

**IDENTIFYING AND ANALYSING RECURRING INCIDENTS IN IT
SERVICE MANAGEMENT USING UNSUPERVISED MACHINE
LEARNING: A CASE STUDY FROM PT XYZ**

By

Muhammad Al Fayed
22451048

MASTER'S DEGREE
in

MASTER OF INFORMATION TECHNOLOGY
FACULTY OF ENGINEERING AND INFORMATION TECHNOLOGY



SWISS GERMAN UNIVERSITY
The Prominence Tower
Jalan Jalur Sutera Barat No. 15, Alam Sutera
Tangerang, Banten 15143 - Indonesia

Revision after the Thesis Defense on 5 February 2026

STATEMENT BY THE AUTHOR

I hereby declare that this submission is my own work and to the best of my knowledge, it contains no material previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any other degree or diploma at any educational institution, except where due acknowledgement is made in the thesis.



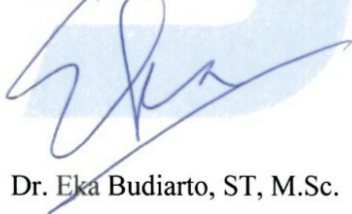
Muhammad Al Fayed

Student

14.2.2026

Date

Approved by:



Dr. Eka Budiarto, ST, M.Sc.

Thesis Advisor

14.2.2026

Date



Dr. Maulahikmah Galinium, S.Kom., M.Sc.

Thesis Co-Advisor

14.2.2026

Date



Dr. Maulahikmah Galinium, S.Kom., M.Sc.

Dean

14.2.2026

Date

Muhammad Al Fayed

ABSTRACT

By

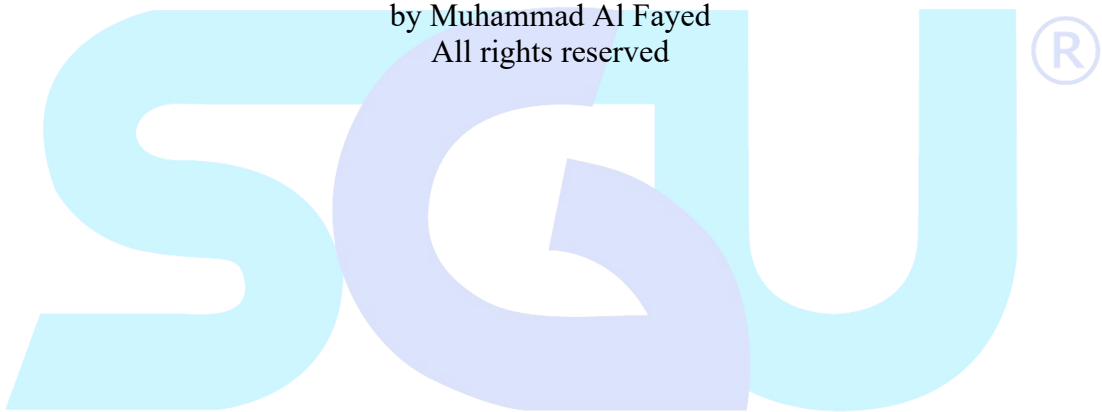
Muhammad Al Fayed
Dr. Eka Budiarto, ST, M.Sc., Advisor
Dr. Maulahikmah Galinium, S.Kom., M.Sc., Co-Advisor

SWISS GERMAN UNIVERSITY

This study addresses the operational inefficiency of reactive IT Service Management (ITSM) at PT XYZ, where the accumulation of 82,328 unstructured incident records (2018–2022) impedes root cause analysis. The research objective is to transition from reactive firefighting to Proactive Problem Management by applying Unsupervised Machine Learning techniques. Guided by the CRISP-DM methodology, a Text Mining pipeline utilizing TF-IDF vectorisation was developed to benchmark symptom-based features against resolution-enriched features. The findings demonstrated that the resolution-enriched K-Means Clustering model ($k=88$) yielded superior semantic distinctiveness, achieving a Silhouette Coefficient of 0.084 and successfully isolating systemic defects from routine noise. Expert validation confirmed the model's efficacy, achieving 100% consensus on high-priority clusters such as Oracle Logic Errors and K2 Instability. Consequently, a "Bifurcation Strategy" was formulated, routing high-volume incidents to automation and high-risk defects to engineering via seven expert-validated Standard Operating Procedures (SOPs). This study confirms that combining unsupervised learning with human-in-the-loop governance effectively translates raw historical data into an actionable, proactive operational strategy.

Keywords: IT Service Management, Unsupervised Machine Learning, K-Means Clustering, TF-IDF, Proactive Problem Management.

© Copyright 2026
by Muhammad Al Fayed
All rights reserved



DEDICATION

I dedicate this work to my beloved family, whose unconditional love and prayers have been my constant source of strength. To my lecturers, I offer this as a testament to your guidance and the wisdom you have shared. Ultimately, I dedicate this research to the future of the country I love, Indonesia, in the hope of contributing to its technological advancement.



ACKNOWLEDGEMENT

First and foremost, I would like to express my deepest praise and gratitude to Allah, the Almighty God, for His blessings, mercy, and guidance, which provided me with the strength and health to complete this thesis.

I would like to extend my sincere appreciation to my thesis advisor for his invaluable mentorship, patience, and critical insights throughout this research. His guidance was instrumental in shaping the methodology and direction of this study. I also express my gratitude to my co-advisor for his strategic advice and constructive feedback that greatly refined the quality of this work.

My sincere thanks go to the management and engineering team at PT XYZ for granting me access to the dataset and for their active participation in the Focus Group Discussions. Their domain expertise was essential in validating the findings of this research.

I am deeply grateful to my wife, my parents and my entire family for their endless love, prayers, and moral support, which have been my rock during this academic journey.

Finally, I thank my fellow master's students and colleagues for their encouragement, discussions, and shared struggles that made this journey memorable.

TABLE OF CONTENTS

	Page
STATEMENT BY THE AUTHOR.....	2
ABSTRACT.....	3
DEDICATION.....	5
ACKNOWLEDGEMENT	6
TABLE OF CONTENTS.....	7
LIST OF FIGURES	14
LIST OF TABLES.....	16
CHAPTER 1 - INTRODUCTION.....	18
1.1 Background.....	18
1.2 Problem Statement.....	19
1.3 Research Objectives.....	22
1.4 Research Questions.....	23
1.5 Hypothesis.....	23
1.6 Significance of the Study	24
1.7 Scope and Limitations.....	25
1.8 Research Structure	26
CHAPTER 2 - LITERATURE REVIEW	28
2.1 Company Background and Context	28
2.2 IT Service Management (ITSM) and IT Infrastructure Library (ITIL)	30
2.2.1 ITIL as Dominant ITSM Framework.....	31
2.2.2 Challenges in ITIL Implementation.....	32
2.2.3 Evolution to ITIL 4	34
2.2.4 Relevance to This Research	35
2.3 Incident Management and Recurring Issues	35
2.4 Problem Management and Root Cause Analysis.....	38
2.4.1 Traditional RCA Techniques in ITSM	38
2.4.2 Traditional vs. Data-Driven RCA Approaches	39

2.4.3	Why Unsupervised Machine Learning is Promising for RCA	40
2.4.4	Relevance to This Research	40
2.5	Data Analytics in IT Service Management.....	41
2.6	Cross-Industry Standard for Data Mining (CRISP-DM)	43
2.6.1	Overview of CRISP-DM.....	43
2.6.2	Phases of CRISP-DM in the Context of ITSM	45
2.6.3	Integrating CRISP-DM with ITIL and the Research Framework.....	46
2.7	Natural Language Processing (NLP) and Dimensionality Reduction	47
2.7.1	Comparative Analysis of Vectorisation Techniques.....	50
2.7.2	Justification for Selecting TF-IDF	53
2.7.3	Feature Extraction: TF-IDF	54
2.7.4	Dimensionality Reduction: UMAP	54
2.8	Uniform Manifold Approximation and Projection (UMAP)	55
2.8.1	Mathematical Formulation of UMAP	56
2.8.2	Comparative Analysis of Dimensionality Reduction Techniques	57
2.8.3	Justification for Selection	59
2.9	Machine Learning Foundations for ITSM Analytics.....	60
2.9.1	Overview of Machine Learning (ML) Paradigms	60
2.9.2	Unsupervised Learning (ML) and Clustering Techniques.....	61
2.10	Comparative Analysis of Clustering Algorithms.....	63
2.10.1	Evaluation of Clustering Methodologies	63
2.10.2	Evaluation of Distance Metrics.....	66
2.10.3	Analytical Strategy Overview	69
2.10.4	Selection of Clustering Algorithms	70
2.10.5	Mathematical Formulation of Selected Algorithms.....	72
2.11	Measuring Metrics for Clustering Validity	75
2.11.1	Silhouette Coefficient	75
2.11.2	Davies-Bouldin Index	76
2.11.3	Calinski-Harabasz Index (Variance Ratio Criterion).....	77
2.11.4	Rationale for Multi-Metric Evaluation	78
2.12	Validation and Policy Development in Data-Driven ITSM	78
2.12.1	The Role of Validation in Data Analytics.....	78
2.12.2	From Analytical Insights to Policy Development.....	79
2.12.3	Integrating Validation and Policy into the Research Framework.....	80
2.13	Research Design & Approach.....	81

2.13.1	Research Paradigm and Philosophy	81
2.13.2	Research Design.....	81
2.13.3	Integration of CRISP-DM and ITIL	82
2.13.4	Rationale for Design Selection	82
2.13.5	Research Framework Flow	83
2.14	Research Framework Operationalisation	83
2.14.1	Business Understanding.....	84
2.14.2	Data Understanding	84
2.14.3	Data Preparation.....	85
2.14.4	Modelling.....	85
2.14.5	Evaluation	86
2.14.6	Deployment and Policy Development	88
2.15	Research Gap and Conceptual Synthesis	89
2.15.1	Literature Synthesis	89
2.15.2	Identified Research Gaps	92
2.15.3	Conceptual Synthesis and Study Contribution	93
2.15.4	Critical Review of Prior Studies	94
CHAPTER 3 - RESEARCH METHODOLOGY		97
3.1	Research Framework	97
3.2	Data Source and Description	97
3.2.1	Data Source	98
3.2.2	Data Characteristics	98
3.2.3	Data Quality & Considerations	99
3.2.4	Relevance to Research Objectives	100
3.3	Ethical Considerations	100
3.3.1	Data Privacy and Anonymisation	100
3.3.2	Compliance with Legal and Organisational Policies	101
3.3.3	Responsible Use of Machine Learning	101
3.3.4	Ethical Value and Organisational Impact	102
3.4	Tools and Implementation Environment	102
3.4.1	Software and Libraries	102
3.4.2	Hardware Environment.....	104
3.4.3	Implementation Workflow and Reproducibility	104
3.4.4	Technical Implementation Architecture	105
3.5	NLP Text Extraction and Dimensionality Reduction	106

3.5.1	Data Cleaning and Integration	107
3.5.2	Text Preprocessing	108
3.5.3	Natural Language Processing (NLP) and Feature Engineering	109
3.5.4	Dimensionality Reduction	111
3.5.5	Toolset and Implementation	112
3.6	Clustering, Modelling, and Analytical Techniques	112
3.6.1	Algorithmic Optimisation Workflows	112
3.6.2	Parameter Tuning and Model Optimisation	115
3.6.3	Cluster Evaluation Metrics	118
3.6.4	Post-Clustering Interpretation and Labelling	119
3.7	Clusters Evaluation Phase and Focused Group Discussion (FGD)	121
3.7.1	Quantitative Evaluation of Clustering Models	121
3.7.2	Qualitative Validation Through Expert Review	122
3.7.3	Focus Group Discussion (FGD) Objectives and Outputs	123
3.7.4	Integration of Quantitative and Qualitative Results	124
3.7.5	FGD Indicators and Question Mapping	125
3.8	Policy Development and Recommendation Derivation	129
3.8.1	Framework for Policy Translation	129
3.8.2	Risk Assessment and Prioritisation Matrix	131
3.8.3	Justification of ITIL Practice Alignment	133
3.8.4	Design of Standard Operating Procedures (SOPs)	135
3.9	Expert Selection Criteria	136
3.9.1	Expert Role Alignment with Research Framework	137
3.9.2	Inclusion Criteria for Expert Selection	137
3.9.3	Required Competencies for Validation	138
3.9.4	Sampling Strategy and Expert Composition	139
3.9.5	Ethical and Confidentiality Considerations	140
3.10	Research Success Criteria	140
3.10.1	Technical Validity of Analytical Models	141
3.10.2	Interpretability and Semantic Coherence	142
3.10.3	Expert Validation and Acceptance	142
3.10.4	Operational Actionability	143
3.10.5	Policy Development and Governance Alignment	144
3.10.6	Success Metrics and Measurement Protocols	145
CHAPTER 4 - RESULTS AND ANALYSIS		148

4.1	Overview of the Analysis.....	148
4.2	Data Exploration and Descriptive Statistics	149
4.2.1	Dataset Overview	149
4.2.2	Incident Volume and Temporal Patterns	150
4.2.3	Distribution by Service and Category	151
4.2.4	Distribution by Team	152
4.2.5	Characteristics of Unstructured Text Fields	152
4.2.6	Summary of Exploration Findings	153
4.3	Data Preprocessing Results.....	153
4.3.1	Text Cleaning and Normalisation	153
4.3.2	Processed Text Variants for Model 1 and Model 2	154
4.3.3	Description Text Statistics	155
4.3.4	Structured Data Cleaning and Feature Engineering.....	156
4.3.5	TF-IDF Vectorisation.....	157
4.4	Feature Space Construction and Vectorization Analysis.....	159
4.4.1	Comparative Feature Engineering: Model 1 vs. Model 2.....	159
4.4.2	Vocabulary Selection Strategy.....	160
4.4.3	Vectorization Walkthrough: Sample Ticket #100131	163
4.4.4	Top Detected Features	164
4.5	KMeans Clustering Results.....	167
4.5.1	Selection of Optimal k: The Granularity Trade-off	167
4.5.2	Model 1 Clustering Findings: Symptom-Based Grouping	170
4.5.3	Model 2 Clustering Findings: Process and Outcome-Based Grouping	172
4.5.4	Comparative Analysis of K-Means Model Variants.....	174
4.5.5	Cluster Volume Distribution and Pareto Analysis.....	175
4.5.6	Deep Dive Analysis of Dominant Clusters	177
4.6	DBSCAN Clustering Results & UMAP Dimensionality Reduction	179
4.6.1	UMAP Visualisation and Data Structure	179
4.6.2	Model 1 DBSCAN Findings: Symptom Isolation	180
4.6.3	Model 2 DBSCAN Findings: High-Precision Anomaly Detection	182
4.6.4	Comparative Analysis of DBSCAN Model Variants	183
4.6.5	Anomaly Detection and The “Long Tail”	185
4.7	Comparative Analysis and Strategic Interpretation	185
4.7.1	Hyperparameter Tuning and Sensitivity Analysis	186
4.7.2	Evaluation: Partitioning (K-Means) vs. Density (DBSCAN).....	189

4.7.3	Cluster Interpretation Methodology	191
4.7.4	Outlier Management and Strategic Prioritisation	193
4.7.5	Findings Synopsys and Remediation Roadmap.....	195
4.7.6	Policy Implications and Remediation Roadmap.....	196
4.8	Clusters Validation and Evaluation Through FGD.....	199
4.8.1	Expert Panel Composition and Competency Alignment	199
4.8.2	Interpretability and Categorisation Alignment	200
4.8.3	Operational Validity and Root Cause Consistency.....	201
4.8.4	Actionability and Knowledge Management Relevance.....	202
4.8.5	Risk Assessment Framework for Cluster Prioritisation.....	203
4.8.6	Prioritisation, Impact, and SLA Implications	206
4.8.7	Synthesis: Model Trustworthiness	206
4.8.8	Evaluation Results per Success Indicator	207
4.9	Policy Development.....	210
4.9.1	Overview of Strategic Recommendations	210
4.9.2	SOP Formulation Methodology and Cluster Mapping	212
4.9.3	Problem Management Protocols (Systemic Defect Resolution).....	215
4.9.4	Service Request Standardisation (Operational Efficiency).....	218
4.9.5	Knowledge Management and Shift-Left Strategy	220
4.9.6	Proactive Asset Management.....	223
4.9.7	Implementation Roadmap	224
4.9.8	Participant Demographics and Session Protocol	226
4.10	Expert Validation and Focus Group Discussion (FGD) Results	228
4.10.1	Participant Demographics and Session Protocol	228
4.10.2	Validation of Cluster Interpretability (Phenomenological Confirmation). 230	
4.10.3	Consensus Building and Deliberation Protocol	232
4.10.4	Assessment of Actionability and SOP Feasibility	233
4.10.5	Correction of Bias and Strategic Refinements.....	235
4.10.6	Evaluation Results by Success Indicator	236
CHAPTER 5 - CONCLUSION AND RECOMMENDATIONS.....		242
5.1	Chapter Overview	242
5.2	Conclusions Derived from Research Questions	242
5.2.1	Optimization of Unsupervised Learning for ITSM (Answering RQ1).....	243
5.2.2	Operationalisation through Expert Governance (Answering RQ2).....	244
5.3	Strategic Recommendations for PT XYZ	244

5.3.1	The "Bifurcation Strategy"	245
5.3.2	Implementation Roadmap Summary	246
5.4	Contributions to Knowledge	247
5.4.1	Theoretical Contribution	247
5.4.2	Practical Contribution	247
5.5	Research Limitations	248
5.6	Suggestions for Future Work	249
5.6.1	Transition to Supervised Learning for Real-Time Automation	249
5.6.2	Integration of Generative AI (LLMs) for Knowledge Synthesis	249
5.6.3	Advanced Outlier Analysis using Statistical Metrics	250
5.7	Closing Remarks	250
REFERENCES		252
APPENDICES		262
Appendix A: Curriculum Vitae		262
Appendix B: Python Code for the Thesis Analysis		263
Appendix C: Data Result of Finding Best k Algorithm		285
Appendix D: DBSCAN Hyperparameter Tuning Logs		289
Appendix E: FGD 01 Documentation		291
Appendix F: FGD 02 Documentation		296

LIST OF FIGURES

Figures	Page
Figure 1.1 Fishbone diagram of the problem statement	21
Figure 2.1 Proactive problem management with machine learning	30
Figure 2.2 Visual diagram of ITIL 4 evolution	34
Figure 2.3 CRISP-DM framework process	44
Figure 3.1 Research Framework	97
Figure 3.2 Implementation Architecture Diagram	105
Figure 3.3 Pseudocode for NLP Preprocessing and Feature Engineering	110
Figure 3.4 Pseudocode for Dimensionality Reduction	111
Figure 3.5 K-Means Optimisation Workflow	113
Figure 3.6 DBSCAN Optimisation Workflow	115
Figure 3.7 Pseudocode for Finding Best K Algorithm	117
Figure 3.8 Pseudocode for Hierarchical Cluster Drill-Down	120
Figure 3.9 Policy Translation Flow Diagram	130
Figure 3.10 Risk Matrix Diagram	132
Figure 3.11 ITIL Practice Mapping Decision Tree	134
Figure 4.1 Incident by Year	150
Figure 4.2 Dataset's top categories	151
Figure 4.3 Model 1 K-Means Cluster Distribution	170
Figure 4.4 Model 1 K-Means Cluster Scatter Plot	171
Figure 4.5 Model 2 K-Means Cluster Distribution	172

Figure 4.6 Model 2 K-Means Cluster Scatter Plot	173
Figure 4.7 Model 2 Top Clusters Volume Distribution	175
Figure 4.8 Model 1 DBSCAN UMAP Projection	181
Figure 4.9 Model 1 DBSCAN Cluster Distribution	181
Figure 4.10 Model 2 DBSCAN UMAP Projection	182
Figure 4.11 Model 2 DBSCAN Cluster Distribution	183
Figure 4.12 Strategic Recommendation Map (Bifurcation Strategy)	212
Figure 4.13 SOP-01 Workflow – Fast-Tracking Oracle LPN Errors (Cluster 43)	216
Figure 4.14 SOP-02 Workflow – K2 Platform Stabilisation (Cluster 58)	217
Figure 4.15 Shift-Left Operational Strategy (Cluster 74 & 19)	220
Figure 4.16 Knowledge-Centred Service (KCS) Workflow Lifecycle	222
Figure 4.17 Hardware Reliability Bathtub Curve	224
Figure 4.18 Strategic Implementation Roadmap (Phased Approach)	226
Figure 4.19 Focus Group Discussion (FGD) Validation Protocol	229

LIST OF TABLES

Table	Page
Table 2.1 Key challenge in ITIL implementation	33
Table 2.2 Key challenges of recurring incidents in ITSM	37
Table 2.3 Recent applications of machine learning in ITSM	42
Table 2.4 Comparison of Text Vectorisation Methodologies	51
Table 2.5 Comparison of Dimensionality Reduction Techniques	57
Table 2.6 Comparative analysis of clustering algorithms	64
Table 2.7 Summary of selected prior studies and gaps	67
Table 2.8 Summary of selected prior studies and gaps	89
Table 3.1 Features of the dataset categorised based on field group	98
Table 3.2 Features of the dataset categorised based on field group	103
Table 3.3 Validation approach of the research	124
Table 3.4 FGD Indicators and Question Mapping	126
Table 3.5 Standard Operating Procedure (SOP) Design Template	136
Table 3.6 Summary of Success Metrics and Measurement Protocols	145
Table 4.1 After preprocessing text description result	155
Table 4.2 Comparison of Input Features for Ticket #100131	160
Table 4.3 Sample of TF-IDF Vector Representation for Ticket #100131	163
Table 4.4 Sample of Top 20 High-Importance Features in the Vector Space	165
Table 4.5 Top 20 k clustering result based on metrics performance	168
Table 4.6 Comparison of the K-Means clustering metrics	174

Table 4.7 Distribution of Top 5 Dominant Clusters	176
Table 4.8 Comparison of the DBSCAN clustering metrics	184
Table 4.9 Proposed N-Tier Incident Management Model	194
Table 4.10 Comprehensive Cluster Analysis and Proposed Solutions	195
Table 4.11 Proposed Mapping Dominant Clusters to Strategic Policies	196
Table 4.12 Cluster Risk Assessment Matrix	205
Table 4.13 Summary of Expert Validation Results by Indicator	208
Table 4.14 Strategic Recommendation Portfolio	211
Table 4.15 Mapping of Clusters to Standard Operating Procedures (SOPs)	213
Table 4.16 FGD Competency Mapping	228
Table 4.17 Expert Labelling Alignment	230
Table 4.18 SOP Feasibility and Governance Validation	234
Table 4.19 Evaluation of Research Success Indicators (Expert Validation)	237