

FOREWORD

Praise be unto God for without His grace and guidance, the writing of this thesis would not have been possible. This thesis, entitled "IMPUTATION USING STATISTICAL AND MACHINE LEARNING METHODS IN LIFE EXPECTANCY DATA" is written as completion to the academic requirement of obtaining Sarjana Matematika in Universitas Pelita Harapan, Tangerang.

Furthermore, the author wishes to give thanks to the many people that have helped and supported the finishing of this thesis. Many of which have given both physical, spiritual and emotional support. With that said, the author wants to express his gratitude to every person who has helped him in the completion of this thesis.

1. Mr. Eric Jobiliong, Ph.D., as Dean of Faculty of Science and Technology.
2. Mrs. Dela Rosa, S.Si., M.M., M.Sc., Apt., as Acting Vice Dean of Faculty of Science and Technology.
3. Mr. Laurence, S.T., M.T., as Administration & Student Affairs Director of Faculty of Science and Technology.
4. Mr. Kie Van Ivanky Saputra, Ph.D., the Head of Mathematics Department of Universitas Pelita Harapan, as well as the thesis advisor, who gave much help and guidance during the writing process and have helped the author develop as a student.
5. Mr. Ferry V. F., S.Si., S.Inf., M.Pd., M.M., as one of the Applied Mathematics study program lecturer that has helped guide the author in constructing this thesis, as a mentor that teaches knowledge and experience beyond what can be taught in a classroom, lastly, as a friend that has become an inspiration in many ways.
6. Mrs. Lina Cahyadi, S.Si., M.Si., as examiner, thesis advisor and lecturer during the years of study.
7. Mr. Ukur Arianto Sembiring, S.Si, M.Si, as the author's academic advisor during years of study.
8. Lecturers and staffs in the Mathematics Department who have shared their knowledge and taught much to the author during the years of study.
9. Katrin Revina, Melissa Susanto, Debby Nugroho, Ribka Maya, Amanda Priscilla and others who have greatly contributed in helping the author learn, understand and pass several classes throughout the years of study.
10. The group Moyung, who have brought great joy through the years of study.

11. Informatics students from Nicky Logan's group and Shella Lolitha's group for aiding the author in understanding the class topics and building projects.
12. Family members, who have encouraged, paid for, prayed for, guided, taught and believed in the author in the pursuit of greater knowledge. Special mention to the author's father and mother for the great love and support, the author's sisters and their husbands who have been kind enough to aid in funding the author's food expenses, the author's aunt who have help in aiding the purchase of the author's laptop, in which the author wouldn't have been able to pass classes and build this thesis.
13. Close friends and classmates, who gave the author support throughout the study and didn't hesitate to offer help whenever needed.
14. Vincent Hartanto Utomo, who have been a great friend, a great teacher and a great inspiration. Special thanks for being the reason that the author has passed several milestones whom without which, the author will not have the opportunity to construct this thesis.
15. Stella Priscilla Wongkor, the author's childhood friend who have known each other since kindergarten and have been in the same class ever since. Special thanks for providing emotional support, educational support, prayers and words of encouragement without which the author may not have survived the years of study.
16. Reinetha Desnilea Candra, the author's partner who have provided joy, great support, encouragement, dedication and aid in constructing this thesis.
17. All other people who have helped the author either directly or indirectly during the completion of this thesis.

It is without doubt that this thesis is still far from perfection and not without limitations. With that in mind, the author welcomes critics and advice that further improves it. Hopefully, this thesis is useful for every one who reads it.

Tangerang, June 27th, 2019

(Sergius Tadao Hayashi)

TABLE OF CONTENTS

	page
TITLE PAGE	
STATEMENT OF THESIS AUTHENTICITY	
APPROVAL BY THESIS SUPERVISORS	
APPROVAL BY THESIS EXAMINATION COMMITTEE	
ABSTRACT	v
ABSTRACT	vi
FOREWORD	vii
TABLE OF CONTENTS	ix
LIST OF FIGURES	xii
LIST OF TABLES	xiv
LIST OF APPENDICES	xv
CHAPTER I INTRODUCTION	1
1.1 Background	1
1.2 Problem Statement	5
1.3 Objectives	5
1.4 Restrictions and Assumptions	5
1.5 Benefits	6
1.6 Thesis Structure	6
CHAPTER II STATISTICAL IMPUTATION METHODS	7
2.1 Missingness	7
2.1.1 Missing Completely at Random (MCAR)	9
2.1.2 Missing at Random (MAR)	10
2.1.3 Missing Not at Random (MNAR)	12
2.2 Mean Imputation	13
2.3 Hot-Deck Imputation	14
2.4 Multiple Imputation	15
2.4.1 Bootstrapping	16
2.4.2 Predictive Mean Matching	17
2.5 Literature Review	17
CHAPTER III MACHINE LEARNING IMPUTATION METHODS	19
3.1 Multi Layer Perceptron	19
3.2 Self-Organizing Maps	24
3.3 K-Nearest Neighbors	26
3.4 Model Evaluation Metrics	28
3.4.1 Mean Squared Error (MSE)	29
3.4.2 Mean Absolute Error (MAE)	29
3.4.3 Mean Absolute Percentage Error (MAPE)	30
3.4.4 T-test	30
3.5 Literature Review	33
CHAPTER IV METHODOLOGY	35
4.1 Data	36

4.1.1	The World Health Organization's Life Expectancy Dataset	36
4.1.2	The World Bank's Life Expectancy Dataset	37
4.2	Data Assumption and Characteristic	37
4.3	Implementation of Statistical Methods	38
4.3.1	Implementation of Mean Imputation	38
4.3.2	Implementation of Hot-Deck Imputation	38
4.3.3	Implementation of Multiple Imputation	39
4.4	Implementation of Machine Learning Methods	40
4.4.1	Implementation of MLP Imputation	40
4.4.2	Implementation of SOM Imputation	41
4.4.3	Implementation of KNN Imputation	41
4.5	Method Evaluation	41
CHAPTER V	EXPERIMENT AND RESULTS	43
5.1	The Original Dataset	43
5.1.1	Features and Characteristic of The Original Dataset	43
5.1.2	Missingness of The Original Dataset	48
5.2	The Artificial Dataset	52
5.2.1	Creating The Artificial Dataset	53
5.2.2	Missingness of The Artificial Dataset	54
5.3	Imputation Results	56
5.3.1	Imputed Artificial Dataset	57
5.3.2	Imputed Original Dataset	76
CHAPTER VI	CLOSING	84
6.1	Conclusion	84
6.2	Future Works	85
REFERENCES	90
APPENDIX A	A-1
APPENDIX B	B-1

LIST OF FIGURES

	page
Figure 2.1	Patterns of nonresponse in rectangular data sets: (a) univariate pattern, (b) monotone pattern, and (c) arbitrary pattern. In each case, rows correspond to observational units and columns correspond to variables [16]. 7
Figure 2.2	An example of a dataset with n cases and m features. 8
Figure 3.1	A multilayer perceptron with two hidden layers [23]. 19
Figure 3.2	The logistic sigmoid function [23]. 20
Figure 3.3	Neighborhoods (N_c) for a rectangular matrix of cluster units. Black brackets represent the BMU, or $N_c = 0$. The red brackets represent $N_c = 1$, blue brackets represent $N_c = 2$, and the pattern goes until the edges [27]. 25
Figure 3.4	Location of the rejection region of the alternative hypothesis [33]. 31
Figure 3.5	Example of decision making using p-value [33]. 32
Figure 4.1	A flowchart diagram depicting the process of the thesis. 35
Figure 5.1	The unique missingness patterns of the original dataset along with frequency of occurrence. 48
Figure 5.2	A nullity bar chart of the original dataset visualizing the missingness of each variable and frequency of the observed values. 49
Figure 5.3	A nullity matrix that displays the density of the original dataset's missingness. 50
Figure 5.4	A correlation heatmap of the original dataset with nullity correlation of each variable. 52
Figure 5.5	A dendrogram of the original dataset showing the nullity correlation of each variable. 52
Figure 5.6	A nullity bar chart of the artificial dataset visualizing the missingness of each variable. 55
Figure 5.7	A nullity matrix that displays the density of the artificial dataset's missingness. 55
Figure 5.8	A correlation heatmap of the artificial dataset with nullity correlation of each variable. 56
Figure 5.9	A dendrogram of the artificial dataset showing the nullity correlation of each variable. 56

Figure 5.10 A graph MLP's loss curve for every MLP model associated with each missingness pattern in the artificial dataset. The X-axis depicts the iteration count and the Y-axis depicts the cost. 73

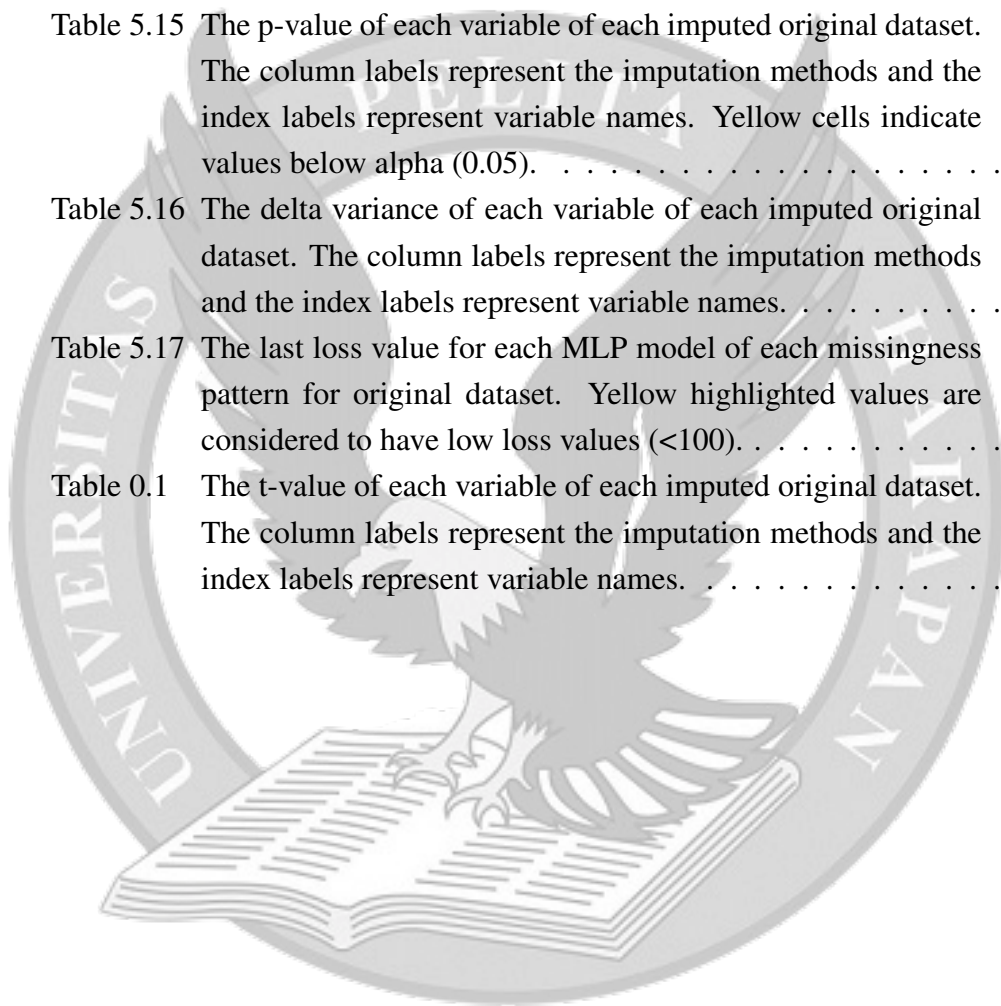
Figure 5.11 A graph MLP's loss curve for every MLP model associated with each missingness pattern in the original dataset. The X-axis depicts the iteration count and the Y-axis depicts the cost. 81



LIST OF TABLES

	page
Table 5.1 Description of the labels and values within the WHO's Life Expectancy dataset.	46
Table 5.2 Characteristic of the concatenated life expectancy dataset containing records from 194 countries. Range, mean, median, mode, and missingness percentage are shown for all variable labels.	47
Table 5.3 The Mean Squared Error of each variable of each imputed artificial dataset. The column labels represent the imputation methods and the index labels represent variable names.	60
Table 5.4 The Mean Absolute Error of each variable of each imputed artificial dataset. The column labels represent the imputation methods and the index labels represent variable names.	61
Table 5.5 The Mean Absolute Percentage Error of each variable of each imputed artificial dataset. The column labels represent the imputation methods and the index labels represent variable names.	62
Table 5.6 Example A with statistical method estimates. Black cells indicates missing values. Blue cells indicate actual value. Yellow cells indicate estimates.	64
Table 5.7 Example A with machine learning method estimates. Black cells indicates missing values. Blue cells indicate actual value. Yellow cells indicate estimates.	65
Table 5.8 Example B with statistical method estimates. Black cells indicates missing values. Blue cells indicate actual value. Yellow cells indicate estimates.	66
Table 5.9 Example B with machine learning method estimates. Black cells indicates missing values. Blue cells indicate actual value. Yellow cells indicate estimates.	67
Table 5.10 Example C with statistical method estimates. Black cells indicates missing values. Blue cells indicate actual value. Yellow cells indicate estimates.	68
Table 5.11 Example C with machine learning method estimates. Black cells indicates missing values. Blue cells indicate actual value. Yellow cells indicate estimates.	69

Table 5.12	Brief summary of the observed result.	71
Table 5.13	The last loss value for each MLP model of each missingness pattern for the artificial dataset. Yellow highlighted values are considered to have low loss values (<100).	74
Table 5.14	The delta mean of each variable of each imputed original dataset. The column labels represent the imputation methods and the index labels represent variable names.	78
Table 5.15	The p-value of each variable of each imputed original dataset. The column labels represent the imputation methods and the index labels represent variable names. Yellow cells indicate values below alpha (0.05).	79
Table 5.16	The delta variance of each variable of each imputed original dataset. The column labels represent the imputation methods and the index labels represent variable names.	80
Table 5.17	The last loss value for each MLP model of each missingness pattern for original dataset. Yellow highlighted values are considered to have low loss values (<100).	82
Table 0.1	The t-value of each variable of each imputed original dataset. The column labels represent the imputation methods and the index labels represent variable names.	B-1



LIST OF APPENDICES

Appendix A	Codes	A-1
Appendix B	T-Value Table	B-1

