

ACKNOWLEDGEMENT

Praise and thank God Almighty because for His blessings and mercy, thesis report entitled "USING DECISION TREE-BASED DATA MINING TO PREDICT TYPES OF APPARELS" can be completed properly and on time.

This thesis report is structured to meet the requirements for students to be taken in accordance with the curriculum of Mathematics Study Program, Faculty of Science and Technology, Universitas Pelita Harapan, Jakarta. This thesis is also useful for the author to apply the knowledge that has been gained and also gain new experiences that cannot be obtained from lectures.

This thesis report was successfully completed with the support of many parties. Therefore, the authors would like to thank:

1. Eric Jobiliong, Ph.D., as Dean of the Faculty of Science and Technology.
2. Dr. Nuri Arum Anugrahati, as Vice Dean of the Faculty of Science and Technology.
3. Laurence, M.T., as the Director of Administration and Student Affairs, Faculty of Science and Technology.
4. Kie Van Ivanky Saputra, Ph.D., as Head of the Mathematics Study Program and thesis supervisor, also as academic supervisor who continues to provide guidance, direction, and support in the making of this thesis and during the lecture period.
5. Dion Krisnadi, S.Inf., S.Si., M.T.I., M.Act.Sc. as thesis supervisor who has provided input and direction during the finalization of this thesis.
6. Ferry Vincentius Ferdinand, S.Si., S.Inf., M.Pd., M.M., as thesis supervisor and also a lecturer in the Mathematics Study Program who has helped author undergo the world of lectures and provided input in this thesis.
7. All lecturers who have educated and taught various courses to author during the lecture period. Your patience is very much appreciated.
8. Papa, Mama, and families who have given a lot of unconditional support, nudges, and encouragement from the beginning to the end of the author's

lecture and the making of this thesis.

9. Vaniecia Citra Dewi & Pho-Pho Alen, who have supported and a place to confide in about the entire process of lectures and thesis writing.
10. Kevin Aliwarga, who is always at author's side during difficult times, also who provides strict encouragement, support, much needed snacks, and importantly, motivates the author to complete this thesis.
11. Anomali, which is Felicia Sofian and Jessica Sutedja, who have been the author's closest friends during and after the lecture period, and hopefully they will remain in the future.
12. Meidiana Metta Hendryani, author's best friend since kindergarten, who is possibly the reason for author's late graduation but also who patiently encourages author to complete the degree.
13. Mathies 2014 who have made author's day throughout the lectures.
14. All parties that cannot be mentioned one by one who have provided assistance directly or indirectly to the author to be able to complete this thesis.

Finally, the author realizes that this thesis report is still very far from perfect. Therefore, the author is very open to criticism and suggestions from readers that can help make this thesis report even better. Hopefully, this report can be useful for all readers.

Tangerang, 17 June 2021

(Viola Citra Dewi)

LIST OF CONTENTS

	page
TITLE PAGE	
FINAL PROJECT UPLOAD STATEMENT AND APPROVAL	
APPROVAL BY THESIS SUPERVISORS	
APPROVAL BY THESIS EXAMINATION COMMITTEE	
ABSTRACT	v
ACKNOWLEDGEMENT	vi
LIST OF CONTENTS	viii
LIST OF FIGURES	x
LIST OF TABLES	xi
LIST OF LISTINGS	xii
LIST OF APPENDICES	xiii
CHAPTER I INTRODUCTION	1
1.1 Background	1
1.2 Problem Statement	2
1.3 Objectives	3
1.4 Restrictions and Assumptions	3
1.5 Benefits	3
1.5.1 Practical Benefit	4
1.5.2 Theoretical Benefit	4
1.6 Thesis Structure	4
CHAPTER II THEORETICAL FRAMEWORK	
2.1 Artificial Intelligence and Machine Learning	5
2.2 Subdivisions of Machine Learning	6
2.2.1 Supervised Learning	7
2.2.2 Unsupervised Learning	7
2.2.3 Semi-Supervised Learning	8
2.3 Bias and Variance Trade-Off	8
2.3.1 Bias Error	9
2.3.2 Variance Error	9
2.3.3 Irreducible Error	10
2.4 Decision Tree	10
2.5 Growing Decision Tree	11
2.5.1 Rule Induction	11
2.5.2 Training Set	11
2.5.3 Confusion Matrix	13
2.5.4 Decision Tree Algorithms	14
2.5.5 Handling Missing Values	14
2.6 Splitting Criteria	15

2.6.1	Gain Ratio	15
2.6.2	Information Gain	15
2.6.3	Gini Index	15
2.6.4	Overfitting and Underfitting	16
2.7	Pruning Trees	17
2.7.1	Cost-Complexity Pruning	17
2.7.2	Error-Based Pruning	18
2.8	Scalability to Large Datasets	19
2.9	Python	19
2.9.1	Python Libraries	20
2.10	Literature Review	21
CHAPTER III METHODOLOGY		
3.1	Data Collection and Extractions	24
3.2	Data Pre-Processing	26
3.3	Data Processing	28
3.4	Data-Set Partition	29
3.4.1	Secondary Partition	30
3.4.1.1	NumpyArray Iteration	32
3.4.1.2	GridSearchCV	33
3.4.2	Main Partition	33
CHAPTER IV RESULT AND ANALYSIS		
4.1	Trial Decision Tree Model	35
4.2	Parameter Tuning Result	36
4.2.0.1	Manual Tuning using NumpyArray Iteration	37
4.2.0.2	Automatic Tuning using GridSearchCV	39
4.3	Final Model	40
4.4	Performance	42
4.4.1	Confusion Matrix	43
4.4.2	Classification Report	48
4.4.3	Feature Importance	49
CHAPTER V CONCLUSION & SUGGESTIONS		
5.1	Conclusion	51
5.2	Suggestions	51

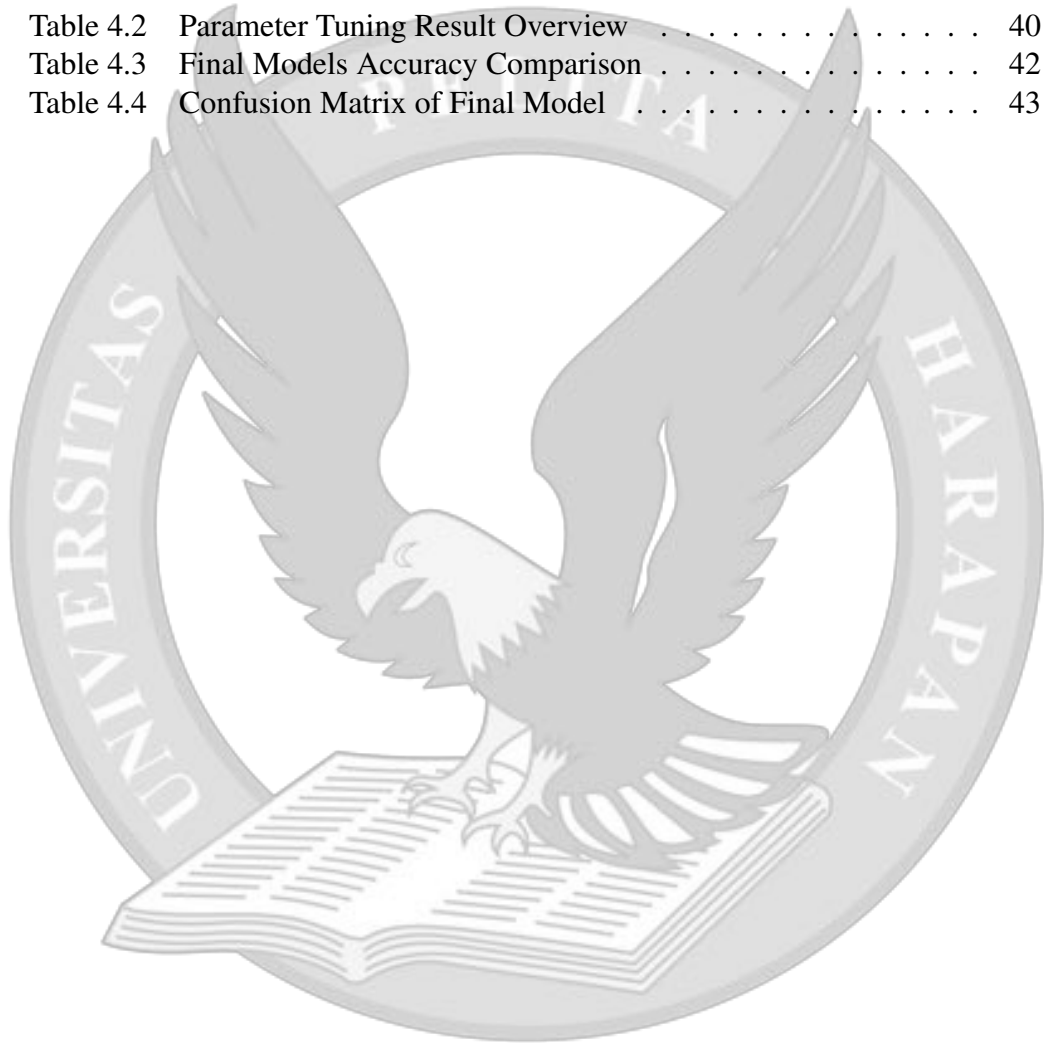
REFERENCES

LIST OF FIGURES

	page
Figure 1.1 Category and Attribute Prediction	1
Figure 2.1 Classical Programming vs Machine Learning	5
Figure 2.2 Overfitting in Decision Trees	16
Figure 2.3 Overfitting in Decision Trees	17
Figure 3.1 System Outline	23
Figure 3.2 DeepFashion Images	24
Figure 3.3 DeepFashion TXT Data	25
Figure 3.4 converting list_attr_img.txt to CSV	27
Figure 3.5 Adding 26 Attribute Names	27
Figure 3.6 Inputting Final Labels	28
Figure 3.7 Importing Data to Python	29
Figure 3.8 Data Information	30
Figure 3.9 Secondary Training Data	31
Figure 3.10 Secondary Testing Data	31
Figure 3.11 Main Training Data	34
Figure 3.12 Main Testing Data	34
Figure 4.1 Trial Decision Tree	35
Figure 4.2 Prediction Result & Accuracy	36
Figure 4.3 Precision, Recall, & f1-Score	36
Figure 4.4 Min Samples Split Tuning	37
Figure 4.5 Min Samples Leaf Tuning	37
Figure 4.6 Max Depth Tuning	38
Figure 4.7 Max Features Tuning	39
Figure 4.8 Automatic Parameter Tuning Result	39
Figure 4.9 Example of A Node	40
Figure 4.10 Manual Parameter Tuning Result	41
Figure 4.11 Automatic Parameter Tuning Result	42
Figure 4.12 Image Indices of Actual Bottom Predicted as Bottom	44
Figure 4.13 Images of Correctly Classified Members	45
Figure 4.14 Images of Actual Bottom Predicted as Top	45
Figure 4.15 Images of Actual Bottom Predicted as Whole	46
Figure 4.16 Images of Actual Top Predicted as Bottom	46
Figure 4.17 Images of Actual Top Predicted as Whole	47
Figure 4.18 Images of Actual Whole Predicted as Bottom	47
Figure 4.19 Images of Actual Whole Predicted as Top	48
Figure 4.20 Precision, Recall & f1-Score	48
Figure 4.21 Feature Importance	49

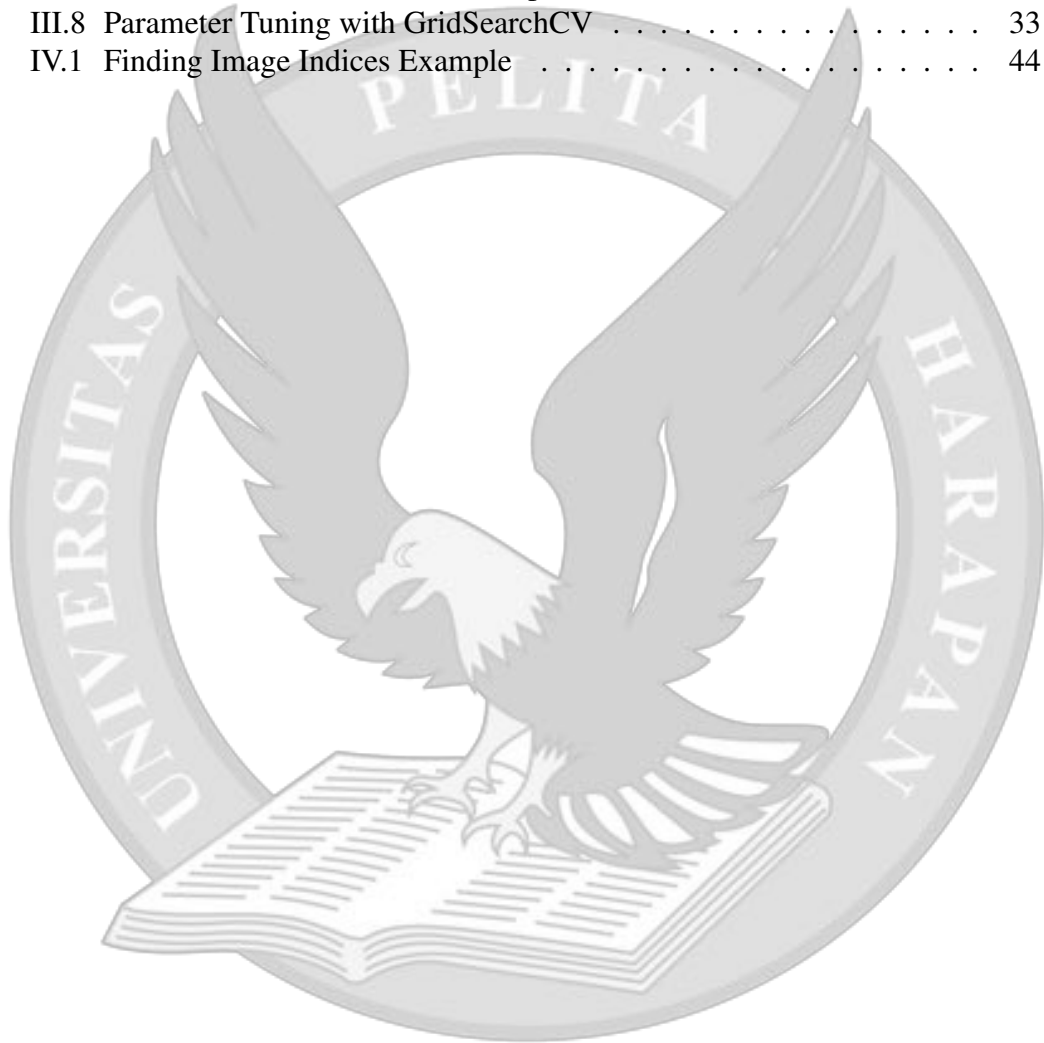
LIST OF TABLES

	page
Table 2.1 Confusion Matrix	13
Table 2.2 Example of 3×3 Confusion Matrix	13
Table 3.1 Clothing Attributes	25
Table 4.1 Confusion Matrix	36
Table 4.2 Parameter Tuning Result Overview	40
Table 4.3 Final Models Accuracy Comparison	42
Table 4.4 Confusion Matrix of Final Model	43



Listings

III.1	Code Example for Importing Data	28
III.2	Splitting Feature and Class Labels in Training Steps	29
III.3	Splitting Data into Training and Testing	30
III.4	Parameter Variation for Max Depth	32
III.5	Parameter Variation for Min Samples Split	32
III.6	Parameter Variation for Max Features	32
III.7	Parameter Variation for Min Samples Leaf	32
III.8	Parameter Tuning with GridSearchCV	33
IV.1	Finding Image Indices Example	44



LIST OF APPENDICES

Appendix A	DATA PROCESSING CODES	
Appendix B	PARAMETER TUNING CODES	
B.1	Parameter Tuning by NumpyArray Iteration	B.1- 1
B.2	Parameter Tuning by GridSearchCV	B.2- 4
Appendix C	CLASSIFICATION AND PERFORMANCE REPORT	
CODES		

