



Mise en garde

La bibliothèque du Cégep de l'Abitibi-Témiscamingue et de l'Université du Québec en Abitibi-Témiscamingue (UQAT) a obtenu l'autorisation de la personne autrice de ce document afin de diffuser, dans un but non lucratif, une copie de son œuvre dans [Depositum](#), site d'archives numériques, gratuit et accessible à tous. L'autrice ou l'auteur conserve néanmoins ses droits de propriété intellectuelle, dont son droit d'auteur, sur cette œuvre.

Warning

The library of the Cégep de l'Abitibi-Témiscamingue and the Université du Québec en Abitibi-Témiscamingue (UQAT) obtained the permission of the author to use a copy of this document for nonprofit purposes in order to put it in the open archives [Depositum](#), which is free and accessible to all. The author retains ownership of the copyright on this document.

UNIVERSITÉ DU QUÉBEC EN ABITIBI-TÉMISCAMINGUE

**MAINTENANCE PRÉDICTIVE BASÉE SUR LES DONNÉES :
UN CADRE À DOUBLE ÉTUDE DE CAS POUR L'EXPLOITATION
MINIÈRE SOUTERRAINE ET LES SYSTÈMES DE POMPES
INDUSTRIELLES**

DATA-DRIVEN PREDICTIVE MAINTENANCE: A DUAL-CASE
FRAMEWORK FOR UNDERGROUND MINING AND INDUSTRIAL PUMP
SYSTEMS

MÉMOIRE
PRÉSENTÉ
COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN INGÉNIERIE

PAR
AYNAZ MOHAMMADI AGHBASH

DÉCEMBRE 2025

Résumé (version française)

Ce mémoire présente un cadre méthodologique fondé sur l'analyse des données pour la maintenance prédictive des systèmes industriels, à travers deux études de cas complémentaires combinant simulation et données industrielles réelles. L'objectif est de démontrer comment les techniques d'apprentissage automatique et d'analytique avancée peuvent améliorer la prise de décision en maintenance et la fiabilité des équipements dans des environnements mécaniques complexes.

La première étude de cas porte sur les opérations minières souterraines, un contexte marqué par une disponibilité limitée de données historiques. Afin de répondre à cette contrainte, un cadre basé sur la simulation est développé pour générer des données opérationnelles synthétiques représentatives des processus miniers. Ces données sont utilisées pour modéliser le comportement des systèmes, identifier des schémas de dégradation et évaluer des stratégies de maintenance prédictive dans des environnements à faible volume de données.

La seconde étude de cas concerne une application industrielle réelle dans le domaine de la fabrication et de la réparation de pompes. L'analyse s'appuie sur des données hétérogènes issues des systèmes ERP, des journaux de maintenance, du contrôle qualité et des transactions commerciales. Des méthodes d'apprentissage non supervisé, incluant K-Means, DBSCAN et les modèles de mélanges gaussiens, sont appliquées afin de réaliser le profilage des clients et d'identifier des segments pertinents pour la planification de la maintenance.

Les résultats conduisent à la proposition d'un cadre intégré reliant la performance des équipements aux profils comportementaux des clients, soutenant une transition vers une maintenance proactive guidée par les données.

Dedication

I dedicate this work to those who supported me, believed in me, and gave me the strength to move forward, this achievement would not have been possible without them.

To my supervisor, thank you for seeing me. You gave me a chance, one that changed everything. You offered me the opportunity to discover myself, to explore my ideas, and to grow. Your kindness, your understanding, and your unwavering support, especially in the most difficult moments, meant more than I can express. You stood by me and made sure I never felt alone in this journey. For that, I will always be deeply grateful.

To my mother, you saw the potential in me before anyone else, with a kind of instinct only a mother can have. You pushed me to reach it, nurtured me with confidence, and guided me toward the right path. Your belief in my abilities, your love, and your support helped shape who I am today.

To my father, your love and support have always been unconditional. By being the hero of my life, you showed me that every impossible could become possible if I had you beside me. Because of you, I was never afraid to dream big, to fight for what I wanted, or to step into the unknown. You made me fearless, and for that, I will always be grateful.

To my husband, my safe place, my constant, my rock. You stood by me in both the best and the hardest moments. You believed in me, often more than I believed in myself. You gave me the space and the support to pursue my goals. Your presence made this journey easier, and I am deeply thankful for everything you've done along the way.

And to every **woman**, and **every person**, who lives under limitation, silence, or oppression: I hope you find your way, your strength, and the chance to be seen. May you have the opportunity to realize your potential, just as I had.

Acknowledgments

I would like to express my deepest gratitude to everyone who contributed to the completion of this thesis and supported me throughout this journey.

First and foremost, I sincerely thank my supervisor, **Hatem Mrad**, for allowing me to explore my ideas and grow as both a researcher and a person. Your trust, constant support, and genuine kindness, especially during the most difficult times, meant more than words can express. Without your guidance, this work would not have been possible.

I would also like to thank my co-supervisor, **Chahid Ahabchane**, for his generosity, patience, and unwavering support. You were always approachable, available, and incredibly helpful throughout every step of the process. I truly appreciated your feedback, openness, and everything you taught me. It was a privilege to work with someone as knowledgeable and encouraging as you.

My heartfelt thanks also go to my **family**, whose encouragement and belief in me laid the foundation for everything I have accomplished. I am especially grateful to my parents for their sacrifices and unwavering support, and to my husband for his constant encouragement and strength throughout this experience.

I extend my sincere appreciation to the **Université du Québec en Abitibi-Témiscamingue (UQAT)** for providing the academic environment and support that made this research possible.

Finally, I wish to thank the pump manufacturing company, **Technosub**, for providing access to their datasets and for their collaboration during this project. Their contribution played a key role in the development and application of this research in a real-world industrial context.

To all those who stood by me, thank you.

Table of Contents

Résumé (version française).....	i
Dedication	ii
Acknowledgments	iii
Table of Contents	iv
List of Figures	ix
List of Tables	xi
List of Abbreviation	xii
List of Symbols	xiv
Summary	xv
CHAPTER 1	1
1. General Introduction	1
1.1. Research Context	3
1.2. Problematic	5
1.3. Research Hypotheses	7
1.3.1. Research Hypothesis (Case Study No. 1: Underground Mining Simulation)	7
1.3.2. Research Hypothesis (Case Study No. 2: Industrial Pump Systems)	8
1.5. Objectives	10
1.5.1. Specific Objectives for Case Study No. 1	11
1.5.2. Specific Objectives for Case Study No. 2	13
CHAPTER 2	16
2. Literature Review.....	16

2.1. Case Study 1: Simulation framework for underground mining systems .	16
2.1.1 <i>Intelligent systems in underground mining operations</i>	16
2.1.2 <i>Predictive maintenance and data-driven logic</i>	17
2.1.3 <i>IoT integration and real-time monitoring</i>	18
2.1.4 <i>Environmental unpredictability and synthetic environmental modeling ...</i>	18
2.1.5 <i>Waste, risk cascades, and predictive simulation</i>	20
2.1.6 <i>Communication reliability and hybrid modeling approaches</i>	21
2.1.7 <i>Predictive maintenance in hazard-prone domains</i>	22
2.2 Case Study 2: Customer profiling and predictive maintenance in industrial pump systems	23
2.2.1 <i>Introduction to data-driven maintenance and client profiling</i>	23
2.2.2 <i>Customer segmentation: From marketing analytics to maintenance contexts</i>	23
2.2.3 <i>Interpretable clustering for industrial client profiling</i>	24
2.2.4 <i>Data scarcity and the value of synthetic augmentation</i>	25
2.2.5 <i>Gaps identified in prior work</i>	25
2.2.6 <i>Machine learning and deep feature models for predictive maintenance</i> ...	26
2.2.7 <i>Health classification in centrifugal pump systems</i>	27
2.2.8 <i>Semantic modeling and narrative data integration</i>	27
2.2.9 <i>Comparative industrial case studies and their strategic relevance</i>	28
2.3 Integrated synthesis and contribution	29
2.4 Summary of gaps and research contributions	31
CHAPTER 3	33

3. General Methodology.....	33
3.1 Methodological approach	33
3.2 Overview of case study selection.....	35
3.3 Data sources and preparation strategies.....	36
<i>Case Study 1: Synthetic data environment.....</i>	<i>36</i>
<i>Case Study 2: Industrial client data</i>	<i>37</i>
3.4 Common tools and analysis techniques used	38
<i>Programming environment and tools.....</i>	<i>39</i>
<i>Data validation and visualization</i>	<i>40</i>
3.5 Validation and Evaluation Criteria	40
<i>Unsupervised clustering evaluation.....</i>	<i>41</i>
<i>Outlier sensitivity and model robustness</i>	<i>41</i>
<i>Practical validation and expert feedback</i>	<i>41</i>
CHAPTER 4	43
Case Study 1: Synthetic Dataset and Automation Framework for Underground Mining	43
4.1 Objective of the Case Study	43
4.2 Synthetic Dataset Design and Contextual Framework.....	44
4.3 Simulation Design and Synthetic Data Generation	47
4.4 Modeling the Process and Building the Framework.....	55
4.5 Outcomes and Observations	58
CHAPTER 5	61
5. Case Study 2: Customer Profiling and Failure Prediction in Pump Systems .	61

5.1 Company Background and Industrial Context.....	61
5.2 Dataset Description and Preprocessing.....	62
5.3. Feature Engineering and Data Normalization	64
<i>Feature Selection and Construction.....</i>	<i>65</i>
<i>Data Normalization and Encoding.....</i>	<i>65</i>
5.4 Clustering and client classification (K-Means, DBSCAN, GMM)	68
<i>Rationale for Algorithm Selection.....</i>	<i>68</i>
<i>Mathematical Foundation</i>	<i>69</i>
<i>K-Means Clustering</i>	<i>69</i>
<i>DBSCAN.....</i>	<i>69</i>
<i>Gaussian Mixture Models (GMM)</i>	<i>70</i>
<i>Determining the Optimal Number of Clusters.....</i>	<i>71</i>
<i>Integrated Clustering Framework.....</i>	<i>72</i>
5.5 Evaluation of Client Segments and Behavior Patterns.....	73
<i>Segments Identified by Clustering Methods.....</i>	<i>73</i>
<i>K-Means Clustering Results</i>	<i>73</i>
<i>Gaussian Mixture Models (GMM) Results.....</i>	<i>79</i>
<i>Comparative Evaluation of Clustering Techniques.....</i>	<i>82</i>
<i>Practical Implications for Predictive Maintenance</i>	<i>83</i>
<i>Strategic Recommendations.....</i>	<i>84</i>
5.6 Predictive Modeling for Failure Detection	84
5.6.1. <i>Objective and Strategic Context.....</i>	<i>84</i>
5.6.2. <i>Contribution of the current Work.....</i>	<i>85</i>

5.6.3. <i>Framework for Future Failure Prediction</i>	85
5.7 Design and Proposal of an Operational Framework for Customer Insight and Maintenance Planning	86
5.7.1. <i>Framework Components and Current Contributions</i>	87
5.7.1.1. <i>Client Classification and Behavioral Segmentation</i>	87
5.7.1.2. <i>Dataset Examination and Data Quality Assessment</i>	87
5.7.1.3. <i>Strategic Framework for Enabling Predictive Modeling</i>	89
5.7.2. <i>Operational and Strategic Value</i>	90
5.8 Data Quality Assessment and Recommendations for Improvement	92
5.9 Comparative Insights from Similar Industrial Case Studies	95
5.10 Strategic Value of the Proposed Approach	98
CHAPTER 6	101
6. Conclusion and Future Work	101
6.1 Summary of Key Contributions	101
6.2 Research Limitations	102
6.3 Practical Recommendations	104
References	110

List of Figures

Figure 1. <i>Conceptual overview of thesis objectives and solution strategies</i>	15
Figure 2. <i>Data flow for Case Study 1: Synthetic mining simulation datasets</i>	37
Figure 3. <i>Data flow diagram for Case Study 2 showing the preprocessing pipeline applied to fragmented industrial datasets, culminating in a structured dataset for unsupervised clustering and client behavior analysis.</i>	38
Figure 4. <i>Conceptual Network Diagram of Interrelated Synthetic Mining Datasets</i> .	46
Figure 5. <i>Sample Synthetic environmental conditions data</i>	48
Figure 6. <i>Sample synthetic equipment maintenance and operational status log</i>	49
Figure 7. <i>Sample synthetic equipment maintenance and operational status log</i>	49
Figure 8. <i>Overview of Sample Synthetic Timeline and details of ground support methods used</i>	Error! Bookmark not defined.
Figure 9. <i>Sample synthetic dimensions of various galleries log</i>	Error! Bookmark not defined.
Figure 10. <i>Overview of Sample Synthetic of rock types and their geological descriptions</i>	51
Figure 11. <i>Sample Synthetic of Summary of significant events, their impacts, and resolutions in mining galleries</i>	51
Figure 12. <i>Sample synthetic of safety metrics log with values and related observations for various mining galleries</i>	52
Figure 13. <i>Sample Synthetic of Time stamped sensor data showcasing various measurements like humidity, voltage, pressure, and distance</i>	52
Figure 14. <i>Sample Synthetic of Sensor readings log with timestamps and metrics</i>	53
Figure 15. <i>Sample Synthetic of Geospatial coordinates with altitude data</i>	53
Figure 16. <i>Network Diagram of Interconnected Operational Datasets</i>	54
Figure 17. <i>Conceptual diagram of dashboard architecture and visualization tools.</i> .	57
Figure 18. <i>Dataset integration process</i>	67
Figure 19. <i>Elbow method score plot</i>	73
Figure 20. <i>Silhouette method score plot</i>	73

Figure 21. <i>Client distribution across K-Means clusters</i>	74
Figure 22. <i>Client distribution across DBSCAN clusters</i>	77
Figure 23. <i>GMM clusters overlapping diagram</i>	79
Figure 24. <i>Comparative clustering coverage across K-Means, DBSCAN, and GMM</i>	83
Figure 25. <i>Example of zero or extremely small cost values</i>	88
Figure 26. <i>Example of negative values in "QTY" and cost columns</i>	89
Figure 27. <i>Example of lack of categorization for item numbers</i>	89
Figure 28. <i>Example of extreme cost values</i>	89
Figure 29. <i>Integrated three-phase predictive maintenance framework for industrial pump systems.</i>	91
Figure 30. <i>Data Quality Challenges, Analytical Impact, and Recommended Solutions.</i>	95

List of Tables

Table 1. <i>Comprehensive Data Suite Provided by Industrial partner</i>	62
Table 2. <i>Datasets quality evaluation results</i>	63
Table 3. <i>Overview of feature engineering</i>	66
Table 4. <i>K-Means clustering result</i>	75
Table 5. <i>DBSCAN Clustering results</i>	78
Table 6. <i>GMM Clustering results</i>	81
Table 7. <i>Overview of algorithm comparison</i>	82
Table 8. <i>Summary of key data quality issues and their impacts</i>	107

List of Abbreviation

Abbreviation	Full Term
AI	Artificial Intelligence
ARIMA	Autoregressive Integrated Moving Average
CNN	Convolutional Neural Network
CRM	Customer Relationship Management
CRISP-DM	Cross-Industry Standard Process for Data Mining
CTGAN	Conditional Tabular Generative Adversarial Network
DBN	Deep Belief Network
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
GMM	Gaussian Mixture Model
GRU	Gated Recurrent Unit
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
IoT	Internet of Things
KPI	Key Performance Indicator
LCA	Life Cycle Assessment
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MLP	Multi-Layer Perceptron
ML	Machine Learning
MOO	Multi-Objective Optimization
MSE	Mean Squared Error
OPTICS	Ordering Points To Identify the Clustering Structure
PCA	Principal Component Analysis
PdM	Predictive Maintenance
R ²	Coefficient of Determination
RFM	Recency, Frequency, Monetary
RF	Random Forest
RNN	Recurrent Neural Network
RMSE	Root Mean Square Error
RTIMS	Real-Time Information Management System
SciPy	Scientific Python (library)
SD	System Dynamics
Seaborn	Python data visualization library
SMOTE	Synthetic Minority Oversampling Technique
SOP	Standard Operating Procedure (unique identifier for repair/service record)
SVC	Support Vector Classification
SVM	Support Vector Machine
TQRNN	Transformer Quantile Regression Neural Network

VSCode	Visual Studio Code
NumPy	Numerical Python
Pandas	Python data manipulation library
Matplotlib	Python plotting library
NetworkX	Python network/graph analysis library

List of Symbols

Symbol	Definition
μ	Mean (average)
σ	Standard deviation
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
MSE	Mean Squared Error
R^2	Coefficient of determination
X	Feature matrix
y	Target variable
k	Number of clusters
n	Sample size or number of observations

Summary

This thesis presents a dual-case analytical framework for predictive maintenance in industrial systems, addressing two distinct data realities: synthetic simulation in underground mining and real-world enterprise data in pump manufacturing. The research applies the CRISP-DM methodology to ensure systematic development, from data preparation to modeling and evaluation.

In the first case study, a modular synthetic data environment was developed to simulate operational subsystems in underground mining, including equipment maintenance, environmental conditions, ground support, and incident communication. The simulation architecture enabled the design and testing of predictive logic, anomaly detection, and safety dashboards in a data-constrained environment, offering a foundation for the future deployment of intelligent automation systems.

The second case study focuses on customer profiling and predictive maintenance using historical repair, sales, and cost data from an industrial pump manufacturer. After rigorous data preprocessing and feature engineering, unsupervised learning algorithms; K-Means, DBSCAN, and Gaussian Mixture Models, were used to segment clients based on behavioral indicators such as repair frequency, monetary impact, and usage patterns. The resulting classifications enabled strategic insights for maintenance planning, resource allocation, and data governance.

This work contributes to the field of intelligent maintenance systems by demonstrating that actionable insights can be derived even in the absence of real-time sensor data or in the presence of fragmented records. The proposed framework emphasizes data quality, interpretability, and scalability, supporting industrial applications where predictive maintenance must be both technically grounded and behaviorally informed.

Keywords

Predictive maintenance; Data processing; Underground mining systems; Customer profiling; K-Means; DBSCAN; Gaussian Mixture Model (GMM); CRISP-DM; Intelligent maintenance systems.

CHAPTER 1

1. General Introduction

The advancement of predictive maintenance (PdM) and intelligent automation systems has become increasingly critical in industrial sectors characterized by high operational complexity, limited data visibility, and costly unplanned downtime. As industries undergo digital transformation, the reliance on machine learning, data engineering, and simulation technologies has shifted from optional innovation to operational necessity. Predictive maintenance has demonstrated significant potential in reducing maintenance costs, extending equipment life, and minimizing service interruptions by enabling early detection of failures based on data trends and behavioral insights. Traditionally, predictive maintenance approaches have been predominantly equipment-centric, focusing on sensor data, component degradation patterns, and failure signatures at the asset level. While effective in well-instrumented environments, such models often overlook the influence of operational context and customer-specific usage behaviors that significantly affect maintenance demand and failure occurrence. In contrast, the client-informed strategy adopted in this study extends beyond asset condition monitoring by incorporating customer behavior, service interaction patterns, and operational practices as predictive variables. This shift enables maintenance analytics to move from purely reactive asset health assessment toward a more contextualized, behavior-aware decision-making framework. However, the implementation of PdM systems in real-world engineering environments remains a challenge due to inconsistent data quality, limited sensor coverage, lack of behavioral modeling, and the absence of structured failure histories. This thesis is built on the recognition that for predictive maintenance to be effective and scalable in real-world applications, two core limitations must be addressed systematically: (1) the scarcity or inaccessibility of high-fidelity operational data in safety-critical environments such as underground mining, and (2) the insufficient integration of customer behavior into maintenance analytics in

industrial settings. The thesis proposes a dual-case framework that tackles both issues using distinct but methodologically aligned approaches. The first case study focuses on developing a simulation-based framework for underground mining systems, where real-time data collection is severely constrained by difficult conditions, regulatory restrictions, and limited system instrumentation. In such settings, access to live data streams for equipment health monitoring or environmental modeling is often unfeasible. To overcome this barrier, a synthetic data environment is constructed, simulating operational behaviors across four critical subsystems: equipment maintenance, environmental monitoring, ground support integrity, and incident communication. These datasets are structured using interdependent logic to reflect real operational causality and support modular system development. This simulation serves as a framework for anomaly detection models, dashboard logic, safety monitoring rules, and predictive maintenance scenarios in a controlled, risk-free environment that mirrors the complexity of underground operations. The second case study addresses the inverse challenge: the availability of industrial data that is fragmented, poorly structured, and underutilized. In collaboration with an industrial pump manufacturing and repair company, this case study focuses on leveraging enterprise datasets, including repair histories, sales transactions, failure logs, and cost records to classify customers based on their maintenance and operational behaviors. The approach employs unsupervised clustering algorithms such as K-Means, DBSCAN, and Gaussian Mixture Models (GMM) to segment clients and uncover hidden patterns that influence equipment failure and service demand. Unlike traditional equipment-centric PdM models, this client-informed strategy introduces behavioral segmentation as a predictive variable, enabling more precise and cost-effective maintenance planning. The clustering framework is further evaluated for its interpretability and strategic utility, offering enterprise's actionable insights for client management and service optimization. Together, the two case studies form a coherent and complementary exploration of how simulation and real-world data can be harnessed to support predictive maintenance in industrial systems. The first case establishes the

methodological foundations and system modeling techniques necessary in data-constrained environments, while the second operationalizes machine learning strategies for customer-specific predictive analytics. Both are implemented within the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, ensuring methodological rigor and facilitating consistent development phases from data preparation to evaluation. This integrated research contributes to the field of intelligent maintenance systems by (i) developing a high-fidelity synthetic simulation framework for underground mining, (ii) introducing interpretable, unsupervised customer profiling for industrial pump maintenance, and (iii) proposing a modular, scalable analytical pipeline that can adapt to different data realities across engineering domains. It addresses both technical and organizational gaps in current PdM research and lays the foundation for future predictive models that are behaviorally informed, operationally grounded, and industrially deployable.

1.1. Research Context

The research presented in this thesis is situated at the intersection of engineering, data science, and digital transformation, with a specific focus on predictive maintenance in safety-critical and operationally demanding sectors. Predictive maintenance (PdM) has gathered increasing attention in recent years due to its capacity to reduce unplanned downtime, optimize resource allocation, and extend asset lifecycles. Despite its proven benefits in theory, the practical adoption of PdM across industrial environments remains fragmented, especially in sectors where either the lack of structured operational data or the underutilization of customer-specific behavioral information limits the effectiveness of data-driven decision-making. For instance, in underground mining, the deployment of intelligent monitoring and predictive systems is slowed down by harsh environmental conditions, sensor reliability issues, safety regulations, and limited access to continuous operational data. These constraints not only restrict the feasibility of direct model training using live datasets but also delay the

development and validation of automated systems for anomaly detection, environmental risk assessment, and equipment failure forecasting. As a result, mining operations are often forced to rely on rule-based or interval-based maintenance strategies that lack adaptability and predictive capability. Moreover, there exists a critical gap in the simulation and testing of these systems under realistic yet controllable conditions, a gap that synthetic data generation and system-level simulation can effectively address. In parallel, manufacturing and service-intensive industries such as industrial pump systems face a different but equally critical challenge: while large volumes of customer and equipment data are collected, ranging from repair records and sales transactions to failure logs and service invoices, these datasets are rarely structured or analyzed in a way that supports predictive maintenance or customer-centric operational optimization. Maintenance strategies are often generalized, ignoring the variation in usage behavior, purchasing patterns, and failure histories across different client types. This leads to inefficiencies in service delivery, increased risk of unexpected breakdowns, and suboptimal allocation of technical resources.

Recognizing these challenges, this thesis is grounded in two distinct yet complementary industrial contexts: one constrained by data scarcity and operational risk (underground mining), and the other characterized by available but structurally fragmented data (industrial pump repair and maintenance). The first context required the creation of a synthetic, modular simulation framework to model environmental, mechanical, and communication subsystems in underground mining. The second context involved a collaborative study with Technosub, a pump manufacturer and repair company, to transform their historical enterprise data into an actionable customer profiling system that supports predictive maintenance planning. To ensure methodological rigor and consistency across these contrasting environments, both case studies were developed using the CRISP-DM methodology. CRISP-DM provides a structured framework that aligns business understanding with data preparation,

modeling, evaluation, and deployment considerations. Its iterative structure was particularly suited for this research, which required a balance between theoretical design (in the case of simulation) and practical, data-driven insights (in the case of industrial profiling). The application of CRISP-DM allowed for the creation of modular, replicate workflows capable of adapting to the distinct requirements of both case studies. These case studies are methodologically unified under a common goal: to develop data-driven maintenance frameworks that are robust, interpretable, and scalable across different operational realities. The dual-case structure provides a unique opportunity to explore predictive maintenance from both ends of the data maturity spectrum: one where simulation compensates for absent field data, and one where machine learning algorithms are used to organize and interpret complex real-world data. By combining synthetic simulation with real-world analytics, it addresses the current disconnect between theoretical maintenance models and their practical deployment in industrial environments. It also reflects the growing need for flexible, domain-specific solutions that incorporate both technical and behavioral dimensions into predictive maintenance strategy, an approach that is especially relevant in asset-intensive, high-variability industries such as mining and manufacturing systems.

1.2. Problematic

Despite significant advancements in artificial intelligence, sensor technology, and data analytics, the practical implementation of predictive maintenance (PdM) in industrial systems continues to face critical challenges. These challenges are particularly critical in environments where operational data is either difficult to obtain or not structured in a way that supports scalable, interpretable, and real-time decision-making. As a result, many industrial actors continue to rely on outdated maintenance strategies, such as time-based or reactive approaches, that are not well-suited to modern equipment complexity and operational variability. In underground mining environments, the primary challenge lies in the absence of high-quality, real-time data. Safety constraints,

environmental unpredictability, and limited sensor coverage make it difficult to monitor system behavior under normal and abnormal conditions. This lack of empirical data delays the development of intelligent automation systems, including real-time dashboards, risk detection modules, and predictive algorithms. Furthermore, without sufficient validation, it becomes impossible to assess the reliability and robustness of PdM models in these settings. Current research lacks simulation frameworks that can realistically mimic underground operational complexity and allow for controlled testing of predictive and decision-support systems. On the other hand, in industrial sectors such as pump manufacturing and repair, data is often available in large volumes, but suffers from fragmentation, inconsistency, and semantic ambiguity. Historical repair logs, cost entries, and sales records are rarely integrated into unified analytical pipelines. As a result, companies miss the opportunity to link customer behavior, such as usage patterns, purchase history, and failure tendencies to maintenance planning. Traditional PdM implementations in such contexts tend to treat all equipment and clients uniformly, without accounting for the intermixture of behavioral risks. This leads to inefficient maintenance scheduling, poor resource allocation, and reduced customer satisfaction. A further limitation in the existing literature is the lack of interpretable models that balance predictive accuracy with industrial transparency. While deep learning and black-box models have demonstrated high performance in controlled environments, they offer limited explainability, an essential requirement for stakeholders in asset-intensive industries. There is also a notable gap in the integration of customer segmentation logic into PdM workflows. Most studies treat failure prediction and customer profiling as separate tasks, missing the strategic value of linking operational behavior to predictive insights. This thesis addresses these problems by proposing two complementary solutions: (1) a modular simulation framework using synthetic datasets to evaluate PdM frameworks in underground mining environments where real data is unavailable, and (2) a client segmentation and profiling system based on unsupervised machine learning to structure and extract value from historical repair and transaction records in an industrial pump context. By

applying the CRISP-DM methodology across both use cases, this research ensures methodological consistency while adapting to distinct data realities. The main problem tackled here is the absence of integrated, interpretable, and domain-adapted predictive maintenance strategies that can operate effectively under both data-scarce and data-fragmented conditions.

1.3. Research Hypotheses

This thesis is based on the principle that predictive maintenance (PdM) systems in industrial settings can be significantly improved by adapting the methodology to the constraints and realities of different operational environments, specifically, using high-fidelity simulation when real-world data is absent and through customer behavior-based segmentation when real data exists but is unstructured or incomplete. The research therefore formulates two specific hypotheses aligned with the two distinct case studies, each of which explores a critical but underserved aspect of PdM in mechanical systems.

1.3.1. Research Hypothesis (Case Study No. 1: Underground Mining Simulation)

Hypothesis:

A simulation-based predictive maintenance framework, built from synthetic datasets that represent interdependent subsystems in underground mining operations (e.g., environmental monitoring, equipment health, ground support, and incident communication), can replicate operational variability with sufficient technical accuracy to enable the testing of diagnostic logic, risk modeling, and decision-support strategies in the absence of real-time sensor data.

This hypothesis is grounded in the understanding that underground mining systems are inherently complex and constrained by factors such as harsh environmental conditions, difficult sensor deployment, and regulatory safety limitations. As a result, there is a

lack of consistent, high-resolution operational data. The assumption here is that by designing a modular and scientifically informed simulation environment, it might become possible to model failure mechanisms, test alert thresholds, evaluate system interdependencies, and visualize responses under controlled, repeatable conditions. The goal is to validate whether this simulation can serve as a functional test bed for predictive algorithms and system behavior analysis, thereby compensating for the data scarcity typically encountered in underground operations.

1.3.2. Research Hypothesis (Case Study No. 2: Industrial Pump Systems)

Hypothesis:

In industrial maintenance settings where data is available but suffers from fragmentation, inconsistent formatting, and semantic ambiguity, it is possible to transform this data into actionable engineering insights by applying a structured preprocessing and feature engineering pipeline. When coupled with unsupervised learning methods such as clustering, this approach can enable the identification of meaningful client behavior profiles that not only support future predictive maintenance strategies but also expose underlying data quality issues and guide improvements in the structure and governance of enterprise datasets.

In this case study, the challenge is not the absence of data, but its lack of analytical readiness. The company possesses repair records, transaction logs, and cost data, yet these datasets are not originally designed for modeling or decision support. Key limitations include missing timestamps, inconsistent item descriptions, lack of structured failure codes, and unlinked records across systems and there is no access to any real-time operational data. These deficiencies inhibit the development of predictive maintenance tools and hinder the company's ability to differentiate between reliable, high-risk, or high-value clients. The hypothesis driving this study is that by applying rigorous preprocessing, including normalization, imputation, reduplication, and

categorical harmonization, these datasets can be reconstructed into a coherent analytical structure. Feature engineering based on customer interactions (e.g., repair frequency, cost accumulation, time intervals, and purchase behavior) then provides a foundation for unsupervised clustering using models such as K-Means, DBSCAN, and GMM. The resulting customer profiles reveal patterns that are not immediately observable in raw data and allow for segmentation strategies that can inform tailored maintenance planning, risk assessment, and client prioritization. Additionally, the analysis serves a dual function: it not only enables client classification but also identifies systemic issues in data management. By highlighting data gaps, redundancies, and structural limitations, the research provides a roadmap for improving data quality over time. The result is both a functional prototype for client-aware predictive maintenance and a set of practical guidelines to improve data utility and decision-making within the organization. Furthermore, this approach supports the secondary goal of diagnosing and documenting the systemic limitations of the current data architecture. Through this analysis, the study aims to propose a framework for improved data governance, including recommendations for standardized field structures, timestamp integration, and unified client identifiers. In doing so, the research provides both a technical foundation for predictive modeling and a strategic roadmap for enhancing data readiness, ensuring that future decision-making processes are informed, client-specific, and aligned with operational realities. This hypothesis addresses the challenge faced by maintenance-intensive industrial sectors where abundant but poorly structured data is generated through repair logs, client transactions, and service records. While conventional PdM approaches typically focus on equipment condition monitoring alone, this hypothesis introduces the idea that client behavior, captured through structured features such as repair frequency, recency, monetary impact, and purchase activity can serve as a predictive variable. It assumes that clustering can classify clients into distinct risk or behavior groups, which in turn can inform strategic maintenance scheduling, resource planning, and even early-warning diagnostics for equipment failures linked to usage profiles. Together, these two

hypotheses challenge the traditional assumptions that predictive maintenance either (a) requires real-time operational data or (b) can rely solely on equipment-level indicators.

1.5. Objectives

The principal objective of this thesis is to investigate and implement technically grounded, data-driven methodologies that enable predictive maintenance planning across two industrial contexts with contrasting data realities. The study aims to demonstrate that, even in the absence of real-time sensor data or the presence of fragmented and inconsistently structured historical data, it is possible to develop actionable maintenance insights through the application of structured engineering analytics. The principal goal is to bridge the gap between theoretical maintenance modeling and operational implementation by tailoring methodological approaches to the specific limitations of each case environment.

In the context of underground mining systems, the objective is to design a high-fidelity simulation framework capable of modeling critical subsystems, including equipment degradation, environmental variability, ground support integrity, and communication reliability, under conditions where empirical data is either inaccessible or non-existent. The intention is to generate a synthetic dataset that preserves realistic operational interdependencies and system-level causality. This framework is intended not as a replacement for real data, but as a controlled environment for testing predictive logic, visualizing system behavior under abnormal conditions, and validating modular decision-support components. The simulation will be evaluated for its ability to represent time-dependent fault propagation, sensor anomalies, and interlinked subsystem responses, providing a foundation for risk-aware maintenance planning in underground operations.

In the second context, focused on industrial pump systems, the objective is to transform existing enterprise data, comprising repair records, customer transactions, and cost

histories into a structured analytical form that supports customer-specific predictive maintenance strategies. This involves developing a preprocessing pipeline to address inconsistencies, missing values, and semantic misalignment in the raw datasets, followed by the application of unsupervised learning techniques to identify clusters of client behavior. The core hypothesis is that meaningful segmentation of clients, based on engineered features such as repair frequency, cost magnitude, and transaction intervals, can reveal underlying maintenance risk profiles. These profiles can then inform strategic decisions about service prioritization, warranty planning, and long-term customer engagement. In parallel, the analysis is used to diagnose systemic weaknesses in data quality, leading to recommendations for data architecture improvements, including standardization protocols, classification development, and identifier consolidation that can support future machine learning deployment.

Overall, this thesis aims to contribute a methodological and practical framework for predictive maintenance in industrial mechanical systems, adapted to both data-constrained and data-fragmented environments. It does so by combining simulation engineering, data preprocessing, behavioral modeling, and machine learning within a unified analytical structure, guided by the CRISP-DM methodology. The result is a validated approach for enabling predictive insight even when conventional maintenance datasets are insufficient, misaligned, or entirely unavailable.

1.5.1. Specific Objectives for Case Study No. 1

The objective of Case Study No. 1 is to develop a modular simulation framework that enables the study and validation of predictive maintenance strategies in underground mining environments where empirical operational data is either unavailable, unreliable, or incomplete. Given the difficult nature of underground systems and the technical difficulties associated with sensor deployment and real-time monitoring, the simulation must be capable of approximating critical operational behaviors across multiple

interdependent subsystems. These include equipment health degradation, environmental parameter fluctuations (such as gas levels, temperature, and humidity), ground support integrity, and communication failures during incident response.

The framework is designed to generate synthetic datasets that are both technically coherent and structurally realistic. These datasets must reflect not only individual parameter distributions but also the dependencies and cascading effects between subsystems that are typical in underground mining operations. For example, a simulated environmental degradation should trigger corresponding synthetic anomalies in equipment status, ground support stress readings, or communication signal integrity. This level of integration ensures that the simulation environment is suitable for testing system logic under conditions that would be difficult or dangerous to replicate in the field.

The objective also includes implementing data preprocessing procedures on the synthetic datasets to replicate the noise, missing values, and formatting inconsistencies commonly observed in real industrial data. This allows for the validation of data preparation pipelines, anomaly detection algorithms, and visualization dashboards under conditions that resemble actual operational constraints. A key focus is the development of a modular architecture where each subsystem (e.g., sensors, maintenance logic, and communication protocols) is designed as an independent yet connected unit, enabling targeted testing of maintenance thresholds, alert logic, and control mechanisms. The simulation output must be compatible with downstream analytics, including statistical modeling and machine learning workflows, to ensure that the platform can serve as a foundation for future predictive algorithm development. Furthermore, the system should incorporate time-resolved data structures, enabling the analysis of sequential patterns and temporal relationships that are essential for failure forecasting.

Ultimately, the goal of this case study is to demonstrate that a well-designed synthetic environment can act as a representation for data-deficient operational systems, providing a scientifically valid platform for evaluating predictive maintenance strategies, designing decision-support tools, and exploring the interplay between mechanical reliability, environmental stressors, and system-level resilience in underground mining operations.

1.5.2. Specific Objectives for Case Study No. 2

The objective of Case Study No. 2 is to transform a fragmented and operationally unstructured industrial dataset, comprising pump repair records, customer transactions, and associated cost data, into a reliable analytical foundation for client behavior modeling and future predictive maintenance deployment. The data originates from a real-world pump manufacturing and service company, and although extensive in volume, it exhibits characteristics that are typical of enterprise environments not originally designed for advanced analytics: inconsistent categorical entries, missing values, ambiguous identifiers, lack of standardized failure categorization, and the absence of real-time operational data. The first objective is to reconstruct analytical continuity within the dataset through systematic preprocessing and feature engineering. This includes identifying and correcting structural issues such as duplicated records, mismatched formats, non-uniform labels, and semantic inconsistencies across datasets. The data is then integrated across multiple sources (e.g., sales logs, repair tracking files, cost summaries) using cross-referencing keys such as customer name, SOP ¹ numbers, and serial identifiers. The aim is to build a coherent, client-level operational history that accurately reflects the lifecycle of service interactions, usage frequency, and financial impact.

¹ Standard Operating Procedure (a unique identifier assigned to each repair or transaction record)

The second objective is to engineer a set of behavioral features that quantify customer interaction with the company in a way that is relevant for predictive maintenance planning. These features include repair frequency, average repair interval, total normalized cost, purchase behavior, and derived relations that capture trends in maintenance intensity versus acquisition activity. These indicators are normalized, transformed, and encoded to allow for compatibility with clustering algorithms.

The third objective is to apply and compare unsupervised learning algorithms, namely K-Means, DBSCAN, and Gaussian Mixture Models (GMM), to segment the customer base into behaviorally distinct profiles. The hypothesis is that these clusters will reveal groups with differing maintenance burdens, spending patterns, reliability levels, and temporal usage characteristics. The interpretability of these segments is central to their utility: rather than treating all clients uniformly, the goal is to provide the company with an evidence-based classification of customers, identifying which ones represent higher risk, lower return, or strategic value.

A further objective is to assess the limitations of the current dataset and to provide clear, technically justified recommendations for improving its long-term analytical value. This includes proposing an optimal structure for data capture and storage, defining field requirements such as timestamps and standardized identifiers, and recommending failure categorization schemes that support both diagnostic precision and model training. By doing so, the research not only delivers immediate operational insight but also contributes to the design of a scalable data infrastructure capable of supporting future machine learning applications.

Finally, the results of the clustering analysis and data quality assessment are used to formulate strategic recommendations for the company. These include guidelines for targeted maintenance planning, client prioritization strategies, and resource allocation frameworks based on data-derived risk profiles. The overall objective is to establish a

technically robust, interpretable, and context-aware foundation that enables the company to evolve from reactive maintenance practices to behavior-informed predictive decision-making.

A visual summary of the main objective, associated challenges, methodological approaches, and specific focus areas of both case studies is presented in Figure 1. This diagram illustrates how the two distinct contexts are addressed through a unified data-driven maintenance framework guided by the CRISP-DM methodology.

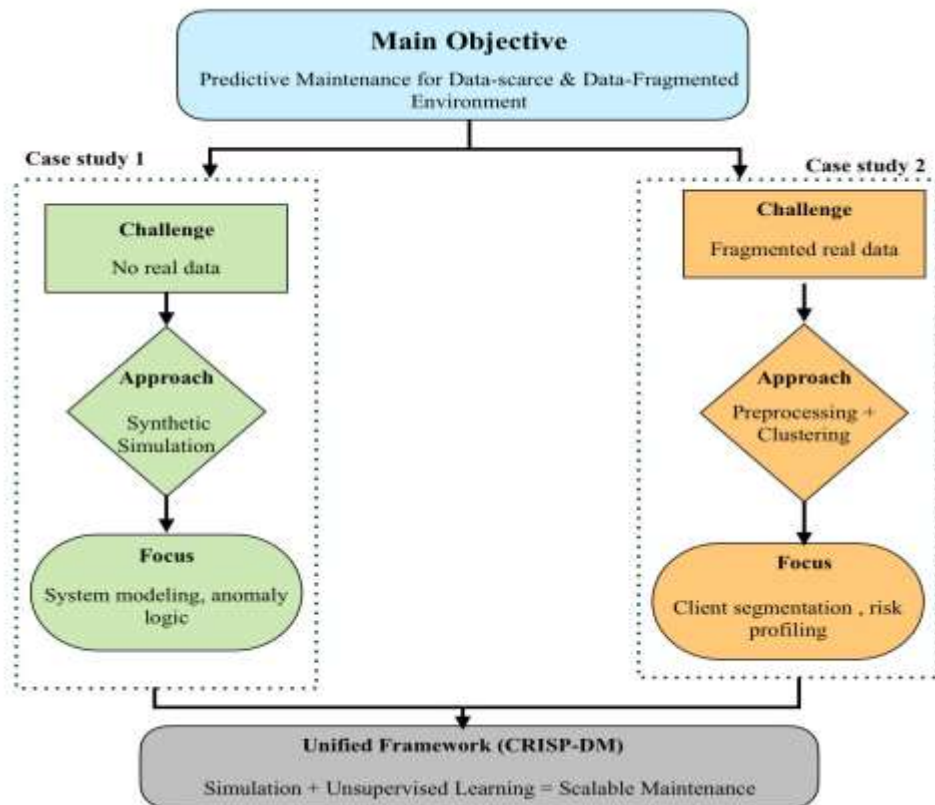


Figure 1. Conceptual overview of thesis objectives and solution strategies

CHAPTER 2

2. Literature Review

2.1. Case Study 1: Simulation framework for underground mining systems

2.1.1 Intelligent systems in underground mining operations

Recent advances in underground mining have increasingly leveraged artificial intelligence (AI), robotics, and modular decision-making systems to address the challenges of hazardous and data-scarce environments. AI is now considered a foundational element across a wide range of mining subsystems, including drilling automation, blast optimization, safety monitoring, and operational control. Liu et al. [1] review this evolution in depth, emphasizing how machine learning and sensor fusion are now embedded in smart drilling rigs, automated roof bolters, and environmental scanning systems. These tools enable the continuous assessment of rock mechanics, microseismic patterns, and atmospheric quality, providing the real-time insight required for both operational safety and cost efficiency. The integration of AI into mining infrastructure is not only a matter of optimizing equipment usage; it also provides the logic required to enable predictive analytics at scale. Raza [2] explores how sensor data from heavy machinery, support systems, and communication networks can be used to identify fault patterns, classify risk levels, and trigger preventive actions. This perspective supports the architectural choice in Case Study 1 to simulate synthetic streams for various underground subsystems; from environmental concentrations and mechanical strain to communication outages, using modular AI components capable of detecting anomalous patterns. The increasing intelligence of mining systems also stems from the integration of Real-Time Information Management Systems (RTIMS). According to Shimaponda-Nawa et al. [3], RTIMS frameworks are central to enabling responsive, data-driven decision-making across the mining value chain. Their maturity model for RTIMS deployment illustrates the importance of sensor interoperability, data

freshness, and user interaction dashboards. In this simulation framework, these concepts are implemented through synthetic indicators that illustrate the scope of dynamic interplay between sensor readings, decision thresholds, and intervention triggers.

2.1.2 Predictive maintenance and data-driven logic

Predictive maintenance (PdM) plays a crucial role in modern mining strategies, especially in environments that the access is hard, where traditional inspection routines are either dangerous or infeasible. Raza [2] highlights the transition from rule-based maintenance (e.g., periodic or reactive) to predictive systems that analyze historical and real-time data to forecast component degradation. These models rely on supervised and unsupervised learning methods applied to vibration signals, temperature variations, and performance metrics, all variables simulated within this synthetic maintenance logic. The technological and organizational sprints to implementing PdM are also substantial. Meyer zu Wickern [4] outlines several key limitations, including data scarcity, inconsistent failure labeling, and the high cost of initial sensor infrastructure. These constraints justify the use of synthetic datasets as a stand-in for operational testing, especially when developing and validating PdM algorithms for underground environments. Further application of PdM principles is demonstrated in Fuente et al. [5], who introduce a hierarchical inference network for mobile mining equipment using TinyML-enabled edge nodes. Their architecture distributes predictive reasoning across sensors, gateways, and cloud components, reducing latency and optimizing energy consumption. As noted in their results, this system achieved more than 90% classification accuracy with minimal computational load. The simulation of embedded AI logic at sensor level, particularly for environmental spikes or equipment malfunction, is directly supported by this design, allowing us to model local threshold-based detection alongside remote anomaly-scoring modules.

2.1.3 IoT integration and real-time monitoring

The convergence of IoT, edge computing, and mining safety systems is central to the operational realities of modern underground sites. Zhang et al. [6] review recent deployments of IoT infrastructure in mines, including gas monitors, geolocation tags, and equipment usage sensors. They identify several persistent barriers, such as “signal weakening and communication loss in confined spaces,” and “sensor failure due to harsh environments, high humidity, and corrosive gases” [6]. These observations justify the decision to simulate sensor degradation and network inconsistencies as probabilistic events within this research synthetic framework. Moreover, their emphasis on the importance of real-time simulation, as they argue: “Simulation tools and digital twins offer an efficient way to test such systems before deployment” [6].

This directly supports the motivation behind Case Study 1, to use synthetic simulation as a safe and analytically rich environment for modeling communication failures, alarm thresholds, falling alerts, and dashboard logic. The work of Bisschoff and Grobbelaar [7] further reinforces the importance of operational data readiness. They assess the implementation of data-driven decision-making across mining companies, revealing major inconsistencies in how sensor data is collected, interpreted, and applied. These findings validate the modular structure of this research simulation engine, which is built to isolate, link, and visualize dependencies between environmental inputs, mechanical systems, and safety alerts.

2.1.4 Environmental unpredictability and synthetic environmental modeling

Environmental unpredictability in underground mines is inherently multidimensional. It includes gas emissions, temperature variations, humidity shifts, and changes in structural integrity, all of which interact nonlinearly. Young et al. [8] provide a formal statistical definition of unpredictability as a combination of residual variance, autocorrelation (referred to as “noise colour”), and stationery of environmental signals.

These parameters translate directly into the synthetic datasets used in Case Study 1. For example, their typology of stochastic processes; white noise (random), pink/red noise (temporal memory), and brown noise (drift), is applied to simulate environmental transitions within different operational states (e.g., stable ventilation vs. degradation during drilling). Furthermore, the authors differentiate between two mechanisms for detecting environmental change: discrete “ancestral cues” and cumulative statistical learning. The simulation architecture in case study 1, incorporates both, embedding discrete triggers (e.g., cave-ins, sensor dropouts) and probabilistic anomaly scoring derived from time-series prediction errors [8]. This theoretical grounding supports the argument that environmental unpredictability cannot be captured through static thresholds. Instead, the model in case study 1 reflects shifting baselines, dynamic risk escalation, and real-time reclassification of threat levels, elements that would be unfeasible to study in operational mines without risking safety or disrupting operations. Widana [9] deepens this argument by reviewing the interconnected environmental effects of mining, including air pollution, structural instability, hydrogeological shifts, and ground deformation. These factors often interact across spatial and temporal scales, creating risk chains that are not immediately visible through isolated sensor readings. The framework in case study 1, by contrast, simulates environmental dependencies explicitly. This system-of-systems modeling fills a critical research gap identified by Widana [9], who critiques the absence of integrated environmental monitoring frameworks in current mining practices. Rahnema et al. [10] further emphasize the necessity of simulation and synthetic modeling when evaluating short-term mine planning. Their review highlights the operational blind spots that emerge from insufficient integration of environmental emissions into daily decision-making. By using Life Cycle Assessment (LCA) and Multi-Objective Optimization (MOO), the authors propose embedding emissions data (such as Scope 1/2 CO₂-equivalent outputs) into the mine scheduling process. However, they acknowledge the lack of high-resolution, real-time emissions data that would make such integration possible. The synthetic simulation in case study 1 directly addresses this challenge by generating

customizable emissions profiles, tracking interrelated KPIs (e.g., energy use, equipment downtime, environmental conditions), and providing a flexible platform for testing optimization algorithms under realistic environmental scenarios [10].

2.1.5 Waste, risk cascades, and predictive simulation

The risks associated with underground operations extend beyond immediate safety concerns and into long-term waste and risk propagation. Tayebi-Khorami et al. [11] argue that waste in mining is no longer a passive byproduct but a dynamic risk factor that can trigger failures through cumulative effects. Their review of investigations, mine water, and waste rock disposal reveals that most failures emerge from “sequential, dependent events” rather than singular causes. Importantly, Tayebi-Khorami et al. [11] identify simulation-based tools as crucial for proactive waste management, especially in underground settings where real-time data is sparse, and risk accumulation is non-obvious. They call for modular decision-support tools capable of integrating multiple variables and temporal feedback, the sort of the architecture implemented in Case Study 1, where environmental, mechanical, and communication subsystems interact via a networked logic engine that simulates delayed effects and trigger thresholds [11]. The argument for using synthetic simulation is further reinforced by Servin et al. [12], who develop a digital twin architecture using distributed particle simulation for mine-to-mill tracking. Although their domain focuses on material flow, their methodology of assigning state variables to modular “particles” mirrors the use of synthetic datapoints to simulate equipment condition, atmospheric status, and alarm status across simulated time intervals. Their emphasis on modularity (“blocks of independently updated systems”) directly validates the proposed approach in case study 1, especially in creating a framework where subsystems such as gas sensors, ventilation, and communication lines can be independently modeled but causally linked.

2.1.6 Communication reliability and hybrid modeling approaches

In the context of underground mining, reliable communication systems are critical for operational safety and coordination. Bazargur et al. [13] offer a rare, quantitative reliability study of a multi-shaft underground communication network. Their findings highlight baseline availabilities as low as 62% in non-redundant setups, increasing to over 85% with dual standby configurations. These figures were instrumental in defining the probabilistic structure of the synthetic simulated communication system, which includes configurable redundancy levels and variable maintenance scenarios. Moreover, the study employs a hybrid modeling framework by combining Markov chains with system dynamics (SD) using tools like Vensim. This methodological precedent supports this thesis' decision to couple discrete state-transition logic (for component status and failure modes) with continuous-time simulations (for recovery and degradation processes). In particular, the authors' use of Weibull and Poisson distributions to model component lifetimes informed the selection of statistical profiles when generating synthetic maintenance and failure logs [13]. Their link between communication failures and evacuation delays, modeled using Dijkstra's algorithm to simulate rerouted escape paths, also aligns with this approach to cascading risk assessment. Scalambrin et al. [14] extend this reliability perspective into energy-efficient IoT networks using LoRa multi-hop communication. Their synchronization mechanism enables ultra-low duty cycles (under 2%) while maintaining transmission accuracy. These insights support this decision to simulate partial data loss, transmission lag, and packet synchronization as dynamic parameters within this thesis synthetic communication subsystem. Their emphasis on real-time synchronization accuracy also justifies the modeling of “alert delay” and “sensor-reporting offset” variables in this thesis dashboarding.

2.1.7 Predictive maintenance in hazard-prone domains

As noted in the broader industrial literature, predictive maintenance has matured across various sectors but remains underutilized in mining. The review by Dayo-Olupona et al. [15] explores this gap, emphasizing that data-driven predictive models, despite their proven utility, are seldom deployed in underground settings due to safety concerns, equipment diversity, and lack of operational transparency. This reinforces the fundamental motivation of this work which tries to create a testable, risk-free simulation environment where PdM strategies can be prototyped using realistic, multi-domain synthetic data. Their classification of PdM approaches into rule-based vs. data-driven categories allows this thesis to position its synthetic framework as an enabler of the latter. By simulating high-resolution, time-indexed logs of equipment wear, environmental drift, and communication disruptions, this platform allows machine learning models to be tested in scenarios that would be either infeasible or dangerous in the field. Additionally, Bisschoff and Grobbelaar [7] observe that even when PdM systems are installed, their effectiveness is limited by poor data integration and lack of feedback mechanisms between operational events and long-term asset strategies. The studies reviewed in this section collectively constitute form a comprehensive foundation for the simulation framework developed in Case Study 1. They validate the theoretical constructs of environmental unpredictability, the operational logic of modular interdependencies, the statistical modeling of component failures, and the architectural need for predictive, testable systems in underground mining. Importantly, the literature also identifies significant implementation gaps, such as the lack of high-resolution environmental modeling [9], the limited integration of emissions in short-term planning [14], which this thesis seeks to address through simulation and synthetic data design.

2.2 Case Study 2: Customer profiling and predictive maintenance in industrial pump systems

2.2.1 Introduction to data-driven maintenance and client profiling

As industrial sectors shift toward digital transformation, there is a growing emphasis on predictive maintenance (PdM) frameworks that are both cost-effective and operationally robust. In this context, combining customer behavior profiling with maintenance analytics has emerged as a novel yet underexplored strategy for enhancing maintenance decision-making. Traditional maintenance methodologies, such as reactive or scheduled preventative maintenance have proven insufficient in highly variable operational environments like mining and manufacturing, where equipment degradation and failure patterns differ significantly among clients. The integration of unsupervised clustering models to segment clients by maintenance behavior and failure risk offers a scalable framework for intelligent maintenance planning. While customer segmentation techniques have been extensively applied in marketing domains [21, 22], their deployment in PdM frameworks remains limited. Clustering algorithms have shown potential in uncovering operational patterns within industrial datasets [23, 24, 25], but these methods are rarely tailored to repair-driven segmentation or integrated with failure prediction models. This literature review explores current gaps and supports the rationale for a unified approach that uses interpretable clustering and predictive modeling to derive client-specific maintenance strategies from large-scale repair, transaction, and failure logs.

2.2.2 Customer segmentation: From marketing analytics to maintenance contexts

The traditional RFM (Recency, Frequency, and Monetary) model has long served as a foundation for customer profiling in business-to-consumer (B2C) environments. Wu et al. [26] extended this framework by integrating it with Principal Component Analysis (PCA) and K-Means clustering, effectively segmenting clients based on transaction

histories. This strategy enabled the identification of high-value, inactive, or at-risk customer groups. Although originally developed for online retail data, the RFM-based segmentation strategy translates well into maintenance-driven contexts, especially where historical repair records, service frequency, and total expenditures offer insights into equipment usage and degradation behavior. In this thesis, RFM-like attributes, such as repair recency, frequency of visits, and normalized service costs, are extracted from transactional datasets to build customer profiles. These profiles are then clustered to detect underlying behavioral patterns. Wu et al.'s [26] framework provides methodological justification for using PCA to balance feature weighting and K-Means to group clients with similar risk profiles.

2.2.3 Interpretable clustering for industrial client profiling

In contrast to black-box AI models, interpretable clustering algorithms are more suitable for industrial decision-making, where model transparency is essential. K-Means, DBSCAN, and GMM have distinct strengths. K-Means provides simple and reproducible cluster boundaries, making it ideal for clients with well-separated behavior patterns [23]. DBSCAN excels at identifying irregular or outlier behaviors in noisy datasets, a property particularly relevant for clients with unpredictable maintenance schedules or high variance in failure records [24, 27]. GMM adds probabilistic flexibility by modeling cluster distributions, allowing for overlap between similar but distinct behavioral classes. Costa et al. [43] demonstrated the value of using clustering as a feature enhancement step before supervised learning. Their hybrid approach, which combined K-Means, GMM, HDBSCAN, and OPTICS with Support Vector Classification (SVC) and Random Forest models, yielded a nearly 5% improvement in accuracy when applied to compressor maintenance datasets. This strategy validates the thesis' decision to build an integrated profiling pipeline that starts with clustering and feeds into downstream failure prediction logic. Additionally, the combined framework in this study by leveraging the interpretability of K-Means, the

outlier detection of DBSCAN, and the flexibility of GMM, creates a multi-layered profile system. This structure reflects the operational diversity found in industrial clients and addresses gaps in prior literature, where single-method clustering was often insufficient to capture complex behaviors.

2.2.4 Data scarcity and the value of synthetic augmentation

In many industrial domains, including pump repair systems, failure events are rare, making it difficult to train robust predictive models using supervised learning alone. Khosravi et al. [44] addressed this challenge by implementing synthetic data augmentation using Conditional Tabular GANs (CTGAN) and SMOTE to enhance class balance and improve model performance in failure prediction. Their results demonstrate that artificially balancing datasets enables more consistent anomaly detection and classification outcomes. This insight reinforces the validity of using synthetic failure data for underrepresented customer profiles in this thesis. In future work, incorporating CTGAN-generated repair scenarios could support failure prediction for clients with sparse histories, thereby expanding the robustness of the client classification model.

2.2.5 Gaps identified in prior work

Despite the substantial contributions from recent studies, several critical gaps persist. First, the integration of customer behavior into predictive maintenance frameworks remains largely unaddressed. Existing studies either emphasize client segmentation for CRM [21, 25] or focus narrowly on equipment-level failure prediction [28, 38], without linking the two domains. Secondly, while clustering models have been applied to detect patterns in operational datasets [23, 24], few works have used combined clustering approaches to strengthen robustness and explainability. Lastly, the lack of structured pipelines for managing repair, service, and transaction logs limits the applicability of ML-driven segmentation in PdM planning which is the issue that this thesis tries to

resolve through a standardized preprocessing and normalization workflow (see Chapter 5).

2.2.6 Machine learning and deep feature models for predictive maintenance

As PdM evolves beyond simple rule-based systems, the role of machine learning in industrial diagnostics has grown considerably. Wuest et al. [47] provided a comprehensive analysis of how machine learning, especially Support Vector Machines (SVM), Random Forests (RF), and clustering methods can improve manufacturing efficiency. They emphasized the necessity of robust data preprocessing (e.g., feature selection, missing value handling, and normalization) before model training. Their findings reinforce this thesis' emphasis on structured data cleaning and normalization to prepare pump-related client data for clustering and downstream predictive modeling. In more advanced fault detection scenarios, deep learning techniques have demonstrated superior accuracy in extracting hidden fault signals from noisy sensor streams. Zhao et al. [49] reviewed the application of autoencoders, deep belief networks (DBNs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs) for industrial equipment health monitoring. While this thesis prioritizes interpretable models due to data format limitations, the insights from Zhao et al. [49] support potential future extensions that incorporate stacked autoencoders or CNN-based encoders once high-frequency time-series data (e.g., vibration sensors, acoustic readings) from pump systems become available. Further refinement of deep sequential modeling was introduced by Zhang et al. [50], who proposed a Gated Recurrent Unit (GRU) model for diagnosing faults in rotating machinery. Their method converts vibration signals into two-dimensional image representations, which are then fed into GRUs to learn temporal dependencies. GRUs were shown to outperform CNNs and MLPs under variable and noisy conditions, making them ideal for sparse datasets typical in maintenance logs. These findings inform the potential transition in this thesis

toward time-series modeling of client repair patterns using GRU-based methods, particularly if chronological repair logs are included in the predictive pipeline.

2.2.7 Health classification in centrifugal pump systems

Given the mechanical complexity and operational criticality of centrifugal pumps, specialized PdM models are necessary. Mallioris et al. [46] designed a tailored predictive framework for centrifugal pump diagnostics using RF, SVM, Naïve Bayes, and XGBoost, leveraging multistage condition monitoring data. Their use of Spearman correlation analysis for feature selection underscores the importance of statistical dependency metrics when prioritizing input variables. In this thesis, a similar strategy has been applied, employing correlation-based filters during preprocessing to ensure that only meaningful features (e.g., cost trend, visit frequency, and downtime) are considered during clustering and classification. Mallioris et al.'s [46] emphasis on class balancing and the need to mitigate majority-class bias directly informs the design of this study's failure/no-failure binary classifier. Their methodology strengthens the argument for ensuring data stratification across various client types to avoid over fitting predictive models to dominant behavior patterns.

2.2.8 Semantic modeling and narrative data integration

Guo et al. [36] explored a novel dimension in predictive intelligence by transforming spoken diagnostic narratives into structured semantic clusters. Their methodology, based on anchor concept detection and lexical consensus, allows for knowledge extraction from unstructured text sources, such as technician speech or repair notes. In practice, this transformation relies on identifying recurring domain-specific anchor concepts (e.g., component names, failure symptoms, or operational conditions) within verbal or textual reports, followed by normalization and grouping of semantically similar expressions. These extracted concepts are then embedded into a shared semantic space and clustered based on similarity, enabling heterogeneous narratives to

be mapped into consistent, structured representations. As a result, informal technician descriptions can be translated into interpretable semantic clusters that reflect underlying maintenance patterns. This approach opens new avenues for semantic enrichment in industrial maintenance datasets. While currently outside the primary scope of this thesis, such techniques could be used in future work to process free-text service descriptions, which are often the richest source of undocumented client behavior and equipment status insights. In the context of this study, where customer segmentation is based on structured fields like repair cost, SOP date, and maintenance frequency, semantic enrichment provides a potential complementary strategy to enhance model interpretability and uncover context-aware maintenance drivers. This can be particularly useful for client-level anomaly detection, where textual logs might reveal underlying causes for irregular repair behaviors not captured in numeric data.

2.2.9 Comparative industrial case studies and their strategic relevance

Four major case studies further reinforce the need for a multi-model segmentation and predictive framework. In the Z-BRE4K project [42], for example, a semantic maintenance ontology is developed to model system components, failure causes, and diagnostic triggers. Their use of a two-layered vocabulary and context-based modeling ensures traceability and system-wide insight, but their approach lacks a behavioral layer to differentiate among client types. This thesis addresses that gap by profiling client behavior alongside system diagnostics, providing a dual-layered intelligence architecture that integrates both semantic structure and unsupervised clustering. Costa et al. [43] built a hybrid clustering-classification system for high-pressure compressors, in which unsupervised models enhanced the input space for supervised classifiers. Their demonstration of a 5% accuracy gain validates this thesis' architectural decision to use clustering as a foundational module, rather than a secondary enhancement in developing intelligent client profiles. Khosravi et al. [44] introduced a synthetic failure event generation strategy using CTGAN and SMOTE, designed to combat dataset

imbalance in rare-event scenarios. The relevance of this work lies in its demonstration that failure augmentation improves recall and model resilience. This is especially relevant when modeling clients who have limited failure histories but might still pose a high maintenance risk. Poland et al. [45] proposed the Transformer Quantile Regression Neural Network (TQRNN) for industrial health monitoring. Their pipeline captured long-range temporal dependencies and nonlinear anomaly thresholds, a design that mirrors this thesis' goal of creating modular, CRISP-DM-aligned pipelines. While the current work does not deploy transformer models, Poland et al.'s [45] structured pipeline logic inspires this study's emphasis on phase-based modeling, explainability, and risk stratification.

2.3 Integrated synthesis and contribution

The convergence of unsupervised clustering, predictive maintenance frameworks, and customer profiling in industrial settings remains a relatively unexplored yet highly promising research frontier. The studies reviewed herein collectively reveal a significant opportunity to bridge the interpretability gap in predictive analytics by introducing explainable, data-driven customer segmentation into maintenance planning workflows. The integration of K-Means, DBSCAN, and GMM in this thesis was strategically chosen based on technical advantages and documented trade-offs in the literature [23, 24, 25, 27]. Each of these models contributes uniquely to client profiling: K-Means offers a fast, interpretable baseline; DBSCAN introduces outlier detection and noise resilience; and GMM enables soft clustering for probabilistic modeling of heterogeneous customer behaviors. In contrast to domain-specific studies that apply these models solely in marketing or retail [21, 22, 25], this work repositions them within a predictive maintenance context for centrifugal pump systems, offering tailored insights into behavioral risk and operational variance. A major limitation in traditional PdM research is the over-reliance on supervised models requiring labeled failure data [27, 28]. Industrial datasets, particularly those involving customer transactions and

equipment repair logs, rarely contain consistent failure annotations or synchronized operational metadata. Studies like those of Wuest et al. [47] and Zhao et al. [49] confirm that data quality and preprocessing are often under-prioritized, leading to poorly generalized models. This thesis directly addresses that issue by designing a preprocessing framework centered on feature normalization, outlier rejection, missing-value strategies, and categorical encoding, prior to executing any clustering or modeling phase. The methodological emphasis on interpretable segmentation also reflects a conscious deviation from black-box deep learning strategies. While GRU-based architectures [50] and transformer-based anomaly models like TQRNN [45] offer superior raw performance, they often lack domain transparency and operational justifiability in industrial settings, where client trust and regulatory traceability are crucial. By comparison, the clustering framework adopted here allows for client-level explainability, helping stakeholders understand why a customer is labeled “high-risk” or “maintenance-prone” based on transparent repair patterns, frequency trends, and cost ratios. From a data augmentation perspective, the use of synthetic repair profiles and potential integration of rare-event simulation strategies, such as CTGAN and SMOTE [44], open a scalable path for addressing imbalance problems in client behavior data. These methods allow for the expansion of minority client segments, ensuring that predictive models are balanced, fair, and capable of generalizing across the full customer spectrum, including those with sparse histories or atypical failure patterns. The semantic structuring introduced by the Z-BRE4K project [42] highlights another conceptual opportunity: incorporating ontology-based tags into clustered customer profiles. While this thesis does not yet integrate ontological hierarchies, it sets a strong foundation for doing so by linking high-level concepts like “urgent repair”, “repeat failure”, or “critical downtime” with structured cluster labels and maintenance risk classes. This could allow for future hybrid systems that blend semantic reasoning with behavioral segmentation. Lastly, the CRM-informed segmentation logic explored in Wu et al. [26], paired with statistical modeling techniques like those in Mallioris et al. [46], and reinforces the thesis’ cross-disciplinary strength: building actionable

maintenance dashboards based on customer repair frequency, cost-normalized patterns, and SOP-linked indicators. These dashboards are designed not only to forecast failure but also to enable strategic resource allocation, such as preventive service campaigns for risk-prone clients or inventory staging based on projected part replacement cycles.

2.4 Summary of gaps and research contributions

Across the reviewed literature, several persistent gaps have been identified that this thesis directly addresses through targeted methodological contributions. First, while predictive maintenance (PdM) frameworks often neglect client-level segmentation, this thesis introduces interpretable, unsupervised customer profiling that is integrated into the PdM logic. In response to the common overreliance on supervised learning despite the scarcity of failure labels, this work constructs predictive models based on unsupervised pattern recognition using clustering methods. The issue of poor data quality and inadequate preprocessing is tackled by developing a modular pipeline for data cleaning, normalization, and outlier rejection. Unlike much of the existing literature, which lacks domain-specific clustering applications in industrial repair contexts, this thesis applies and compares K-Means, DBSCAN, and Gaussian Mixture Models (GMM) to model pump repair behaviors effectively. To bridge the gap between Customer Relationship Management (CRM) insights and predictive maintenance strategies, the study incorporates RFM-style features and SOP-based repair logs to contextualize client risk levels. Although literature on synthetic data generation in maintenance contexts is limited, this thesis outlines a future direction involving CTGAN or SMOTE to augment data for rare repair cases. Finally, the limited interpretability of deep PdM models is addressed by prioritizing explainable clustering outputs, such as centroids, density boundaries, and probabilistic labels ensuring the resulting models remain actionable for industrial stakeholders. By addressing the identified gaps, the second case study contributes to engineering knowledge in three distinct dimensions. Methodologically, it introduces a novel unsupervised

segmentation strategy specifically tailored for pump systems, grounded in explainable clustering logic. From a practical standpoint, it integrates CRM-style customer profiling into maintenance planning, providing a directly applicable framework for B2B industrial settings. Theoretically, it establishes a behavioral foundation for client-aware predictive maintenance, setting the stage for future extensions using semantic metadata and synthetic data augmentation.

CHAPTER 3

3. General Methodology

This chapter presents the general methodological framework adopted throughout the thesis. The research follows a structured, data-driven approach inspired by the CRISP-DM methodology, ensuring consistency across both case studies while allowing for adaptation to different industrial contexts and data availability constraints. The chapter outlines the rationale behind the selected methodology, the criteria for case study selection, the common data preparation strategies, analytical tools, and evaluation methods used. By establishing a unified methodological foundation, this chapter provides the conceptual and procedural link between the research objectives and the detailed analyses presented in the subsequent case study chapters.

3.1 Methodological approach

This research adopts a structured, problem-driven methodology inspired by the Cross-Industry Standard Process for Data Mining (CRISP-DM), a robust framework widely used for industrial and applied machine learning projects. The CRISP-DM model was selected for its flexibility, compatibility with both synthetic and real-world datasets, and its ability to guide end-to-end data science workflows from initial problem definition to deployment readiness. Its phase-based structure allowed this research to manage complexity across two distinct yet complementary case studies: a simulation-based automation framework for underground mining and a customer profiling model for predictive maintenance in pump systems.

The CRISP-DM process consists of six iterative stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. In this study, each phase was tailored to accommodate the specific constraints and opportunities presented by the two case studies. For the simulation-based framework in underground mining, the emphasis was placed on synthetic data generation and

operational modeling. In contrast, the client profiling study prioritized feature engineering, behavioral segmentation, and actionable clustering insights. The business understanding phase clarified the industrial objectives of each case: the need for intelligent decision support tools in hazardous mining environments and the demand for customer informed maintenance planning in the industrial pump sector. These goals shaped the technical and data-driven direction of the research. The data understanding and data preparation phases were approached thoroughly, with an emphasis on data quality assessment, noise reduction, and contextual validation, especially critical in the second case study where data inconsistencies posed significant challenges. In the synthetic case, simulation fidelity and operational practicality were prioritized over simple data volume or diversity. In the modeling phase, distinct strategies were applied based on the nature of each case. For case study 1, modular logical systems and network-based dependency structures were implemented to simulate and validate system behavior. For case study 2, unsupervised machine learning models were developed to extract hidden client behavior patterns and lay the foundation for future predictive modeling. The evaluation stage involved qualitative and quantitative assessments of each model or framework, using appropriate technical validation tools (e.g., network integrity checks, clustering validity indices). The evaluation did not attempt direct cross comparison of the case studies but ensured that each delivered technically coherent and industrially relevant outcomes. While deployment in the strictest sense (e.g., real-time system integration) was beyond the scope of this academic study, deployment readiness was considered in both frameworks through system modularity, scalability of methods, and integration potential with operational data pipelines. The use of CRISP-DM provided a consistent, scalable, and interpretable methodology that bridged theoretical accuracy with applied engineering goals. It enabled the parallel development of simulation tools and client analytics within a unified scientific process, while remaining adaptable to the differing data realities of each case study.

3.2 Overview of case study selection

The case studies addressed in this thesis were defined within the framework of two parallel research initiatives led by the academic supervisor and aligned with ongoing institutional and industrial collaborations.

Case Study 1 builds upon an earlier academic project that explored intelligent automation in underground mining environments. The challenge of limited access to real operational data in such settings necessitated the development of a synthetic data-driven simulation framework. This continuation offered an opportunity to extend and refine a methodological foundation for designing and evaluating automation logic, system modularity, and environmental risk responses within a simulated but realistically structured dataset.

Case Study 2 originated from a collaboration with an industrial partner, a pump manufacturing and repair company (Technosub) seeking to extract more value from its internal data. The project was initiated to address a business-oriented need: developing client profiles to support predictive maintenance strategies. The dataset was provided by the company, and the scope was shaped in accordance with their operational priorities and data availability. This case was assigned as a means of applying machine learning techniques to an authentic industrial dataset, aiming to deliver practical and scalable insights. Although the contrasting nature of topics, one simulation based and exploratory, the other grounded in real world data and industrial application, offered a unique opportunity to explore the methodological spectrum of data-driven engineering. This dual case configuration enabled the thesis to demonstrate the flexibility of applied machine learning and data engineering strategies across different data maturity levels. Importantly, the integration of both cases under a common methodological structure ensures coherence, while their differences highlight the adaptability of the approach. Each case plays a complementary role: one advancing early-stage innovation in

constrained environments, the other operationalizing client informed analytics within an existing enterprise setting.

3.3 Data sources and preparation strategies

This section outlines the origin, structure, and preparatory methods applied to the datasets used across both case studies. Although the data environments were different in the case studies, one being synthetic in Case Study 1 and industrial in Case Study 2, the underlying objective remained consistent: to ensure that all data used in the analysis was reliable, structured, and analytically valid.

Case Study 1: Synthetic data environment

In the absence of accessible real-time mining data due to safety, legal, and operational constraints, a synthetic data environment was developed. The generation process followed a structured approach grounded in domain knowledge derived from previous research, industry documentation, and technical manuals. The simulation targeted four operational domains typical of underground mining: equipment maintenance, environmental conditions, ground support systems, and incident communication logs. Each dataset was designed to mirror realistic behaviors, with built-in interdependencies to reflect real-world operational causality. Data preparation in this context involves raw data cleaning by being focused on designing the synthetic datasets with implementing inconsistencies, missing values, and noise. This approach allowed for the development and testing of data preprocessing workflows that would be used in real-world applications. Standard Python tools were employed to simulate, structure, and validate the datasets, ensuring they represented plausible operational scenarios suitable for modeling and dashboard visualization.

The data flow for the synthetic dataset generation and preparation process is illustrated in figure 2.

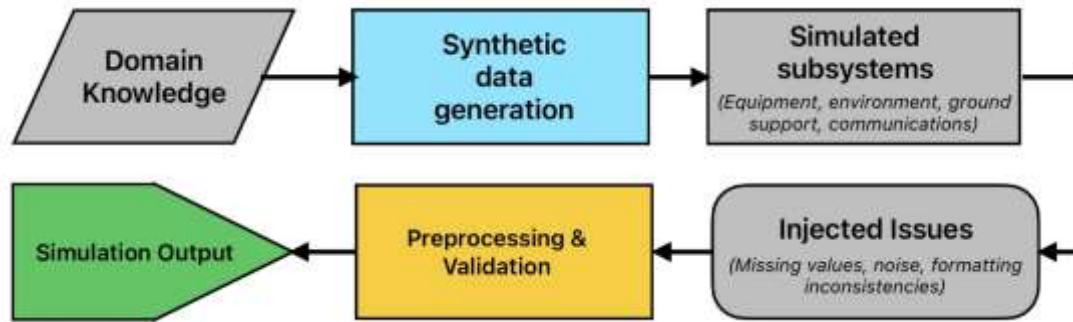


Figure 2. *Data flow for Case Study 1: Synthetic mining simulation datasets*

Case Study 2: Industrial client data

In contrast, the data used in Case Study 2 originated from enterprise-level sources provided by an industrial partner. These datasets encompassed sales records, historical repair logs, failure classifications, and cost tracking entries. As these data sources were not initially structured for analytical purposes, they exhibited significant fragmentation, inconsistencies, and missing values. Addressing these challenges required a systematic, engineering-driven data cleaning and restructuring process. The preparation strategy involved several key steps. First, categorical entries, such as; SOP types, item descriptions, and pump models were normalized to remove redundancy and improve coherence across records. Missing values were handled using context-specific imputation, where median values were applied to numerical fields, and placeholder categories were introduced for absent categorical attributes. Outliers in the data were identified and corrected using statistical methods, including log transformations and quantile-based filters, to ensure that erroneous cost entries did not angle analysis. To consolidate the fragmented information, datasets were integrated using common identifiers such as customer names, pump serial numbers, and SOP references, thereby constructing a unified operational history for each client. The result of this data preparation stage was a validated, multi-source dataset optimized for unsupervised clustering and behavioral pattern analysis. While the synthetic dataset in Case Study 1 prioritized simulation fidelity and controlled variability, the industrial dataset in Case

Study 2 focused on recovering structural integrity and analytic readiness. In both scenarios, data preparation served as a critical enabler, directly influencing the modeling outcomes and revealing key limitations and possibilities related to data quality, operational traceability, and scalability. The adaptability of the preparatory methods to different data environments underscores a core methodological strength of this research.

The industrial data preparation workflow is outlined in figure 3, summarizing key steps from raw dataset integration to clustering readiness.

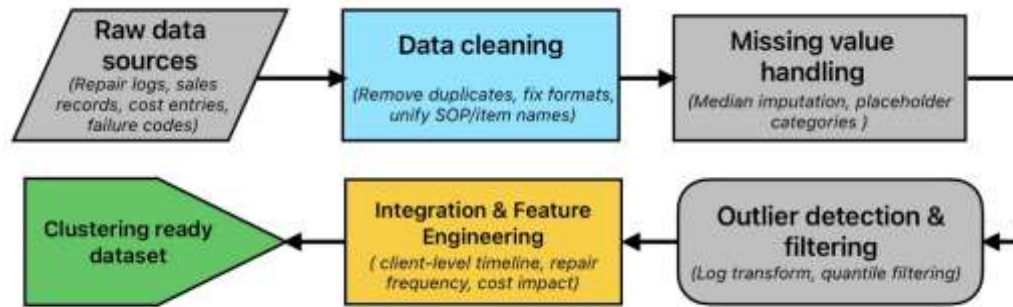


Figure 3. Data flow diagram for Case Study 2 showing the preprocessing pipeline applied to fragmented industrial datasets, culminating in a structured dataset for unsupervised clustering and client behavior analysis.

3.4 Common tools and analysis techniques used

This section presents the software tools, programming environments, and analytical techniques employed across both case studies to preprocess data, perform clustering and modeling, visualize patterns, and interpret operational insights. Despite the contrasting nature of the datasets, synthetic versus industrial, the research maintained a consistent technical load and methodological consistency to ensure reproducibility and reliability of results.

Programming environment and tools

All computational tasks were performed using Python within the Visual Studio Code (VSCode) environment, selected for its flexibility, integration capabilities, and efficient debugging features. Python's broad ecosystem of libraries made it well-suited for the diverse tasks required in both case studies. The primary packages included:

- **Pandas** and **NumPy** for data manipulation and preprocessing.
- **Scikit-learn** for clustering algorithms, classification models, and preprocessing utilities.
- **Matplotlib** and **Seaborn** for data visualization and exploratory data analysis.
- **Plotly** for interactive dashboards and visual reports, especially in Case Study 2.
- **SciPy** for statistical calculations such as outlier filtering, transformations, and hypothesis testing.
- **HDBSCAN** and **Yellowbrick** (when applicable) for enhanced clustering and model evaluation.

This combination of tools allowed for modular, interpretable, and scalable development of analytics pipelines that were tailored to both synthetic simulation and real-world industrial datasets.

Clustering and dimensionality reduction techniques

To identify underlying structures in customer behavior and operational dynamics, a range of unsupervised learning techniques was employed, primarily in Case Study 2. K-Means clustering was selected for its simplicity, interpretability, and computational efficiency in segmenting clients based on normalized metrics such as repair frequency, purchase behavior, and associated costs. DBSCAN was introduced to address the limitations of centroid-based methods by enabling the detection of outliers and capturing irregular groupings, which proved particularly effective in handling noisy

and imbalanced industrial datasets. To further refine segmentation, Gaussian Mixture Models (GMM) were applied, offering a probabilistic framework that accommodated overlapping client profiles and allowed for a more nuanced interpretation of behavioral clusters. Prior to clustering, Principal Component Analysis (PCA) was conducted to reduce the dimensionality of the feature space while preserving key variance components, thereby enhancing both visualization and model performance. These methods were evaluated using quantitative metrics such as silhouette scores, measures of cluster compactness, and qualitative assessments based on operational interpretability. The comparative analysis guided the selection of optimal clustering strategies for both client profiling and the development of predictive maintenance models.

Data validation and visualization

Visualization played a critical role throughout both case studies, serving as a central tool for exploratory data analysis and post-modeling interpretation. It was employed to investigate the distribution of features and detect anomalies or outliers that could distort clustering results or misrepresent client behavior. These visual representations facilitated the assessment of the quality and relevance of clusters, particularly in terms of how well they captured meaningful operational patterns. In addition, visualizations were used to communicate key behavioral trends and failure dynamics across different customer profiles and timeframes. Each plot was designed with the specific analytical objective in mind, enhancing the understanding of complex relationships within the data, supporting hypothesis formulation, and informing the generation of actionable insights tailored to the industrial context of the study.

3.5 Validation and Evaluation Criteria

This section defines the criteria and methodologies used to assess the quality, robustness, and practical relevance of the analytical outputs generated in both case

studies. Given the applied nature of this research, particularly in the context of predictive maintenance and operational profiling, validation was structured around both quantitative metrics and domain-specific relevance.

Unsupervised clustering evaluation

Since clustering algorithms such as K-Means, DBSCAN, and GMM were central to the research, their outputs were evaluated using multiple internal validation metrics, such as Silhouette score or elbow method. These metrics were applied consistently across all models, allowing objective comparison of clustering quality. In addition, visual evaluation provided interpretability and was used to confirm metric-driven decisions.

Outlier sensitivity and model robustness

For algorithms like DBSCAN and HDBSCAN, noise point detection rates were analyzed to understand the algorithm's sensitivity to anomalous patterns. Outliers were validated against known operational anomalies (e.g., unusually high repair costs or inconsistent purchase behaviors) where available. The robustness of clustering results was further tested by varying hyperparameters (e.g., `eps`, `min_samples`, `k`) and observing the stability of cluster assignments across runs. This ensured that the derived client groupings were not artifacts of specific tuning choices.

Practical validation and expert feedback

For the industrial dataset in Case Study 2, the analytical outputs, including client segmentation results and failure predictions, were reviewed by domain experts within the company. Their feedback was influential in validating the interpretability and operational relevance of the identified clusters, ensuring that the groupings reflected meaningful distinctions observed in actual client behavior and maintenance records.

This expert input facilitated the identification of limitations related to data quality, such as incomplete records or missing contextual information that could influence model accuracy. This practitioner-oriented validation step was essential for confirming that the findings were not only statistically sound but also applicable in real-world operational settings, reinforcing the practical value and reliability of the methodology.

This chapter defined the overarching methodological approach guiding the thesis and detailed the common principles applied across both case studies. By structuring the research around a unified data-driven framework, the methodology ensures methodological coherence while accommodating the specific characteristics of simulated and real industrial datasets. The framework established in this chapter serves as the basis for the application-oriented analyses developed in Chapters 4 and 5, where the methodology is implemented and evaluated within distinct industrial contexts.

CHAPTER 4

Case Study 1: Synthetic Dataset and Automation Framework for Underground Mining

This chapter presents the first case study, which focuses on the development of a simulation-based framework for predictive maintenance in an underground mining context. The objective of this case study is to investigate how synthetic data can be generated and exploited to support maintenance analysis in environments where access to real operational data is limited. The chapter introduces the industrial context, describes the simulation design and data generation process, and outlines the analytical steps applied to extract meaningful patterns from the synthetic dataset. This case study serves as a controlled experimental environment to validate the feasibility of data-driven maintenance strategies prior to real-world deployment.

4.1 Objective of the Case Study

The primary objective of this case study is to design, develop, and evaluate an intelligent automation system tailored for underground mining environments, with a particular focus on tasks related to equipment management and ground support operations. Given the challenges of obtaining real-world operational data, such as privacy concerns, operational safety constraints, and data unavailability, this study employs a carefully constructed synthetic dataset to simulate realistic underground mining conditions.

The case study specifically aims to achieve the following goals:

- Develop a modular and scalable intelligent system framework capable of handling real-time decision-making and automation tasks.

- Create realistic synthetic datasets representing key operational categories, including equipment maintenance logs, ground support conditions, environmental sensor readings, and incident communication reports.
- Use synthetic data to carefully test and refine system logic, ensuring readiness for eventual integration of actual sensor and operational data.
- Design conceptual diagrams and dashboard prototypes to effectively visualize data interactions, system behavior, and user interactions in critical operational scenarios.
- Provide foundational insights and methodological guidelines for future developments involving real-world data integration, predictive analytics, and machine learning capabilities.

By achieving these objectives, this case study sets a robust foundation for transitioning from simulation-based evaluations to practical deployments, ensuring system adaptability, operational reliability, and enhanced safety protocols in underground mining contexts.

4.2 Synthetic Dataset Design and Contextual Framework

Given the lack of access to authentic operational datasets from underground mining operations, this case study focuses on building a technically sound, scientifically structured synthetic data environment. The goal is to create a representative foundation for simulating underground mining scenarios and validating intelligent support systems. The development of this synthetic framework is rooted in a detailed study of literature, previous research, and documented industry practices. The synthetic data generated for this study is not just a placeholder, but a functional simulation of the types of data required for real-time analytics and operational decision-making in high-risk mining contexts. This work emphasizes designing a modular and interdependent data structure that reflects realistic operational complexity. By simulating these

conditions with high fidelity, the framework lays the groundwork for future implementation and testing once live sensor data becomes available.

The synthetic datasets were modeled across four primary operational domains that are critical for underground mining systems: equipment maintenance, environmental monitoring, ground support integrity, and incident communication. Instead of being designed in isolation, these datasets were carefully linked together to reflect the interdependencies observed in active mining sites. The operational structure for each domain is summarized below:

- **Equipment Maintenance Logs** simulate maintenance cycles, error codes, operational timelines, and performance degradation patterns. These were designed based on industry equipment profiles and manuals.
- **Environmental Parameters**, such as temperature, gas concentrations, and humidity, were generated using realistic baseline conditions and stress scenarios (e.g., gas spikes, ventilation drops). Although no physical sensors were used, the values reflect documented ranges from mining studies.
- **Ground Support Integrity** data includes structural properties such as bolt tension, geological layer classification, and support type. The data reflects the interaction between geological characteristics and reinforcement behaviors.
- **Incident and Communication Logs** simulate operational communications during incident escalation and resolution phases. These were designed to simulate the common safety scenarios in mining environments.

A major advancement of this study is the deliberate structuring of datasets to mirror real-world operational logic. For instance, degraded environmental conditions (e.g., rising gas levels) affect the synthetic generation of safety scores and trigger maintenance tasks in the equipment dataset. Similarly, unresolved ground support anomalies influence incident log generation. This logic was formalized using a network

graph model to visualize dependencies and validate systemic interactions, as illustrated in Figure 4. The diagram highlights how environmental parameters influence safety metrics and incident reporting, how safety conditions propagate into maintenance actions, and how ground support integrity is linked to both incident escalation and equipment-level responses.

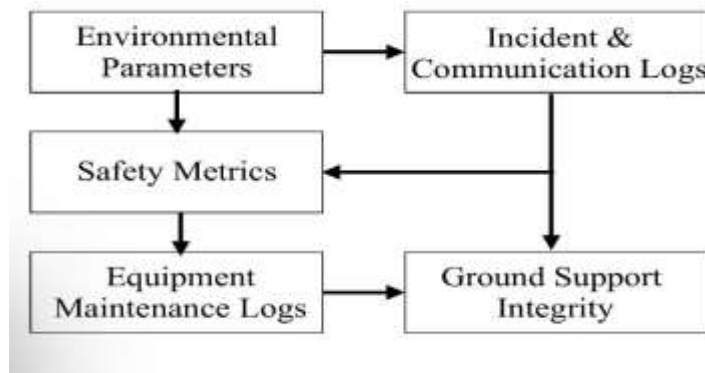


Figure 4. Conceptual Network Diagram of Interrelated Synthetic Mining Datasets

This approach ensures that future analytics or simulations performed on this data are grounded in realistic assumptions about operational interconnectivity. It also supports meaningful development of predictive maintenance, risk modeling, and operator decision support systems.

The synthetic dataset design presented here offers three key contributions:

- It introduces a structured methodology for simulating operational data in safety-critical industrial domains.
- It provides a ready to use environment for testing algorithms and decision support systems without relying on sensitive or inaccessible real-world data.
- It formalizes operational interdependencies, elevating the practicality of the datasets compared to isolated synthetic approaches.

The framework developed here is transferable to other industrial contexts where data scarcity, safety constraints, or data confidentiality present barriers to innovation. It sets a precedent for future work that requires robust early-phase testing in the absence of field data.

4.3 Simulation Design and Synthetic Data Generation

The simulation design and synthetic data generation are central to creating a reliable testing ground, capable of supporting advanced predictive models, system logic validations, and real-time operational simulations. The initial phase of this process involved systematically identifying the essential categories of data required to realistically simulate underground mining conditions. Extensive literature reviews, analysis of industry standard maintenance practices, technical manuals, geological evaluations, and discussions with domain experts were foundational in this identification process. As a result, a comprehensive framework emerged, clearly outlining the primary operational areas including equipment maintenance, environmental monitoring, ground support integrity, and operational incident communications. The simulation design was structured to ensure that all synthetic datasets-maintained consistency, logical coherence, and realistic interdependencies. The multidimensional nature of these data categories was explicitly addressed, recognizing that real-world mining operations do not exist in isolated database but rather in complex, dynamic, and interconnected contexts. The synthetic data generation process employed Python scripting to methodically create structured datasets reflective of real-world conditions. For example, equipment behavior data were accurately made to capture operational timestamps, error codes reflective of documented equipment failure modes, scheduled maintenance intervals, and realistic component degradation patterns. These synthetic maintenance logs, designed with technical accuracy, provide critical input for testing predictive maintenance algorithms and operational simulations. Environmental condition datasets were generated to reflect realistic

variations commonly encountered in mining operations, despite the absence of physical sensor deployments. Parameters such as gas concentrations, humidity, temperature fluctuations, and ventilation conditions were algorithmically varied to simulate normal operational baselines, gradual environmental shifts, and potential anomaly scenarios like sudden gas leaks or temperature anomalies. These simulations allow for robust validation of the system’s reaction and operational safety measures. Representative samples of the generated synthetic datasets are illustrated in Figures 5 through 15. These figures illustrate the structure, temporal resolution, and variability of the simulated data across environmental conditions, equipment maintenance records, incident communication logs, geological characteristics, sensor measurements, and spatial positioning. Together, they demonstrate how the synthetic data captures both routine operational patterns and abnormal conditions required for validating predictive maintenance algorithms, safety logic, and system-level simulations.

1	Date	Temperature (C)	Humidity (%)	Visibility (m)	Comments
2	4/25/2023	23	68	88	
3	9/10/2023	34	32	86	
4	3/16/2023	11	33	31	
5	12/3/2022	34	32	28	
6	5/28/2023	23	24	10	
7	5/21/2023	38	26	83	
8	3/28/2023	34	13	98	
9	5/9/2023	39	11	68	
10	10/24/2022	-3	25	36	

Figure 5. *Sample Synthetic environmental conditions data*

	A	B	C	D
1	Date	Type	Duration	Message
2	11/6/2022	Pager	20 minutes	
3	2/16/2023	Cell Phone	47 minutes	Equipment malfunction reported in B2 shaft. Engineers dispatched.
4	9/27/2022	Satellite Phone	16 minutes	Communication down in C5.
5	9/21/2023	Walkie-Talkie	26 minutes	Switch to backup channel.
6	3/29/2023	Intercom	9 minutes	Rockfall detected near D6 entrance.
7	1/3/2023	Intercom	53 minutes	Proceed with caution.
8	4/21/2023	Walkie-Talkie	33 minutes	Fire drill scheduled for tomorrow at 10 AM Power outage in B7.
9	12/10/2022	Satellite Phone	34 minutes	Backup generators initiated.
10	4/20/2023	Pager	50 minutes	Water leakage in C2. Pumping teams on-site.
11	1/17/2023	Cell Phone	15 minutes	Heavy rain forecasted. Monitor external exits for flooding.
12	3/30/2023	Intercom	39 minutes	Scheduled maintenance for conveyors in D1.
13	10/3/2022	Cell Phone	5 minutes	Elevated methane levels detected in A1. Ventilation increased.
14	12/24/2022	Cell Phone	11 minutes	Safety inspection for galleries C3 and C4 next Medical team dispatched to B3 for an injured worker.
15	8/4/2023	Pager	45 minutes	Elevator system in A4 offline for 2 hours.
16	10/16/2022	Cell Phone	1 minutes	Air quality reduced in B5. Use masks until further notice.
17	1/18/2023	Walkie-Talkie	10 minutes	Emergency assembly point activated for D2.
18	7/16/2023	Walkie-Talkie	13 minutes	Power restored in C1. Resume operations.
19	12/24/2022	Satellite Phone	45 minutes	New shift timings for D3 crew starting next week.
20	5/17/2023	Pager	2 minutes	Drill in A2 completed successfully.

Figure 6. Sample synthetic equipment maintenance and operational status log

	A	B	C
1	Equipment	Last Maintenance Date	Operational
2	Excavator	12/2/2022	No
3	Bulldozer	8/26/2023	No
4	Truck	11/11/2022	No
5	Drill	9/10/2023	No
6	Excavator	4/16/2023	No
7	Crane	5/7/2023	Yes
8	Bulldozer	7/5/2023	Yes
9	Excavator	9/30/2022	Yes
10	Drill	12/14/2022	No
11	Bulldozer	3/19/2023	No
12	Bulldozer	4/14/2023	No
13	Bulldozer	9/6/2023	No
14	Bulldozer	8/8/2023	Yes
15	Crane	2/4/2023	No
16	Excavator	8/16/2023	Yes

Figure 7. Sample synthetic equipment maintenance and operational status log

	A	B	C
1	Date	Method	Duration
2	11/4/2022	Shotcrete	3 hours
3	12/9/2022	Mesh	3 hours
4	5/26/2023	Mesh	4 hours
5	4/3/2023	Rock Bolts	2 hours
6	8/14/2023	Rock Bolts	1 hours
7	9/1/2023	Shotcrete	1 hours
8	7/4/2023	Rock Bolts	3 hours
9	2/11/2023	Mesh	4 hours
10	8/19/2023	Mesh	3 hours
11	3/6/2023	Shotcrete	3 hours
12	11/12/2022	Mesh	2 hours
13	12/27/2022	Shotcrete	1 hours
14	1/22/2023	Steel Sets	2 hours
15	10/14/2022	Rock Bolts	3 hours

Figure 8. Overview of Sample Synthetic Timeline and details of ground support methods used.

	A	B	C	D	E
1	Gallery Name	Length (m)	Width (m)	Height (m)	
2	Gallery_0	325	19	10	
3	Gallery_1	485	6	11	
4	Gallery_2	901	16	11	
5	Gallery_3	963	17	13	
6	Gallery_4	677	8	16	
7	Gallery_5	647	13	8	
8	Gallery_6	517	13	20	
9	Gallery_7	814	19	20	
10	Gallery_8	415	14	12	
11	Gallery_9	891	15	12	
12	Gallery_10	718	20	6	
13	Gallery_11	266	17	5	
14	Gallery_12	626	17	10	
15	Gallery_13	622	8	5	
16	Gallery_14	824	18	18	
17	Gallery_15	996	17	7	
18	Gallery_16	354	17	10	
19	Gallery_17	220	15	11	
20	Gallery_18	840	18	18	

Figure 9. Sample synthetic dimensions of various galleries log

	A	B	C
1	Date	Rock Type	Description
2	4/23/2023	Igneous	Cooled from magma or lava; interlocking crystals.
3	5/14/2023	Sedimentary	Accumulated and compacted sediments; may contain fossils.
4	10/10/2022	Igneous	Solidified from molten material.
5	11/13/2022	Sedimentary	Result of sediment accumulation and compaction.
6	12/28/2022	Sedimentary	Derived from sediment deposition.
7	5/30/2023	Metamorphic	Recrystallized due to heat and pressure; altered parent rock.
8	8/23/2023	Igneous	Crystallized from magma/lava.
9	4/12/2023	Igneous	Cooled molten material.
10	5/4/2023	Sedimentary	Layers of compacted sediments.
11	1/4/2023	Metamorphic	Modified by heat, pressure, or chemically active fluids.
12	1/15/2023	Igneous	Solidified molten origin.
13	9/21/2023	Sedimentary	Laminated; sediment origin.
14	4/8/2023	Sedimentary	Compressed sediments.
15	8/5/2023	Metamorphic	Heat and pressure-transformed parent rock.

Figure 10. Overview of Sample Synthetic of rock types and their geological descriptions

	A	B	C	D
1	Date	Event	Impact	Resolution
		Elevated methane levels detected, Gallery A3	Risk of explosion increased, Gallery evacuation advised.	Ventilation systems enhanced, Monitoring ramped up.
2	5/18/2023			
		Ground vibration spike, Gallery B2	Possible seismic activity, Potential collapse risk.	Gallery temporarily closed, Structural assessment conducted.
3	12/8/2022			

Figure 11. Sample Synthetic of Summary of significant events, their impacts, and resolutions in mining galleries

	A	B	C	D	E	F	G	H	I	J	K
1	Date	Metric Type	Value	Comments							
2	2/24/2023	Air Quality		90	Noted occasional spikes in ground vibrations in Gallery B2. Might need closer monitoring.						
3	4/11/2023	Noise Levels		82	Temperature rise in Gallery C5 was a temporary anomaly; systems reacted efficiently.						
4	4/3/2023	Vibration Levels		46	Elevated radiation levels in Gallery D1 likely due to natural uranium deposits. Needs frequent checks.						
5	6/17/2023	Noise Levels		63	Oxygen levels in Gallery A7 fluctuating. Possible leak in the ventilation?						
6	2/2/2023	Vibration Levels		22	Moisture levels in Gallery C1 a bit high. Check drainage systems.						
7	12/17/2022	Noise Levels		42	Carbon monoxide trace detected in Gallery B5. Check for malfunctioning equipment.						
8	12/3/2022	Humidity		47	Pressure in Gallery A2 stable. Earlier fluctuation resolved.						
9	5/18/2023	Vibration Levels		58	Ambient light in Gallery C4 restored after short failure. Backup systems operational.						
10	11/19/2022	Temperature		69	Noise levels in Gallery D6 consistent. Good sign of equipment health.						
11	5/31/2023	Noise Levels		11	Dust particle count in Gallery A1 is high. Enhance air filtering systems.						
12	6/10/2023	Temperature		28	Electrical conductivity spike in Gallery D3 suggests water intrusion. Inspect seals.						
13	3/18/2023	Humidity		97	Barometric pressure stable throughout. Positive sign for gallery integrity.						
14	1/1/2023	Humidity		76	UV sensor malfunction in Gallery C2. Needs immediate replacement.						
15	10/29/2022	Air Quality		54	Noticing occasional drop in airflow in Gallery B4. Check fans and air ducts.						

Figure 12. Sample synthetic of safety metrics log with values and related observations for various mining galleries

	A	B	C	D
1	Timestamp	Sensor Type	Value	
2	9/19/2023 4:03	Humidity	43.53	
3	9/19/2023 19:02	Humidity	54.03	
4	9/4/2023 19:31	Humidity	7.48	
5	9/13/2023 6:38	Voltage	53.53	
6	9/19/2023 17:00	Pressure	43	
7	9/14/2023 0:56	Pressure	4.99	
8	9/26/2023 8:20	Humidity	39.03	
9	9/23/2023 2:03	Humidity	54.14	
10	9/18/2023 2:10	Temperature	0.66	
11	8/29/2023 19:33	Distance	75.81	
12	9/4/2023 18:40	Pressure	35.21	
13	9/3/2023 19:22	Voltage	12.52	
14	8/29/2023 10:54	Temperature	34.05	
15	9/23/2023 19:25	Pressure	87.52	

Figure 13. Sample Synthetic of Time stamped sensor data showcasing various measurements like humidity, voltage, pressure, and distance

	A	B	C	D
1	Timestamp	Sensor Type	Value	
2	9/19/2023 4:03	Humidity	43.53	
3	9/19/2023 19:02	Humidity	54.03	
4	9/4/2023 19:31	Humidity	7.48	
5	9/13/2023 6:38	Voltage	53.53	
6	9/19/2023 17:00	Pressure	43	
7	9/14/2023 0:56	Pressure	4.99	
8	9/26/2023 8:20	Humidity	39.03	
9	9/23/2023 2:03	Humidity	54.14	
10	9/18/2023 2:10	Temperature	0.66	
11	8/29/2023 19:33	Distance	75.81	
12	9/4/2023 18:40	Pressure	35.21	
13	9/3/2023 19:22	Voltage	12.52	
14	8/29/2023 10:54	Temperature	34.05	
15	9/23/2023 19:25	Pressure	87.52	
16	9/16/2023 19:08	Distance	94.23	
17	9/16/2023 23:03	Temperature	75.48	
18	9/24/2023 14:37	Temperature	44.16	
19	9/26/2023 10:29	Voltage	7.15	
20	9/19/2023 21:39	Pressure	87	

Figure 14. Sample Synthetic of Sensor readings log with timestamps and metrics

	A	B	C	D
1	Latitude	Longitude	Altitude	
2	34.648965	96.794454	3795	
3	4.115151	70.210082	8200	
4	21.442891	-136.535744	6852	
5	-61.117753	-164.186528	1363	
6	82.241977	-97.100187	816	
7	-62.071766	-179.002551	4523	
8	72.379054	-163.33707	7066	
9	88.243953	121.297727	6997	
10	26.839147	18.759606	6814	
11	43.795152	-94.372734	3808	
12	2.941334	-0.45584	8605	
13	23.622408	-87.488902	1631	
14	9.261091	-92.692349	2082	
15	-18.059089	-141.761684	8502	
16	64.991258	-40.882298	620	

Figure 15. Sample Synthetic of Geospatial coordinates with altitude data

A significant novelty in this study's synthetic data approach lies in its intentional interconnected design. Conceptual network diagrams were developed using Python libraries such as NetworkX to visually represent and validate these interdependencies (Figure 16), clarifying data flow and interaction points within the synthetic simulation environment.

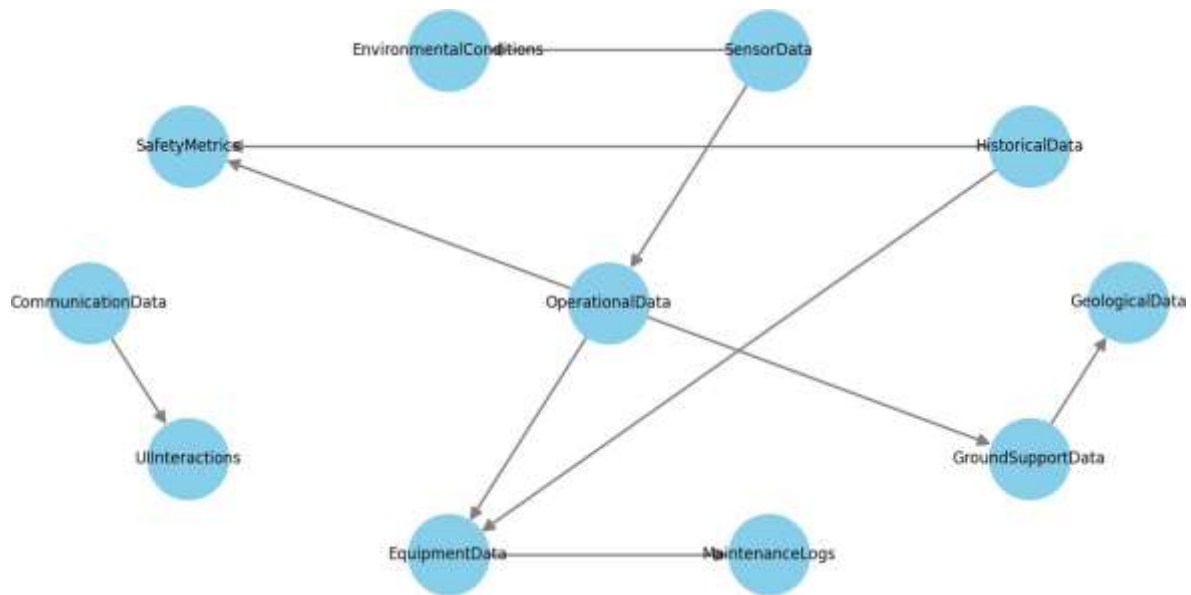


Figure 16. *Network Diagram of Interconnected Operational Datasets*

The synthetic datasets were systematically validated against real-world operational references documented in technical manuals, maintenance reports, and published literature. This validation step ensured that the synthetic data not only captured realistic operational scenarios but also maintained technical credibility and practical relevance. The structured design facilitated iterative refinement, allowing scenario adjustments and fine tuning of data characteristics based on feedback from theoretical and domain-specific reviews. The synthetic data generation framework established in this study is scientifically significant because it addresses the critical gap in data availability in underground mining research, providing a comprehensive, realistic, and practically applicable foundation for future predictive and operational analytics. Unlike traditional

simplified synthetic datasets, the current approach is distinguished by its realistic, detail, and explicit operational interconnectedness, significantly enhancing the potential accuracy and operational relevance of subsequent analytics and system deployments. This method sets a robust precedent for future studies facing similar challenges, demonstrating how synthetic data can effectively support rigorous scientific inquiry, realistic operational testing, and meaningful predictive analytics even when access to authentic operational datasets is severely restricted.

4.4 Modeling the Process and Building the Framework

Following the establishment of the synthetic datasets and simulation design, the following step involved developing a robust analytical framework capable of effectively handling these datasets. The goal of this phase was to create a linked, adaptable system designed to simulate real-world data processing workflows and decision-making protocols applicable to underground mining environments. Effective preprocessing is vital for ensuring the analytical readiness of synthetic datasets, thereby facilitating meaningful analyses and realistic simulations. This stage involved three interconnected steps: data cleaning, transformation, and normalization. Data transformation at this stage focused on structuring raw synthetic outputs into analysis-ready formats through temporal aggregation, feature construction, categorical encoding, and scaling operations appropriate to each dataset type. Detailed descriptions of these preprocessing steps are provided in the following subsections.

Data Cleaning

For demonstrating a realistic scenario, the generated synthetic datasets included intentional inconsistencies, such as missing values, duplicate entries, and simulated outliers. The cleaning process employed standard Python libraries (Pandas and NumPy) to systematically identify and address these issues. Missing data points were managed through context appropriate imputation techniques, such as median and mode

replacement methods. Duplicate entries were identified via timestamps and unique identifiers and subsequently removed to preserve data integrity. Additionally, outliers, especially in critical parameters like gas concentrations or equipment status indicators, were detected through statistical methods and carefully adjusted or removed.

Data Transformation

Following data cleaning, transformation procedures standardized the datasets for consistent analysis. This included standardizing timestamp formats, aligning numerical codes across datasets, and converting categorical text into numerical representations suitable for analytical processing. Furthermore, structural transformations, such as merging separate date and time columns into unified timestamp fields, were performed to streamline analytical tasks and data visualization procedures.

Data Normalization

Normalization was applied to scale numerical values consistently across the datasets, ensuring comparability and analytical coherence. Techniques such as Min-Max scaling and Z-score normalization were utilized for parameters like environmental sensor readings, equipment condition metrics, and safety indicators. This step was essential to effectively visualize data, detect subtle operational anomalies, and lay groundwork for potential future predictive modeling applications.

The preprocessed datasets formed the foundation for developing the modular system architecture. Python scripts facilitated this architecture, allowing datasets to be processed through independent logic modules. This modularity was essential, enabling continuous system functionality even if individual data sources were temporarily unavailable, accurately reflecting operational realities. The designed system architecture comprised modules dedicated to equipment health monitoring, environmental data analysis, and operational safety evaluations. Each module

independently processed input datasets, applied predefined logic conditions (e.g., safety threshold breaches), and generated structured outputs alerts, data visualizations, and status updates. Visual representation of system outputs and interactions was a critical component of the framework. Conceptual diagrams and dashboard prototypes provided simulation platforms. The dashboard was developed with visualization principles, clearly communicating critical alerts, equipment health statuses, and environmental conditions to potential users, matching realistic user system interactions. To support decision-making, a modular dashboard was implemented to centralize data streams and link them to visual outputs (see Figure 17 for an overview of the dashboarding architecture and its connections).

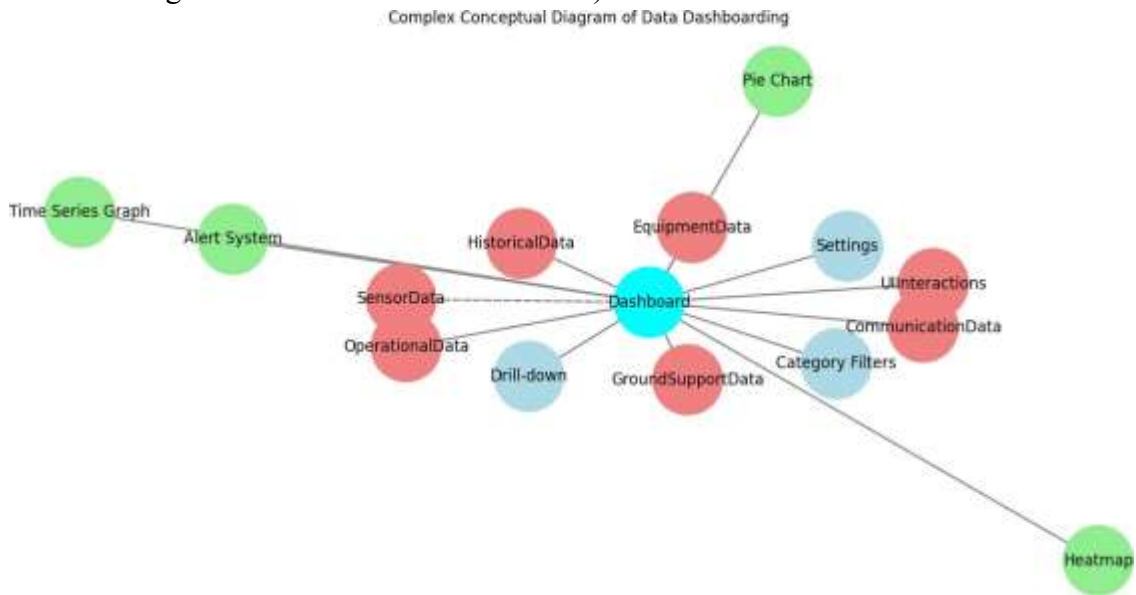


Figure 17. *Conceptual diagram of dashboard architecture and visualization tools.*

This diagram illustrates the central role of the dashboard in aggregating data streams such as sensor readings, equipment data, ground support logs, and communication records. The links to visual components like heatmaps and alert systems reflect how raw data is transformed into operational insights and decision-support mechanisms.

This structured preprocessing and modeling framework significantly enhances the practical relevance of the synthetic datasets, demonstrating robust methodology and

facilitating realistic operational simulations. The modular and scalable architecture enables seamless integration of future real-world datasets, ensuring analytical consistency and practical utility. Furthermore, this approach provides valuable methodological insights applicable beyond the mining sector, highlighting practices in synthetic data handling, system modularity, and interactive visualization within constrained operational environments.

4.5 Outcomes and Observations

The outcomes of this case study reflect the successful development and testing of a modular, simulation-based framework capable of visualizing and processing synthetic data representative of underground mining environments. The system integrates preprocessed datasets, applies logical decision rules, and outputs meaningful insights through visual interfaces. While this implementation remains within the scope of simulated inputs, the outputs demonstrate the technical feasibility and analytical depth necessary for eventual real-world deployment. The dashboard serves as the primary user interface for the intelligent support system. Designed in Python using data visualization libraries such as Matplotlib and Seaborn, the dashboard displays integrated outputs including sensor status, environmental conditions, equipment health indicators, and system alerts. Each panel on the interface reflects the output of its corresponding module, updating in response to simulated changes in the synthetic datasets. Scenarios such as abnormal gas levels, equipment degradation, and ground support failures were tested using tailored variations in the synthetic data. The system responded appropriately to these inputs, triggering alerts, and recalculating safety indices. These tests confirmed that the designed architecture behaves in alignment with predefined thresholds and operational logic.

To validate data interdependencies and logic flow, several conceptual diagrams were developed. A key outcome was the creation of a network diagram that maps the

interactions between datasets. This visualization clarified which datasets influence others and revealed feedback pathways within the system, especially between environmental readings, equipment performance, and safety metrics. Another diagram illustrated the interaction between user actions and system responses. This was especially important for planning future system upgrades, such as incorporating reinforcement learning to optimize user interface behavior or customizing alert thresholds based on operator feedback patterns.

Several important insights emerged during system testing and simulation:

- **System Modularity Enhances Reliability:** The flexible architecture ensured that each module could operate independently, making the system resilient to partial data availability, an expected scenario in real-world mining environments.
- **Visualization Affects Usability:** Interface layout and data prioritization significantly affected ease of use. When alert signals were presented too frequently or with low differentiation, the interface became overwhelming. This led to iterative redesigns that improved signal clarity and operational relevance.
- **Feedback Mechanisms Are Essential:** Simulated user behavior, such as repeated dismissal of alerts, suggested the need for adaptive logic in future iterations. The project outlined potential mechanisms to incorporate user behavior as a variable in system recommendations.
- **Time Resolution Matters:** The importance of precise timestamps and time series data continuity was reinforced throughout the simulation. Minor inconsistencies in time formatting or resolution had significant impacts on data synchronization and event reconstruction.

These observations highlight the maturity and applicability of the simulation framework. While real-time data integration remains a future step, this study has

established a foundational infrastructure upon which predictive analytics, adaptive interface systems, and real-time safety protocols can be reliably built.

This chapter demonstrated the application of a simulation-based approach to generate and analyze synthetic operational data for predictive maintenance purposes. The results illustrate how synthetic datasets can be effectively used to model system behavior, explore degradation trends, and test analytical workflows in data-scarce environments. The findings from this case study provide methodological insights and serve as a conceptual reference for the second case study, which applies similar analytical principles to real industrial data within a production and maintenance.

CHAPTER 5

5. Case Study 2: Customer Profiling and Failure Prediction in Pump Systems

This chapter presents the second case study, which focuses on a real industrial application in the pump manufacturing and repair sector. The objective of this case study is to analyze heterogeneous operational datasets in order to develop customer profiles and support predictive maintenance decision-making. The chapter describes the industrial context, the structure and characteristics of the available datasets, and the data preprocessing and feature engineering strategies applied. Unsupervised machine learning techniques are then employed to identify customer segments and behavioral patterns relevant to maintenance planning and service optimization. This case study represents the practical implementation of the data-driven framework established in Chapter 3.

5.1 Company Background and Industrial Context

Technosub is an industrial pump manufacturing and repair enterprise, operates primarily within the mining, municipal, and industrial sectors. The company's expertise lies in delivering robust pump solutions and maintenance services designed to meet the demanding requirements of heavy-duty applications. Technosub manages a significant volume of pumps across a diverse customer base, including companies such as Vale Canada Limited, IAMGOLD Corporation, and Goldcorp Canada Ltd., each presenting distinct usage and repair behaviors. The industry in which Technosub operates is characterized by high operational costs associated with pump maintenance, frequent equipment downtime, and challenging environmental conditions. These complexities have necessitated a shift from traditional reactive maintenance practices toward proactive, data-driven predictive maintenance strategies. However, implementing these advanced maintenance models requires comprehensive insights into customer behaviors, purchasing patterns, and equipment usage; an area that traditional analytical methods have typically overlooked. By recognizing these challenges, Technosub

initiated a strategic project aimed at integrating customer profiling insights with predictive maintenance planning. The goal was to develop a robust, scalable analytical framework capable of segmenting clients based on their maintenance behaviors and optimizing service interventions proactively. This integration of customer segmentation with predictive maintenance is expected to improve resource allocation, minimize downtime, and reduce operational expenditures. This case study explores the methodological and practical dimensions of the project, detailing how unsupervised clustering algorithms, such as: K-Means, DBSCAN, and Gaussian Mixture Models (GMM) can help us to identify meaningful customer segments. These insights can help the company to facilitate targeted maintenance strategies and find better solutions for improvements in overall operational efficiency and customer service effectiveness.

5.2 Dataset Description and Preprocessing

The analysis utilized comprehensive datasets provided by Technosub, each serving specific analytical purposes: their key fields and analytical roles are summarized in Table 1.

Datasets	Description
Repair Cost Dataset	Contains pump ID, Repair Cost, pump Model, Item Number, and Item Description. Used for tracking normalized repair costs.
Repair Tracking Dataset	Includes pump ID, Serial Number, Repair Dates, and Customer Name. Used to monitor repair history and tracking.
Sales Transactions Dataset	Contains Client Names, Sales Records, pump Models, and pump Serial Numbers. Used for client sales tracking and analysis.
Failure Classification Dataset	Includes historical repair causes and failure classifications.

Table 1. *Comprehensive Data Suite Provided by Industrial partner*

A structured, Python-based validation process was implemented to assess dataset completeness, consistency, and uniqueness. The evaluation revealed quality concerns affecting reliability and accuracy, including inconsistent naming conventions (e.g.,

varied SOP type formats), missing values in customer identifiers, pump models, serial numbers, and timestamps, as well as unrealistic values such as negative quantities and implausible costs. A validation workflow quantified fitness-for-use across completeness (40%), consistency (30%), and uniqueness (30%); dataset-level scores and inclusion decisions are reported in Table 2.

Dataset	Quality Score (/100)	Identified Issues	Actions Taken
Maintenance and Repair Data	99.99	Near-perfect quality	Directly used in analysis
Failure Classification Logs	87.22	Missing pump ID references	Excluded from primary models
Client Sales Data	95.98	Minor duplicates detected	Cleaned and utilized
Repair Tracking Logs	93.34	Missing fields in specific columns	Processed post-refinement

Table 2. *Datasets quality evaluation results*

The actions reported in Table 2 were guided by quality score ranges rather than absolute cutoffs. In general, datasets scoring above 95 were deemed suitable for direct use, scores between 90 and 95 underwent targeted refinement, and datasets below 90 were excluded from primary modeling due to elevated risk of inconsistency or missing critical identifiers.

To address these issues, a comprehensive preprocessing framework was applied:

- **Standardization and Formatting:** Harmonized naming conventions, applied consistent units, and categorized item descriptions to enable aggregation across datasets.
- **Data Imputation:** Numerical features such as repair costs and sales amounts were imputed using median values to reduce the impact of outliers while preserving the distribution’s central tendency. Missing categorical values (e.g.,

pump models, failure types) were assigned a new category (“Unknown”) to avoid bias in clustering.

- **Time-Series Integrity:** Missing timestamps were estimated through linear interpolation to preserve temporal continuity in repair and operational logs.
- **Outlier Management:** Domain knowledge and statistical thresholds were used to remove implausible entries. High-cost maintenance records were retained but normalized via log transformation to mitigate skewness.
- **Anomaly Preservation:** Infrequent but valid customer behaviors were kept for downstream analysis using anomaly detection methods such as DBSCAN, recognizing their potential business value.
- **Categorical Feature Processing:** One-hot encoding was applied to nominal variables such as client type (industrial, municipal, mining), while label encoding was used for ordinal features like pump model categories.
- **Feature Scaling:** Applied min-max normalization to frequency-based metrics (purchase counts, maintenance frequency) and log transformation to monetary values to optimize performance in clustering algorithms.
- **Quality Scoring Framework:** Dataset readiness was evaluated using a weighted scoring approach: completeness (40%), consistency (30%), and uniqueness (30%). This ensured continuous monitoring of data integrity throughout the project lifecycle.

These measures produced a unified, high-quality dataset, ensuring analytical robustness for the clustering and predictive maintenance modeling phases.

5.3. Feature Engineering and Data Normalization

Feature engineering is a critical step in preparing datasets for effective clustering and predictive modeling. In this project, careful attention was given to selecting,

constructing, and normalizing features that accurately reflect customer behavior and maintenance patterns.

Feature Selection and Construction

Initially, raw transactional and maintenance datasets contained numerous potential features. However, to maintain clarity and ensure computational efficiency, a subset of features was selected based on their relevance to customer segmentation and predictive maintenance. The selected features were:

- **Recency (R):** Time (days) since the customer's last repair or purchase. This captures recent customer interaction and urgency of maintenance needs.
- **Frequency (F):** The number of repair and purchase transactions per customer, indicating activity levels and potential wear-and-tear on pumps.
- **Monetary (M):** Total normalized cost of repairs per customer, highlighting financial impact and resource allocation needs.
- **Repair Frequency:** Total number of repairs within a specified period, used to classify maintenance intensity.
- **Average Repair Interval:** Mean time interval (days) between consecutive repairs per pump, providing insights into maintenance predictability.
- **Buying Frequency:** Number of pump-related purchases per customer, offering insights into customer purchasing patterns and product lifecycle management.

Additionally, derived features such as ratios of repair-to-purchase frequencies were calculated to capture nuanced customer behaviors.

Data Normalization and Encoding

Given the diversity of data types and scales, a structured normalization and encoding strategy was essential:

- **Numerical Features (Monetary Values, Repair Frequency, Buying Frequency):**
 - **Log Transformation:** Applied to monetary values (repair costs) to reduce skewness and handle extreme outliers, ensuring more stable clustering results.
 - **Min-Max Scaling:** Utilized for frequency-based measures (repair frequency, buying frequency) to scale values within a $[0,1]$ range, equalizing their influence in clustering algorithms.
- **Categorical Features (Pump Models, Failure Types, Customer Categories):**
 - **One-Hot Encoding:** Implemented for nominal categorical features such as customer types and distinct pump models, converting these into binary vectors to facilitate algorithmic compatibility.
 - **Label Encoding:** Applied to ordinal features such as severity of failures (if applicable) to preserve inherent order and significance in clustering.

These normalization methods effectively mitigated biases due to variable scale and improved algorithm performance, accuracy, and interpretability.

Feature Name	Type	Description	Normalization Technique
Recency (days)	Numerical	Days since last transaction	Min-Max Scaling
Frequency (transactions)	Numerical	Total repair/purchase transactions	Min-Max Scaling
Monetary (total cost)	Numerical	Cumulative normalized repair costs	Log Transformation, Min-Max Scaling
Repair Frequency	Numerical	Total number of repairs	Min-Max Scaling
Average Repair Interval	Numerical	Average time interval between repairs	Min-Max Scaling
Buying Frequency	Numerical	Frequency of purchases	Min-Max Scaling
Pump Model	Categorical	Specific pump models used by clients	One-Hot Encoding
Failure Type	Categorical	Type of failure experienced	One-Hot Encoding

Table 3. *Overview of feature engineering*

Final Feature Set for Clustering

Log transformation for monetary variables, min max scaling for numerical features, and one hot encoding for categorical variables ensure comparability across features and produce a representation suitable for distance based clustering and subsequent predictive modeling, as summarized in Table 3. The outcome of the feature engineering and normalization process was a refined dataset, optimized for accurate clustering analysis. The final feature set included:

This systematic approach in feature engineering and normalization ensured the dataset was appropriately structured, providing a robust foundation for accurate and meaningful clustering and predictive modeling in subsequent analyses.

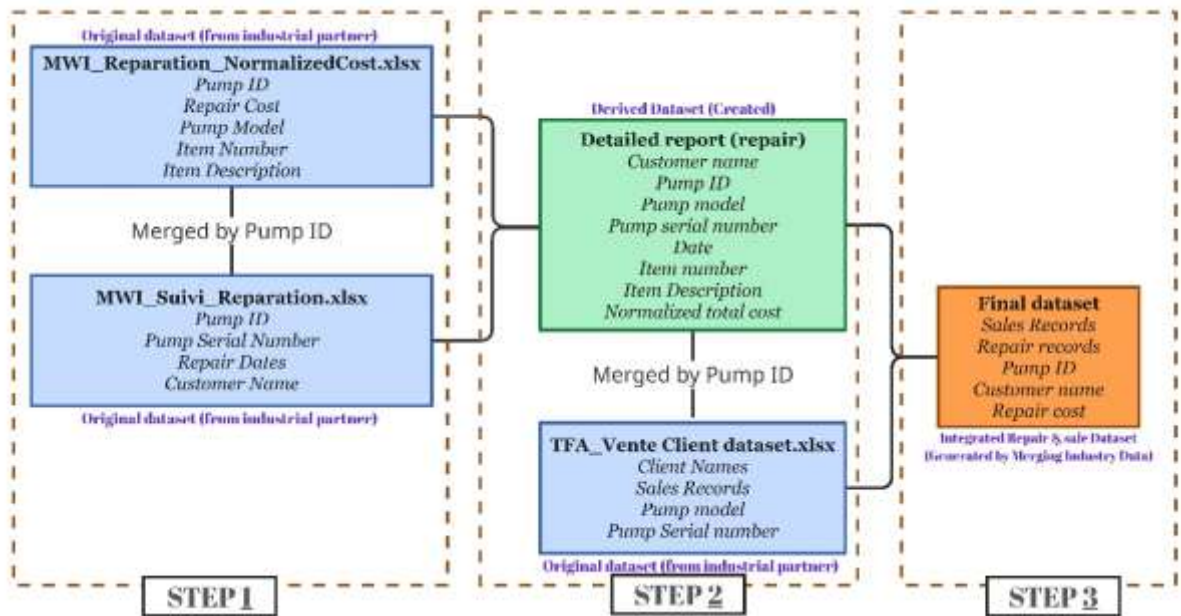


Figure 18. Dataset integration process

5.4 Clustering and client classification (K-Means, DBSCAN, GMM)

This study leveraged three distinct unsupervised clustering algorithms: K-Means, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), and Gaussian Mixture Models (GMM), each selected for its unique strengths in handling industrial datasets and customer behavior complexities. The combined use of these methods aimed to achieve robust and comprehensive customer profiling to support predictive maintenance strategies.

Rationale for Algorithm Selection

Each clustering technique was strategically chosen based on its capability to address specific aspects of the data:

- **K-Means:** Ideal for structured clustering due to its computational efficiency, ease of interpretability, and capability to produce clearly defined, spherical clusters. K-Means was particularly useful in generating baseline segments that highlight standard maintenance behaviors.
- **DBSCAN:** Selected for its superior performance in detecting anomalies, handling noise, and identifying clusters of arbitrary shapes. This method is especially valuable in identifying customers with irregular or irregular maintenance patterns, often overlooked by conventional clustering algorithms.
- **Gaussian Mixture Models (GMM):** Chosen for its probabilistic approach that assigns data points to multiple clusters based on varying probabilities. GMM effectively captures overlapping customer segments, thereby providing subtle classifications that are essential in complex industrial environments.

Mathematical Foundation

K-Means Clustering

K-Means partitions the dataset into K clusters by minimizing intra-cluster variance. The objective function for K-Means clustering is defined as:

Given a dataset $X=\{x_1, x_2, \dots, x_n\}$, the objective function of K-Means is:

$$J = \left| \sum_{i=1}^k \sum_{x_j \in C_i} \|x_j - \mu_i\| \right|^2 \quad \text{Equation 1}$$

Where:

- **J** is the total variance within clusters,
- **K** is the number of clusters,
- x_j is a data point assigned to cluster C_i ,
- μ_i is the centroid of cluster C_i ,
- $\|x_j - \mu_i\|^2$ Represents the squared Euclidean distance between a point and its centroid.

K-Means was applied to group clients based on repair frequency, total repair costs, and purchase behavior. The algorithm's simplicity and efficiency made it suitable for large-scale customer segmentation, providing distinct client groups with stable purchasing patterns and varying maintenance needs.

DBSCAN

DBSCAN groups points based on local density, identifying clusters with varying shapes and distinguishing noise effectively. The algorithm depends on two primary parameters, ϵ (epsilon) and MinPts:

- **Core point:** A point that has at least MinPts points within a radius ϵ .
- **Density reachable:** A point p is reachable from another point q if there is a path of core points leading from q to p .

A point p is considered density-reachable if there exists a sequence of core point's $p_1, p_2, \dots, p_n$ such that:

$$\forall_i d(p_i, p_{i+1}) \leq \epsilon \quad \text{Equation 2}$$

Where:

- $\mathcal{P}$ denotes a data point in the feature space,
- p_i and p_{i+1} represent consecutive core points in the sequence connecting point $\mathcal{P}$ to a core point,
- n is the total number of core points in the density-reachable chain,
- $d(\cdot, \cdot)$ is the distance metric used to measure similarity between two data points (Euclidean distance in this study),
- ϵ is the neighborhood

DBSCAN was used to identify high-maintenance clients and outliers with irregular repair patterns. It effectively detected clients with unpredictable repair needs, allowing for further analysis of their maintenance behaviors.

Gaussian Mixture Models (GMM)

GMM assumes the dataset is generated from a mixture of several Gaussian distributions. The model assigns probabilistic cluster memberships, allowing data points to belong to multiple clusters with different probabilities. The Probability Density Function (PDF) of a GMM for data point x is given by:

$$P(x) = \sum_{i=1}^k \pi_i \cdot N(x|\mu_i, \Sigma_i) \quad \text{Equation 3}$$

Where:

- $\mathbf{x}$ denotes an observed data point in the feature space,
- k is the number of Gaussian components (clusters) in the mixture model,
- π_i represents the mixing coefficient of the i -th Gaussian component, satisfying $\sum_{i=1}^k \pi_i = 1$,
- $N(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ is the multivariate Gaussian probability density function associated with the i -th component,
- $\boldsymbol{\mu}_i$ is the mean vector of the i -th Gaussian distribution,
- $\boldsymbol{\Sigma}_i$ is the covariance matrix of the i -th Gaussian distribution.

GMM was applied to model overlapping customer segments, identifying clients exhibiting mixed behaviors across different clusters. This approach provided a subtle segmentation strategy, highlighting customers with similar maintenance needs.

Determining the Optimal Number of Clusters

To determine the most appropriate cluster count, two robust validation techniques were applied:

- **Elbow Method:**

Evaluates the Within-Cluster Sum of Squares (WCSS) across various cluster counts (K). The optimal number is identified at the inflection point ("elbow") where additional clusters provide diminishing returns in variance reduction.

$$WCSS = \sum_{i=1}^K \sum_{x_j \in C_i} \|x_j - \boldsymbol{\mu}_i\|^2 \quad \text{Equation 4}$$

where:

- K is the number of clusters,

- C_i denotes the set of data points assigned to cluster i ,
 - x_j represents the j -th data point belonging to cluster C_i ,
 - μ_i is the centroid (mean vector) of cluster i ,
 - $\|x_j - \mu_i\|^2$ denotes the squared Euclidean distance between a data point and its cluster centroid.
- **Silhouette Score:**

Measures cohesion (how close points are within clusters) and separation (distance between clusters), providing a score ranging from -1 to +1. Values closer to +1 indicate well-defined and distinct clusters.

$$S_i = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad \text{Equation 5}$$

Where:

- $a(i)$ represents the average intra-cluster distance,
- $b(i)$ represents the average nearest-cluster distance.

Using these validation methods, the optimal number of clusters was selected for each algorithm, ensuring robust and meaningful segmentation.

Integrated Clustering Framework

Integrating K-Means, DBSCAN, and GMM into a comprehensive analytical framework allowed for:

- Clear definition of typical customer behaviors (K-Means).
- Identification of irregular and outlier clients (DBSCAN).
- Enhanced understanding of nuanced, overlapping client segments (GMM).

This integrated approach provided Technosub with a detailed and versatile customer classification, effectively supporting targeted predictive maintenance planning.

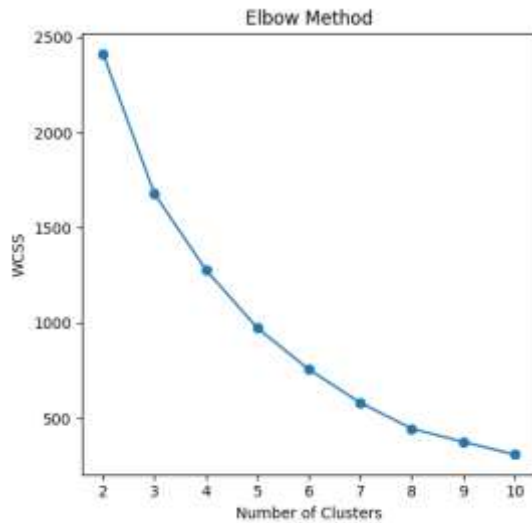


Figure 19. *Elbow method score plot*

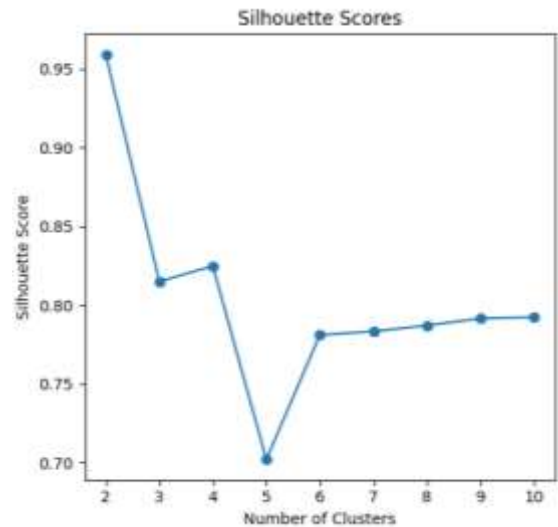


Figure 20. *Silhouette method score plot*

5.5 Evaluation of Client Segments and Behavior Patterns

Following the application of clustering algorithms, the customer segments were systematically evaluated to identify distinctive behavior patterns. Each algorithm provided insights into varying facets of customer behavior, collectively offering a comprehensive understanding crucial for predictive maintenance strategies.

Segments Identified by Clustering Methods

K-Means Clustering Results

To establish a baseline segmentation, K-Means was applied as a structured clustering approach, assuming relatively distinct behavioral groupings within the dataset. This method provides an efficient way to categorize customers based on maintenance frequency, cost trends, and service history, forming the foundation for further analysis.

The distribution of clients across the identified clusters is shown in Figure 21, while detailed characteristics of each group are summarized in Table 4.

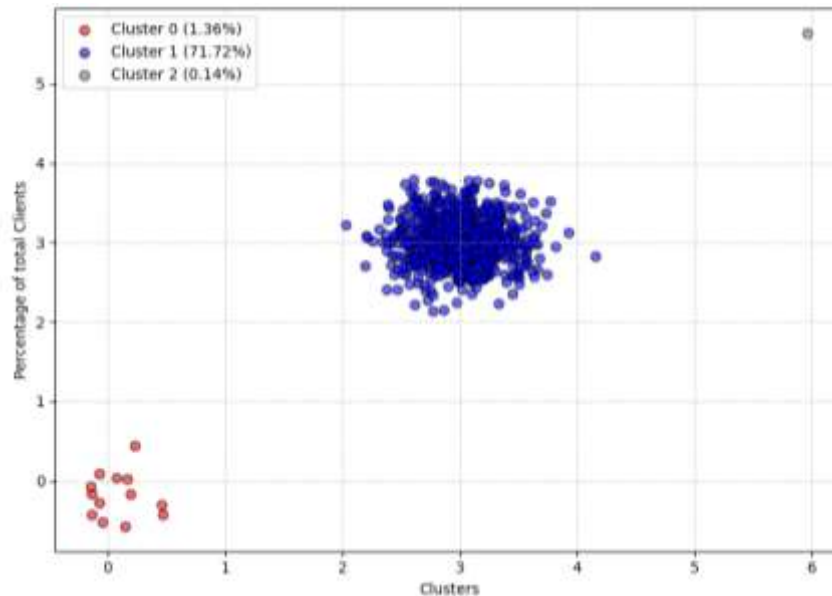


Figure 21. *Client distribution across K-Means clusters*

- **Cluster 1: Reliable Clients (71.72%)**
 - **Characteristics:** Low repair frequency, minimal normalized maintenance costs, and consistent buying patterns.
 - **Implications:** Well-suited possible candidates for long-term maintenance contracts due to predictable service needs and lower operational risk.

- **Cluster 0: High-Maintenance Clients (1.36%)**
 - **Characteristics:** Very high frequency of repairs, moderate to high maintenance costs, irregular purchasing patterns, frequent urgent interventions.

- **Implications:** Targeted for enhanced predictive maintenance strategies and proactive monitoring systems to reduce operational disruptions and resource allocation inefficiencies.

- **Cluster 2: Anomalous High-Repair Client (0.14%)**

- **Characteristics:** Extremely high repair frequency and very high repair cost indicators, with irregular or non-representative purchasing patterns.
- **Implications:** Treated as an anomaly case; important for risk monitoring and exception-handling strategies rather than general segmentation.

Cluster Number	No. of Clients	Perc. of Total Clients	Total cost of Repairs	Repair Freq.	Buying Freq.	Ave. repair Interval	Buying Interval
Cluster 0	10	1.36%	Moderately high (106 units)	Extremely high (15,294 times)	Moderate	Very low	Non-existent
Cluster 1	1092	71.72%	Low (2.75 units)	Low (237 times)	High	Regular	Regular
Cluster 2	1	0.14%	Very high (421.5 units)	Extremely high (85,611 times)	Moderate	Non-existent	Non-existent –

Table 4. *K-Means clustering result*

Percentages reported in the clustering results reflect relative proportions within the analytically retained client set following preprocessing and activity-based filtering. As a result, reported cluster shares reflect analytical relevance rather than population-level distribution. K-Means segmentation allows businesses to prioritize maintenance

resources effectively. Reliable Clients (Cluster 1) can be placed on long-term maintenance contracts, ensuring consistent revenue without frequent interventions. High-Maintenance Clients (Cluster 0) should be flagged for preventive maintenance programs, reducing unexpected failures and costly emergency repairs. Anomalous high-repair clients (Cluster 2) should be treated as exception cases, informing risk monitoring and anomaly-handling strategies rather than standard service optimization.

DBSCAN Clustering Results

While K-Means effectively categorized structured customer groups, it assumes spherical clusters and struggles with irregular or noisy data points. To address this limitation, DBSCAN was employed to identify high-maintenance clients with unpredictable behavior patterns and detect outliers, offering a more flexible segmentation approach. The distribution of clients across the identified clusters is shown in Figure 22, while detailed characteristics of each group are summarized in Table 5.

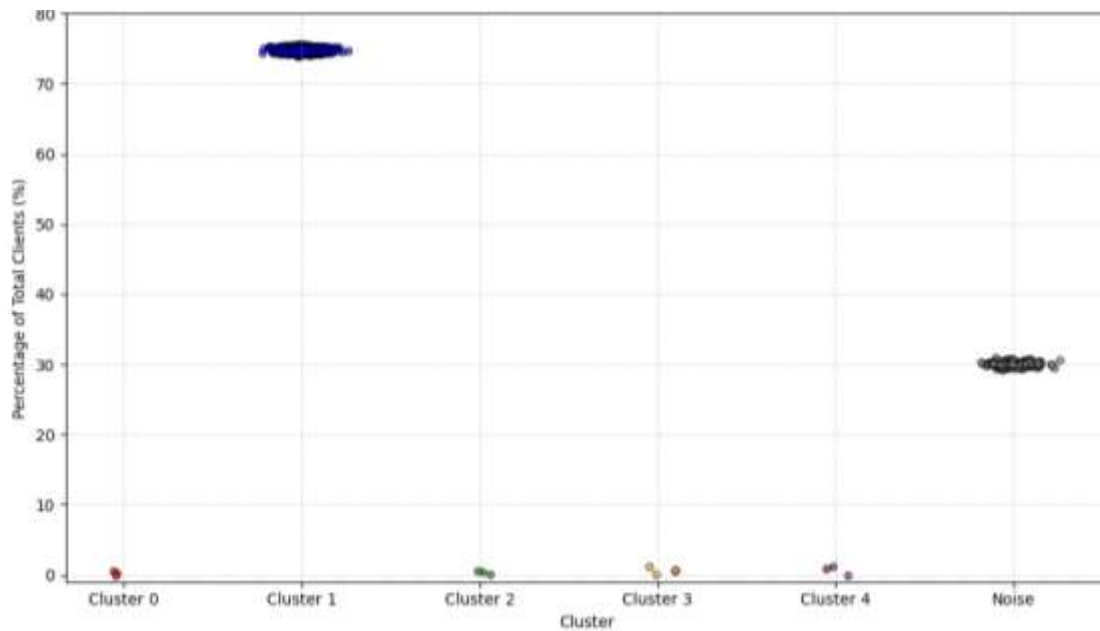


Figure 22. *Client distribution across DBSCAN clusters*

- **Cluster 0 (0.41%):** Clients in this cluster have low repair frequency and cost, low buying frequency, and high days since the last purchase. These clients exhibit minimal activity, with a high gap between their last purchase and repairs.
- **Cluster 1 (74.86%):** This is the majority cluster, representing customers with moderate repair frequency and moderate buying frequency, but with low repair cost. These clients are ideal candidates for optimized preventive maintenance, as their repair needs and purchase behaviors are predictable, making them easier to manage and service.
- **Cluster 2 (0.41%):** Clients here exhibit high repair frequency, moderate repair costs, and high purchasing activity. These clients require consistent service attention and have active engagement in purchasing and repairs.
- **Cluster 3 (0.55%):** This cluster includes clients with low repair frequency and cost, but they engage with a moderate variety of products. Their repair needs

are minimal, but they explore different product options, suggesting potential for service optimization based on their product variety preference.

- **Cluster 4 (0.41%):** Clients in this group show very low repair costs and frequency and have low buying activity. These clients require minimal attention but should still be monitored for any future service opportunities.
- **Noise (-1, 30.05%):** A substantial portion of the client base (30.05%) falls into the outlier or noise category. These customers have irregular or inactive maintenance behaviors, making it challenging to predict their needs. Further analysis is required to determine their impact on resource allocation and how they should be factored into service planning.

Cluster Number	Number of Clients	Perc. of Total Clients	Ave. Total Cost of Repairs	Repair Freq.	Buying Freq.	Buying Interval
Cluster 0	3	0.41%	Low (0.09)	Low	Low	High (6636.0)
Cluster 1	549	74.86%	Low (0.41)	Moderate	Moderate	Low
Cluster 2	3	0.41%	Moderate (7.62)	High	High	Low
Cluster 3	4	0.55%	Low (0.098)	Low	Low	Low
Cluster 4	3	0.41%	Very Low	Low	Low	Low
Noise (-1)	220	30.05%	N/A	N/A	N/A	N/A

Table 5. DBSCAN Clustering results

For DBSCAN, percentages are reported relative to the full client population, including noise points (label -1), which represent clients not assigned to any dense cluster. DBSCAN's ability to detect irregular maintenance behaviors and high-risk outliers makes it ideal for proactive failure management. Outlier clients (30.05%) may indicate equipment that are close to failure or clients requiring customized service plans.

Businesses can leverage these insights to introduce targeted intervention strategies, such as real-time monitoring for high-risk customers or adjusted maintenance schedules for clients with unpredictable service patterns. This flexible clustering approach enhances cost efficiency by ensuring that high-risk clients are identified early, reducing unscheduled maintenance and optimizing workforce deployment.

Gaussian Mixture Models (GMM) Results

Although DBSCAN improved outlier detection, it lacks soft clustering capabilities, which are crucial for customers with overlapping behaviors. To address this, GMM was used as a probabilistic clustering approach, with final cluster assignments determined by maximum posterior probability. The distribution of clients across the identified clusters is shown in Figure 23, while detailed characteristics of each group are summarized in Table 6.

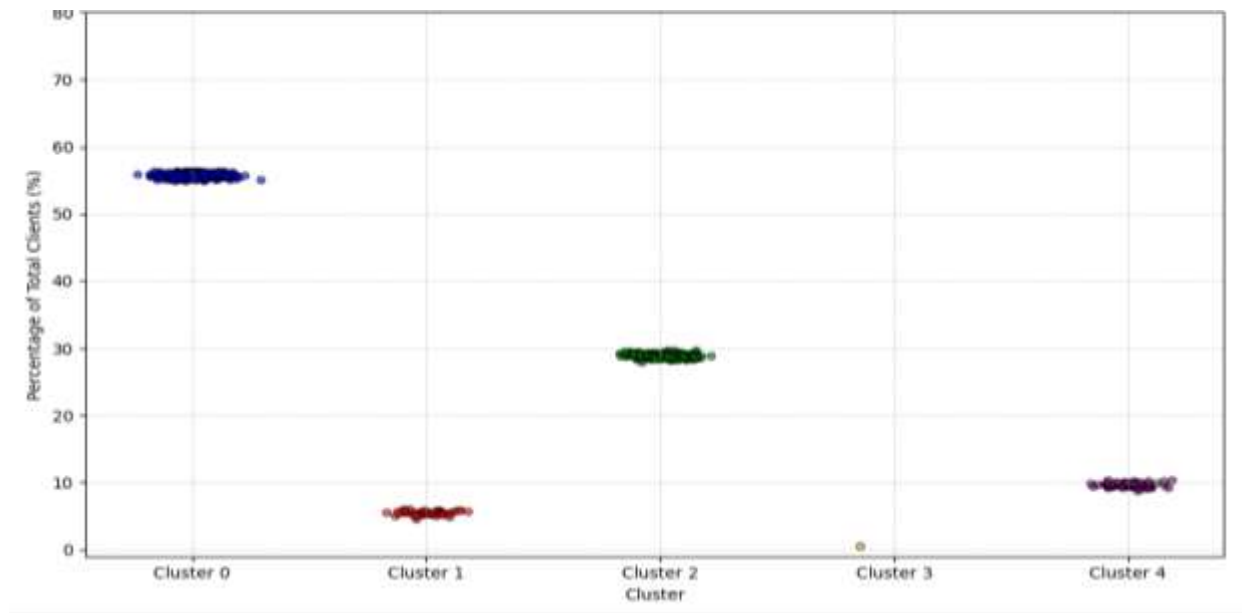


Figure 23. *GMM clusters overlapping diagram*

- **Cluster 0 (55.76%):** This cluster represents high-value clients who have frequent repairs and high spending. They are the core group for predictive maintenance, requiring high engagement and more attention to their service needs.
- **Cluster 1 (5.50%):** These are occasional users who have moderate repair costs but low engagement. They tend to make fewer purchases and have low repair frequency.
- **Cluster 2 (28.91%):** These clients have frequent repairs but lower total spending. They are likely cost-conscious customers, actively managing their maintenance expenses. This group requires service plans that are financially optimized, with a moderate buying frequency and moderate buying intervals.
- **Cluster 3 (0.13%):** A very small group with infrequent use, minimal repair frequency, and high days since last engagement. These clients likely require minimal attention but should be monitored for potential future needs.
- **Cluster 4 (9.68%):** These are potential high-engagement clients, exploring services with moderate repair costs and a preference for a wider variety of products. Their buying behavior indicates they could become more engaged with the right service offerings.

Cluster Number	No. of Clients	Perc. of Total Clients	Ave. Total Cost of Repairs	Repair Freq.	Buying Freq.	Buying Interval
Cluster 0	436	55.76%	High	High	High	Low
Cluster 1	43	5.50%	Moderate	Low to Moderate	Low	Low
Cluster 2	226	28.91%	Low	High	Moderate	Moderate
Cluster 3	1	0.13%	Very Low	Very Low	Low	High
Cluster 4	76	9.68%	Moderate	Low	Moderate	Moderate

Table 6. *GMM Clustering results*

The GMM clustering analysis provided a probabilistic segmentation of customers, allowing for overlapping cluster memberships, as illustrated in Figure 23. Unlike K-Means and DBSCAN, which assign customers to a single cluster, GMM enables soft assignments, capturing clients who exhibit mixed behavioral patterns across multiple groups. Additionally, overlapping regions reveal clients with shared characteristics. For instance, 50 clients belong to both Cluster 0 and Cluster 2, indicating customers who exhibit variable spending behaviors. Similarly, 20 clients belong to both Cluster 2 and Cluster 4, suggesting hybrid maintenance approaches that blend cost efficiency with strategic spending. GMM's probabilistic clustering provides a dynamic and adaptable approach to customer maintenance management. Clients overlapping multiple clusters can be offered customized service bundles that balance predictive maintenance with cost-efficient interventions.

For example, clients transitioning between high-maintenance and cost-conscious clusters can be targeted for discounted service plans or loyalty programs, improving retention. Unlike rigid classification models, GMM helps businesses adapt service

offerings based on customer behavior trends, ensuring sustained client engagement and optimized maintenance costs.

Comparative Evaluation of Clustering Techniques

Each clustering algorithm demonstrated unique strengths, offering complementary perspectives on customer behavior patterns. To support strategic decision-making, Table 6 summarizes the comparative attributes of K-Means, DBSCAN, and GMM, highlighting their respective strengths, limitations, and relevance for business applications. This multi-faceted evaluation provides a foundation for selecting appropriate segmentation techniques based on operational priorities whether interpretability, anomaly detection, or probabilistic modeling.

Algorithm	Strengths	Limitations	Business Relevance
K-Means	High interpretability, clear and actionable segments.	Assumes spherical clusters, sensitive to outliers.	Ideal for straightforward client classification and service optimization.
DBSCAN	Effective anomaly detection, handles irregular clusters.	Sensitive to parameter tuning, less effective on large homogeneous segments.	Crucial for identifying problematic customers requiring specialized interventions.
GMM	Captures overlapping behaviors, probabilistic membership for nuanced segmentation.	Computationally complex, less immediately interpretable.	Supports nuanced predictive analytics and flexible resource allocation.

Table 7. *Overview of algorithm comparison*

To further illustrate the segmentation outcomes, Figure 24 presents a visual comparison of clusters generated by K-Means, DBSCAN, and GMM. Rather than displaying the full set of cluster labels from each model, this figure selectively highlights a single representative cluster per algorithm to demonstrate how different clustering techniques interpret and isolate client behavior patterns within the same dataset. By mapping one cluster per algorithm onto the same PCA-transformed space, the figure enables a qualitative comparison of cluster coverage, shape, and spatial distribution. This visualization underscores the core idea that each algorithm captures different aspects

of structure, K-Means enforcing rigid partitions, DBSCAN focusing on dense regions, and GMM modeling overlapping distributions. The final panel overlays all three clusters, emphasizing the complementary strengths of the algorithms. This visual synthesis supports the use of ensemble or comparative clustering strategies in operational contexts, particularly where maximizing insight coverage is critical to decision-making.

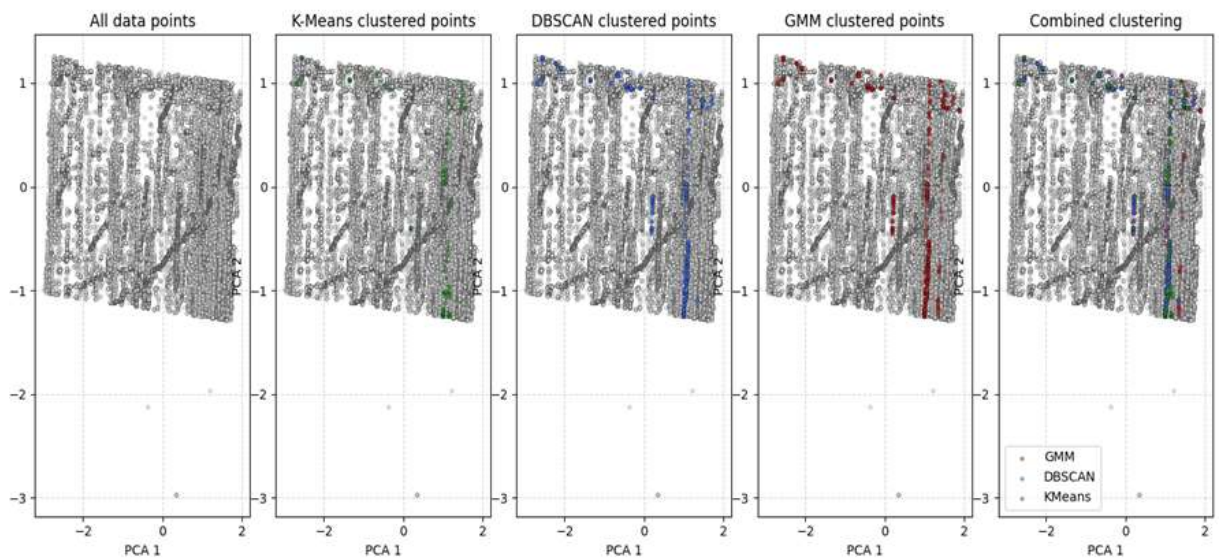


Figure 24. *Comparative clustering coverage across K-Means, DBSCAN, and GMM*

Practical Implications for Predictive Maintenance

The evaluation of segments identified clear operational implications:

- **Resource Allocation:** Efficiently allocate resources by prioritizing maintenance efforts towards high-risk (high maintenance) clients identified through clustering.
- **Maintenance Scheduling:** Improve service scheduling precision by aligning maintenance activities with client-specific behavior patterns, reducing downtime and costs.

- **Risk Management:** Implement targeted monitoring and early-warning systems for clusters identified as high-risk, significantly reducing the likelihood of operational disruptions.

Strategic Recommendations

Based on clustering evaluation, several recommendations emerged:

- **Proactive Interventions:** Establish proactive maintenance programs tailored specifically for high maintenance and irregular client clusters.
- **Customized Service Plans:** Develop customized service agreements reflecting cluster specific requirements and resource demands.
- **Enhanced Monitoring:** Implement continuous monitoring for high-risk customers identified through DBSCAN's outlier detection.

5.6 Predictive Modeling for Failure Detection

While predictive modeling was not yet implemented at this stage of the project, a core objective of this case study was to lay the groundwork for developing such models by first establishing a reliable framework for customer classification. The insights derived from clustering client behavior provide a strategic foundation for future failure prediction efforts.

5.6.1. Objective and Strategic Context

The novelty of this work lies in its approach which is rather than jumping directly into failure prediction using generalized assumptions, the study begins by identifying and analyzing client specific patterns, repair frequency, maintenance cost, and purchase behavior. This allows for a structured understanding of how different customer profiles interact with pump systems, which is essential before attempting any predictive modeling. Traditional predictive maintenance often focuses solely on equipment

monitoring data or time-based metrics and also overlooking the variability introduced by customer-specific behaviors. Our approach fills this gap by recognizing that clients themselves, through how they operate, maintain, and purchase can directly impact the probability and type of equipment failure.

5.6.2. Contribution of the current Work

This phase of the project achieved the following:

- **Customer Segmentation Framework:** Using K-Means, DBSCAN, and GMM, customers were classified into operationally meaningful groups, such as reliable, high-maintenance, and irregular clients.
- **Behavioral Feature Mapping:** Key behavioral indicators such as repair intervals, cost trends, and purchase frequency were engineered and normalized to prepare the dataset for future modeling.
- **Data Requirements and Quality Assessment:** Based on the limitations identified in the datasets (e.g., missing timestamps, lack of failure codes), clear recommendations were made to improve the structure and content of future data collection processes, particularly for developing robust failure prediction models.

These steps form the essential prerequisites for any future application of machine learning or AI-based predictive maintenance.

5.6.3. Framework for Future Failure Prediction

Based on the current work, a conceptual framework for future predictive modeling is proposed. It consists of:

1. **Client-Based Input Features:** Incorporating cluster membership, historical repair frequency, average maintenance intervals, and normalized cost behavior into future predictive models.
2. **Enhanced Dataset Requirements:** Adding time-resolved data (e.g., timestamps of failure events, cause codes, duration of downtime) to support more accurate and temporally aware models.
3. **Modeling Strategy:** Use of interpretable machine learning models (e.g., decision trees or probabilistic models) that allow integration of behavioral, transactional, and temporal inputs.

This framework not only anticipates future developments in predictive maintenance but also ensures that those models will be deeply grounded in client specific realities. Although the predictive modeling phase has not yet been executed, this study provides a critical stepping stone by organizing the data, identifying behavioral patterns, and classifying clients in a way that informs and enhances future modeling efforts. By recognizing the operational significance of customer behavior in pump maintenance, this research bridges the gap between client profiling and data driven predictive maintenance offering a scientifically grounded and industry relevant framework for future implementation.

5.7 Design and Proposal of an Operational Framework for Customer Insight and Maintenance Planning

The long-term objective of this project is to enable predictive maintenance capabilities through a comprehensive and client-informed analytical framework. While full predictive modeling was not completed due to the absence of real-time operational data, this phase of the study delivered critical foundational components client segmentation, data quality analysis, and strategic recommendations that form the backbone of such a framework.

5.7.1. Framework Components and Current Contributions

The proposed operational framework is composed of three core pillars:

1. Client Classification and Behavioral Segmentation
2. Comprehensive Data Assessment and Integration Strategy
3. Recommendations for Future Predictive Modeling Enablement

Each of these components was addressed through detailed, methodical analysis using Technosub's internal datasets, providing both immediate operational insights and long-term strategic value.

5.7.1.1. Client Classification and Behavioral Segmentation

The first completed section involved unsupervised clustering of clients using three algorithms: K-Means, DBSCAN, and GMM, to identify operational behavior patterns among customer groups. These segments included:

- **Reliable Clients** (low maintenance needs, stable usage)
- **High-Maintenance Clients** (frequent breakdowns, high repair costs)
- **Irregular Clients** (unpredictable behavior, atypical patterns)

This segmentation offers Technosub actionable insight into where to allocate service resources, how to tailor maintenance plans, and where to anticipate risk. It also serves as a crucial feature layer for future failure prediction models.

5.7.1.2. Dataset Examination and Data Quality Assessment

A critical barrier to predictive accuracy is data quality. This study conducted a detailed assessment of four major datasets provided by Technosub:

- **MWI-suividereparationNormalizedCost** (repair transactions and cost tracking)
- **TFA-Vente Client** (client purchasing log)
- **SUIVI REPARATION** (repair tracking and history)
- **CAUSE DE BRISE** (failure classification logs)

The review identified several recurring issues: (see figures 25 – 28)

- Inconsistent categorical naming (e.g., SOP types, item descriptions)
- Missing values (pump serial numbers, failure causes, customer IDs)
- Structural gaps (lack of timestamps, absence of failure code linkage)
- Cost-related anomalies (zero or negative values, abnormal outliers)

These issues degrade the reliability of both real-time insights and historical trend analysis, especially in the context of predictive modeling. As a response, this work proposed concrete data governance strategies, including:

- Implementation of standardized naming conventions
- Categorization of pump models and components
- Mandatory field enforcement and intelligent imputation strategies
- Timestamp inclusion for all service lifecycle events
- Failure code taxonomy integration

These recommendations directly support the technical feasibility of future modeling, reducing data inconsistency and increasing interpretability.

SOP Type	SOP Number	Item Number	Item Description	QTY	Normalized unit	Normalized total	Customer Number
Invoice	MWV100002	002-103-26	Cable Protection Tube	1	0.000015 \$	0.000015 \$	T-MIN10

Figure 25. Example of zero or extremely small cost values

SOP Type	SOP Number	Item Number	Item Description	QTY	Normalized unit	Normalized total	Customer Number
Invoice	MW114399	900-083-60	Electric Cable # 4/0 - 3C	-150	(0.046572) \$	6.985755 \$	T-ELK10

Figure 26. Example of negative values in "QTY" and cost columns

Invoice	MW104328	008-156-19	# 1-3 / Rubber Boot	1	0.000052 \$	0.000052 \$	T-DET00
Invoice	MW104511	008-117-18	# 1-2 / Rubber Boot	1	0.000061 \$	0.000061 \$	T-DET00

Figure 27. Example of lack of categorization for item numbers

SOP Type	SOP Number	Item Number	Item Description	QTY	Normalized unit	Normalized total cost	Customer Number
Invoice	MW100002	99	ELECTRIC SHIELDED C	1	0.000001 \$	0.000001 \$	T-MIN10
SOP Type	SOP Number	Item Number	Item Description	QTY	Normalized unit	Normalized total cost	Customer Number
Invoice	MW108643	900-083-51	# 4 - 3C SHD-GC / Elect	500	0.079600 \$	39.800001 \$	T-ASS00

Figure 28. Example of extreme cost values

5.7.1.3. Strategic Framework for Enabling Predictive Modeling

While the implementation of failure prediction algorithms was not completed due to the absence of real-time operational data, the analytical foundation created here is essential for enabling such models in the future.

Problems Identified:

- No access to live pump sensor data or time-series failure logs
- Incomplete linkage between maintenance actions and real operational context (e.g., runtime hours, usage load)
- Scarce time-based tracking of repair and downtime events

Without these dimensions, any predictive modeling would risk low generalizability and high false positive/negative rates. Rather than pursuing a model with weak reliability, this project focused on designing the architecture and identifying the data structures required for robust, client-informed predictive models.

Proposed Enhancements to Technosub’s Data Ecosystem Include:

- Integration of IoT sensor feeds from deployed pumps (e.g., vibration, temperature, flow rate)
- Real-time repair logging systems with full lifecycle timestamps
- Consistent failure categorization linked to client usage context
- Cross-referencing Enterprise Resource Planning (ERP), Computerized Maintenance Management System (CMMS), and Quality Control (QC) systems to create unified equipment histories

These enhancements would allow for the development of robust predictive models that not only forecast failure probability but do so based on client usage patterns, something rarely addressed in current industrial applications.

5.7.2. Operational and Strategic Value

By separating the foundational work (client classification, data integrity assessment) from the implementation of predictive models, this project ensures a modular, scalable roadmap for Technosub’s digital transformation. It is scientifically grounded and operationally realistic, recognizing that predictive models cannot be effective without a high-fidelity, structured data environment. Moreover, the framework uniquely emphasizes customer behavior as a key predictive variable, a novel contribution compared to standard equipment-only modeling approaches. This positions Technosub to not only predict failures but to understand *why* they occur more frequently in certain segments, and to intervene pre-emptively.

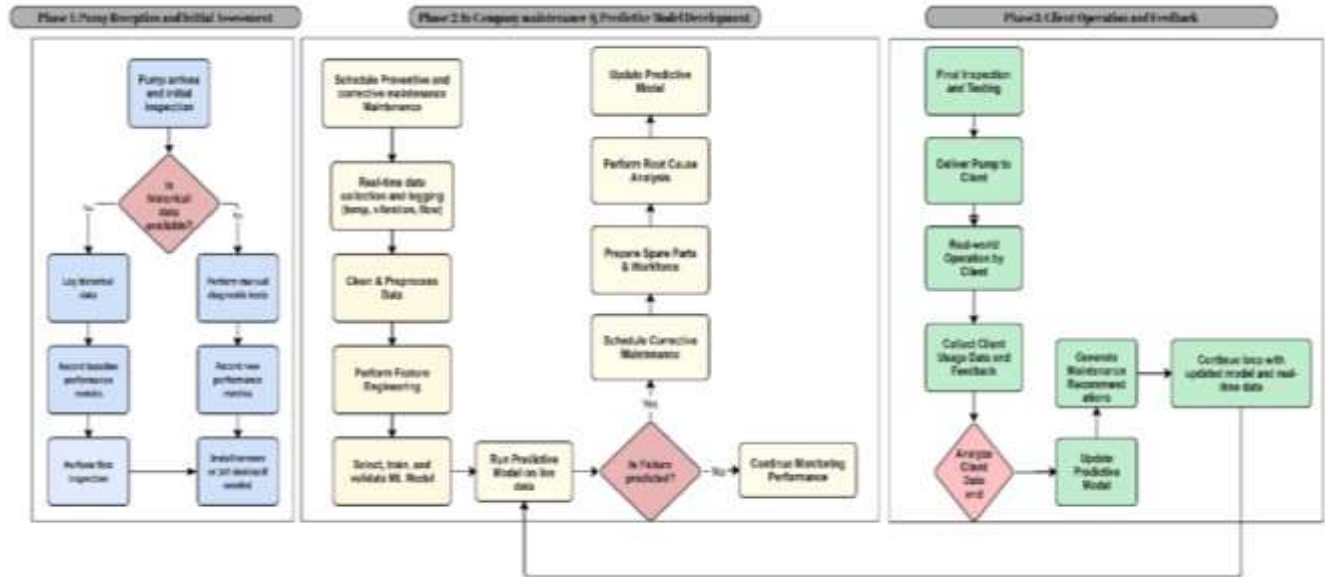


Figure 29. Integrated three-phase predictive maintenance framework for industrial pump systems.

The diagram presented in Figure 29 illustrates the sequential structure of the proposed three-phase predictive maintenance framework. It outlines the full operational lifecycle, from the initial reception and assessment of the pump to machine learning-based failure prediction, and finally to the integration of continuous client feedback.

Phase 1 (pump Reception and Initial Assessment): The process initiates upon the arrival of the pump, where a decision is made based on the availability of historical operational data. If such data exist, they are logged and used to establish baseline performance metrics. If not, manual diagnostic procedures are conducted to record new baseline data. IoT sensors are installed as needed to enable real-time data collection. This phase ensures that each unit enters the predictive pipeline with complete performance context.

Phase 2 (In-Company Maintenance and Predictive Model Development): In this phase, internal operations are supported by machine learning for predictive maintenance. Real-time data, such as vibration, temperature, and flow, is logged and

preprocessed. Feature engineering is then applied to extract key attributes for failure prediction. These features inform the development and validation of machine learning models. Once deployed, the model continuously analyzes live data to detect potential failure events. If a failure is predicted, the system schedules corrective maintenance, prepares spare parts, and performs root cause analysis. If no failure is detected, normal performance monitoring continues. This phase encapsulates the dynamic decision-making core of predictive maintenance.

Phase 3 (Client Operation and Feedback): After a final inspection, the pump is delivered to the client. During its operational life, client-specific data, usage patterns, environmental factors, and performance metrics are collected and analyzed. This feedback loop enables continuous model refinement, supports the generation of tailored maintenance recommendations, and ensures alignment with actual usage conditions. Updated models and insights are reintegrated into earlier phases, creating a closed-loop system.

5.8 Data Quality Assessment and Recommendations for Improvement

The reliability and precision of any data-driven industrial strategy depend on the quality of the data supporting it. In the context of this project, where the goal is to build a behavior informed framework for predictive maintenance, poor data quality would not only hold back model accuracy but also increase misleading or inapplicable insights. As such, this section details by evaluation of the datasets provided by Technosub, identifies critical limitations affecting analysis, and offers a set of strategically aligned recommendations to enhance data quality for advanced modeling applications. The datasets were analyzed, ranging from cost-tracking repair logs and sales records to historical failure classifications were assessed through both quantitative metrics and qualitative business aligned logic. These two layered approaches enabled the detection of not only surface-level issues like null values or duplicates, but also deeper structural

problems affecting relational integrity. One of the key findings during the assessment was the disintegration of information across datasets. For example, pump serial numbers appeared inconsistently or were missing altogether in certain repair logs, making it difficult to trace the lifecycle of a specific asset across sales, repairs, and failures. Similarly, failure records existed in isolation, without links to the associated client, product, or SOP. This disconnection significantly undermines the possibility of extracting causal relationships between customer behavior and maintenance outcomes, a prerequisite for meaningful predictive analysis. Another critical barrier to analytical depth was the absence of time resolved data. Most datasets only included basic date entries without repair duration, downtime length, or timestamps. Without time-based resolution, it becomes impossible to reconstruct usage intervals, detect failure progression, or perform sequence modeling. Moreover, the lack of structured failure codes and the inconsistency in descriptive fields (e.g., “Invoice” vs. “invoice”) created ambiguity in categorical analysis. In industrial contexts, where interpretation drives action, such ambiguity translates into operational noise. Efforts were made to diminish these issues at the preprocessing level. Data cleaning approach included standardizing labels, imputing missing values using contextual logic, and applying transformations to normalize skewed distributions. Nonetheless, these corrective actions, while essential, were not sufficient to overcome structural deficiencies. The absence of a unified identifier across all data sources remained a major limitation, making full data integration more difficult. This division not only restricts the scope of descriptive analytics but also compromises the foundation required for future supervised or unsupervised learning models. Beyond technical fixes, the assessment also addressed the procedural dimension of data quality. The lack of enforced standardization at the point of data entry, whether through manual logs or unstructured exports, was identified as a recurring theme. For example, customer names appeared under multiple spellings, or with varying suffixes referencing internal divisions. These inconsistencies compromise grouping logic and weakened the accuracy of clustering algorithms. Moreover, certain essential fields, such as “PO Client” in sales records or failure causes

in service reports, were often incomplete or omitted altogether, indicating a gap in data governance protocols. The novelty of this work lies in how the data quality evaluation was aligned with the analytical and strategic goals of the project. Rather than treating completeness and consistency as abstract metrics, the assessment was grounded in the functional purpose of the data: to enable client segmentation, inform predictive models, and support operational decision making. This purpose driven lens allowed for a more subtle prioritization of interventions, such as recognizing that the absence of timestamps poses a more immediate threat to predictive modeling than cosmetic label mismatches. The recommended improvements are therefore not isolated data cleaning tasks, but foundational steps toward architectural readiness. These include introducing a standardized classification for failure codes, enforcing field validation at the point of entry, consolidating client identities through a controlled naming system, and expanding data capture to include real-time operational variables such as pump usage hours, runtime anomalies, and environmental load conditions. Importantly, the inclusion of timestamps for repair starts and end times would enable calculation of downtime and repair efficiency, two critical variables for lifecycle cost modeling and failure probability estimation. The data quality assessment was a core analytical phase of the project. It revealed what was missing, and what is essential: temporal context, structural linkage, semantic standardization, and behavioral traceability. Addressing these gaps is not simply a matter of improving database cleanliness, it is a prerequisite for enabling the kind of intelligent, client-informed predictive maintenance system that this framework ultimately aims to support.

To contextualize the findings of this study, it is instructive to examine comparable industrial case studies that have implemented predictive maintenance (PdM) strategies. These cases highlight the critical role of data quality, integration, and strategic planning in the successful deployment of PdM systems.

A structured summary of the main data quality issues, their analytical implications, and the aligned recommendations is presented in Figure 30.

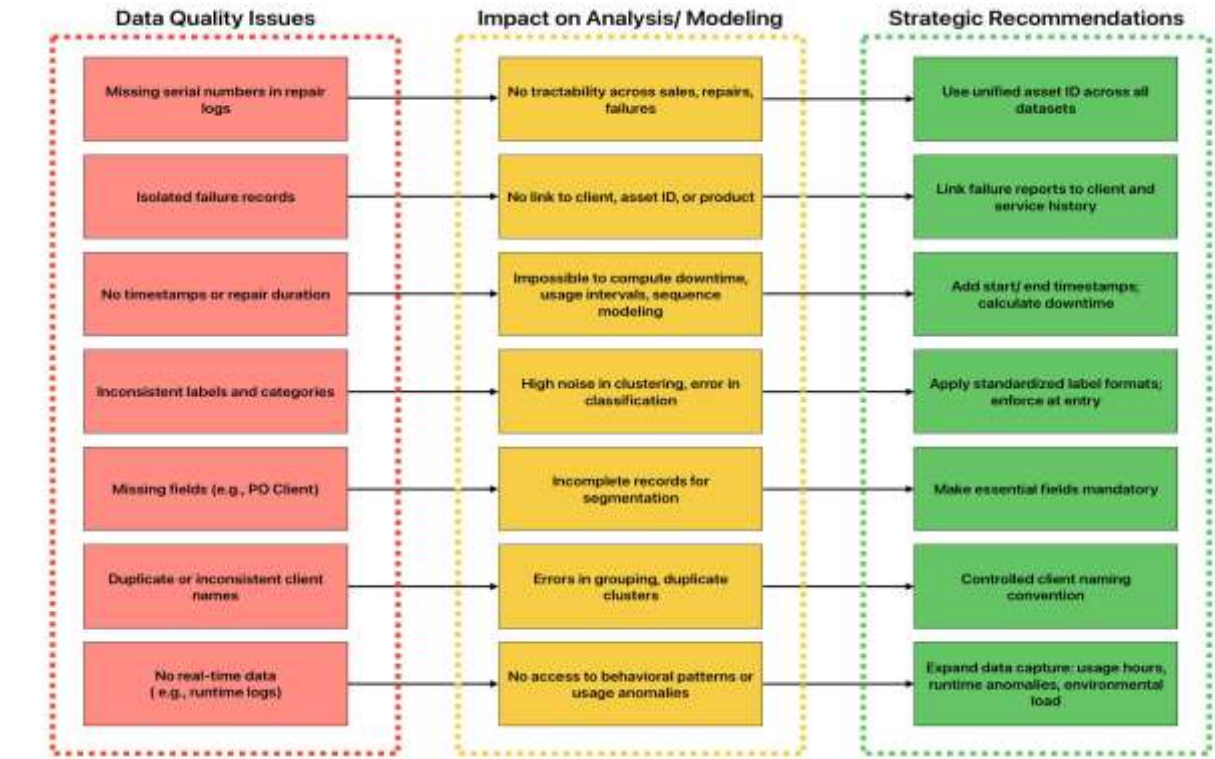


Figure 30. Data Quality Challenges, Analytical Impact, and Recommended Solutions.

This diagram summarizes the key data quality issues identified in the industrial dataset provided by Technosub, their impact on client segmentation and predictive modeling, and the corresponding strategic recommendations proposed to enable reliable, interpretable, and scalable predictive maintenance systems.

5.9 Comparative Insights from Similar Industrial Case Studies

To contextualize the contribution of this study within the broader field of predictive maintenance, it is essential to examine several key industrial case studies that explore complementary or alternative methodologies. Each of the following studies, while differing in scope, provides important insights that either reinforce the validity of this thesis's framework or suggest valuable directions for future integration. In particular, the comparative review includes a semantic modeling approach developed within the

Z-BRE4K (Zero-Breakdown Manufacturing Process Based on Cognitive Predictive Maintenance and Equipment Optimization) project, a hybrid clustering-classification model for compressor systems, a synthetic data augmentation strategy using Conditional Tabular Generative Adversarial Network (CTGAN) for rare failure prediction, and a transformer-based anomaly detection framework for industrial health prognosis. The Z-BRE4K project, presented by May et al. [42], introduces a semantic model for predictive maintenance rooted in a two-cycle ontology construction process that emphasizes vocabulary standardization and contextual modeling. This approach enables the semantic integration of heterogeneous maintenance data, allowing for structured reasoning over system components, failure causes, and diagnostic processes. By leveraging ontologies based on established standards and applying principles of Failure Mode and Effects Analysis (FMEA), their model improves data traceability and system-wide decision-making. However, while the Z-BRE4K framework excels in system-level structuring and interoperability, it lacks mechanisms to capture behavioral variability across customers. The present study addresses this gap by profiling customer behavior patterns using clustering algorithms, enabling client-specific maintenance strategies. A hybrid approach that enriches customer profiles with semantic metadata could allow future systems to achieve both interpretability and behavioral precision.

A second relevant contribution is provided by Costa et al. [43], who developed a hybrid clustering-classification framework for predictive maintenance of high-pressure industrial compressors. Their study focuses on using unsupervised clustering algorithms, such as K-Means, GMM, HDBSCAN, and OPTICS, not just for segmentation, but to generate synthetic features that improve the performance of supervised models like SVC and Random Forest. By embedding clustering derived features into the predictive pipeline, the authors reported accuracy improvements of nearly 5% on average, validating the value of hidden pattern recognition prior to classification. This approach directly supports the rationale of this thesis, in which

unsupervised clustering is used not as a secondary step but as the foundation for customer profiling. Moreover, their methodology strengthens the case for clustering even in systems with complex failure modes. A third case study by Khosravi et al. [44] addresses a persistent limitation in predictive maintenance: the rarity of failure events. Their work introduces a synthetic data augmentation framework using Conditional Tabular GANs (CTGAN) and Synthetic Minority Over-sampling Technique (SMOTE) to balance class distributions and improve model recall in predicting paper break events in the pulp and paper industry. The significance of this study lies in its demonstration that synthetic data can meaningfully enhance model training where real-world failure data is sparse or imbalanced. This insight has clear relevance to the present research, particularly when modeling rare or extreme client behaviors with few historical examples. In future developments, integrating CTGAN-generated data to simulate underrepresented client segments could enhance the robustness of predictive models developed in this thesis. Finally, Poland et al. [45] propose an advanced deep learning architecture, Transformer Quantile Regression Neural Network (TQRNN) that combines quantile regression neural networks with transformer models for machine-level health prognosis. Deployed in a beverage manufacturing facility, their system processes sensor data from 43 features across multiple machines and predicts failure events up to one hour in advance with over 70% accuracy. The integration of quantile estimation enables the identification of anomalies in different operating conditions, while the transformer layer captures long-range dependencies in temporal data. Although their focus is primarily on equipment prognopolandsis rather than customer behavior, the architectural design of their pipeline offers strategic inspiration for modeling irregular or sporadic client behavior. Notably, their approach to outlier detection through quantile estimation parallels this study's use of DBSCAN for behavioral anomaly identification. Moreover, the structured modularity of their pipeline reflects the logic of CRISP-DM followed in the present research. In synthesizing these case studies, it becomes clear that while each explores a different aspect of predictive maintenance from semantic knowledge modeling and feature-

enhanced clustering to synthetic data generation and deep anomaly detection. They collectively support the design and direction of this thesis. Specifically, the current work distinguishes itself by combining interpretable unsupervised learning with structured data preprocessing and segmentation strategies to model customer-specific maintenance needs. The potential to integrate elements from each of these studies semantic tagging for interpretability, CTGAN for rare behavior simulation, and transformer-based modeling for temporal prediction, suggests a clear roadmap for future research that aims to build scalable, intelligent, and client-aware predictive maintenance systems.

5.10 Strategic Value of the Proposed Approach

The strategic value of the approach developed in this case study lies in its ability to bridge the gap between traditional equipment centered predictive maintenance models and client-specific analytical framework. Unlike standard predictive maintenance systems, which focus solely on machine level parameters and time based weakening, the proposed methodology integrates customer behavioral data, such as repair frequency, monetary cost patterns, and purchase behavior into the maintenance planning process. This dual focus model aligns both operational and business intelligence, enhancing decision making across technical and managerial dimensions. Also this framework introduces a scalable and modular data driven architecture that begins with robust feature engineering, progresses through unsupervised learning based client segmentation, and concludes in strategic recommendations that are operationally actionable. By using K-Means, DBSCAN, and GMM together, we captured clear customer profiles, identified irregular behaviors, and effectively handled overlapping patterns, making the results practically useful. These insights enable tailored maintenance interventions and the prioritization of resource allocation, which is not achievable through generic, asset only models. One of the most strategically valuable aspects of the framework is its focus on behavior informed predictive

maintenance. By analyzing client-specific behavioral trends prior to building failure prediction models, the approach ensures that any future supervised learning tasks are grounded in the real-world variability of customer interaction with assets. This significantly improves the explanatory power of predictive models and reduces the risk of over fitting to generalize patterns that do not reflect the operational context of specific client segments. In addition, the data preprocessing and integration strategy directly enhances the long-term utility of enterprise data. By identifying and correcting inconsistencies, such as missing serial numbers, misaligned identifiers, and unstructured failure descriptions, the proposed method not only facilitates accurate clustering and modeling but also lays the foundation for a unified equipment history across repair, sales, and failure datasets. This reinforces the role of data governance as a pillar of predictive maintenance readiness and ensures that future analytic applications can leverage clean, context rich datasets. Another strategic benefit is the operational adaptability of the framework. Rather than being limited to static historical analysis, the clustering results can be used to design dynamic customer dashboards, automated flagging of high risk clients, and simulation environments to test the effect of different maintenance policies on specific client groups. This opens pathways for continuous improvement and real-time decision support, particularly when the framework is integrated with ERP, CMMS, or IoT systems in future deployments. Moreover, by including client segmentation as a precondition for predictive modeling, the framework addresses a critical oversight in conventional PdM systems: the tendency to treat all clients and equipment contexts as homogeneous. This approach corrects that assumption, introducing granularity that increases predictive reliability and business relevance. The result is a predictive maintenance paradigm that is not only more accurate, but more economically viable, as it aligns maintenance schedules and service intensity with both equipment health and client specific risk profiles. The proposed framework aligns with industry wide digital transformation goals, particularly in asset heavy sectors such as mining and industrial manufacturing. As companies increasingly seek to implement AI and machine learning for operational

excellence, the need for interpretable, modular, and behaviorally grounded analytics is critical. This framework provides a clear roadmap toward that vision, balancing academic rigor with industrial feasibility.

The strategic value of this approach lies in its multidimensional impact: it enhances analytical depth, improves predictive reliability, optimizes maintenance resources, and aligns data science initiatives with practical business objectives. By integrating behavioral segmentation with technical diagnostics, the proposed framework represents a foundational step toward intelligent, customer aware predictive maintenance systems, positioning Technosub to transition from reactive operations to proactive, data driven service leadership.

This chapter demonstrated the application of clustering and data analytics techniques to real industrial data in order to support customer profiling and predictive maintenance strategies. The analysis highlighted the value of integrating technical, operational, and transactional data to uncover actionable insights related to equipment behavior and customer usage patterns. The proposed framework illustrates how data-driven segmentation can enhance proactive maintenance planning and customer relationship management. The results and limitations identified in this chapter form the basis for the global conclusions and future research directions discussed in Chapter 6.

CHAPTER 6

6. Conclusion and Future Work

6.1 Summary of Key Contributions

This thesis introduces a dual-case analytical framework designed to address predictive maintenance challenges in both simulated and real-world industrial environments. A key contribution lies in the structured application of the CRISP-DM methodology, which provided a systematic foundation for the client profiling and behavior analysis conducted in Case Study 2. Although the simulation-based study in Case Study 1 did not adopt the CRISP-DM model explicitly, its development process reflected the same foundational logic; beginning with a clearly defined operational problem, progressing through structured data design and preprocessing, and concluding in a modular framework designed for eventual deployment with real-world data. This alignment ensured methodological consistency while respecting the different data realities of each case.

In Case Study 1, the research produced a technically rigorous simulation framework for underground mining operations, based on synthetic data generated across several interdependent domains, such as: equipment maintenance, environmental monitoring, ground support integrity, and incident communication. These datasets were intentionally structured with realistic interdependencies to reflect operational causality, enabling the development and evaluation of intelligent system logic. A modular system architecture was designed to process these inputs, detect anomalies, and trigger system-level alerts through dashboard visualizations, offering a practical framework for testing automation protocols in constrained environments.

In Case Study 2, the study analyzed enterprise data from an industrial pump manufacturer to segment clients based on their maintenance behaviors. Through

systematic feature engineering, variables such as repair frequency, average repair intervals, monetary cost, and buying frequency were derived, normalized, and used to classify customers using three clustering algorithms: K-Means, DBSCAN, and Gaussian Mixture Models. These unsupervised learning models revealed distinct client segments (reliable, high-maintenance, and irregular) providing a foundation for proactive service planning and resource optimization. The project further demonstrated the importance of data preprocessing tailored to industrial datasets. Missing value imputation, outlier filtering, label normalization, and dataset integration were executed to consolidate fragmented repair, sales, and failure records into a unified analytical foundation. This preprocessing phase was especially critical given the structural inconsistencies and semantic ambiguities identified during the data quality assessment. Additionally, the thesis proposed a strategic operational framework that links behavioral client segmentation with future predictive modeling. While supervised failure prediction models were not yet implemented, the groundwork laid through segmentation, feature extraction, and structural assessment defines a clear pathway for integrating such models. The study also included expert validation of cluster results to ensure operational relevance and business applicability. By combining simulation innovation with real-world data application, this research demonstrates a scalable and adaptable methodology for predictive maintenance, grounded in both technical precision and industrial relevance. The ability to bridge these two domains, synthetic data modeling and enterprise data analytics, marks a distinctive contribution to the field of intelligent maintenance systems.

6.2 Research Limitations

While this thesis presents a comprehensive analytical framework combining synthetic simulation and real-world industrial data, several limitations must be acknowledged. These limitations are related to both methodological constraints and the inherent characteristics of the datasets used in each case study. In Case Study 1, the use of

synthetic data, while necessary due to the unavailability of operational mining datasets, introduces limitations in terms of external validity. Although significant effort was made to design realistic simulation conditions grounded in domain literature and operational logic, the absence of real-time sensor input and field-validated events constrains the extent to which the findings can be generalized to live mining environments. Moreover, the simulation logic, while modular and interdependent, was manually encoded and lacked adaptive behavior or feedback learning mechanisms, limiting its dynamic response capabilities. In Case Study 2, the industrial datasets provided by the pump manufacturer were characterized by fragmentation, missing fields, inconsistent identifiers, and structural gaps. These issues restricted the analytical depth of certain components, particularly the predictive modeling phase. The lack of high-resolution temporal data (e.g., detailed timestamps, downtime duration, or usage cycles) slowed down the development of time-aware models and sequence-based analysis. Additionally, the absence of unified failure codes and limited linkage between transactional records and equipment condition metrics prevented a full reconstruction of asset lifecycles. Across both cases, another constraint was the limited availability of ground-truth validation. In the industrial case, although clustering results were reviewed by domain experts, there were no labeled failure outcomes or confirmed customer classifications to quantitatively assess clustering accuracy or prediction validity. Similarly, in the synthetic case, evaluation relied primarily on logical consistency and scenario testing rather than empirical comparison to real-world outcomes.

Finally, while the methodological design emphasized scalability and modularity, the actual deployment of the developed systems, either the simulation framework or the clustering pipeline, was not conducted in a live industrial setting. As a result, questions of real-time integration, system performance under operational load, and user interface effectiveness remain untested. These limitations do not undermine the validity of the findings but rather define the scope within which they should be interpreted. They also

highlight clear directions for refinement and expansion, which are addressed in the subsequent sections on practical recommendations and future research.

6.3 Practical Recommendations

Based on the findings of both case studies, several practical recommendations are proposed to enhance the implementation of predictive maintenance strategies, improve data governance, and support decision-making in industrial settings. For organizations operating in data-scarce or safety-critical environments, such as underground mining, simulation frameworks like the one developed in Case Study 1 offer a valuable early-stage testing platform. It is recommended that companies lacking real-time operational data adopt similar modular and domain-informed synthetic environments to validate system logic, train decision-making algorithms, and refine interface designs before live deployment. To increase realism and analytical utility, future implementations should integrate sensor installation, dynamic feedback loops, and adaptive behavior modeling into the simulation environment.

In the context of industrial operations with existing datasets, as demonstrated in Case Study 2, the immediate priority should be the standardization of data structures and the enforcement of consistent data entry protocols. Incomplete identifiers, inconsistent naming conventions, and missing timestamps significantly reduce the effectiveness of analytical methods. Enterprises can implement controlled vocabularies (which is a foundational requirement for analytical integrity, model accuracy, system automation, and ultimately, business decision-making.), mandatory field validation at data entry points, and periodic audits of data quality. Specifically, the introduction of structured failure codes, time stamped maintenance events, and customer identifiers that are consistent across systems will directly enhance the feasibility of supervised predictive modeling. The clustering results from Case Study 2 revealed distinct customer segments with varying maintenance behaviors and operational risks. Organizations are

advised to operationalize these segments through tailored service plans, differentiated monitoring strategies, and customized maintenance intervals. High-maintenance or irregular clients, for example, should be prioritized for proactive interventions and continuous monitoring. Reliable clients may be candidates for extended service contracts with reduced frequency but data-driven reassessment points. From a system design perspective, both studies emphasize the importance of modularity, traceability, and interpretability. Any future predictive maintenance or automation solution should be constructed with loosely coupled components to ensure adaptability, particularly in environments where data availability or system connectivity may be inconsistent. Visual interfaces, such as dashboards, should prioritize clarity, alert differentiation, and end-user usability to avoid cognitive overload and support operational decision-making. Finally, a phased implementation strategy is recommended for transitioning from exploratory analytics to real-time deployment. Organizations should begin by validating segmentation and insights using historical data, followed by pilot integration of dashboards or alert systems in controlled operational zones. Feedback from domain users should inform iterative improvements to both the logic and presentation of predictive models. This staged approach ensures that predictive systems are not only technically robust but also aligned with operational realities and stakeholder needs. In addition to the analytical findings, the data quality assessment conducted during this study identified critical improvements required in the company's datasets to support future predictive capabilities. Specific recommendations were provided to standardize naming conventions (e.g., SOP types, customer names, failure causes), implement structured classification systems for pumps and parts, and introduce failure codes with taxonomies. It was also recommended that Technosub enhance data completeness by ensuring all service records include timestamps, pump serial numbers, and customer purchase orders. Furthermore, redundant entries and duplicates across repair logs and transaction records should be removed to maintain analytical integrity. These foundational data governance measures are necessary precursors to achieving traceability, comparability, and reliability in any future AI-driven maintenance

systems. In practice, implementing these recommendations requires some operational modifications in the company's data workflows. For example, item descriptions such as "Pump Seal (OLD)" and "pumpseal" were recorded in different formats across datasets, making it difficult to aggregate repair statistics. Standardizing such entries through controlled input lists or reference tables significantly improves classification accuracy. Similarly, introducing structured failure codes and timestamped repair stages enables the transition from descriptive reporting to predictive modeling. These changes are not just technical; they require coordination across data entry and quality control processes.

To support the transition from descriptive analytics to predictive maintenance, a comprehensive assessment of the company's datasets was conducted. This analysis identified several recurring data quality issues that directly affect traceability, model accuracy, and the operational usability of analytical outputs. The table below summarizes these issues, their impact on predictive capabilities, and the corresponding recommendations provided to the industrial partner. These corrective actions are essential to establish a reliable and scalable data infrastructure that can support future predictive modeling, anomaly detection, and behavior-based decision-making.

Issue Identified	Impact on Analysis or Modeling	Recommended Action
Inconsistent SOP types, item descriptions, customer names	Misclassification of records, inaccurate grouping, duplication in analysis	Implement controlled vocabularies and standardize using reference tables
Missing pump serial numbers and PO Client fields	Inability to link repairs to assets or transactions, weak traceability	Enforce mandatory fields at data entry, infer missing values where applicable
Lack of timestamps in repair and transaction logs	Prevents lifecycle tracking, repair duration analysis, and sequence-based modeling	Add start_time and end_time fields; enable time tracking in systems
No structured failure codes	Hinders failure type analysis, limits predictive modeling of breakdowns	Introduce standardized failure taxonomy and train staff on code usage

Duplicate records in repair and sales datasets	Distorts transaction counts and inflates cost/failure rates	Implement duplicate detection scripts and maintain unique constraints (e.g., SOP ID)
Unclassified pump models and parts	Limits component-level insights and behavior trend analysis	Categorize pump models and components; link to type definitions via lookup tables
Zero or unrealistic cost values	Twists financial summaries and invalidates cost-based features	Set validation rules for numeric fields; flag and review outliers using statistical filters

Table 8. *Summary of key data quality issues and their impacts*

6.4 Directions for Future Research

This research establishes a robust analytical foundation for predictive maintenance and customer profiling using both synthetic and industrial datasets. However, several areas remain open for future investigation, particularly in advancing the predictive modeling capabilities, enhancing dataset structure, and extending system deployment. A primary direction for future work involves the development and implementation of failure prediction models based on the segmented client profiles established in Case Study 2. While this thesis laid the groundwork by engineering relevant features, such as repair frequency, average maintenance intervals, and normalized cost metrics, actual predictive modeling was not performed due to the absence of labeled failure events and real-time operational data. Future studies should focus on collecting time-resolved datasets that include precise timestamps for failures, repair durations, operational hours, and downtime events. This temporal granularity is essential for training supervised learning models capable of capturing failure progression and accurately estimating time-to-failure. Additionally, integrating usage context variables, such as load conditions, runtime cycles, and environmental stressors, would enable a more comprehensive modeling framework that reflects real-world failure mechanisms. Another area of future research is the incorporation of real-time IoT sensor data into the customer profiling and predictive maintenance pipeline. While this thesis operated on static historical datasets, integrating live telemetry (e.g., vibration data, temperature readings, and flow rates) would allow dynamic monitoring and improve the precision

of anomaly detection. Such integration requires the development of data pipelines that synchronize sensor inputs with maintenance records and customer behavior profiles, ensuring time-aligned and context-rich input to predictive algorithms. Furthermore, sensor-based failure indicators can be used to supplement behavioral features, improving the robustness of both classification and prediction models. The data architecture itself also requires further enhancement to support scalable and interpretable predictive analytics. As demonstrated in Section 5.8, a major limitation was the lack of structural linkage across repair, sales, and failure records. Future work should aim to establish a unified equipment history by implementing relational identifiers (e.g., persistent pump IDs, standardized customer keys) across all data tables. This would allow for longitudinal analysis of equipment performance at the customer level, enabling survival modeling, root cause analysis, and predictive risk scoring. Moreover, the inclusion of structured classifications for failure modes, maintenance actions, and part types would facilitate semantic modeling and enable reasoning-based diagnostics. Methodologically, future research can explore the use of advanced machine learning architectures such as recurrent neural networks (RNNs), transformer-based models, or hybrid systems that combine clustering outputs with probabilistic classification. These models are capable of processing sequential data and learning complex dependencies across time, making them suitable for modeling cumulative damage, delayed failure effects, and irregular maintenance intervals. However, their effectiveness is conditional upon the availability of high-quality, time-series data, which must be prioritized in future data collection strategies. Another promising direction is the integration of explainability tools into the modeling pipeline. A critical area for future development is the operational deployment of the proposed framework. While the current study evaluated clustering results and dashboard mock-ups offline, the next step involves real-time integration into enterprise systems. This would require building automated data ingestion routines, deploying clustering or prediction models as callable services, and embedding outputs into user interfaces for maintenance engineers or operations managers. Pilot deployments in controlled

environments can help evaluate system responsiveness, usability, and alignment with real operational workflows. In summary, future research should focus on enabling predictive modeling through improved temporal and structural data, incorporating sensor streams, applying advanced algorithms, enhancing model interpretability, and operationalizing the framework through system-level integration. These directions are essential to transform the current analytical foundation into a fully functional, intelligent maintenance system capable of delivering measurable operational value. While this study was grounded in the mining and industrial pump sectors, the underlying methodology, particularly the integration of behavior-based segmentation with data-driven maintenance planning is applicable to other asset-intensive industries such as manufacturing, energy, and transportation, where similar operational challenges and data fragmentation issues exist.

References

1. Liu, X., Zhang, X., Wang, L., Qu, F., Shao, A., Zhao, L., Wang, H., Yue, X., Li, Y., Yan, W., & He, J. (2024). Research progress and prospects of intelligent technology in underground mining of hard rock mines. *Green and Smart Mining Engineering*, 1(1), 12–26. <https://doi.org/10.1016/j.gsme.2024.03.007>
2. Raza, F. (2023). AI for predictive maintenance in industrial systems. *Cosmic Bulletin of Business Management*, 2(1). <http://doi.org/10.13140/RG.2.2.27313.35688>
3. Shimaponda-Nawa, M., Nwaila, G. T., Zhang, S. E., & Bourdeau, J. E. (2023). A framework for measuring the maturity of real-time information management systems (RTIMS) in the mining industry. *The Extractive Industries and Society*, 16, 101368. <https://doi.org/10.1016/j.exis.2023.101368>
4. Meyer Zu Wickern, Vincent. (2019). Challenges and Reliability of Predictive Maintenance. <http://doi.org/10.13140/RG.2.2.35379.89129>.
5. de la Fuente, R., Radrigan, L., & Morales, A. S. (2024). Enhancing predictive maintenance in mining mobile machinery through a TinyML-enabled hierarchical inference network. <https://arxiv.org/abs/2411.07168>
6. Zhang, H., Li, B., Karimi, M., Saydam, S., & Hassan, M. (2023). Recent advancements in IoT implementation for environmental, safety, and production monitoring in underground mines. *IEEE Internet of Things Journal*, PP, 1–1. <https://doi.org/10.1109/JIOT.2023.3267828>
7. Bisschoff, R. A. D. P., & Grobbelaar, S. (2022). Evaluation of data-driven decision-making implementation in the mining industry. *South African Journal of Industrial Engineering*, 33(3), 218–232. <http://dx.doi.org/10.7166/33-3-2799>
8. Young, E. S., Frankenhuys, W. E., & Ellis, B. J. (2020). Theory and measurement of environmental unpredictability. *Evolution and Human Behavior*, 41(6), 550–556. <https://doi.org/10.1016/j.evolhumbehav.2020.08.006>

9. Widana, Anura. (2019). *Environmental impacts of the mining industry: A literature review*. <http://doi.org/10.13140/RG.2.2.11702.86083>
10. Rahnema, M., Amirmoeini, B., & Moradi Afrapoli, A. (2023). Incorporating environmental impacts into short-term mine planning: A literature survey. *Mining*, 3(1), 163–175. <https://doi.org/10.3390/mining3010010>
11. Tayebi-Khorami, M., Edraki, M., Corder, G., & Golev, A. (2019). A review of current and emerging approaches for mining waste management. *Minerals*, 9(5), 286. <https://doi.org/10.3390/min9050286>
12. Servin, M., Vesterlund, F., & Wallin, E. (2021). Digital twins with distributed particle simulation for mine-to-mill material tracking. *Minerals*, 11, 524. <https://doi.org/10.3390/min11050524>
13. Bazargur, B., Bataa, O., & Budjav, U. (2023). Reliability study for communication system: A case study of an underground mine. *Applied Sciences*, 13(2), 821. <https://doi.org/10.3390/app13020821>
14. Scalabrini, L., Zanella, A., & Vilajosana, X. (2023). LoRa multi-hop networks for monitoring underground mining environments. In *2023 IEEE Globecom Workshops (GC Wkshps)* (pp. 696–701). IEEE. <https://doi.org/10.1109/GCWkshps58843.2023.10464954>
15. Dayo-Olupona, O., Genc, B., Celik, T., & Bada, S. (2023). Adoptable approaches to predictive maintenance in mining industry: An overview. *Resources Policy*, 86(Part A), 104291. <https://doi.org/10.1016/j.resourpol.2023.104291>
16. Zhang, J., Wang, J., Zhang, Q., Wang, N., Wang, C., & Wang, L. (2022). Design and workspace analysis of drilling arm of mining anchor drilling robot. *Journal of Physics: Conference Series*, 2202(1), 012057. <https://doi.org/10.1088/1742-6596/2202/1/012057>
17. Bisset, C., Venter, P. V. Z., & Coetzer, R. (2023). A systematic literature review on machine learning applications at coal-fired thermal power plants for improved energy efficiency. *International Journal of Sustainable Energy*, 42(1), 845–872. <https://doi.org/10.1080/14786451.2023.2244618>

18. Liu, P., Hao, J., Hu, H., Luan, X., & Meng, B. (2024). Design optimization of underground mining vehicles based on regenerative braking energy recovery. *Applied Sciences*, 14(1), 467. <https://doi.org/10.3390/app14010467>
19. Sherifdeen, K. (2024). *Improving operational efficiency through automation*, EasyChair Preprint 13532. <https://easychair.org/publications/preprint/S16h>
20. Kozłowski, A., & Bołoz, Ł. (2021). Design and research on power systems and algorithms for controlling electric underground mining machines powered by batteries. *Energies*, 14(1), 4060. <https://doi.org/10.3390/en14134060>
21. Kasem, M. S., Hamada, M., & Taj-Eddin, I. (2024). Customer profiling, segmentation, and sales prediction using AI in direct marketing. *Neural Computing and Applications*, 36, 4995–5005. <https://doi.org/10.1007/s00521-023-09339-6>
22. Ryhänen, A. (2023). *Designing customer segmentation model for analyzing consumer data* (Master's thesis, Tampere University). <https://urn.fi/URN:NBN:fi:tuni-202305115612>
23. John, J. M., Shobayo, O., & Ogunleye, B. (2023). An exploration of clustering algorithms for customer segmentation in the UK retail market. *Analytics*, 2, 809–823. <https://doi.org/10.3390/analytics2040042>
24. Rivera-Castro, R., Pilyugina, P., Pletnev, A., Maksimov, I., Wyz, W., & Burnaev, E. (2019). *Topology-based clusterwise regression for user segmentation and demand forecasting*. In **2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)** (pp. 326–336). IEEE. <https://doi.org/10.1109/DSAA.2019.00048>.
25. Abrokwah-Larbi, K. (2023). The role of generative artificial intelligence (GAI) in customer personalisation (CP) development in SMEs: A theoretical framework and research propositions. *Industrial Artificial Intelligence*, 1(11), 1–11. <https://doi.org/10.1007/s44244-023-00012-4>
26. Wu, Jun, Shi, Li, Lin, Wen-Pin, Tsai, Sang-Bing, Li, Yuanyuan, Yang, Liping, Xu, Guangshu, [Retracted] An Empirical Study on

- Customer Segmentation by Purchase Behaviors Using a RFM Model and K-Means Algorithm, *Mathematical Problems in Engineering*, 2020, 8884227, 7 pages, 2020. <https://doi.org/10.1155/2020/8884227>
27. Talaat, F. M., Aljadani, A., Alharthi, B., Farsi, M. A., Badawy, M., & Elhosseini, M. (2023). A mathematical model for customer segmentation leveraging deep learning, explainable AI, and RFM analysis in targeted marketing. *Mathematics*, 11(3930). <https://doi.org/10.3390/math11183930>
 28. Pawełoszek, I., & Korczak, J. (2016). An approach to discovery of customer profiles. In *10th International Conference on Research and Practical Issues of Enterprise Information Systems (CONFENIS)* (pp. 88–99). Vienna, Austria. https://doi.org/10.1007/978-3-319-49944-4_7
 29. Turkmen, B. (2022). Customer Segmentation with Machine Learning for Online Retail Industry. *The European Journal of Social & Behavioural Sciences*, Volume 31(Issue 2), 111-136. <https://doi.org/10.15405/ejsbs.316>
 30. Jagabathula, S., Subramanian, L., & Venkataraman, A. (2017). A model-based projection technique for segmenting customers. *Operations Research*. <https://doi.org/10.48550/arXiv.1701.07483>
 31. Mach-Król, M., & Hadasik, B. (2021). A certain research gap in big data mining for customer insights. *Applied Sciences*, 11, 6993. <https://doi.org/10.3390/app11156993>
 32. Maryani and D. Riana, "Clustering and profiling of customers using RFM for customer relationship management recommendations," *2017 5th International Conference on Cyber and IT Service Management (CITSM)*, Denpasar, Indonesia, 2017, pp. 1-6, [doi: 10.1109/CITSM.2017.8089258](https://doi.org/10.1109/CITSM.2017.8089258).
 33. Giri, C., Thomassey, S., & Zeng, X. (2019). Customer analytics in fashion retail industry. In *Functional Textiles and Clothing* (pp. 349–361). https://doi.org/10.1007/978-981-13-7721-1_27

34. Kit, T. C., & Azmi, N. F. M. (2023). Diving into data: An analysis of customer profiling. *Vizologi Report*, December. <https://vizologi.com/diving-into-consumer-segmentation-analysis/>
35. Lukens, S., Naik, M., Saetia, K., & Hu, X. (2019). Best practice framework for improving maintenance data quality. In *PHM Society Annual Conference, 2019*. <http://doi.org/10.36001/phmconf.2019.v11i1.836>
36. Guo, X., Yu, Q., Alm, C. O., Calvelli, C., Pelz, J. B., Shi, F., & Haake, A. R. (2014). From spoken narratives to domain knowledge: Mining linguistic data for medical image understanding. *Artificial Intelligence in Medicine*, 62, 79–90. <https://doi.org/10.1016/j.artmed.2014.08.001>
37. Saucedo-Dorantes, J. J., Arellano-Espitia, F., Delgado-Prieto, M., & Osornio-Rios, R. A. (2021). Diagnostic methodology based on deep feature learning for the identification of defects in metallic, hybrid and ceramic bearings. *Sensors*, 21, 5832. <https://doi.org/10.3390/s21175832>
38. Ahmad, Z., Nguyen, T. K., Ahmad, S., Nguyen, C. D., & Kim, J. M. (2022). Multi-stage centrifugal pump fault diagnosis using informative principal component analysis. *Sensors*, 22, 179. <https://doi.org/10.3390/s22010179>
39. Rajakarunakaran, S., Venkumar, P., Devaraj, D., & Rao, K. S. P. (2008). Artificial neural network approach for fault detection in rotating system. *Applied Soft Computing*, 8, 740–748. <https://doi.org/10.1016/j.asoc.2007.06.002>
40. Lewaaelhamd, I. (2024). Customer segmentation using machine learning model: An application of RFM analysis. *Journal of Data Science and Intelligent Systems*, 2(1), 29–36. <https://doi.org/10.47852/bonviewJDSIS32021293>
41. Ali, A. O. D., & Elaswad, H. M. A. (2025). Predictive maintenance approaches in Industry 4.0: A systematic literature review. *Journal of Industrial Information Integration*, 30, 100–115. <http://doi.org/10.1109/WETICE57085.2023.10477802>
42. May, G., Cho, S., Majidirad, A., & Kiritsis, D. (2022). A semantic model in the context of maintenance: A predictive maintenance case study. *Applied Sciences*, 12(12), 6065. <https://doi.org/10.3390/app12126065>

43. Costa, A., Mastriani, E., Incardona, F., Munari, K., & Spinello, S. (2024). Predictive maintenance study for high-pressure industrial compressors: Hybrid clustering models.
<https://doi.org/10.48550/arXiv.2411.13919>
44. Khosravi, H., Farhadpour, S., Grandhi, M., Raihan, A.S., Das, S., & Ahmed, I. (2023). Strategic Data Augmentation with CTGAN for Smart Manufacturing: Enhancing Machine Learning Predictions of Paper Breaks in Pulp-and-Paper Production.
<https://doi.org/10.48550/arXiv.2311.09333>
45. Poland, D. J., Puglisi, L., & Ravi, D. (2024). Industrial machines health prognosis using a transformer-based framework. <https://doi.org/10.48550/arXiv.2411.14443>
46. Mallioris, P., Diamantis, E., Bialas, C., & Bechtsis, D. (2024). Predictive maintenance framework for assessing health state of centrifugal pumps. *IAES International Journal of Artificial Intelligence*, 13(1), 850–862.
<https://doi.org/10.11591/ijai.v13.i1.pp850-862>
47. Wuest, T., Weimer, D., Irgens, C., & Thoben, K.-D. (2016). Machine learning in manufacturing: Advantages, challenges, and applications. *Production & Manufacturing Research*, 4(1), 23–45.
<https://doi.org/10.1080/21693277.2016.1192517>
48. Tabianan, K., Velu, S., & Ravi, V. (2022). K-means clustering approach for intelligent customer segmentation using customer purchase behavior data. *Sustainability*, 14(7243), 1–15. <https://doi.org/10.3390/su14127243>
49. Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., & Gao, R. X. (2019). Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing*, 115, 213–237. <https://doi.org/10.1016/j.ymssp.2018.05.050>
50. Zhang, Y., Zhou, T., Huang, X., Cao, L., & Zhou, Q. (2021). Fault diagnosis of rotating machinery based on recurrent neural networks. *Measurement*, 171, 108774.
<https://doi.org/10.1016/j.measurement.2020.108774>