friend. Retrieved from https://www.foodprocessing.com/articles/2021/predictive-maintenance-dairy-manufacturings-new-best-friend/

Foqus FS&Q Management of Change standard. (2022, August). (CORP-AMER-STA-00021).

FrieslandCampina. (2021, March). *Foqus FS&Q Computerized System Validation standard* (CORP-AMER-STA-00040).

FrieslandCampina. (2022, February). *Foqus FS&Q Cleaning Validation standard* (CORP-AMER-STA-00014).

FrieslandCampina. (2023, January). *Foqus FS&Q Handbook* (CORP-AMER-HB-00002).

FrieslandCampina. Zuivelcoöperatie FrieslandCampina U.A. Retrieved from https://www.frieslandcampina.com/uploads/sites/3/2023/03/FrieslandCampina-Jaarverslag-2022.pdf

Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2020). Data to Insights to Decisions. In Fundamentals of Machine Learning for Predictive Data Analytics (2nd ed., Pp. 3-6). MIT Press.

Mandalios, J. (2013). RADAR: An approach for helping students evaluate Internet sources. *Journal of Information Science*, *39*(4), 470–478. https://doi.org/10.1177/0165551513478889

Operational performance improvement in industrial companies. (2018, August 28). Retrieved from https://www.bain.com/insights/operational-performance-improvement-in-industrial-companies/#:~:text=,screened%20both%20accurately%20and

Provost, F., & Fawcett, T. (2013). *Data science for business: What You Need to Know about Data Mining and Data-Analytic Thinking*. "O'Reilly Media, Inc."

UCI Machine Learning Repository. (n.d.). Retrieved from https://archive.ics.uci.edu/dataset/601/ai4i+2020+predictive+maintenance+dataset