

**A MACHINE LEARNING MODEL FOR PREDICTING PRE-  
ECLAMPSIA FOR RESOURCE-CONSTRAINED REGIONS.**

**ABDUSHAKUR JAMWA ARINA**

**A THESIS SUBMITTED TO THE INSTITUTE OF COMPUTING  
AND INFORMATICS IN PARTIAL FULFILLMENT OF THE  
REQUIREMENTS FOR THE AWARD OF DEGREE OF MASTER  
OF SCIENCE IN INFORMATION TECHNOLOGY OF  
TECHNICAL UNIVERSITY OF MOMBASA**

**2023**

## DECLARATION

This thesis is my original work and has not been presented for academic award in any other university.

-----

**ABDUSHAKUR JAMWA ARINA**

**MSIT/0005/2020**

-----

**Date**

This thesis has been submitted with our approval as university supervisors.

-----

**DR. MGALA MVURYA**



-----

**Date**

-----

**DR. ANTONY LUVANDA**



-----

**Date**

-----

**DR. PAMELA T. KIMETO**

-----

**Date**

## **DEDICATION**

To my beloved wife Salwa Lali Shee and my daughter Fatma Abdushakur Jamwa.

## ACKNOWLEDGEMENT

First, I would like to express my gratitude to the divine power of God, who is the ultimate creator of everything visible and invisible, the supreme ruler of life's precious gift, and the driving force behind my passion for seeking knowledge.

Secondly, I would like to convey my sincere gratitude to my supervisors, namely Dr. Mgala Mvurya, Dr. Antony Luvanda, and Dr. Pamela Kimeto, for their guidance and support throughout all the phases of this thesis. Their assistance has been crucial and significant in facilitating the completion of this work.

Thirdly, my sincere gratitude goes to my classmates and friends Titus, George, Joseph, Ester Asenath, Athman, Amran, Michael, and Peter for always motivating my thought process by always challenging me to be my best.

Lastly, I wish to acknowledge all those who gave their input either through prayers, words of wisdom, and any kind of contribution. I cannot thank you enough for the sacrifices you afforded me throughout this period. Only God can repay you.

## TABLE OF CONTENTS

<b>DECLARATION</b> .....	<b>ii</b>
<b>DEDICATION</b> .....	<b>iii</b>
<b>ACKNOWLEDGEMENT</b> .....	<b>iv</b>
<b>TABLE OF CONTENTS</b> .....	<b>v</b>
<b>LIST OF TABLES</b> .....	<b>viii</b>
<b>LIST OF FIGURES</b> .....	<b>ix</b>
<b>ACRONYMS AND ABBREVIATIONS</b> .....	<b>x</b>
<b>ABSTRACT</b> .....	<b>xi</b>
<b>CHAPTER ONE</b> .....	<b>1</b>
<b>INTRODUCTION</b> .....	<b>1</b>
1.1 Background of the Study.....	4
1.2 Statement of the Problem.....	7
1.3 General Objective .....	8
1.4 Specific Objectives .....	8
1.5 Research Questions .....	8
1.6 Significance of the Study.....	8
1.7 Limitations of the Study.....	9
1.8 Scope of the Study .....	9
1.9 Organization of the Study.....	9
<b>CHAPTER TWO</b> .....	<b>11</b>
<b>REVIEW OF LITERATURE</b> .....	<b>11</b>
2.1 Introduction.....	11
2.2 Pre-eclampsia and its Impact.....	11
2.3 Related Work on Pre-eclampsia in the Developed World.....	12
2.4 Related Work on Pre-eclampsia in the Developing World .....	13
2.5 Prediction of Pre-eclampsia Using Machine Learning .....	14
2.6 Lessons Learnt From Literature Review .....	21
2.7 Research Gap Identification.....	22
2.8 Conceptual Framework.....	23
2.9 Chapter Summary .....	24

<b>CHAPTER THREE.....</b>	<b>25</b>
<b>RESEARCH METHODOLOGY .....</b>	<b>25</b>
3.1 Introduction.....	25
3.2 Research Approach and Design.....	25
3.3 Research Location.....	26
3.4 Study Population.....	26
3.4.1 Sampling.....	26
3.5 Data Mining Process.....	27
3.5.1 Domain Understanding Pre-Eclampsia .....	29
3.5.2 Data Understanding.....	29
3.5.3 Data Collection.....	30
3.5.5 Data Pre-processing .....	30
3.6 Modeling Techniques.....	34
3.8 Evaluation.....	41
3.8.1 Evaluation matrix.....	42
3.9 Deployment of the Model .....	43
3.10 Ethical Considerations.....	44
3.11 Chapter Summary .....	44
<b>CHAPTER FOUR.....</b>	<b>46</b>
<b>RESULTS AND DISCUSSION .....</b>	<b>46</b>
4.1 Introduction.....	46
4.2 Identify Optimal Features for Building Pre-Eclampsia Model .....	46
4.3 Determining the Best Model for Predicting Pre-Eclampsia Using the Optimal Features.....	46
4.3.1 Logistic Regression(LR) .....	47
4.3.2 Random Forest(RF) .....	49
4.3.3 Support Vector Machine (SVM) .....	51
4.3.4 Naïve Bayes (NB) .....	53
4.3.5 Linear Discriminant Analysis (LDA).....	55
4.4 Discussion of Model Performance Findings .....	57
4.5 Validating the Best Model with the Test Data .....	60

4.6 Chapter Summary .....	63
<b>CHAPTER FIVE.....</b>	<b>65</b>
<b>CONCLUSIONS AND RECOMMENDATIONS.....</b>	<b>65</b>
5.1 Introduction .....	65
5.2 Synthesis of the Research Process.....	65
5.2.1 What Are The Optimal Features for Predicting Pre-Eclampsia Among Pregnant Women? .....	65
5.2.2 How will the Best Machine Learning Model Be Built that will Predict Pre-Eclampsia Accurately among Pregnant Mothers? .....	66
5.2.3 What is the Performance of the Proposed Machine Learning Model in Predicting Pre-Eclampsia in Comparison to Existing Models? .....	67
5.3 The Adopted Process.....	68
5.4 Contribution to Knowledge .....	69
5.5 Limitations of the Study .....	69
5.6 Recommendations for Future Research .....	70
<b>REFERENCE.....</b>	<b>72</b>
<b>APPENDICES .....</b>	<b>94</b>
Appendix I: SGS Authorization.....	94
Appendix II: Nacosti License.....	95
Appendix III: County Authorization .....	96
Appendix IV: Hospital Authorization .....	97

## LIST OF TABLES

Table 3.1: Data Description from the Maternity Bio-Data Card.....	31
Table 4.1: Logistic Regression Performance.....	49
Table 4.2 : Random Forest Performance.....	51
Table 4.3: Support Vector Machine Performance.....	53
Table 4.4: Naïve Bayes Performance.....	55
Table 4.5: Linear Discriminant Analysis Performance.....	57
Table 4.6: Summarized Model for Pre-Eclampsia Prediction Models.....	59



## LIST OF FIGURES

Figure 3.1: The Cross Industry Standard Process for Data Mining (Kristoffersen et al., 2019). .....	27
Figure 3.2: Modeling Process (Hou et al., 2019) .....	35
Figure 3.3: Random Forest classifier (Jackins et al., 2021).....	38
Figure 3.4: Confusion Matrix (Luque et al., 2019) .....	42
Figure 4.1 Modeling With Extracted Features.....	47
Figure 4.2: Code Snapshot for Logistic Regression Model Building and Results...	48
Figure 4.3 Logistic Regression Confusion Matrix .....	49
Figure 4.4 Code Snapshot for Random Forest Model Building and Results .....	50
Figure 4.5 Random Forest Confusion Matrix .....	51
Figure 4.6 Code Snapshot for Support Vector Machines Model Building and Results .....	52
Figure 4.7 Support Vector Machines Confusion Matrix .....	52
Figure 4.8: Code Snapshot for Naïve Bayes Model Building and Results .....	54
Figure 4.9 Naïve Bayes Confusion Matrix.....	54
Figure 4.10: Code Snapshot for Linear Discriminant Analysis Model Building and Results .....	56
Figure 4.11 Linear Discriminant Analysis Confusion Matrix.....	57
Figure 4.12 Summarized Model Performance .....	59
Figure 4.13 Accuracy Using Line Graph.....	60
Figure 4.14 Precision Using Line Graph.....	61
Figure 4.15 Recall Using Line Graph .....	62

Figure 4.16 F1 Score Using Line Graph .....62  
Figure 4.17 AUC-ROC Using Line Graph .....63

### ACRONYMS AND ABBREVIATIONS

<b>ACOG</b>	American College of Obstetricians and Gynecologists
<b>AI</b>	Artificial Intelligence
<b>ANC</b>	Ante Natal Care
<b>ANN</b>	Artificial Neural Network
<b>ICT</b>	Information Communication & Technology
<b>KDH</b>	Kenya Demographic and Health Survey
<b>KNBS</b>	Kenya National Bureau of Statistics
<b>LDA</b>	Linear Discriminant Analysis
<b>LR</b>	Logistic Regression
<b>MDG</b>	Millennium Development Goals
<b>ML</b>	Machine Learning
<b>MCH</b>	Maternal and Child Health
<b>MoH</b>	Ministry of Health
<b>NB</b>	Naïve Bayes
<b>PE</b>	Pre-eclampsia
<b>RF</b>	Random Forest
<b>SDG</b>	Sustainable Development Goals
<b>SVM</b>	Support Vector Machines

**UNCTAD**

United Nations Conference on Trade and  
Development

**WB**

World Bank

**WHO**

World Health Organization

### **ABSTRACT**

Pre-eclampsia is globally recognized by the World Health Organization as a significant contributor to high rates of morbidity and mortality among infants and mothers. It accounts for approximately 3% to 5% of all reported pregnancy-related complications worldwide. However, in developing nations like Kenya, particularly in the sub-Saharan region, the prevalence of pre-eclampsia is notably higher, ranging from 5.6% to 6.5% of reported pregnancies. Key risk factors associated with pre-eclampsia include sudden elevation in blood pressure, increased protein levels in urine, chronic kidney disease, and the presence of either Type 1 or Type 2 diabetes. This research developed a predictive model for pre-eclampsia utilizing supervised machine learning techniques on socio-demographic data gathered from Kilifi County. A total of 500 secondary data records gathered from Kilifi county hospital were pre-proposed to train and test the machine learning models. To train and test the models, the study employed five supervised machine learning algorithms, namely Logistic Regression, Random Forest, Naïve Bayes, Linear Discriminant Analysis, and Support Vector Machines. Maternal age, marital status, gravida, education level, and ANC attendance were identified as the optimal extracted features using PCA. The logistic regression model outperformed other supervised machine learning models in the study, achieving a high accuracy rate of 0.96 in predicting pre-eclampsia. The results show that Logistic Regression can accurately predict pre-eclampsia within the first trimester of pregnancy. Future research will involve collecting more data from different regions to improve performance and building a mobile application that will improve MCH accessibility in resource-constrained regions in the country.