



Cairo University

# **MLRMUD: A MULTI LINEAR REGRESSION APPROACH FOR MISSING VALUES PREDICTION WITH UNKNOWN DEPENDENT VARIABLE**

By

**Ahmed Karama Mahboab Alhebshi**

A Thesis Submitted to the  
Faculty of Engineering at Cairo University  
in Partial Fulfillment of the  
Requirements for the Degree of

**MASTER OF SCIENCE  
in  
Computer Engineering**

**FACULTY OF ENGINEERING, CAIRO UNIVERSITY  
GIZA, EGYPT  
2019**

**MLRMUD: A MULTI LINEAR REGRESSION APPROACH  
FOR MISSING VALUES PREDICTION WITH  
UNKNOWN DEPENDENT VARIABLE**

By  
**Ahmed Karama Mahboab Alhebshi**

A Thesis Submitted to the  
Faculty of Engineering at Cairo University  
in Partial Fulfillment of the  
Requirements for the Degree of  
**MASTER OF SCIENCE**  
in  
**Computer Engineering**

Under the Supervision of

**Prof. Dr. Samir I. Shaheen**  
Professor  
Computer Engineering Department,  
Faculty of Engineering, Cairo  
University

**Prof. Dr. Amir F. Atiya**  
Professor  
Computer Engineering Department,  
Faculty of Engineering, Cairo  
University

**Dr. Mona F. Ahmed**  
Assistant Professor  
Computer Engineering Department,  
Faculty of Engineering, Cairo  
University

**FACULTY OF ENGINEERING, CAIRO UNIVERSITY  
GIZA, EGYPT  
2019**

**MLRMUD: A MULTI LINEAR REGRESSION APPROACH  
FOR MISSING VALUES PREDICTION WITH  
UNKNOWN DEPENDENT VARIABLE**

By  
**Ahmed Karama Mahboab Alhebshi**

A Thesis Submitted to the  
Faculty of Engineering at Cairo University  
in Partial Fulfillment of the  
Requirements for the Degree of  
**MASTER OF SCIENCE**  
in  
**Computer Engineering**

Approved by the  
Examining Committee

---

**Prof. Dr. Samir I. Shaheen**

Thesis Main Advisor

---

**Prof. Dr. Ihab Elsayed Talkhan**

Internal Examiner

---

**Prof. Dr. Reda Abd-Alwahab Ahmed**

External Examiner

**FACULTY OF ENGINEERING, CAIRO UNIVERSITY  
GIZA, EGYPT  
2019**



**Engineer's Name:** Ahmed Karama Mahboab Alhebshi  
**Date of Birth:** 20/5/1985  
**Nationality:** Yemeni  
**E-mail:** [engcmp505@gmail.com](mailto:engcmp505@gmail.com)  
**Phone:** 01091655230  
**Address:** Giza-Main Mahattah street  
**Registration Date:** 1/10/2014  
**Awarding Date:** ....../....../.....  
**Degree:** Master of Science  
**Department:** Computer Engineering



**Supervisors:**

Prof. Dr. Samir I. Shaheen  
Prof. Dr. Amir F. Atiya  
Dr. Mona F. Ahmed

**Examiners:**

Prof. Dr. Samir I. Shaheen (Thesis Main Advisor)  
Prof. Dr. Ihab Elsayed Talkhan (Internal Examiner)  
Prof. Dr. Reda Abd-Alwahab (External Examiner)  
Professor in faculty of computers and information  
Cairo university

**Title of Thesis:**

MLRMUD: A Multi Linear Regression Approach for Missing Values Prediction with  
Unknown Dependent Variable

**Key Words:**

missing values; splitting algorithm; dependent variable; multi linear regression; regression coefficients

**Summary:**

The Missing Value problem (MV) is the problem of predicting the missing value in the data set while achieving accurate values. An Additional attribute has been imposed on the missing value problem which is an unknown dependent variable.

In this work, a new approach, MLRMUD, based on Multiple Linear Regression is used to predict Missing values for a data set with an Unknown Dependent variable if complete rows are at least 20%. If they are less than that the Mean method is used to fill some rows until the complete rows reach 20%, after that MLRMUD can be applied normally. This approach is composed of three algorithms; splitting algorithm, dependent variable selection algorithm and multi linear regression algorithm.

MLRMUD is compared to other counterparts in the literature where it was proved that it outperforms them all in the accuracy of missing values computation determined in terms of the Root Mean Square Error (RMSE) and Mean Standard Error (MSE). A method to determine the unknown dependent variable from the training set is proposed. The results show that the proposed method can successfully select the dependent variable with an accuracy of 83% overall the data sets examined

## **Disclaimer**

I hereby declare that this thesis is my own original work and that no part of it has been submitted for a degree qualification at any other university or institute.

I further declare that I have appropriately acknowledged all sources used and have cited them in the references section.

Name: Ahmed Karama Mahboab Alhebshi

Date:10/7/2019

Signature:

## **Dedication**

I'd like to dedicate this thesis to my wife and my family for supporting me during my work

## Acknowledgments

I would have never gone through this work without the honest help and support of many people in my life to whom I dedicate this section.

First of all, I would like to express my gratitude to **Allah**, who is the reason of my strength and who always raises us over limits beyond our expectations, and who can do everything immeasurably far beyond our thoughts.

I would like to thank my supervisor **Dr. Mona Farouk** for suggesting the research path and for her consistent guidance and advice. He has assisted me in different aspects of my work and has been very keen on my progress in this thesis.

I would like to thank my supervisor **Prof. Samir I. Shaheen and Prof. Amir F. Atiya** for their continuous efforts and concern. They have provided me with important scientific information and aided me throughout my thesis.

I would like to thank my **beloved family** for always finding them by my side, and for their sincere encouragement throughout my work and throughout my entire life.



# Table of Contents

<b>LIST OF TABLES .....</b>	<b>V</b>
<b>LIST OF FIGURES .....</b>	<b>VI</b>
<b>NOMENCLATURE.....</b>	<b>VII</b>
<b>ABSTRACT.....</b>	<b>VIII</b>
<b>CHAPTER 1: INTRODUCTION</b>	<b>1</b>
1.1.    OVERVIEW OF MISSING VALUE PROBLEM (MVP).....	1
1.2.    MOTIVATION.....	1
1.2.1.    Missing Value Approaches .....	1
1.2.2.    Real Applications of Unknown Dependent Variable.....	2
1.2.2.1.    NodesTraffic Problem.....	2
1.2.2.2.    Weather Prediction.....	2
1.3.    PROBLEM STATEMENT.....	2
1.4.    THESIS OBJECTIVE .....	4
1.5.    ORGANIZATION OF THE THESIS.....	5
<b>CHAPTER 2: BACKGROUND.....</b>	<b>7</b>
2.1.    BIG DATA DEFINITION .....	7
2.2.    BIG DATA CONSIDERATIONS.....	7
2.3.    BIG DATA ANALYTICS TECHNOLOGIES AND TOOLS.....	8
2.4.    HOW BIG DATA ANALYTICS WORKS.....	8
2.5.    AWS MANAGEMENT CONSOLE.....	9
<b>CHAPTER 3 : RELATED WORK.....</b>	<b>12</b>
3.1.    INTRODUCTION.....	12
3.2.    STANDARD ALGORITHMS .....	12
3.2.1.    k-Nearest Neighbor Imputation(KNN) .....	12
3.2.2.    Listwise or Case Deletion.....	16
3.2.3.    Mean Substitution.....	16
3.2.4.    Expectation-Maximization .....	16
3.2.5.    Multiple Imputation .....	16
3.3.    NEW APPROACHES TO ESTIMATE MISSING VALUES.....	17
3.3.1.    Simple Linear Regression.....	17
3.3.2.    Statistical Models of Multiple Linear Regressions .....	18
3.3.3.    Incomplete Data Recovery Using Linear Regression .....	20
3.3.4.    Linear Regression using A Matrix .....	22
3.3.5.    Decision Tree Induction.....	22
3.3.6.    Selecting Scalable Algorithms: C4.5 Algorithm and K-Means .....	24
3.3.7.    Incomplete Data Hierarchical Clustering.....	25
3.3.8.    Identify the Missing Data Algorithms (IMDA) .....	25
3.3.9.    Fuzzy Possibilistic C Means Based on Support Vector Regression and Genetic Algorithm... ..	25
3.3.10.    Multiple Imputation Approaches .....	28
3.3.11.    Imputed Data Using The Classifier.....	30

3.3.12.	Jaccard Dissimilarity Coefficients for The Missing Value in The Text .....	33
3.3.13.	Fuzzy C means Clustering Algorithm.....	35
3.3.13.1.	FCM Impute.....	36.
3.3.14.	Bayesian Genetic Algorithm.....	38
3.3.14.1.	Genetic Algorithm .....	38
3.3.14.2	Bayesian Theorem.....	39
<b>CHAPTER 4 : PROPOSED APPROACH.....</b>		<b>42</b>
4.1.	OVERVOEW OF MULTI LINEAR REGRESSION .....	42
4.2.	CORRELATION MATRIX.....	42
4.3.	OUR PROPOSED ALGORITHM.....	43
4.3.1.	Splitting Algorithm.....	45
4.3.2.	Dependent Variable Selection Algorithm.....	47
4.3.2.1.	Steps of Dependent Variable Algorithm.....	47.
4.3.3.	Regression Model Estimation.....	49
4.3.4.	Complete Proposed Algorithm .....	51
<b>CHAPTER 5 : EXPERIMENTAL RESULTS.....</b>		<b>52</b>
5.1.	EXPERIMENTAL SETUP.....	52
5.2.	MISSING VALUE ESTIMATION.....	53
5.2.1.	The Experiment 1 .....	53
5.2.1.1.	5% missing values.....	53
5.2.1.2.	10% missing values .....	53
5.2.1.3.	15% missing values .....	54
5.2.1.4.	20% missing values.....	55
5.2.2.	The Experiment 2 .....	55
5.2.2.1.	20% Missing values.....	56
5.3.	DEPENDENT VARIABLE EXPERIMENTS .....	56
5.3.1.	Dependent variable selection algorithm for Brainsize data set.....	57
5.3.2.	Dependent variable selection algorithm for data set 1.....	57.
5.3.3.	Dependent variable selection algorithm for data set 2.....	58
5.3.4.	Dependent variable selection algorithm for data set 3.....	59
5.4.	REGRESSION COEFFICIENTS ESTIMATION.....	60
5.5.	COMPARISON OF THE PROPOSED WORK AND FUZZY POSSIBILISTIC C MEANS OPTIMIZED WITH SUPPORT VECTOR REGRESSION AND GENETIC ALGORITHM.....	65
5.6.	COMPARISON OF ACTUAL DEPENDENT VARIABLE AND THE PROPOSED WORK DEPENDENT VARIABLE.....	63
5.7.	COMPARISON OF THE PROPOSED WORK AND SIX DIFFERENT APPROACHES	64
5.8.	COMPARISON OF THE PROPOSED WORK AND BGA) .....	64
<b>CHAPTER 6 : CONCLUSION AND FUTURE WORK .....</b>		<b>66</b>
6.1.	CONCLUSION .....	66
6.2.	FUTURE WORK.....	67
<b>REFERENCES.....</b>		<b>68</b>
<b>APPENDIX A: DEFINITIONS .....</b>		<b>71</b>

## List of Tables

Table I.1: Nomenclature .....	VII
Table 3.1: Employee data set with missing values .....	18
Table 3.2: Data set before and after applying an unsupervised filter .....	20
Table 3.3: Accuracy of imputed missing value on day 8.....	21
Table 3.4: Accuracy of imputed missing value on day 8.....	22
Table 3.5: Iris data set properties.....	27
Table 3.6: Comparison between six different approaches with respect to mean standard error.....	30
Table 3.7: Different missing value methods with k-means.....	31
Table 3.8: Comparing the results with k mix clustering.....	31
Table 3.9: Different Missing Imputation Methods with J48 Classification.....	31
Table 3.10: Different Missing Imputation Methods with K-NN Classification .....	32
Table 3.11: Different Missing Imputation Methods with Fuzzy Rule Induction Algorithm ...	32
Table 3.12: highest sensitivity value found with each of the imputation method...	32
Table 3.13: Text files and Corresponding Frequent Item Sets Algorithm.....	34
Table 3.14: Similarity Matrix Obtained By Jaccard Similarity Measure over Frequent Item Sets Obtained From Each Text File .....	34
Table 3.15: Similarity matrix obtained after step 2.....	35
Table 3.16: The information of four data sets in UCI.....	35
Table 4.1: The main data set for apply the proposed splitting algorithm.....	46
Table 4.2: Training data set after applied splitting algorithm .....	46
Table 4.3: Test data set after applied splitting algorithm .....	46
Table 4.4: Show different splitting algorithm .....	46
Table 4.5: Training data set after applied another splitting algorithm .....	46
Table 4.6: Test data set after applied another splitting algorithm .....	46
Table 4.7: The sample data set for explaining dependent variable selection algorithm	48
Table 4.8: Show the variable and the variables which related to it after applied dependent variable selection algorithm .....	48
Table 4.9: The sample data set for more than one dependent variable .....	48
Table 4.10: Show the variable and the variables which related to it after applied dependent variable selection algorithm for more than one dependent variable...	49
Table 5.1: The proposed Work computes predicted values and RMSE to measure the accuracy of the missing value ratio 5%.....	53
Table 5.2: The proposed work computes predicted values and RMSE to measure the accuracy of the missing value ratio 10% .....	54
Table 5.3: The proposed work computes predicted values and RMSE to measure the accuracy of the missing value ratio 15% .....	54
Table 5.4: The proposed work computes predicted values and RMSE to measure the accuracy of missing value ratio 20% .....	55
Table 5.5: The detailed results to compute the mean standard error of the approach.	56
Table 5.6: Sample data to test the proposed work in selecting the dependent variable.	56

Table 5.7: Results of comparing the actual dependent variable and the proposed dependent variable (brain data set). .....	57
Table 5.8: Results of comparing the actual dependent variable and the proposed dependent variable (data set 1) .....	58
Table 5.9: Results of comparing the actual dependent variable and the proposed dependent variable (data set 2)....	58
Table 5.10:Results of comparing actual dependent variable and the proposed dependent variable (data set 3).....	59
Table 5.11: The regression coefficients of our approach for a number of different missing value cases (data set 1) .....	60
Table 5.12: The regression coefficients of our approach for a number of different missing value cases (data set 2)....	61
Table 5.13: The regression coefficients of our approach for a number of different missing value cases (data set 3).....	62
Table 5.14: The results of comparing our approach to Fuzzy Possibilistic C Means (FPCM-SVRGA) respect to RMSE.....	63
Table 5.15: Comparing our approach to another six approach for missing value ratio 20% of iris data set respect to mean standard error.....	64
Table 5.16: Comparing our approach to BGA respect to mean standard error .....	65

## List of Figures

Figure 1.1: Missing value problem with unknown dependent variable .....	3
Figure 1.2: Sample data contains missing values and unknown dependent variable.....	4
Figure 3.1: K-Nearest Neighbor Imputation (KNN) (Liqiang Pan, Jianzhong Li,2010)12	
Figure 3.2: RMSE vs. Sampling Interval (Liqiang Pan, Jianzhong Li,2010) ..	13
Figure 3.3: RMSE vs. Number of Neighbor Nodes(Liqiang Pan, Jianzhong Li,2010). 13	
Figure 3.4: RMSE vs. Number of Missing Data (Liqiang Pan, Jianzhong Li, 2010)....	14
Figure 3.5: RMSE vs. Sampling Interval (Liqiang Pan, Jianzhong Li,2010).....	14
Figure 3.6: RMSE vs. A Number of Neighbor Nodes (Liqiang Pan, Jianzhong Li).....	15
Figure 3.7: RMSE vs. Number of Missing Values .....	15
Figure 3.8: Statistical Models of Multiple Linear Regressions (Mr.M.B.Shelke,2013)	18
Figure 3.9: Graphical Representation of Employee Data (Mr.M.B.Shelke, 2013).....	19
Figure 3.10: The Values of Salary Attribute After Prediction All The Missing Values (Mr.M.B.Shelke, Mr.K.B.Badade, 2013).....	19
Figure 3.11: The Approach Diagram (Hailin , 2014) .....	21
Figure 3.12: Listwise Deletion, Top, Removal of Corrupted Spectra(Yasser Beyad, Marcel Maeder, 2013).....	22
Figure 3.13: Decision Tree Induction (Raju Dara and Dr.Ch.Satyanarayana, 2015) ..	23
Figure 3.14: Diagram of Decision Tree For Age(Raju Dara..... 2015).....	24
Figure 3.15: Diagram of Decision Tree for Empid (Raju Dara, 2015).....	24
Figure 3.16: Membership Function Estimation(p.saravanan, p.sailakshmi, 2015)	26
Figure 3.17: Fuzzy Possibilistic C Mean Based Support Vector Regression And Genetic Algorithm.....	26
Figure 3.18: FPCM Algorithm (p.saravanan, p.sailakshmi, 2015) .....	27
Figure 3.19: Comparison between FCM-SVRGA and FPCM-SVRGA.....	28
Figure 3.20: Distribution Missing Values in Four Attributes; Petal Width, Petal Length, Sepal Length and Sepal Width(Geeta Chhabra , 2017).....	29
Figure 3.21: Algorithm of Imputed the Missing Values Using Classifier(Raju Dara1 and Dr.Ch.Satyanarayana2, 2015) .....	30
Figure 3.22: Jaccard Dissimilarity Coefficients Algorithm for Missing Value In Text (Dr. Ch. Satyanarayana, Raju Dara, Dr. A. Govardhan Professor, 2017).....	33
Figure 3.23: Three Dimensional Shaded Surfaces Obtained by Using Jaccard Similarity Measure over Frequent Item Sets Obtained from Each Text File .....	34
Figure 3.24: Three Dimensional Shaded Surfaces Obtained After Applying Step 2 ...	35
Figure 3.25: Comparison of The Accuracy of Fcmimpute, Knnimpute, and Sknnimpute Methods for Data Set 1 over 1 and 20% Data Missing(JiaWei Luo, TaoYang .....	37
Figure 3.26: Comparison of The Accuracy of Fcmimpute, Knnimpute, and Sknnimpute Methods for Data Set 2 over 1 and 20% Data Missing.....	37
Figure 3.27: Comparison of The Accuracy of Fcmimpute, Knnimpute, and Sknnimpute Methods for Data Set 3 over 1 and 20% Data Missing .....	38

Figure 3.28: Comparison of The Accuracy of Fcmimpute, Knnimpute, and Sknnimpute Methods for Data Set 4 over 1 and 20% Data Missing.....	39
Figure 3.29: Structure of Chromosome (R. Devi Priya and S. Kuppuswami , 2015). 39	
Figure 3.30: Structure of Bayesian Genetic Algorithm .....	34
Figure 3.31: RMSE for Abalone Data set by Implementing Mean, KNN Imputation, MI and BGA at Different Missing Rates.....	40
Figure 3.32: RMSE for Wine Data set by Implementing Mean, KNN Imputation, MI, And BGA At Different Missing Rates .....	40
Figure 3.33: RMSE for Automobile Data set by Implementing Mean, KNN Imputation, MI And BGA at Different Missing Rates .....	41
Figure 3.34: RMSE for housing data set by implementing mean, kNN imputation, MI and BGA at different missing rates .....	41
Figure 4.1: MLRMUD Flowchart .....	44
Figure 4.2: Splitting Algorithm .....	45
Figure 4.3: Dependent Variable Selection Algorithm .....	47
Figure 4.4: Complete Proposed Algorithm .....	51

# Nomenclature

**Table I.1: List of Acronyms**

<b>Acronym</b>	<b>Definition</b>
AKE	Applying K-nearest neighbor Estimation
BGA	Bayesian Genetic Algorithm
Bi	Regression Coefficients
CART	Multiple Classification And Regression Tree
CD	Complete Data
CDKDV	Complete Data with Known Dependent Variable
CDUDV	Complete Data with Unknown Dependent Variable
FPCM-SVRGA	Fuzzy Possibilistic C Means optimized with Support Vector Regression and Genetic Algorithm
MAR	Missing at Random
MCAR	Missing Completely at Random
MD	Missing Data
MLR	Multi Linear Regression
MLRMUD	Multi Linear Regression for Missing Value and Unknown Dependent Variable
MSE	Mean Standard Error
MV	Missing Value
MVPUDV	Missing Value Problem with Unknown Dependent Value
RMSE	Root Mean Square Error
TS	Training Set
TSz	Test Set
UDV	Unknown Dependent Variable
$X_i$	Independent Variables
Y	Dependent Variable