

**MACHINE LEARNING MODEL FOR PREDICTION OF POSTPARTUM
DEPRESSION, A CASE OF MOMBASA COUNTY**

GEORGE MONGARE KIMWOMI

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

2023

DECLARATION

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

George Mongare Kimwomi

MSIT/0004/2020

Signature: _____ Date: _____

This thesis report has been submitted with our approval as University Supervisors.

Dr. Mvurya Mgala

Signature: _____ Date: _____

Dr. Fullgence Mwakondo

Signature: _____ Date: _____

Dr. Pamela Kimeto

Signature:  Date: JUNE 7, 2023

DEDICATION

I dedicate this thesis to my wife, children and dear parents for their motivation as I undertook the research.

ACKNOWLEDGEMENT

I wish to appreciate all the parties whose contribution and support enabled me to undertake this research. I acknowledge the guidance by Dr Mvurya Mgala, Dr Fullgence Mwakondo and Dr Pamela Kimeto right from shaping the research concept all through to the production of this thesis report which was a great learning experience for me. Much appreciation to this team for sacrificing their time and other resources to educate and guide me, and to check the document for errors to which I pray the almighty God to shower them with blessings.

I acknowledge the panel members of the various seminars I attended at the Technical University of Mombasa and Kabarak University for the positive critique and advice to the research work which contributed to the quality of this thesis report. I also recognize the support of Dr. Obadiah Musau, Dr. Kevin Tole and the staff of the Institute of Computing and Informatics at the Technical University of Mombasa for the immense support extended to me as I undertook the study.

I also acknowledge the role played by Dr Moses Thiga (Kabarak University) and Dr Mgala Mvurya (Technical University of Mombasa) in my selection as a beneficiary for research funding with the Kenya National Research Fund which facilitated my studies.

Many thanks to my wife Linet and children for the motivation and finally to the almighty God for the opportunity to undertake the research.

TABLE OF CONTENTS

DECLARATION	ii
DEDICATION.....	iii
ACKNOWLEDGEMENT	iv
LIST OF TABLES	ix
LIST OF FIGURES	x
ACRONYMS AND ABBREVIATIONS.....	xii
DEFINITION OF KEY TERMS.....	xiv
ABSTRACT.....	xv
CHAPTER ONE.....	1
INTRODUCTION	1
1.1 General Introduction	1
1.2 Background of the Study.....	3
1.3 Statement of the Problem.....	4
1.4 Objectives of the Study.....	5
1.4.1 General Objective.....	5
1.4.2 Specific Objectives	5
1.5 Research Questions	5
1.6 Significance of the Study	5
1.7 Limitations of the Study.....	6
1.8 Scope of the Study	6
1.9 Overview of the Research Process.....	6
CHAPTER TWO	7
LITERATURE REVIEW	7
2.1 Introduction.....	7
2.2 Postpartum Depression.....	7
2.3 Causes of Postpartum Depression.....	8
2.4 Machine Learning.....	10
2.5 Evaluation of PPD Prediction Models	12
2.6 Prediction Models for Postpartum Depression.....	15
2.6.1 Global Focus	15

2.6.2 Focus on Kenya.....	20
2.7 Experiential Learning Theory for PPD Prediction Modeling.....	23
2.8 Conceptual Model	25
2.9 Summary: Lessons Learnt.....	26
2.10 Research Gap	27
2.11 Chapter Summary	28
CHAPTER THREE.....	29
RESEARCH METHODOLOGY	29
3.1 Introduction.....	29
3.2 Research Approval.....	29
3.3 Research Design	29
3.4 Research Framework.....	31
3.5 Research Methods	32
3.5.1 Study Population.....	32
3.5.2 Target Population.....	32
3.5.3 Sampling Procedure and Sample Size	33
3.5.4 Inclusion and Exclusion Criteria.....	33
3.6. Operationalization of CRISP-DM Process.....	34
3.6.1 Business Understanding.....	34
3.6.2 Data Understanding	35
3.6.2.1 Preparation of Data Collection Tools.....	35
3.6.2.2 Reliability of Data Collection Tool.....	36
3.6.2.3 Data Collection.....	36
3.6.2.4 Data Digitization.....	37
3.6.3 Data Preparation.....	38
3.6.3.1 Data Cleaning for Errors and Missing Values	38
3.6.3.2 Feature Scaling.....	38
3.6.3.3 Feature Selection	39
3.6.3.4 Data Separation.....	40
3.6.4 Modelling	41
3.6.4.1 Logistic Regression Model	41

3.6.4.2 Support Vector Machine (SVM) Model	45
3.6.4.3 Extremely Randomized Forests (XRT) algorithm.	47
3.6.4.4 Random Forest Model	47
3.6.4.5 Adaptive Boosting Model.....	48
3.7 Model Evaluation	50
3.8 Chapter Summary	52
CHAPTER FOUR.....	53
MODELLING, FINDINGS AND DISCUSSION	53
4.1 Introduction.....	53
4.2 Distribution of Respondents by County of Residence.....	53
4.3 Features for Model Development.....	53
4.4 Reliability of Data Collection Tool.....	54
4.5 Data Collection and Digitization	55
4.6 Imputation of Missing Values.....	55
4.7 Results of Feature Selection	55
4.8 Dataset Separation.....	59
4.9 Result of Model training	60
4.9.1 Logistic Regression Model Training.....	60
4.9.2 Sequential Minimal Optimization (SMO) Model Training.....	61
4.9.3 Extremely Randomized Trees (XRT) Model Training	63
4.9.4 Random Forest Training Model.....	64
4.9.5 Ada Boosting Model Training.....	66
4.9.6 Selection of the Best Model	68
4.10 Evaluation of Selected Model	69
4.11 Discussion of Research Findings	71
4.11.1 Discussion of Most Predictive Features for PPD.....	71
4.11.2 Discussion of Model Development.....	73
4.11.3 Discussion of Model Evaluation	75
4.12 Chapter Summary	76
CHAPTER FIVE.....	77
SUMMARY, CONCLUSION AND RECOMMENDATION	77

5.1 Introduction.....	77
5.2 Summary of Research Findings in Line with Research Objectives	77
5.3 Conclusion.....	79
5.4 Contribution to Knowledge	79
5.5 Recommendation.....	80
REFERENCES.....	81
Appendix I: Research Schedule.....	95
Appendix II: Research Budget.....	96
Appendix III: Research Approval Documents	97
Appendix IV: National Commission for Science, Technology and Innovation Research Licence	98
Appendix V: Authorization to Conduct Research in Mombasa County.....	100
Appendix VII: Features for Building the PPD Prediction Model	102
Appendix VIII: Data Collection Tool	108
Appendix IX: Cross-Section of Pilot Data (Source : SPSS)	113
Appendix X: Cross-Section of Digitized Dataset (Source : SPSS)	114
Appendix XI: Univariate Analysis of Collected Data (Source: SPSS).....	115
Appendix XII: Descriptive Statistics for Scaled Dataset (source: SPSS)	118
Appendix XIII: Rotated Component Matrix (source: SPSS).....	121

LIST OF TABLES

Table 2.1: Analysis of evaluation methods for different models used	15
Table 2.2: Analysis of studies on PPD prediction models reviewed	21
Table 3.1: Research design and expected outcome	30
Table 4.1: Distribution of respondents by county of residence	53
Table 4.2: Analysis of attribute distribution by category	54
Table 4.3: Results of Cronbach's Alpha reliability test (source: SPSS)	55
Table 4.4: Results of KMO and Bartlett's Test	56
Table 4.5: Interpretation of components and Assigned labels.....	58
Table 4.6: Results of modelling highlighting the best score for each matrix (source: WEKA)	68
Table 4.7: Result of model evaluation with different data sizes (source: WEKA) ..	70
Table 4.8: Analysis of selected features by category	72
Table 4.9: Table comparing performance results of this study with reviewed studies	75
Table 4.10: Comparison of training and evaluation results for selected model (Random Forest).....	75

LIST OF FIGURES

Figure 2.1: Basic structure of Machine Learning cycle (H. Wang et al., 2009)	11
Figure 2.2: Standard machine learning process (Alzubi et al., 2018).....	12
Figure 2.3: Conceptual model of the proposed PPD prediction model	26
Figure 3.1: CRISP-DM Data mining model (Plotnikova et al., 2020).....	31
Figure 3.2: Research framework for postpartum depression prediction model (Plotnikova et al., 2020)	35
Figure 3.3: Front page photograph of “MOTHER & CHILD HEALTH HANDBOOK – 2020” (source: CGTRH).....	37
Figure 3.4: Cross-section of SPSS Statistics variable view of data editor screen settings (source: SPSS Statistics)	38
Figure 3.5:K-fold cross validation process of training the prediction model for postpartum depression (k=10).....	41
Figure 3.6: Sigmoid function graph showing predictive features and the probability of postpartum depression (“Understanding Logistic Regression,” 2017).....	43
Figure 3.7: Support vector machine for a two-dimensional classifier (Joloudari et al., 2020)	46
Figure 3.8: Illustration of AdaBoost modelling process.....	49
Figure 4.1: Scree plot graph for eigenvalue against the component number (source: SPSS).....	57
Figure 4.2: Cross-section of discretized features for the selected dataset (key: “-inf” = infinity. source: WEKA)	59
Figure 4.3: Configuration screen of the resample filter screen for the training dataset (source : WEKA)	60
Figure 4.4: Result of Logistic Regression training model (source: WEKA)	61
Figure 4.5: Visualized Plot of ROC Curve for Logistic Regression for class value 1 (where class value 1=Depressed, ROC Area =0.8031). (Source: WEKA)	61
Figure 4.6: Result of sequential minimal optimization modelling (source: WEKA).	62

Figure 4.7: Visualized ROC Curve of Sequential Minimal Optimization model for class value 1 (where class value 1=Depressed, ROC Area =0.7881). (Source: WEKA) ..	63
Figure 4.8: Result of Extremely Randomized Trees model (source: WEKA).....	64
Figure 4.9: Visualized ROC Curve of Extremely Randomized for class value 1 (where class value 1=Depressed, ROC Area =0.7509). (Source: WEKA).....	64
Figure 4.10: Result of Random Forest training model (source: WEKA)	65
Figure 4.11: Visualized ROC Curve of Random Forest for class value 1 (where class value 1=Depressed, ROC Area =0.8626). (Source: WEKA)	66
Figure 4.12: Results of Ada Boosting model training (source: WEKA).....	67
Figure 4.13: Visualized ROC Curve of AdaBoostM1 for class value 1 (where class value 1=Depressed, ROC Area =0.8437). (Source: WEKA).....	67
Figure 4.14: Comparative graph showing results of model training by performance matrix	69
Figure 4.15: Results of evaluation for selected model (Random Forest) (source: WEKA)	70
Figure 4.16: Evaluation of random forest model with different data sizes.....	71
Figure 4.17: Analysis of selected features by category.....	73
Figure 4.18: Empirical representation of the PPD prediction model.....	74

ACRONYMS AND ABBREVIATIONS

AUC-ROC	Area under receiver operating characteristic curve
BDI	Beck Depression Inventory -II
CGTRH	Coast General Teaching and Referral Hospital
CRISP-DM	Cross-Industry Standard Process for Data Mining
EHR	Electronic Health Records
EPDS	Edinburgh Postnatal Depression Scale
KDD	Knowledge Discovery for Databases
KMO	Kaiser-Meyer-Olkin
ML	Machine Learning
MDD	Major Depressive Disorder
NYCCDRN	New York City Clinical Data Research Network data
PCA	principal component analysis
PHQ-9	Patient Health Questionnaire-9
PPD	Postpartum Depression
RF	Random Forest
ROC	Receiver operating characteristic curve
SEMMA	Simple, Explore, modify, Model and Asses
SDG	Sustainable Development Goals
SFS	sequential forward selection

SPSS	Statistical Package for Social Sciences
SVM	Support Vector Machine
TSH	Tudor Sub-County Hospital
USPSTF	United States Preventive Services Task Force
UK	United Kingdom
UN	United Nations
WEKA	Waikato Environment for Knowledge Analysis
WCM	Weill Cornell Medicine
XRT	Extremely randomized trees

DEFINITION OF KEY TERMS

Area Under Receiver Characteristic Curve (AUC-ROC)	Performance matrix used to measure how well an algorithm can correctly discriminate output classes for all possible classifications
Binary Cross Entropy	the negative average of the log of corrected predicted probabilities for each class
Machine Learning:	Equipping computers with ability to learn from experience without directly reprogramming them
Model:	A conceptual representation of reality with a set of selected elements of the target system
Postpartum Depression:	An incapacitating but curable mental illness which affect mothers after delivery

ABSTRACT

Postpartum depression is a medical condition which affects many mothers. The condition exposes the mother and the newborn baby to illnesses that can lead to death. Management of the condition requires heavy expenditure incurred by the family, government, and stakeholders. The condition is also a source of many social problems. Manual systems are currently used to predict the condition which is slow and inconsistent. Machine learning technology which has reliably been used in prediction modelling in other domains can also be employed to build a model to predict mothers at risk of postpartum depression during pregnancy for primary prevention. In this study, perinatal records were collected from 324 mothers attending postnatal healthcare clinics at the Coast General Teaching and Referral Hospital and Tudor Sub-County Hospital in Mombasa County. The filter feature in WEKA was used to split the data into 70% and 30% for model training and testing respectively. Models were built on WEKA machine learning platform using logistic regression, support vector machine, extremely randomized trees, random forest and adaptive boosting algorithms which were identified from literature. A positive case of postpartum depression was defined as diagnosis or treatment of major depression within one year after delivery. Random forest model produced the best performance with a receiver operating characteristic (ROC) curve area of 0.863867 which is comparable within the bracket of high performing models. With this level of performance, the model can be used by healthcare staff to make quick and consistent prediction for early mitigation measures. Further research could be done with more data collected from other counties in Kenya.

Key words: Machine Learning, postpartum depression, prediction, mitigation measures.