

TABLE OF CONTENTS

STATEMENT	2
TABLE OF CONTENTS	4
LIST OF TABLES	6
LIST OF FIGURES	7
ABSTRACT	8
CHAPTER 1 INTRODUCTION	1
1.1 Research Background	1
1.2 Research Problem	2
1.3 Research Boundary	3
1.4 Research Aim / Objective	4
1.5 Research Advantage	4
CHAPTER 2 LITERATURE REVIEW	5
CHAPTER 3 THEORETICAL BASIS	9
3.1 Machine Learning Models	9
3.1.1 Logistic Regression	9
3.1.2 Decision Trees	12
3.1.3 Random Forest	15
3.1.4 Extra-Trees Classifier	17
3.1.5 Gradient Boosting	19
3.1.6 XGBoost	21
3.1.7 LightGBM	25
3.1.8 Cat Boost	28
3.1.9 K Nearest Neighbors (KNN)	30
3.1.10 Multi-layer Perceptron (MLP)	33
3.1.11 Naive Bayes	35

3.2 Pareto Principle	37
CHAPTER 4 RESEARCH METHODOLOGY	38
4.1 Research Description	38
4.2 Research Dataset	38
4.3 Research Model	40
4.4 Research Evaluation	42
CHAPTER 5 IMPLEMENTATION	44
5.1 Development Tools and Environment	44
5.2 Programming Code and Explanations	45
5.3 Identified Limitation	51
CHAPTER 6 RESULTS AND DISCUSSION	53
6.1 Models Results Using Pre-Selected Features	53
6.2 Pareto Analysis on Best Model	56
6.3 Evaluating XGBoost with Full Features versus Pareto Features	58
CHAPTER 7 CONCLUSION	63
7.1 Summary of Findings	63
7.2 Implications	63
7.3 Limitations	63
7.4 Recommendations for Future Research	64
REFERENCES	65

LIST OF TABLES

Table 2.1 Literature Review Content/Description	7
Table 4.1 Beneficiary and Fraud Label Data	39
Table 4.2 Inpatient and Outpatient Data	40
Table 5.1 Development Environment Specifications	46
Table 5.2 Python Libraries and Modules	47

LIST OF FIGURES

Figure 4.1 Research Flow	42
Figure 4.2 Evaluation Flow	44
Figure 5.1 Notebook Error During One-Hot Encoding Transformation	54
Figure 6.1 F1 Score Across Machine Learning Models	55
Figure 6.2 AUC Score Across Machine Learning Models	56
Figure 6.3 Precision versus Recall among Different Models	56
Figure 6.4 Training Times of Machine Learning Models	57
Figure 6.5 Performance Metrics Heatmap	58
Figure 6.6 Top Feature Importances	59
Figure 6.7 Pareto Analysis of Feature Importance	59
Figure 6.8 Performance and Resource Usage Comparison	61
Figure 6.9 Sorted Feature Importance in Pareto-Optimized Model	61
Figure 6.10 SHAP Waterfall Plot for Full Feature Set XGBoost Model	62
Figure 6.11 SHAP Waterfall Plot for Pareto-Optimized XGBoost Model	62
Figure 6.12 SHAP Waterfall Plot for Full Feature Set XGBoost Model	63
Figure 6.13 SHAP Waterfall Plot for Pareto-Optimized XGBoost Model	63