



PREDICTING ANTENATAL CARE (ANC) VISITS USING MACHINE LEARNING ALGORITHMS

Ngaji Calvince Otieno

I56/32981/2019

Department of Mathematics
Faculty of Science and Technology

A research project submitted for the partial fulfillment of the requirement for the
degree of Master of Science in Social Statistics.

University of Nairobi

November 2022

Predicting Antenatal Care (ANC) Visits using Machine Learning Algorithms

Research Report in Social Statistics, Number XX, 2022

Ngaji Calvince

Department of Mathematics
Faculty of Science and Technology
Chiromo, off Riverside Drive
30197-00100 Nairobi, Kenya

Master Thesis

Submitted to the Department of Mathematics in partial fulfilment for a degree in Master of Science in Social Statistics

Submitted to: The Department of Mathematics, University of Nairobi, Kenya

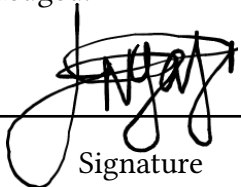
Abstract

In 2016, the World Health Organization (WHO) updated the prior recommendations for Antenatal Care (ANC) visits and suggested 8+ ANC appointments over the course of pregnancy. The aim of this was to address the problem of perinatal deaths and improve maternal health. However, the reality is that the uptake of ANC visits is still very low as pregnant women do not adhere to this guideline. As a result, the initial step in resolving the issue could be to work toward meeting the goal of four ANC visits that was initially set. This can be accomplished by first gaining an understanding of the factors that cause women to not adhere to the established guidelines, which ultimately results in a low number of pregnant women using the service. The purpose of this study is, therefore, to provide a solution to the problem of understanding why there is such a low uptake of the service by developing a robust machine learning model that can accurately predict whether or not a pregnant woman in Kenya is likely to make at least four visits to an ANC clinic given a set of demographic factors. The research also intends to uncover the determinants connected to at least four ANC visits in Kenya, as well as to describe those pregnant women who are at risk of not obtaining ANC services, using data from KDHS, 2014. The KDHS data is a large-scale dataset provided by the Demographic Health Survey Program. The research looks at different machine learning algorithms, including Artificial Neural Network (ANN), Support Vector Machine (SVM), Generalized Linear Model (GLM) and to see which one surpasses the others when it comes to predicting 4+ ANC visits. The trained models' accuracy was increased using the cross-validation technique, and the model's performance was measured using metrics including accuracy, specificity, and sensitivity. The characteristics associated with at least four ANC visits among Kenyan women were determined using an odds ratio and a p-value at a 5% significance level. The most relevant features were selected using the Random Forest Gini index. When compared to the other algorithms, the Artificial Neural Network (ANN) performed the best in predicting ANC visits, with an accuracy of 82.9%. According to the findings, in Kenya, factors like wealth index, level of education, woman's current age, whether the pregnant woman has a supportive partner, total children ever born, and place of delivery have been revealed to be significant determinants of at least four ANC visits. The study suggests that safer programs for disadvantaged and vulnerable women be implemented, as well as the inclusion of male partners during antenatal care, good media coverage, and promotion of early antenatal care and health promotion programs for pregnant mothers with low education.

Master Thesis in Mathematics at the University of Nairobi, Kenya.
ISSN 2410-1397: Research Report in Mathematics
©Ngaji Calvin, 2022
DISTRIBUTOR: Department of Mathematics, University of Nairobi, Kenya

Declaration and Approval

I undersigned hereby declare that Predicting Antenatal Care (ANC) Visits in Kenya using Machine Learning Algorithms is my original work, that it has never been published before and has not been submitted for consideration for any academic award or examination at any other University, and all sources cited or used in this dissertation have been properly cited and acknowledged.



Signature


23/11/2022

Date

NGAJI CALVINCE OTIENO

Reg No. I56/32981/2019

In my capacity as a supervisor of the candidate's dissertation, I certify that this dissertation has my approval for submission.



Signature

22/11/2022

Date

Dr Timothy Kamanu
Department of Mathematics,
University of Nairobi,
Box 30197, 00100 Nairobi, Kenya.
E-mail: tkamanu@uonbi.ac.ke

Dedication

This dissertation is dedicated to my wife Janet, who has continuously inspired me and supported me during my graduate studies and personal struggles. I appreciate having you in my life. This work is also a tribute to my mother, Nancy Anyango, who has always loved me without condition and inspired me to strive for my goals by setting a good example. Last but not least, I must thank Yasmin, my dear daughter, who has always stood by my side as I prepared this paper.

Contents

Abstract	ii
Declaration and Approval	v
Dedication	viii
Acknowledgments	xi
1 CHAPTER ONE: INTRODUCTION	1
1.1 Background.....	1
1.2 Problem Statement.....	2
1.3 Objectives.....	3
1.4 Significance of the study.....	3
1.5 Conceptual Framework.....	4
2 CHAPTER TWO: LITERATURE REVIEW	5
2.1 Introduction.....	5
2.2 Antenatal Care (ANC).....	5
2.3 Empirical Literature Review.....	6
3 CHAPTER THREE: METHODOLOGY	11
3.1 Introduction.....	11
3.2 Data Source.....	11
3.3 Sample Design.....	11
3.4 Exploratory Data Analysis (EDA).....	11
3.5 Data Pre-Processing Techniques.....	12
3.5.1 Feature Engineering.....	12
3.5.2 Variable transformation.....	14
3.5.3 Train-Test Split.....	15
3.5.4 Feature Importance/Selection using Random Forest.....	15
3.5.5 Correlation.....	16
3.6 Machine Learning Models Used.....	17
3.6.1 Logistic Regression.....	17
3.6.2 Random Forest.....	18
3.6.3 Support Vector Machine (SVM).....	19
3.6.4 Artificial Neural Network (ANN).....	20
3.7 Improving Model Performance.....	22
3.7.1 Cross Validation.....	22
3.8 Model Evaluation Metrics.....	22
3.8.1 Confusion matrix.....	22
3.8.2 ROC Curve.....	24
4 CHAPTER FOUR: RESULTS	25

4.1	Data Gathering	25
4.2	Exploratory Data Analysis (EDA)	27
4.2.1	Participants Characteristics	27
4.2.2	Univariate Analysis	29
4.2.3	Bivariate Analysis	32
4.2.4	Correlation	34
4.2.5	Missing Value Analysis	35
4.3	Data pre-processing	37
4.3.1	Feature Engineering	37
4.4	Binary Logistic Regression Model	38
4.5	Model Training	40
4.5.1	Model Comparison and Evaluation Metrics	40
5	CHAPTER FIVE: CONCLUSIONS AND RECOMMENDATIONS	44
	Bibliography	45

Acknowledgments

Let me begin by thanking the almighty God for the gift of life He bestowed upon me during the time I was working on this project. I would also like to thank my supervisor, Dr. T Kamanu, for his constant support and guidance during this time. Throughout the research effort, Kamanu was a constant source of encouragement and was always eager to help in any way he could. I am really thankful for your knowledge and inspiration that I received from you during class and this project. Finally, I would want to thank Mr. Krish Naik, the co-founder of iNeuron. Krish has over ten years of industry experience and is a mentor, lecturer, and expert in computer vision, deep learning, and machine learning. On his YouTube channel, Krish explains several concepts related to Artificial Intelligence (AI), Machine Learning (ML) and Deep Learning, (DL) using numerous examples of real-world problems that inspires many data scientists. At several meetings, scientific institutions, and community-organized forums, he has presented more than thirty technical talks on data science, machine learning, and AI.

Ngaji Calvince

Nairobi, 2022.

1 CHAPTER ONE: INTRODUCTION

1.1 Background

Antenatal Care (ANC) is a plan that was adopted by the World Health Organization (WHO) to ensure the health of women and their unborn children are well taken care of. It is a preventive health care measure where pregnant women are educated by skilled health personnel about health behaviors during pregnancy, receive teachings on family planning and breastfeeding. Through ANC, women can learn and understand warning signs or complications that may arise during pregnancy and childbirth. They also stand a better chance of receiving support related to social life, psychological and emotional. Through this care they can also get treatment of hypertension to help them prevent eclampsia, as well as immunization to prevent tetanus. It is during this period when women can be tested and provided with medication to prevent mother-to-child HIV transmission and be provided with insecticide-treated mosquito nets especially for those coming from areas where malaria prevalence is high. Where specialized treatments are necessary, mothers are referred to other facilities and they are also encouraged to use skilled birth attendant during delivery.¹

It is worth noting that ANC utilization increased in middle- and low-income countries since the WHO FANC model was launched in 2002. The model not only focus on the ANC utilization but also ensures that women get all the related maternal treatments. In FANC, health care professionals place a lot of attention on individualized assessments and the steps required to decide on Antenatal Care with the pregnant woman.

In the traditional system of Antenatal Care, women made ANC visits as a routine activity, and health care professionals categorized pregnant women based on routine risk indicators. This led to higher rates of complications throughout pregnancy. While in FANC system, specific conditions of each woman is taken into account when providing services. With this strategy, the care of the pregnant woman becomes a family duty. The Woman's husband and the medical professional discuss any difficulties the woman might encounter, help her prepare for childbirth, and discuss with her about postpartum care and future childbirth concerns. The outcomes for pregnant women and their unborn children are improved when they receive basic care both at home and at a medical institution, when problems are rapidly diagnosed by the family and a healthcare professional, and when interventions are started promptly.²

¹ https://www.who.int/health-topics/maternal-health#tab=tab_1

² <https://www.open.edu/openlearncreate/mod/oucontent/view.php?id=44&printable=1>

Despite the increase in ANC utilization brought about by FANC model, only 64% of pregnant women worldwide attended the minimum of four ANC visits between 2007 and 2014 (Oshinyemi et al., 2018), showing that much more work needs to be done to improve ANC use and quality. As a result, the WHO has produced a comprehensive recommendation for pregnant women and adolescent girls about frequent ANC. These recommendations are aimed to supplement the most recent WHO recommendations for treating various pregnancy-related conditions. The guidelines are also meant to reflect and respond to the complexities of the challenges surrounding the practice and delivery of ANC. The new guidelines require women to seek the services of ANC at least eight times during their pregnancy. Despite the World Health Organization (WHO) making all these efforts to introduce a new recommendation on ANC, the usage rate among pregnant women is still quite low, which calls for urgent intervention. The plan would be to work toward achieving the objectives of the initial recommendation, which required pregnant women to attend at least four ANC appointments throughout their pregnancy. This can begin with an investigation of the factors that led to the low usage of the initially planned 4 ANC visits as a starting point.

1.2 Problem Statement

Since 1990, there has been a dramatic decrease in the number of maternal deaths around the world. If the condition is not improving or worsening, it suggests that efforts to solve this issue are either insufficient or improperly implemented. Maternal mortality is way too high around the world, as evidenced by the fact that 295 000 women died in 2017. In low-resource areas, 94% of these deaths occurred, and 95% of them might have been prevented (World Health Organization, 2019).

In 2017, more than two-thirds of all maternal deaths or 254 000 maternal deaths occurred in Sub-Saharan Africa and Southern Asia. Most maternal deaths (196 000) happened in Africa, whereas just 58% (58 000) occurred in Asia, Southern Europe and the Middle East included (Bongaarts, 2019). As a result, maternal and perinatal mortality are serious public health issues that require investigation. There are an estimated 362 maternal fatalities for every 100,000 live births in Kenya at this time. On the other hand, there were 23 stillbirths in 1,000 live births, which falls far short of the target of 12 stillbirths and 147 maternal deaths for every 1,000 live births (World Health Organization, 2015). Despite the fact that this practice has been shown to reduce maternal and perinatal morbidity and death, only 61.8% of deliveries in Kenya are attended by a qualified health care provider. The low usage of specialized care during pregnancy and childbirth has been linked to the high cost of maternity care (Chou et al., 2015). In Kenya, a pregnant woman should see a doctor at least four times during her pregnancy. The aim of this study is to identify the factors that put pregnant women at risk of not using ANC services and to develop an accurate machine learning model for determining whether a pregnant woman will complete the advised 4+ ANC visits.

1.3 Objectives

The objective of this study is to identify demographic characteristics related with at least four ANC visits among pregnant women in Kenya and use these characteristics to create a robust Machine Learning model that can effectively predict whether a pregnant woman in Kenya is likely to seek at least four ANC visits based on demographic characteristics. The specific objectives are:

1. To build a robust Machine Learning Model capable of predicting the likelihood of a pregnant woman in Kenya utilizing 4+ ANC visits given demographic characteristics using KDHS dataset.
2. To identify factors associated with at least four ANC visits among pregnant women in Kenya using KDHS data.
3. To select the optimal Machine Learning algorithm in predicting the likelihood of ANC utilization by pregnant women in Kenya using KDHS dataset.

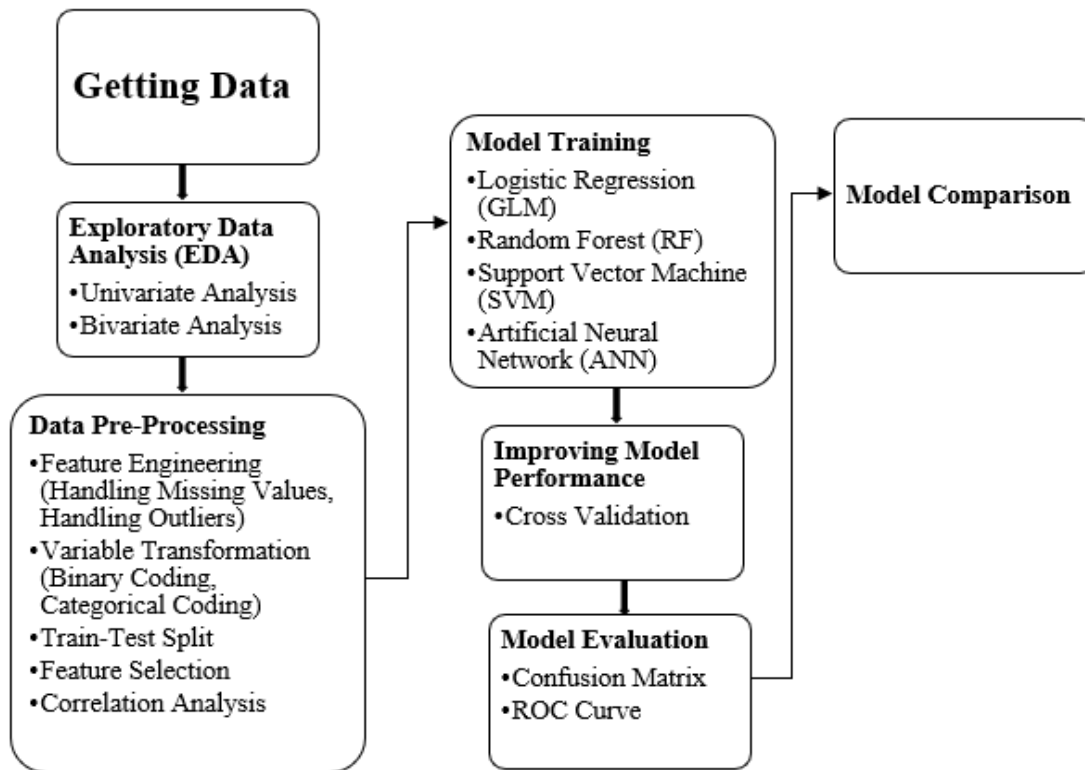
1.4 Significance of the study

Antenatal Care enables the management of a pregnancy, as well as the detection and treatment of any complication that may arise during childbirth and the promotion of good health. Women, on the other hand, rarely considers childbearing as a problem and hence do not seek care from health professional. This has an effect on the usage of maternal health services in places where poverty and illiteracy are prevalent. However, the chance of complications occurring is high, and therefore frequent examinations are highly recommended. The study combines multiple methods of analysis, including exploratory analysis, adopting Machine Learning life cycle, and inferential analysis, to produce unique and thorough conclusions about specific features of Kenyan women and their households that are linked to ANC utilization patterns. In this study, the chosen model will aid health professionals in evaluating the chance of pregnant women using 4+ ANC visits.

The following is how rest of the thesis is organized: The second chapter offers a comprehensive literature review of contemporary studies on the subject. First, the section starts by looking at comprehensive understanding of Antenatal Care (ANC). Then, the chapter discusses similar work in Antenatal Care usage modeling. The technique and methods utilized in this thesis is detailed in Chapter 3, including a description of the data source and the sample design that was used. The methodology discussed here includes Exploratory Data analysis (EDA) process, data pre-processing techniques like Feature Engineering, Train Test Split, Feature Selection method, the machine learning algorithm

that was used, as well as the metrics that were used to evaluate the models. Chapter 4 presents the research findings. Finally, discussions and recommendations are presented in Chapter 5.

1.5 Conceptual Framework



2 CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

This chapter offers a thorough overview of the relevant literature for this project. First, the study discusses theoretical literature by looking into the concept of Antenatal Care (ANC) and its utilization in Kenya. The research then summarizes relevant Antenatal Care utilization modeling past studies, as well as the findings and conclusions obtained from these studies.

2.2 Antenatal Care (ANC)

ANC is described as the treatment given to teenage girls and pregnant women by trained medical specialists (WHO, 2016). Important components of ANC include health promotion and education, risk detection, prevention, and management for diseases associated with pregnancy and their coexisting conditions. ANC can reduce maternal and perinatal mortality by identifying pregnant women and girls who are more likely to have issues during labor and delivery and then referring them to the appropriate level of treatment (Otundo Richard, 2019). In order to set the number of ANC visits to 4, the WHO created the Focused Antenatal Care (FANC) model, which concentrated on treatments that have been demonstrated to benefit the health of pregnant women. This technique is crucial to a healthy birth since it allows women to become aware of the signs of potentially fatal disorders such pre-eclampsia, malaria infection, and obstructed labor in an effort to reduce maternal mortality.

The World Health Organization recommends that a healthy pregnant woman receive four prenatal visits at the very least. When expecting mothers start prenatal treatment in the first trimester and keep it up throughout the pregnancy, they are more likely to have a healthy baby (Heaman et al., 2008). However, there is still more to be done to safeguard the health of women everywhere in the world during pregnancy. Preventing life-threatening risks and mortality during pregnancy and childbirth requires easy access to high quality maternal health care delivered by trained health professionals. As a result, efforts to promote safe motherhood must be comprehensive in scope; even when excellent health treatments are readily available, social, economic, and cultural barriers may hinder women from taking advantage of them (Hapsara, 2005). Prenatal care coverage, according to the World Health Organization (WHO), is the proportion of women between the ages of 15 and 49 who gave birth to a live child and received antenatal care at least four times during pregnancy. Pregnancy care, birth planning, and emergency

preparedness are the focus of new FANC guidelines from the Kenyan Ministry of Health (MOH). Allowing for a more comprehensive and integrated care delivery, these visits are increasingly employed as a gateway to various reproductive health treatments (MOH, 2019). Every pregnant woman and her baby will receive high quality care throughout their pregnancy as a result of this intervention. Despite attempts to reduce maternal mortality, less than half of pregnant women in Kenya receive the necessary four ANC visits (Micronutrient Initiative Kenya, 2016).

2.3 Empirical Literature Review

A lot of study has been done on ANC visits modeling, especially the statistical modeling technique approach. Only a few papers concentrating on the application of Machine Learning and Data Science to Antenatal Care utilization have been found. In this study, both of these elements were explored, and the results are presented below.

A research conducted in Bangladesh (Bhowmik et al., 2020) aimed at finding the best count regression model for predicting the number of ANC visits by pregnant women while taking into consideration over dispersion, intra-cluster correlation and zero inflation. The study also sought to establish the factors for ANC utilization. The study also sought to establish the factors for ANC utilization. The study employed data from the Demographic Health Survey 2014 in Bangladesh, which revealed that 4493 women (or 22% of the study participants) did not attend an ANC visit. The traditional one-part count regression model and a two-part zero-inflated hurdle regression were both taken into account. The researchers were able to account for rectification among response values by integrating cluster-specific random effects in the model. As a diagnostic test tool, the likelihood ratio and uniformity test were employed, and the results revealed that the most effective model was a hurdle negative binomial regression model with a cluster-specific random intercept in both the zero and count parts. The study discovered that women with a low educational background, no access to the media, and who live in a low-income home use ANC less frequently. When modeling ANC visits in Bangladesh, the study recommended employing over dispersion, zero-inflation, and cluster-specific random effects. The study further recommended safer programmes to disadvantaged and vulnerable women.

The researcher also took into account a cross-sectional community-based study conducted in Ethiopia (Ftwi et al., 2020), which combined quantitative and qualitative methods. The study's objective was to identify the prevalence and contributing factors among mothers who had given birth during the previous six months and had received the necessary four or more antenatal care (ANC) visits. The qualitative data was gathered from key informants in eight mothers who were chosen using a purposive sampling strategy. The quantitative portion entailed employing simple random sampling to present a systematic questionnaire that was trialed on 466 mothers. To establish the relationships between input and outcome variables and draw conclusions based on the results of logistic

regression, bivariate and multivariate analysis were used. According to the findings, 63% of the study participants achieved at least four ANC visits. The findings also found that 9.9% of the population attained 4+ ANC visits. Membership in a community health insurance plan, having a full-time job at home, having a supportive husband, and commuting a long distance to the hospital were discovered to be connected to making complete four ANC visits. The study suggested that community-based interventions be integrated.

Another study conducted in Ethiopia (Tizazu et al., 2020) aimed to determine the uptake of 4+ ANC visits in Ethiopia, as well as the factors that influence it. This was a community based cross sectional study design, which recruited 390 women who had recently given birth in the six months before the research using a simple random sample technique. The data was collected using a questionnaire which had been pre-tested. The characteristics associated with the usage of four or more ANC visits were identified using a both bivariate and multivariate logistic regression model. 78% of the women, was discovered to have attained at least four ANC visits. The key factors of at least four ANC visits were identified as education status, husband support, and early commencement of ANC visits. The study recommended involvement of male partners during antenatal care, promotion of early initiations of antenatal care and health promotion programs to pregnant mothers with low education.

In another study undertaken in Ethiopia (Abegaz, 2018) the goal was to investigate the trend and identify the impediments to Ethiopian mothers utilizing ANC. The data mining technique was used in this work, which is a part of big data research that tries to identify patterns and information in enormous amounts of data. The whole Ethiopian Demographic Health Survey dataset from 2000 to 2016 was used in this study. The pooled cross-sectional study included steps in the knowledge discovery process such as cleaning, selection, transformation, integration, and data mining approaches. The algorithms employed include classification, grouping, association rules, and attribute ranking with pattern prediction. The results showed that in the year 2000, 2005, 2011, and 2016, the proportion of people that used ANC was 27.6%, 28.2%, 34.5%, and 62.9%, respectively. The total number of pregnant mothers in the study was 28,631, according to the pooled statistics. More than half of these women (56.09 percent) did not use ANC during their pregnancy. Pregnancy problems, mothers' and husbands' educational status, mothers' residence, economic position, and media exposure all had a confidence level of 95 percent or greater in relation to ANC utilization. According to the author, even if there are no pregnancy complications, pregnant mothers should seek ANC services. Second, enhancing women's understanding of ANC throughout pregnancy requires addressing critical areas such as education and poverty reduction. Third, in order to boost ANC service utilization, infrastructure expansion in rural communities with good media coverage should be prioritized.

Research conducted in Sub-Saharan Africa (Okedo-Alex et al., 2019) had the aim of determining the characteristics that affect prenatal care utilization in Sub-Saharan Africa. Past research from online databases such as PubMed, CINAHL, EMBASE, OVID, and Web of Science were used in the study, which used a systematic review method. The research looked at papers published in English between 2008 and 2019 that looked at the factors of ANC use in Sub-Saharan Africa using a multi-variate approach. Using a data extraction form, information such as author, publication year, location of study, study subjects, sample size, study design, and determinants were all collected. The study selection and eligibility were guided by meta-analysis protocol and the Preferred Reporting Items for Systematic Reviews. The studies were assessed using Quality Assessment Tool for Observational Cohort and Cross-Sectional Studies, and the findings were reported using the Andersen framework. A total of 74 studies were found to meet the criteria for inclusion and were thoroughly examined. Factors such as type of dwelling, socioeconomic position, age, woman's level of education, having a partner who is educated, work, marriage, and religion had all been identified as predictors of ANC utilization in several research. Awareness of warning signs, adequate scheduling and quantity of antenatal checkups, media exposure, and a positive attitude toward ANC usage were all factors that contributed to attendance and the start of ANC in the first trimester. An unanticipated pregnancy, lack of support from the husband, previous pregnancy issues, lack of health insurance, the considerable distance to the health facility, reduced autonomy, and the high cost of services all connected to ANC uptake. Inter-sectorial cooperation to improve female empowerment and education, as well as geographic diversity access, and reinforcing ANC policies with active community participation were all proposed in this study.

(Wairoto et al., 2020) used the KDHS, 2014 dataset with the aim of estimating the coverage of 4+ visits in Kenya. This was conducted at the sub-county level. The research also aimed at investigating characteristics related with ANC use in Kenya. Small Area Estimation was used to construct sub-county estimates of 4+ ANC using data from the KDHS 2014. This method relies on geographical relatedness to provide exact and reliable estimates for each of the 295 sub-counties. Factors influencing ANC use were determined using hierarchical mixed-effect logistic regression. To show discrepancies, sub-county estimates of characteristics strongly linked to ANC use were created and plotted using SAE methodologies. ANC coverage varied greatly between sub-counties, starting from 17percent in Mandera West to above 77percent in Ruiru and Nakuru Town West. Coverage was less than half in 31percent of the 295 sub counties. ANC use was linked to marital status, wealth index, age at first marriage, maternal education, place of delivery, and birth order. Low ANC utilization rates were found in areas with fewer educated women, low socioeconomic status, and a small number of deliveries at health facilities. According to the study, national and county governments should rely on initiatives that rely on data to guide policy formulation and budget allocation in order to enhance equity in health service access and utilization.

With the aim of determining the use and determinants of at least four ANC visits in 12 East African nations, (Tessema & Minyihun, 2021), conducted a study. Data between 2008 to 2018 from 12 East African countries' Demographic and Health Surveys were used in the study. To collect information that is comparable across countries, the DHS program normally uses standardized methods including consistent questionnaires, manuals, and field procedures. A multivariable logistic regression model was used to determine the factors that were predictive of receiving at least four antenatal care services. In the East African area, the pooled utilization of 4+ ANC visits was 52.44%, with Zimbabwe (75.72%) having the greatest use and Ethiopia the least (31.82%). The results showed that determinants of completing 4+ ANC visits were marital status, age category, education levels, birth order, planned pregnancy, contraceptive utilization, wealth index, and having no problem accessing health care. The study proposed that inter-sectorial collaboration be used to increase female education and empowerment, increase geographical access to health care, and boost antenatal care policy implementation with active community participation. Furthermore, a favorable atmosphere for impoverished women to engage in entrepreneurial activity is required.

The last study reviewed was conducted in Guinea (Shibre et al., 2020). The goal of this study was to learn more about the factors that determine the use of trained ANC personnel in Guinea. The participants in this study were 7812 women who had never married before. To investigate factors linked to the use of ANC, researchers employed multivariable logistic regression. The odds ratio (adjusted) and the corresponding 95% CI were utilized to present the results of the multivariate logistic regression. Having decision-making power, work status, media exposure, mother education, husband or partner education, household economic position, place of residence, and ethnicity were found to have significant associations with usage of ANC skilled personnel services in Guinea. The report recommends Guinean officials to make the necessary measures to minimize disparities and gaps in the utilization of ANC services among various categories of women.

According to this empirical literature study, most research use logistic regression, bivariate and multivariate analysis to identify the characteristics that substantially influence the utilization of at least four ANC visits in the studied geographical areas. Most of the studies also reveal that low education, access to media, safer, involvement of male partners, lack of health insurance, significant distance to the health institution, restricted autonomy, and high cost of services are the main predictors of ANC consumption. The gap identified in these studies is that each study uses either multivariate statistical modeling or data mining techniques, whereas this research uses a combination of the two methodologies. This paper is more organized because it starts by understanding the factors associated with at least four ANC visits, then filters these factors and uses them to build a robust machine learning model capable of predicting whether a pregnant woman is likely to complete the minimum four ANC visits based on her demographic character-

istics. The study also makes effective use of the cross-sectional KDHS dataset, which is a well-known large-scale dataset covering Antenatal Care information.

3 CHAPTER THREE: METHODOLOGY

3.1 Introduction

This section covers all of the methodologies, stages, and approaches used in this study to reach all of the research's specific goals. The following is a step by step breakdown of this section.

3.2 Data Source

Data was obtained from mothers who had a live birth five years preceding the KDHS 2014 program. The dataset is available in their website³ upon request through writing a formal e-mail. There are new indicators that were not captured in prior KDHS surveys but were included in the 2014 survey. For instance, it is the first nationwide survey that provided county level figures for demographic and health factors.

The KDHS gather information to estimate fertility, adult mortality, assess childhood, track changes in fertility and contraceptive use, estimate women's and children's nutritional status, and look at basic indicators of maternal and child health, and describe patterns of HIV and other disease transmission knowledge and behavior. In addition to this, KDHS aims to determine the extent and pattern of domestic violence and female genital mutilation.

3.3 Sample Design

The KDHS (2014) sample was derived from the Kenya National Bureau of Statistics' Fifth National Sample Survey and Evaluation Programme (NASSEP V), a master sampling frame used to conduct household-based surveys across the country.

3.4 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is type of analysis that focuses on finding, exploring, and empirically recognizing patterns in data (Jebb et al., 2017). EDA is the process whereby the researcher is getting to know or becoming acquainted with the variables contained in the data. EDA is also used to inform model building decisions and to supplement inferential analyses. In this research, univariate and bivariate analysis are the examples of

³ www.measuredhs.com

techniques covered under EDA. Box plots, density plots, and histograms are the univariate data visualization methods explored. On the other hand, bivariate are multivariate techniques such as correlation matrices and scatterplots (Winson-Geideman et al., 2017). In this paper, the researcher explored the relationships among variables in the dataset using correlation matrix.

In this research, the specific goal of EDA was to gain a deeper understanding of the underlying patterns in Antenatal Care visits among women in Kenya and to uncover links between it and the demographic characteristics of the women under study. Under Univariate analysis, this paper focuses on discovering the patterns on the ANC visits among women in Kenya, by examining the distribution of at least four ANC visits and those who used less or did not completely utilize this service. In addition, the research explored the distribution of each variable in the dataset using descriptive and visualization techniques. Under bivariate the research explored the distribution of Antenatal care utilization with respect to demographic characteristics such as region, type of residential area, level of education, religion, wealth index and marital status of pregnant women in Kenya.

3.5 Data Pre-Processing Techniques

3.5.1 Feature Engineering

Feature preprocessing is one of the most significant steps in the creation of a Machine Learning model, and it is considered one of the most important aspects of machine learning project. The approach can be described as a technique which involves extracting and refining features from raw data, and it can be used to develop adequate input data for the model and to improve the model's performance. Feature engineering was one of the pre-processing techniques explored. According to (Nargesian et al., 2017), Feature Engineering involves process of modifying a dataset's feature space to improve predictive modeling performance. Feature engineering's goal is to get the highest prediction accuracy possible. It is easier to interpret a model with fewer features, it is more likely to generalize better (less chance of over fitting), and it is much faster (more efficient). Feature Engineering includes task such as dealing with missing values, feature transformation, feature selection and many more. In this research, the Feature Engineering techniques employed are as follows.

Handling missing values

In the real world the researchers are often faced with the challenge of missing data. This may be as a result of corruption in data, privacy concerns, human error, and failure of equipment used to collect data (Tran et al., 2017).

Missing data diminishes the sample's representativeness, which can lead to erroneous conclusions about the population. To perform a task like classification or regression, machine learning algorithms require complete data (Salgado et al., 2016). This emphasizes the importance of missing data handling in the machine learning life cycle. Various strategies, such as deletion, predictive modeling, and imputation, can be used to deal with missing values (Yadav & Roychoudhury, 2018). To treat missing values in the dataset, this study used an imputation method. In this method, missing values in categorical data are replaced with the most frequent category, while missing values in continuous variables are replaced with the mean/median. This research adopted mean value imputation for all continuous variables with missing observations and mode for all the categorical variables. *Hmisc* package in R programming Language was used to accomplish this task.

Handling outliers

Outlier detection is a data mining and machine learning process aimed at identifying unusual observations, often known as outliers (Yang et al., 2019). The occurrence of outliers, which are data points that differ greatly from others, is one of the most challenging concerns in data pre-processing. Outliers can occur as a result of an experimental, measurement, or encoding error. Outliers can be detected using a variety of techniques ranging from simple descriptive statistics such as Maximum, minimum, boxplot, histogram, and percentiles to more formal procedures like the Grubbs, Hampel filter, Dixon, and Rosner tests for outliers. The treatment of outliers is determined by the domain/context of the analysis and the research question, as well as whether the tests to be used are robust to the existence of outliers and how far the outliers are from other observations. This research chose boxplot which is also useful to detect potential outliers in all the continuous variables in the dataset. A boxplot shows five commonly used measures of location such as minimum, maximum, median, 1st and 3rd quartiles, and observations classified as an outlier using the interquartile range (IQR) criterion. In IQR, all observations outside of the following interval will be treated as potential outliers.

$$[q_{0.25} - 1.5IQR; q_{0.75} + 1.5IQR]$$

The IQR criterion classifies observations as potential outliers, which are represented as points in the boxplot. ⁴

⁴ <https://statsandr.com/blog/outliers-detection-in-r/>

3.5.2 Variable transformation

Binary Coding

One of the strategies for increasing the performance of data in a model is variable transformation. In clinical and psychological research, binary coding is frequently employed, either to make interpretation easier or to rule out the possibility of a threshold effect (Liquet & Riou, 2019). Mathematically transformation of a given variable X can be defined as below:

$$X_{(k)} = \begin{cases} 1, & \text{if } X \geq C_k; \\ 0, & \text{if } X < C_k, \end{cases} \quad (1)$$

This technique was used to transform the target variable which is Number of ANC visits into a binary form. The researcher was interested in categorizing the pregnant women who had at least four ANC visits and those who had less than four visits.

Categorical Encoding

The act of turning textual data into numerical values that machine learning approaches can employ to improve model accuracy is known as feature encoding. To turn textual data into numerical values, researchers have utilized a variety of methods, including Label Encoding, One Hot Encoding, and Binary Encoding (Jackson & Agrawal, 2019).

A numerical feature matrix is required as input for many statistical learning techniques. Feature engineering is required when data comprises categorical variables, the reason is to encode the numerous categories into a vector that is suitable (Cerda et al., 2018). In this research one-hot encoding techniques was used to convert string levels of variables in the dataset into numerical to improve the accuracy of our model.

One-hot encoding has been identified as simple technique that is commonly used. Given a categorical variable with categories A, B, and C, the following 3-dimensional feature vectors can be encoded: $[1, 0, 0]$, $[0, 1, 0]$, $[0, 0, 1]$. In the resulting vector space, each category is orthogonal and equally spaced from the others, which is consistent with common understanding of nominal categorical variables (Cerda et al., 2018).

In One-hot encoding, we choose a variable's base level, and then introduce dummy variables for each categorization that isn't the base level. For observations that fall into the related category, a dummy variable equals one, and for those that do not, it equals zero. Assume we have four-level categorical variables $Z : A, B, C, \text{ and } D$. We'll start with A

as the basic level and add dummy variables Z_B , Z_C , and Z_D . The values of these three dummy variables are shown in the table below for each conceivable value of Z .

Table 1 : One-Hot encoding illustration

Z	Z_B	Z_C	Z_D
A	0	0	0
B	1	0	0
C	0	1	0
D	0	0	1

3.5.3 Train-Test Split

It is good practice in a machine learning workflow to split the research data into training and test set. The idea is to test the performance of the trained model by subjecting it to the test set. For decades, the gold standard of machine learning has been the randomized split of training and testing sets (Tan et al., 2021).

This is the most basic way of evaluation, and it's commonly used in Machine Learning applications. The entire dataset (population) is split into two groups: the train set and the test set. Depending on the use case, the data can be divided into 70%-30% or 60%-40%, 75%-25% or 80%-20%, or even 50%-50%. Generally speaking, the ratio of training data to test data must be higher. In this study, the train-test split threshold was 70%-30%.

3.5.4 Feature Importance/Selection using Random Forest

We often have hundreds, if not millions, of features in a dataset, and we want a technique to develop a model that only incorporates the most significant ones. Feature selection is the term for this procedure. Random Forests are frequently used in data science workflows to choose features. The rationale for this is that in random forests, tree-based techniques are naturally ordered by how efficiently they increase node purity. There are a number of advantages to this. For one, we make the model easier to understand. Second, this can minimize the model's variance, resulting in over fitting. Finally, it can lower the computational cost (and time) of model training. Variable importance technique of Random Forests is commonly used to rank variables according to their importance in predicting the outcome variable both in a classification task and a regression task, hence reduces the number of features used to train the model which ideally enhances computing efficiency (Behnamian et al., 2017).

Random forest uses implicit feature selection as a classifier, using only a small group of "strong variables" for classification, as a result, improved performance on high-dimensional data is achieved. The "Gini importance" can be used to illustrate the result of the random forest's implicit feature selection, and it can be used as a broad indicator of the importance of a feature.

To discover the ideal split at each node within the binary trees T of the random forest, the Gini impurity $i(\tau)$ is utilized, which is an effective assessment of the entropy that reflects how effectively a prospective split divides the samples of the two classes in a given node τ .

With $P_K = \frac{n_k}{n}$ representing proportion of the n_k samples, the Gini Impurity can be determined with class $k = \{0, 1\}$ out of a total of n samples at the node τ as:

$$i(\tau) = 1 - P_1^2 - P_0^2 \quad (2)$$

Decrease denoted Δ_i as a result of separating the samples and sending them to two sub-nodes τ_l and τ_r by a threshold t_θ on variable θ is given by (with the corresponding sample proportions $P_l = \frac{n_l}{n}$ and $P_r = \frac{n_r}{n}$)

$$\Delta i(\tau) = i(\tau) - P_l i(\tau_l) - P_r i(\tau_r) \quad (3)$$

The reduction in Gini impurity as a result of the optimal split $\Delta i_\theta(\tau, T)$ is recorded and summed up for all nodes τ in all trees T in the forest, individually for all θ :

$$I_G(\theta) = \sum_T \sum_\tau \Delta i_\theta(\tau, T) \quad (4)$$

This score represents the frequency with which a given characteristic was chosen for a split and the magnitude of its overall discriminative value for the classification task in question.

3.5.5 Correlation

Correlation is a statistical technique which measures the strength and nature (positive or negative) of association between two variables. In correlation when one variable changes the other variable also changes, in other words a change in one variable causes a change in the other variable. This researcher used a correlation matrix to investigate the variables which are highly correlated in the dataset. A correlation matrix comes in form of a Table, and it displays the values of coefficients of correlation between variables. Hence the

value of correlation between two variables is displayed in each cell of the table. Pearson correlation technique was used:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2} \quad (5)$$

Where r is the Pearson product moment correlation coefficient, x_i , y_i are values of the first and second variables in the dataset respectively. \bar{x} and \bar{y} are the means of first and second variables in the dataset respectively.

3.6 Machine Learning Models Used

3.6.1 Logistic Regression

Logistic regression, which is a generalized version of linear model (described in this research as GLM), is a model that combines a set of predictor variables to estimate the likelihood of a specific event occurring. Logistic Regression Model is now a vital part of every statistician's arsenal due to its simplicity and effectiveness (Bertsimas & King, 2017). It calculates the chances of a subject belonging to a particular category of interest. Because of its high computing speed and the output of a model that lends itself to rapid data scoring, it is a popular machine learning method. The model is based on the chance that a label equals 1. i.e.

$$P = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_q X_q \quad (6)$$

The predictors are given a logistic response function, which we use to model P.

This guarantees that P is between 0 and 1. We use odds ratio, which is basically the ratio of success (1) to non-success, to generate exponential expression of the denominator (0). The odds ratio is the chance of an event divided by the probability that it will not occur.

$$\text{Odds ratio } (Y = 1) = \frac{P}{1 - P} \quad (7)$$

$$P = \frac{\text{Odds}}{1 + \text{Odds}} \quad (8)$$

By combining with the logistic response function, we obtain

$$\text{Odds}(Y = 1) = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_q X_q} \quad (9)$$

Taking logarithm,

$$\log(\text{Odds}(Y = 1)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_q X_q \quad (10)$$

For an odds ratio whose value is between 0 and 1, it is interpreted by saying that the outcome of interest is $100 \times (1 - \text{Odds Ratio})$ % less likely to occur for every unit increase in the predictor. If the value of an odds ratio is a value greater than 1 but less than 2, then it is interpreted by saying that the outcome of interest is $100 \times (\text{Odds Ratio} - 1)$ % more likely to occur for every unit increase in the predictor. For an odds ratio with value greater than 2, it implies that the outcome of interest is Odds Ratio times more likely to occur for every unit increase in the predictor. All this is applicable if the predictor variable is continuous.

For a binary category predictor, we begin by establishing a dummy variable for each of the two categories after picking one as a reference. For an odds ratio between 0 and 1, the result of interest is 100 percent less likely to occur for individuals in the category with the value 1 compared to those in the reference category. For an odds ratio larger than 1 but less than 2, the result of interest is 100 percent more likely to occur for those in the category assigned the value 1 than those in the reference category. The outcome of interest is Odds Ratio times more likely to occur when the odds ratio is more than 2 than when the odds ratio is less than 2.

3.6.2 Random Forest

Random Forest (RF) is a supervised tree-based classification model. It's a decision tree-based ensemble classification algorithm that generates a series of decision trees from which the best estimate is selected. Even when predictive characteristics incorporate irrelevant variables, RF has shown to perform well in both classification and regression task (Ding et al., 2015). It makes multiple trees, one for each feature in the training data. All of the decision trees' outputs are combined, and the best estimator is used (Ambikavathi et al., 2020).

Random forest's adaptability is one of its most enticing features. It may be used to solve both regression and classification problems, and the relative value of the input characteristics is clearly displayed. It is implemented by bagging decision trees and sampling variables in addition to records, as is the case with decision trees. The algorithm chooses a variable and split point by minimizing a criterion such as Gini impurity while producing sub-partitions of a partition M . It's worth mentioning that the variables available are limited to a random subset. In contrast to decision trees, the technique includes two additional steps: bagging and bootstrap sampling of variables. The random forest training algorithm employs the technique of bootstrap aggregation (bagging).

Given a training set $X = x_1, x_2, \dots, x_n$

with response variable $Y = y_1, y_2, \dots, y_n$

bagging repeatedly (B times) selects a sample at random with replacement of the training dataset and fits trees to these selected samples.

For $b = 1, 2, \dots, B$ By Taking sample, with replacement, n training examples from X, Y ; call these X_b, Y_b

The model trains a classification or regression tree f_b on X_b, Y_b

To obtain predictions for unknown samples x' , the algorithm combine the predictions from all the individual regression trees on x' or take the majority vote in the case of classification trees after training.

$$\hat{f} = \frac{1}{B} \sum_{b=1}^B f_b(x') \quad (11)$$

This bootstrapping strategy enhances model performance because it minimizes model variance without increasing bias.

3.6.3 Support Vector Machine (SVM)

The Support Vector Machine (SVM) technique creates a flat border called a hyper plane that produces a fair homogeneous split of data on both sides. This approach combines components of instance-based closest neighbor learning, such as K-Nearest Neighbor (KNN), with a regression model, resulting in a very strong algorithm that is used to model high-complexity connections in data. It can be used for both regression and classification problems.

Assuming the training data are linearly separable, we can select two parallel hyper planes with the greatest practical distance between them to divide the two groups of data. The maximum margin hyper plane is the hyper plane that is located halfway between these two hyper planes, while the margin hyper plane is located exactly halfway between them. These hyper planes can be described by the equations below:

$$\vec{W} \cdot \vec{x} - b = 1 \text{ and} \quad (12)$$

$$\vec{W} \cdot \vec{x} - b = -1 \quad (13)$$

Ideally the distance between these two hyper planes is equal

$\frac{2}{\|\vec{W}\|}$ We wish to minimize $\|\vec{W}\|$ to increase the distance between the planes. Hence, we can add the following constraint to each i either to avoid data points slipping into the margin.

$$\vec{W} \cdot \vec{x}_i - b \geq 1 \text{ if } y_i = 1 \text{ or} \quad (14)$$

$$\vec{W} \cdot \vec{x}_i - b \leq -1 \text{ if } y_i = -1 \quad (15)$$

Each data point must be on the correct side of the margin, according to these requirements.

This can be presented mathematically as:

$$y_i(\vec{W} \cdot \vec{x}_i - b) \geq 1 \text{ for all } 1 \leq i \leq n \quad (16)$$

Putting this together, we can obtain optimization problem:

Minimize $\|W\|$ subject to $y_i(\vec{W} \cdot \vec{x}_i - b) \geq 1$ for $i = 1, \dots, n$

Where \vec{W} and b that solve this problem determine classifier $\vec{x} \mapsto \text{sign}(\vec{W} \cdot \vec{x} - b)$

3.6.4 Artificial Neural Network (ANN).

Artificial Neural Networks are designed to mimic neural networks in the human brain and have a structure that is similar to that of biological neural networks. In other words, Neural networks are a class of algorithms that make an effort to identify patterns, correlations and information in data by employing a method that is modeled after and operates similarly to the biology and human brain. The inputs/information from the outside world that the model utilizes to learn and generate conclusions are included in the input layer, also known as the input nodes. Using input nodes, the data is transmitted to the hidden layer, the following layer. A collection of neurons in the hidden layer performs all calculations on the input data. A neural network can use any number of hidden layers. The most fundamental network, known as the perceptron, consists of a single hidden layer. The output layer contains all of the model's results and conclusions from calculations. The output layer may contain one or more nodes. A multilayer neural network, sometimes referred to as an artificial neural network or multilayer perceptron, is made up of several neurons that are put together.

In the first stage, the input units are passed. Since each hidden layer is made up of neurons and each neuron is connected to all the inputs, data is transmitted to the hidden layers with certain weights attached. Following the transmission of the inputs, all calculations are completed in the hidden layer. In hidden layers, the computation is divided into two steps. All of the inputs are first multiplied by the corresponding weights. The weight of each variable is its gradient or coefficient. It illustrates the effectiveness of a particular input. After the weights have been assigned, a constant called bias is added. This constant called bias helps the model fit as accurately as it can. Utilization of the activation function happens in the second phase. The activation function corrects the input nonlinearly before it is sent on to the next layer of neurons. Therefore, for the model to be nonlinear, the activation function is essential. The process moves on to the final layer, the

output layer, where the final result is obtained, after traversing through all of the hidden levels, this entire process is called Forward Propagation. The error is then identified as the discrepancy between the actual and expected output after gathering the predictions from the output layer. Back propagation is a technique used to eliminate big or significant errors. Back propagation is the process of updating and choosing the best weights or coefficients for a model in order to minimize the error, or the discrepancy between actual and predicted values.

The process can be demonstrated mathematically as follows;

Weights w_1, w_2, \dots, w_n , and thresholds for parameters and α are set to random integers. Perceptron is active by feeding it inputs $x_1(p), x_2(p), \dots, x_n(p)$ and the desired output $Y_d(p)$, where iteration p refers to the perceptron's p^{th} training example ($p = 1, 2, \dots$). At iteration $p = 1$, the process calculates the actual output $Y(p)$.

$$Y(p) = step \left[\sum_{i=1}^n X_i(p) W_i(p) - \theta \right], \quad (17)$$

i.e.

$$Y(X) = step[X] = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases}$$

In this case step represent step activation function, and n represent number of perceptron inputs.

The weights of the perceptron are then updated.

$$W_i(p+1) = W_i(p) + \Delta W_i(p)$$

Where $\Delta W_i(p)$ which is calculated by delta rule, represent weight correction at iteration p

The delta rule is given by: $\Delta W_i(p) = \alpha \times X_i(p)(p)$

Where α represent learning rate, $\alpha \subseteq (0, 1]$; $e(p) = Y_d(p) - Y(p)$, where $p = 1, 2, 3, \dots$

Iteration p is increased by 1 each time until a convergence is achieved i.e., until error $e = 0$

3.7 Improving Model Performance

3.7.1 Cross Validation

K-fold Cross-Validation

In this technique, observations are divided randomly into k groups known as folds of equal sizes. The first set is used to test the model performance while the remaining $k-1$ folds are used to train the model. MSE_i is then calculated on the observations that fall in the test set in the case of a regression problem. This procedure is repeatedly performed k times, each time using a new set of observations as a validation/test set. In the case of regression, the k -fold CV estimate is calculated by finding the average of these values. When it comes to regression, this is given by.

$$CV_k = \frac{1}{k} \sum_{k=1}^k MSE_i \quad (18)$$

For the case of classification.

$$Training\ error\ rate = \frac{1}{n} \sum_{i=1}^n I(y_i \neq \hat{y}_i) \quad (19)$$

Where \hat{y}_i is actually the predicted class label for the i^{th} observation and $I(y_i \neq \hat{y}_i)$ denotes indicator variable which is equals to 1 if $y_i \neq \hat{y}_i$ and 0 if $y_i = \hat{y}_i$ If $I(y_i \neq \hat{y}_i) = 0$ then i^{th} observation was classified correctly otherwise misclassified. Hence the equation calculates the proportion of incorrect classification.

In practice, researchers using k -fold CV usually set $k = 5$ or $k = 10$ for ease of computations. Since this method is rooted in re-sampling, clearly, it prevents the problem of under fitting and over fitting and hence improves the accuracy of the model (Nordhausen, 2009).

3.8 Model Evaluation Metrics

3.8.1 Confusion matrix

Figure 2 depicts the confusion matrix in a two-class classification task. Four distinct result predictions can be made, as illustrated in the diagram. True positive and true negative outcomes are accurate classifications, but false positive and false negative outcomes are errors. False positive examples are negative instances of classes that have been classified wrongly as positive, whereas false negative examples are positive examples of classes

that have been classified incorrectly as negative. In the context of this investigation, the following definitions apply to the entrance to confusion matrix.

Table 2: Confusion Matrix illustration

		Predicted class	
		Negative	Positive
Actual class	Negative	a	b
	Positive	c	d

a This is the number of correct predictions in cases where the outcome is negative, True Negative (TN)

b This is the number of incorrect predictions that cases are positive, False Positive (FP)

c This is the number of incorrect predictions that cases are negative, False Negative (FN)

d This is the number of correct predictions that cases are positive, True Positive (TP)

In this context, accuracy refers to the percentage of true outcomes (including true positives and true negatives) in relation to the total number of cases considered.

The following equation can be used to determine accuracy.

$$Accuracy = \frac{a+b}{a+b+c+d} = \frac{TN+FP}{TN+FP+FN+TP}$$

The fraction of correctly categorized positive instances is known as the true positive rate:

$$True\ Positive\ Rate\ (Recall\ or\ Sensitivity) =$$

$$\frac{d}{c+d} = \frac{TP}{FN+TP}$$

The false positive rate is the percentage of cases that were wrongly labelled as positive when they were actually negative:

$$False\ Positive\ Rate = \frac{b}{a+b} = \frac{FP}{TN+FP}$$

The real negative rate was defined as the proportion of accurately diagnosed negative cases:

$$True\ Negative\ Rate = \frac{a}{a+b} = \frac{TN}{TN+FP} \quad (20)$$

The proportion of positive instances that were wrongly labelled as negative is known as the false negative rate:

$$\text{False Negative Rate} = \frac{c}{c+d} = \frac{FN}{FN+TP} \quad (21)$$

Finally, precision, also known as positive predictive value, shows the percentage of accurate predictive positive cases:

$$\text{Precision} = \frac{d}{b+d} = \frac{TP}{FP+TP} \quad (22)$$

3.8.2 ROC Curve

Another way to assess the performance of a classifier is to use a ROC graph. In the ROC graph, the false positive rate plotted on the X axis, while the true positive rate is plotted on the Y axis. The ROC curve (Novakovi et al., 2017) is a visual tool for evaluating a classifier's ability to identify erroneously classified positive and negative instances.

4 CHAPTER FOUR: RESULTS

4.1 Data Gathering

There were 20,964 observations and 1,130 variables in the dataset. Feature selection approach was used to ensure that only the variables of interest remained for analysis. The first step was to identify all variables with 100 percent missing values, indicating that they lacked information about the target variable. 23.45 percent ($n=265$) of the 1,130 variables had 100% missing values. This task was accomplished using the R programming language's '*DataExplorer*' package, which profiles variables based on the number and percentage of missing data. As a consequence, only 865 features remained to be investigated further.

A threshold of 50% missing observations was then set, so any variables with more than 50% missing values were eliminated from the dataset. As a result, a total of 509 variables were discovered to have more than 50% missing data, and were thus eliminated from further research, leaving behind only 356 variables. These were screened again, and only those factors relevant to mother and child health, namely those variables associated with Antenatal care visits according to the literature review, were chosen for further analysis. As a result, the final study only included 17 features. Using Excel functions, the selected variables were cleaned up even more including variable renaming to display real variable names. The final selected variables are shown in Table 3 below:

Table 3: Selected Features used in the dataset

Feature	Description
Respondent current age	Continuous; The current age of Respondent
Type of place of Residence	Nominal; Either Urban or Rural
Region	Nominal; Coast, Central, Nyanza, Rift Valley, Eastern, Western, Nairobi and North Eastern
Highest Education Level	Ordinal; Higher, Secondary, Primary and No education
Religion	Nominal; Christian, Muslim, No religion, Roman Catholic, Protestants, Other.
Ethnicity	Nominal; All ethnic groups in Kenya such as Kikuyu, Kalenjin, Kisii, Luhya, Boran, Iteso, Mbere, Orma, Rendile, Taveta/Taita, Luo, Embu, Meru, Turkana, Samburu, Somali, Pokomo, Mijikenda, Maasai, Kuria, Kamba, Gabra, Other.
Age of household head	Continuous; Age of head of household, ranging between 15-99
Frequency of reading newspaper	Nominal; At least once a week, less than once a week, not at all
Frequency of listening to radio	Nominal; At least once a week, less than once a week, not at all
Frequency of watching television	Nominal; At least once a week, less than once a week, not at all
Wealth index	Nominal; Middle, Poorer, Poorest, Richer, Richest
Total children ever born	Continuous; Total number of children respondent is having
Age of respondent at first birth	Continuous; Age of respondent at time of first delivery
Current marital status	Nominal; Married, Separated, Widowed, never in union, living with partner, Divorced

Source: Kenya Demographic Health Survey, 2014

4.2 Exploratory Data Analysis (EDA)

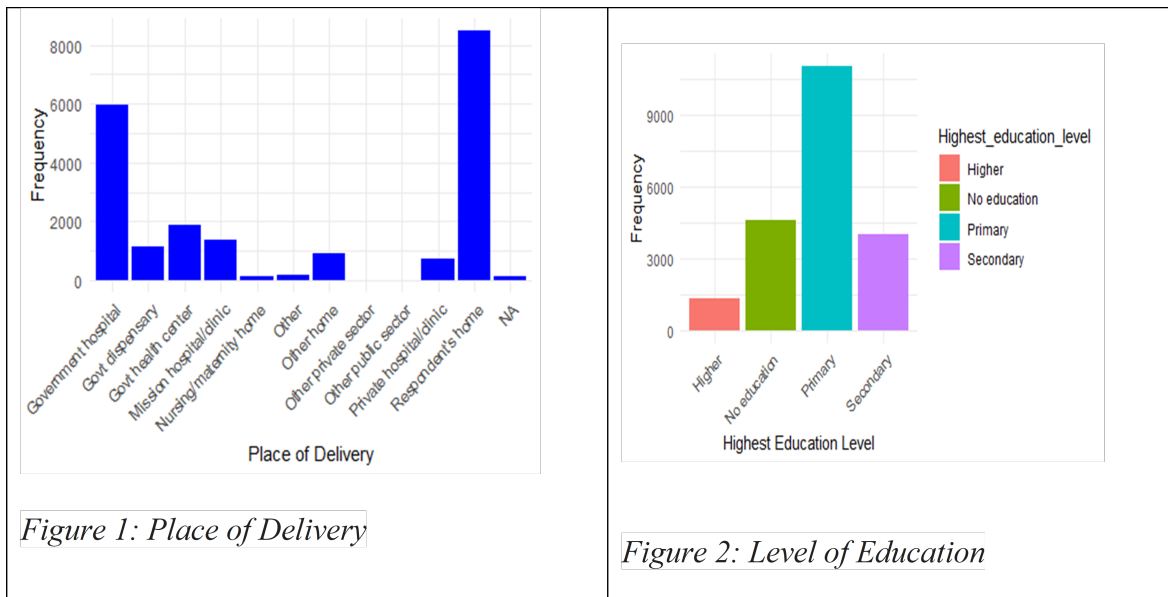
4.2.1 Participants Characteristics

Table 4: Participants Characteristics

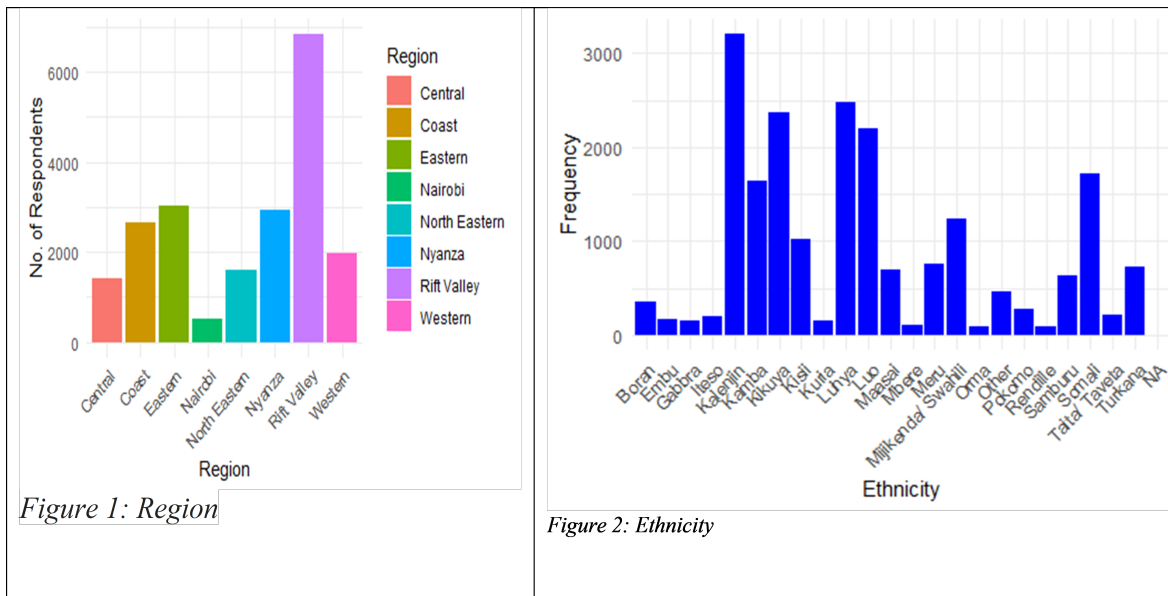
Variable	Category	Frequency	%
Type of place of residence	Rural	14136	67.4%
	Urban	6828	32.6%
	Total	20964	100.0%
Highest Education level	Higher	1321	6.3%
	No education	4585	21.9%
	Primary	11055	52.7%
	Secondary	4003	19.1%
	Total	20964	100.0%
Frequency of reading newspaper	At least once a week	2024	9.7%
	Less than once a week	3101	14.8%
	Not at all	15827	75.5%
	Total	20964	100.0%
Frequency of listening to radio	At least once a week	12291	58.6%
	Less than once a week	2610	12.4%
	Not at all	6056	28.9%
	Total	20964	100.0%
Frequency of watching TV	At least once a week	4844	23.1%
	Less than once a week	2429	11.6%
	Not at all	13677	65.2%
	Total	20964	100.0%
Wealth index	Middle	3497	16.7%
	Poorer	4348	20.7%
	Poorest	7178	34.2%
	Richer	3131	14.9%
	Richest	2810	13.4%
	Total	20964	100.0%

Source: Kenya Demographic Health Survey, 2014

The study included a total of 20,964 women who had experienced at least one delivery event. According to the survey results, most of the study participants were from rural areas, accounting for 67.4% ($n=14136$), while those from urban areas, accounted for 32.6% ($n=6828$). Most of the participants in this survey, 52.7% ($n=11055$), only received primary education and just 6.3% ($n=1321$) completed higher education, indicating that the prevalence of illiteracy is still relatively high among Kenyan women. The findings also revealed that Kenya's high poverty rate remains a problem, as 34.2% ($n=7178$) of the participants fell into the poorest category, with only 13.4% ($n=2810$) claiming to be the wealthiest. The fact that most of the participants only had access to radio as their source of information, with 58.6% ($n=12291$) saying they listen to radio at least once a week, supports the findings of high rate of poverty in Kenya. *See Table 4 above.*



The findings further revealed that most deliveries still take place at home, followed by those that occur in government facilities. See Fig. 2 for more information. The majority of the participants, 52.7% (n=11055) had Primary education, the study reveals. Those who were illiterate formed 21.9% (n=4585) of the total participants. See Fig 3 for more information.



Regionally, Rift Valley region, 32.7% ($n=6850$) had the biggest share of participants, followed by Nyanza 14.4% ($n=3015$) and Eastern, 14.0% ($n=2926$). The same information is represented in *Figure 4*. The largest ethnic groups in Kenya, such as the Kalenjins, 15.3% ($n=3205$), Kikuyus, 11.8% ($n=2482$) Luhyas, 11.3% ($n=2366$), and Luos, 10.5% ($n=2195$) formed the largest proportion of research participants. See *Fig. 5* for more information

4.2.2 Univariate Analysis

From the univariate analysis, it is observed that 6.2% ($n=924$) study participants did not utilize antenatal care, recording counts of 0 during the survey year. This implies that about 6% of women aged 15 to 49 who had a live delivery five years preceding the date of the 2014 KDHS did not receive antenatal care in Kenya. 39.5 percent ($n=5881$) of the participants made 1-3 antenatal care visits during their pregnancy while 48.1% ($n=7168$) made 4-6 ANC visits, 5.6% ($n=839$) between 7-9 and 0.6% ($n=86$) 10+ visits. From the findings ANC coverage is 54.30% ($n=8093$) for the 2014 KDHS that is, those pregnant women who made at least four visits / had 4+ contacts with health professional which is the recommended number of ANC visits by WHO. It's worth noting that 93.8% of pregnant women had at least one encounter with a health care provider. According to WHO, the first ANC visit is very crucial since it provides an opportunity for the health workers to differentiate between pregnant mothers who require standard care (only four visits) those who require special attention (more than four visits).

Table 5: Analysis of Antenatal Care visits patterns

Number of antenatal visits during pregnancy	Frequency (n)	%
No visit (0)	924	6.2
1-3	5881	39.5
4-6	7168	48.1
7-9	839	5.6
10+	86	0.6

Number of antenatal visits during pregnancy	Frequency (n)	%
Less than 4 (<4)	6805	45.7
More than/equals four (4+)	8093	54.3
Total	14898	100

Source: Kenya Demographic and Health Survey, 2014

Participants recruited in the study were aged between 15-49 years with an average age of $M=28.73$ years ($SD=6.56$) while the average age at first birth was found to be $M=19.31$ years ($SD=3.55$). To recognize and manage pregnancy problems, the timing and frequency

of ANC visits are critical. Hence according to WHO, ANC should begin in the first trimester for pregnant women (i.e., first three months of pregnancy). However, many women from Kenya do not utilize ANC according to this guideline as revealed in *Table 6* which eventually could contribute to under-five mortality. The results indicates that women in Kenya begin their first antenatal checkups at the end of the fourth month, $M=4.89$ months ($SD=1.54$) with average number of ANC visits, $M=3.74$ ($SD=1.86$) which is about 4 recommended number. See *Table 6*. The same is presented in *Fig 5-8*.

Table 6: Descriptive Analysis of selected continuous Variables

Variable	Mean	SD	Minimum	Maximum
Respondent current Age	28.73	6.56	15	49
Age of Respondent at first Birth	19.31	3.55	6	44
Number of Antenatal Care Visits during pregnancy	3.74	1.86	0	20
Timing of first Antenatal Check (Months)	4.89	1.54	0	9

Source: Kenya Demographic Health Survey, 2014

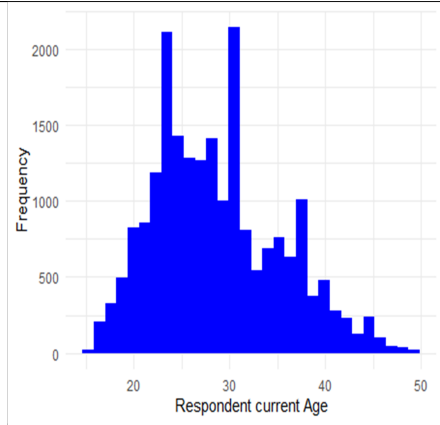


Figure 1: distribution of the Respondent current Age

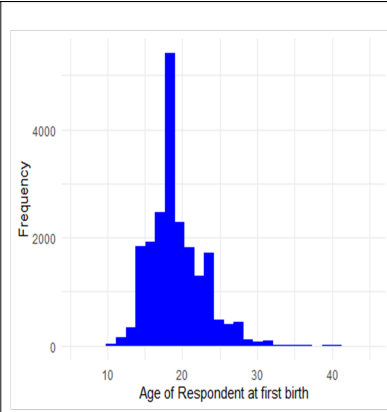


Figure 2: distribution of Respondent age at first birth

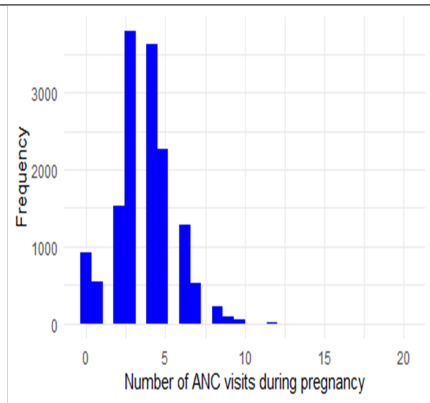


Figure 3: distribution of the number of ANC visits

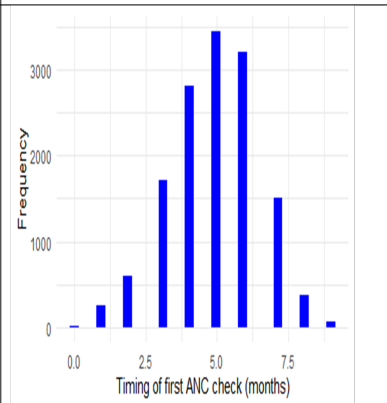


Figure 4: distribution of timing of first ANC checkup (in months)

4.2.3 Bivariate Analysis

Region	Mean	SD
Central Region	4.14	1.87
Coast Region	4.00	1.78
Eastern Region	3.74	1.75
Nairobi Region	4.86	2.17
North-eastern	2.64	2.09
Nyanza Region	3.9	1.63
Rift Valley Reg.	3.58	1.88
Western Region	3.78	1.67
Overall Mean	3.83	0.1915

Table 1: ANC utilization per region

Highest Education Level	Mean	SD
Higher	5.04	1.85
No Education	2.96	2.07
Primary	3.69	1.68
Secondary	4.12	1.73

Table 2: ANC utilization with respect to level of education

Source: Kenya Demographic Health Survey, 2014

Table 7 shows that ANC use is high in the Nairobi region, with $M=4.86$ ($SD=2.17$). This could be attributable to the fact that this region has a large number of health-care institutions. This is followed by the central region, $M=4.14$ ($SD=1.87$), while Northeastern region recorded the least ANC utilization, $M=2.64$ ($SD=2.09$). Women with a higher education, have higher ANC usage, $M=5.04$ ($SD=1.85$) than women without a higher education, $M=2.96$ ($SD=2.07$). Table 9 shows that ANC usage is high in metropolitan areas, $M=4.16$ ($SD=1.84$).

Type of Residence place	Mean	SD
Rural	3.52	1.82
Urban	4.16	1.84

Table 1: ANC utilization according to type of Residence

Source: Kenya Demographic Health Survey, 2014

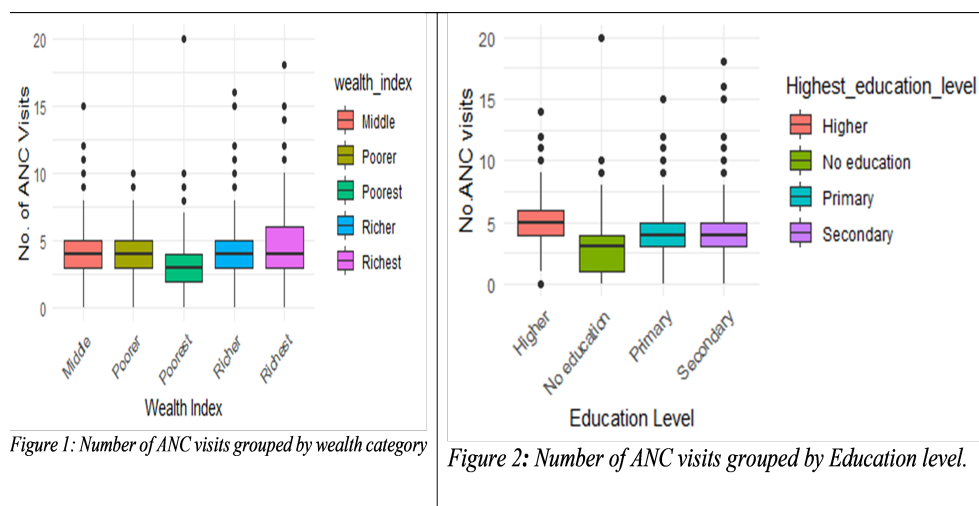
Wealth Index	Mean
Middle	3.84
Poorer	3.61
Poorest	3.11
Richer	4.07
Richest	4.70

Table 2: ANC utilization according to wealth index

Source: Kenya Demographic Health Survey, 2014

Those participants in the Richest category were found to have the maximum number of ANC visits on average, $M=4.70$ ($SD=1.91$). The poorest category recorded the lowest ANC utilization on average, $M=3.11$ ($SD=1.91$), see Table 10. This is a clear indication that

women with poor wealth index are more at risk of not having access to ANC services compared to those who are rich. It was further revealed that women with low education are at higher risk of not utilizing ANC services compared to women who have education. The information is further presented using a boxplot in *Fig 10-11*.



Wealth Index	Mean	SD
Middle	4.9266	1.4953
Poorer	5.0567	1.5169
Poorest	5.1137	1.4939
Richer	4.7908	1.4923
Richest	4.3824	1.6099

Education	Mean	SD
Higher	4.04087	1.5518
No education	5.01744	1.5518
Primary	5.04313	1.5021
Secondary	4.74809	1.5000

Place of Residence	Mean	SD
Rural	5.00879	1.5009
Urban	4.68787	1.5831

Source: Kenya Demographic Health Survey, 2014

Table 11: Timing of first ANC Checkup versus other demographic factors

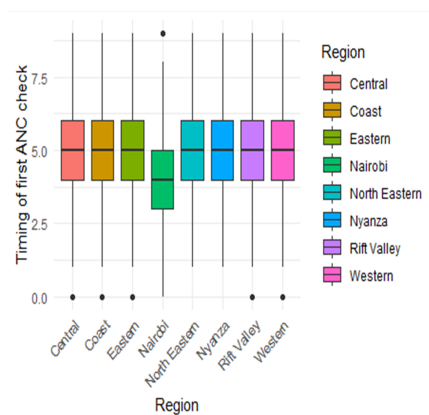


Figure 1: distribution of Timing of first ANC checkup per region

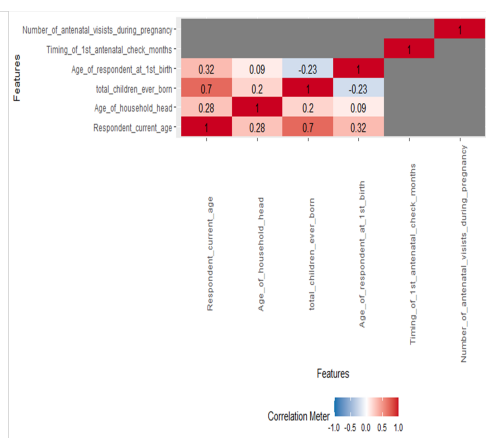


Figure 2: Correlation Analysis for the continuous variables

Early inclusion in ANC has been shown in previous research to allow health care providers to provide timely information and services based on the mother's gestational age and health state. Pregnant women who arrive late for ANC, on the other hand, miss out on important health information and interventions including early HIV testing, malaria prevention, and anemia prophylaxis, as well as the prevention and management of complications. According to the findings of this study, Nairobi women begin ANC visits earlier than women in other locations. Take a look at *Figure 12*. Pregnant women should begin getting ANC services in the first trimester of their pregnancy, or during the first three months, according to the WHO. *Table 11* shows that women from low-income backgrounds are more likely to seek ANC services later than the recommended period, $M=5.11$ ($SD=1.49$), compared to those from higher-income backgrounds, $M=4.38$ ($SD=1.61$). This clearly demonstrates that a pregnant woman's ability to obtain early ANC services is influenced by her finances. Further findings indicates that women without education begin seeking ANC services later, $M=5.02$ months ($SD=1.55$), than women with higher education, $M=4.04$ months ($SD=1.55$). Rural inhabitants begin ANC care later, $M=5.01$ ($SD=1.50$), than women in urban areas, $M=4.69$ ($SD=1.58$), according to the survey.

4.2.4 Correlation

The continuous features that are highly correlated to the target variable were examined using a correlation matrix in this study. See *Fig 13*. The coefficients of correlation between variables are displayed in a correlation matrix where each table cell displays the correlation between two variables. Although there exist correlation between these variables, the results shows the relation is not strong.

4.2.5 Missing Value Analysis

Table 12: Missing value Analysis

Feature	No. missing (<i>n</i>)	% missing
Respondent current age	0	0
Region	0	0
Type of place of Residence	0	0
Highest Education Level	0	0
Religion	38	0.181
Ethnicity	5	0.0239
Age of household head	0	0
Reading newspaper	12	0.0572
Listening to radio	7	0.0334
Frequency of watching television	14	0.0668
Wealth index	0	0
Total children ever born	0	0
Age of respondent at first birth	0	0
Current marital status	0	0
Timing of first antenatal check in months	6968	33.2
Number of ANC visits during pregnancy	6066	28.9
Place of Delivery	114	0.544

Source: Kenya Demographic Health Survey, 2014

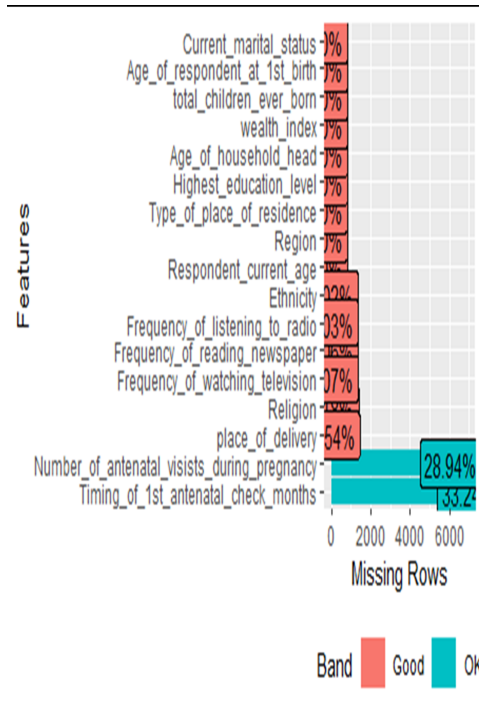


Figure 1: Graphical Missing Value Analysis

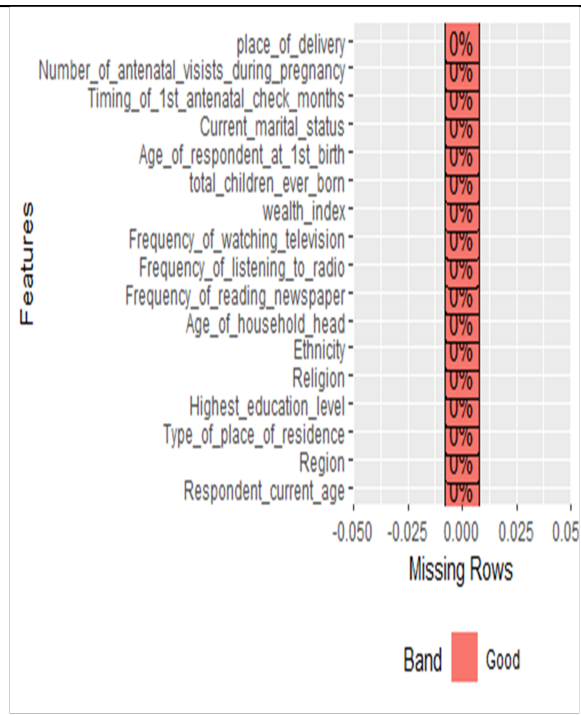


Figure 2: Information of Missing values after imputation

From the findings above (See Fig 14), it is observed that *Religion* has only 0.181% ($n=38$) missing values, *Age of household head*, 0.0239% ($n=5$), *Frequency of reading newspaper*, 0.0572% ($n=12$), *Frequency of watching television* has 0.0668% ($n=14$), *Frequency of listening to radio* has 0.0334% ($n=7$). The results show that the variable *Timing of first antenatal check (in months)* has the highest number of missing values 33.2% ($n=6968$) followed by *Number of antenatal visits during pregnancy* 28.9% ($n=6066$) which is the variable of interest. Lastly, *place of delivery* also has a significant missing value, 0.544% ($n=114$). This was achieved by use of 'DataExplorer' package in R.

4.3 Data pre-processing

4.3.1 Feature Engineering

Handling missing Values

This study employed an imputation method to deal with missing values in the dataset. In categorical data, missing values were replaced with the most common category, and missing values in continuous variables were replaced with the mean. *Hmisc* package in R programming Language was used to accomplish this task. The results is presented in *Figure 15*.

Variable Transformation

Because WHO advises at least four ANC visits throughout pregnancy, this was chosen as the cutoff point for a more accurate prediction model. As a result, the variable Number of Antenatal Visits during pregnancy was transformed to binary, with all women who made at least four visits falling into the positive group and those who had fewer visits falling into the negative category (0). As a result, our model shifted from regression to classification. As a result, the dataset's variable *Number of ANC visits throughout pregnancy* was removed from the dataset. This resulted in 67.5% ($n=14159$) of pregnant women who made at least four ANC visits and 32.5% ($n=6805$) who made less than four visits. This is a reasonably well-balanced dataset.

Handling Outliers

This study used boxplot, which is excellent for detecting outliers. All the observations which were found to be potential outliers as demonstrated in *Fig 16* were all replaced with the corresponding mean of the variable under consideration. This was performed in all the continuous variables in the dataset which were found to contain outliers such as *Respondent current age, Age of household head, total children ever born, Age of respondent at first birth* and *Timing of first antenatal visits*.

Categorical Encoding

For purposes of improving the performance of the fitted model all the non-numeric variables were converted to numeric through one-hot encoding. This resulted in 79 variables. A numerical feature matrix is required as input for many machine learning algorithm techniques.

Feature Selection using Random Forest

Random Forests typically uses variable importance measures to rank variables according to their relevance to a classification task, thereby reducing the number of model inputs in high-dimensional data sets hence boosting processing performance. Gini Index, is used to determine the relevance of variables. The score indicates how frequently a specific attribute was chosen for a split and how important its overall discriminative value was for the classification task at hand. According to the *Fig 17*, Timing of first antenatal checkup in months emerged to be the best predictor of ANC utilization among women in Kenya. This is followed by age of the household head, respondent current age, age of respondent at first birth, total number of children ever born, place of delivery, religion, highest level of education, frequency of listening to radio and wealth index in that order.

4.4 Binary Logistic Regression Model

Using the KDHS 2014 dataset, the first phase in the model training method was to construct a multiple binary logistic regression with the goal of determining the factors that significantly influence ANC utilization in Kenya. The target variable was the number of Antenatal Care (ANC) visits, which was divided into two categories: at least four (≥ 4) and less than four (< 4) visits. The predictors were chosen based on the results from the Feature selection Process using Random Forest. As a result, the following variables were considered in fitting the model; *Timing of first ANC check (months)*, *Total children ever born*, *wealth index*, *Highest education level*, *Respondent current age*, *place of delivery* and *current marital status*. The result is presented in *Table 13*

Table 13: Logistic Regression output

Variable	OR	LCL	UCL	P-value
(Intercept)	4009.679	2811.314	5749.233	< 2e-16
Timing of first antenatal check (months)	0.23322	0.220461	0.246446	< 2e-16
Total children ever born	1.096978	1.067793	1.127061	1.85E-11
Wealth index (Poorest)	0.795276	0.721226	0.876956	4.36E-06
Highest education level (Higher)	1.897003	1.538258	2.352148	3.36E-09
Highest education level (No education)	0.765268	0.686131	0.85358	1.57E-06
Respondent current age	0.979408	0.971074	0.987823	1.84E-06
place of delivery (Mission hospital clinic)	1.210927	1.014874	1.448481	0.0349
Current marital status (Married)	1.537943	1.394325	1.696049	< 2e-16

Source: Kenya Demographic Health Survey, 2014

High-quality, timely prenatal care (ANC) is an essential part of efforts to enhance the health of mothers and newborn children. Antenatal consultations help to diagnose and address pregnancy concerns early on. The study sought to find out how timing of the first antenatal check influences the number of ANC visits. The results were significant, $OR=0.23322$, $p = 0.000$, $95\%CI (0.220461, 0.246446)$. According to the findings of this research, for every unit increase in the number of months during pregnancy, a woman is 76.7% less likely to make at least four prenatal appointments. This is a clear indicator that scheduling the first prenatal check early in the pregnancy increases the likelihood of multiple visits during the pregnancy. The number of ANC visits is directly proportional to the total number of children ever born, $OR=1.096978$, $p = 0.000$, $95\% CI (1.067793, 1.127061)$. With each successive child, a woman's chances of having at least four ANC visits increase by 9.7%, demonstrating that women who have previously interacted with doctors are better informed and more willing to have several prenatal visits during their pregnancy. The findings also showed that poorest wealth index is a significant predictor of ANC visits, $OR= 0.795276$, $p = 0.000$ $95\% CI (0.721226, 0.876956)$. According to the findings, a woman in the poorest category is 20.5% less likely to visit at least four times during her pregnancy, indicating that poverty makes it difficult for women to get antenatal care services. The study discovered that pregnant women with a higher level of education had a greater number of ANC visits during their pregnancy, $OR = 1.897003$, $p=0.000$, $95\% CI (1.538258, 2.352148)$. A pregnant woman with a higher education is 89.7% more likely than one without to attend at least four ANC visits. No education showed a significant effect as well, with an OR of 0.765268 , $p=0.000$, and a $95\% CI (0.686131, 0.85358)$.

According to the findings, a woman with no education is 23.5% less likely to have at least four ANC visits. Respondent current age, $OR= 0.979408$, $p=0.000$, 95% CI (0.971074, 0.987823) was also observed to influence ANC visits. For every unit increase in age, a pregnant woman is 2.1% less likely to make at least four ANC visits. Delivery in Hospital clinic was also found to have a significant association with the number of ANC visits, $OR=1.210927$, $p=0.0349$, 95% CI (1.014874, 1.448481). A pregnant woman who delivers in a Hospital clinic facility is 21.1% more likely to make 4+ ANC visits during her pregnancy compared to one delivers at home, according to the findings. Finally, marital status was found to have a significant impact on the number of ANC visits, $OR= 1.537943$, $p = 0.000$, 95% CI (1.394325, 1.696049). This means that a woman who has a spouse is 53.8% more likely than a woman who does not have a spouse to visit the doctor at least four times during pregnancy.

4.5 Model Training

GLM, SVM, and ANN were chosen as the machine learning models for this study. The models were then assessed using a variety of metrics such as Accuracy, Recall, and Precision.

4.5.1 Model Comparison and Evaluation Metrics

Confusion matrix report

The research used a Confusion matrix to assess the performance of the Machine Learning models trained in this research. The results of the model predictions is presented by a confusion matrix in *Table 14* while the results of the each evaluation metric expressed as a percentage is shown in *Table 15*. The results demonstrated that, while all of the models performed well in terms of predicting ANC consumption, ANN outperformed the others, attaining accuracy score of 82.9%. Both the GLM and SVM closely followed this, with 82.2% and 82.7% accuracy, respectively. Specificity refers to how precise the positive class assignment is, in this case, at least four ANC utilization. Because the ANN model has a specificity of 62%, this implies that 38% of all ANC visits were mistakenly predicted as at least four ANC visits. Similarly, the model's sensitivity or recall reflects how well it can detect positive occurrences. The sensitivity score for the Generalized Linear Model was 93.5%, which was quite high.

Table 14: Confusion matrix output

			Reference
Artificial Neural Network (ANN)	Prediction 0	0	1
	0	1038	246
	1	663	3294
General Linear Model (LR)	Prediction 0	0	1
	0	1002	231
	1	669	3308
Support Vector Machine (SVM)	Prediction 0	0	1
	0	1038	246
	1	663	3294

Table 15: Performance metrics

Sr.#	Model	Accuracy (%)	Kappa (%)	Sensitivity (%)	Specificity (%)
1	GLM	82.2	56.4	93.5	58.9
2	ANN	82.9	57.9	93.0	62.0
3	SVM	82.7	57.4	93.1	61.0

ROC Curve

The AUC which is simply the area under the ROC indicates how good the model is capable of predicting negative and positive classes correctly. According to the AUC-ROC Curve, ANN achieved a score of 83.33% while both GLM and SVM both achieved a score of 83.04% which is an indication that ANN is still leading in terms of prediction of ANC utilization. This means that ANN has an 83.33% chance of differentiating between positive and negative classes, whereas GLM and SVM both have an 83.04% chance. *See Fig 18*

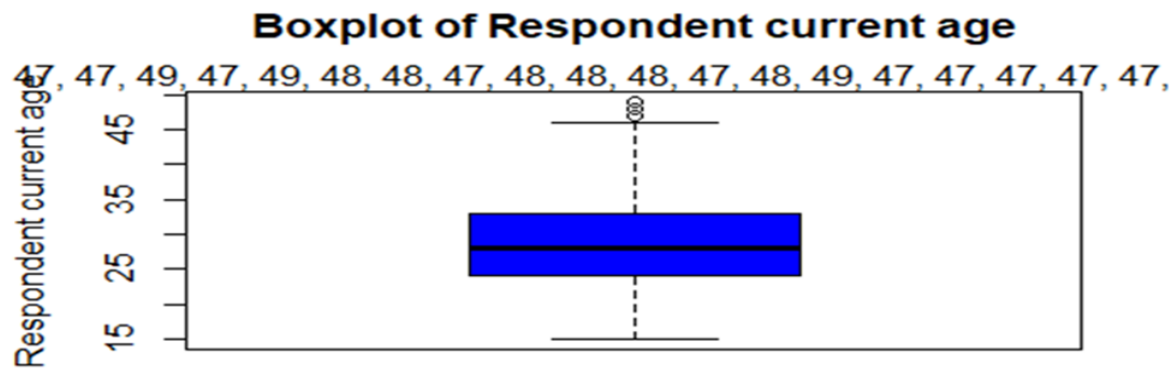


Fig 16 15: Outlier identification using boxplot

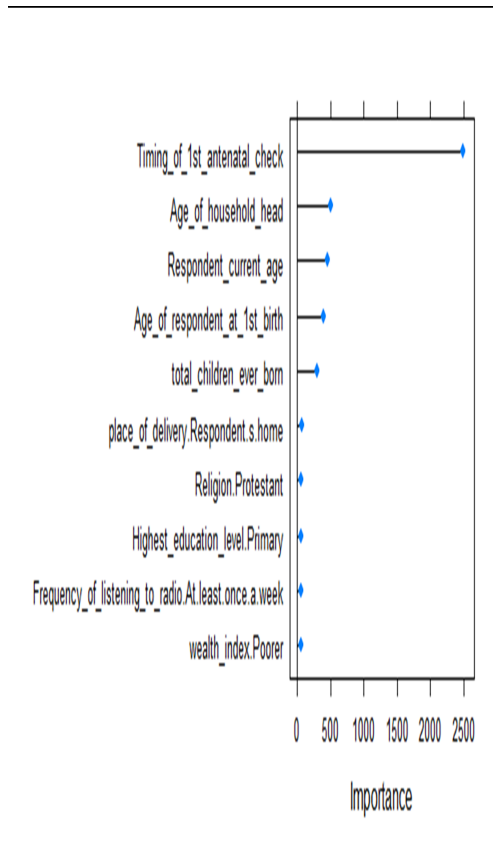


Figure 1: Variable Importance according to Random Forest

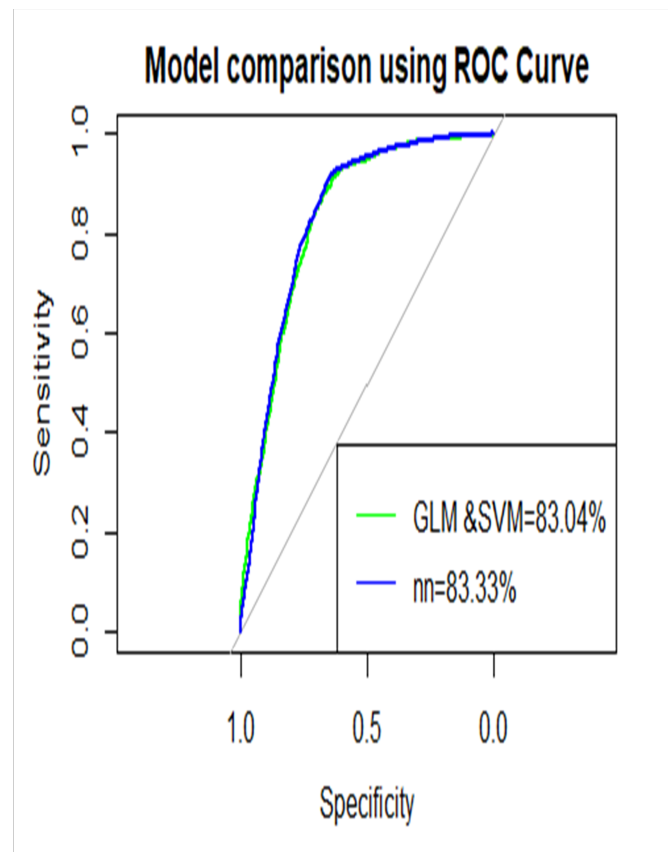


Figure 2: ROC Curve Results

5 CHAPTER FIVE: CONCLUSIONS AND RECOMMENDATIONS

Based on the outstanding demographic characteristics retrieved from the KDHS dataset (2014), a machine learning model was developed in this study to predict Antenatal Care utilization in Kenya. The research focused on data mining techniques such as EDA, data preparation, modeling and evaluation. A feature selection strategy was adopted using Random Forest classifier to rank the demographic characteristics associated with the ANC utilization among women in Kenya. A binary logistic regression model was then fitted with the best selected features from Random Forest to find out how each one of them affects ANC consumption. Machine Learning models including ANN, SVM and GLM was then trained with the selected features. When the performance of the classifiers was compared, it is clear that the ANN classifier is the best choice, as it obtained an accuracy of 82.9%. According to the research, key factors that determine ANC visits in Kenya include wealth index, level of education, woman's present age, whether the pregnant woman has a supportive partner, total children ever born, and place of delivery. The discovered risk variables, according to the conclusions of this study, will be useful in determining whether a pregnant woman is at risk of not using ANC services.

The created model will aid in the maternity care decision-making process by detecting and notifying pregnant women who are at risk of not accessing ANC services, hence averting birth difficulties and, eventually, reducing the incidence of maternity-related mortality. The model built in this study has limitations in that it lacks a user interface because it is not deployed or hosted anywhere. As a result, the researchers believe that this could be a promising field for further research, such as deploying the generated model on the cloud so that health care professionals can use it while providing ANC services to pregnant women during their visits. The study further suggests that safer programs for disadvantaged and vulnerable women be implemented, as well as the inclusion of male partners during Antenatal Care, good media coverage, and promotion of early Antenatal Care and health promotion programs for pregnant mothers with low education.

Bibliography

- [1] Lang'at, E., Mwanri, L. & Temmerman, M. Effects of implementing free maternity service policy in Kenya: an interrupted time series analysis. *BMC Health Serv Res* 19, 645 (2019). <https://doi.org/10.1186/s12913-019-4462-x>
- [2] Chou D, Daelmans B, Jolivet RR, Kinney M, Say L. Ending preventable maternal and newborn mortality and stillbirths. *BMJ*. 2015;351
- [3] Ganchimeg T, Ota E, Morisaki N, et al. Pregnancy and childbirth outcomes among adolescent mothers: a World Health Organization multicountry study. *BJOG* 2014;121 Suppl 1:40–8
- [4] Trends in maternal mortality: 2000 to 2017: estimates by WHO, UNICEF, UNFPA, World Bank Group and the United Nations Population Division. Geneva: World Health Organization; 2019
- [5] Mehboob, R., Ahmad, F. J., Gilani, S. A., Hassan, A., Khalid, S., & Akram, J. (2020). Maternal mortality Ratio in low income developing countries-focusing on Pakistan.
- [6] Otundo Richard, M. (2019). WHO recommendations on antenatal care for a positive pregnancy experience in Kenya. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3449460>
- [7] Heaman, M. I., Newburn-Cook, C. V., Green, C. G., Elliott, L. J., & Helewa, M. E. (2008). Inadequate prenatal care and its association with adverse pregnancy outcomes: A comparison of indices. *BMC Pregnancy and Childbirth*, 8. <https://doi.org/10.1186/1471-2393-8-15>
- [8] Hapsara, H. R. (2005). World Health Organization (WHO): Global health situation. *Encyclopedia of Biostatistics*. <https://doi.org/10.1002/0470011815.b2a17156>
- [9] Bhowmik, K. R., Das, S., & Islam, M. A. (2020). Modelling the number of antenatal care visits in Bangladesh to determine the risk factors for reduced antenatal care attendance. *PLOS ONE*, 15 e0228215. <https://doi.org/10.1371/journal.pone.0228215>
- [10] Ftwi, M., Gebretsadik, G. G., Berhe, H., Haftu, M., Gebremariam, G., & Tesfau, Y. B. (2020). Coverage of completion of four ANC visits based on recommended time schedule in northern Ethiopia: A community-based cross-sectional study design. *PLOS ONE*, 15, e0236965. <https://doi.org/10.1371/journal.pone.0236965>

-
- [11] Tizazu, M. A., Asefa, E. Y., Muluneh, M. A., & Haile, A. B. (2020). Utilizing a minimum of four antenatal care visits and associated factors in Debre Berhan town, north Shewa, Amhara, Ethiopia, 2020. *Risk Management and Health-care Policy*, 13, 2783-2791. <https://doi.org/10.2147/rmhps.s285875>
- [12] Abegaz, K. H. (2018). Exploring trend and barriers of antenatal care utilization using data mining:evidence from EDHS of 2000 to 2016. <https://doi.org/10.1101/351858>
- [13] Okedo-Alex, I. N., Akamike, I. C., Ezeanosike, O. B., & Uneke, C. J. (2019). Determinants of antenatal care utilisation in sub-Saharan Africa: A systematic review. *BMJ Open*, 9, e031890. <https://doi.org/10.1136/bmjopen-2019-031890>
- [14] Wairoto, K. G., Joseph, N. K., Macharia, P. M., & Okiro, E. A. (2020). Determinants of subnational disparities in antenatal care utilisation: A spatial analysis of demographic and health survey data in Kenya. *BMC Health Services Research*, 20. <https://doi.org/10.1186/s12913-020-05531-9>
- [15] Tessema, Z. T., & Minyihun, A. (2021). Utilization and determinants of antenatal care visits in East African countries: A Multicountry analysis of demographic and health surveys. *Advances in Public Health*, 2021, 1-9. <https://doi.org/10.1155/2021/6623009>
- [16] Shibre, G., Zegeye, B., Idriss-Wheeler, D., & Yaya, S. (2020). Factors affecting the utilization of antenatal care services among women in Guinea: A population-based study. *Family Practice*, 38, 63-69. <https://doi.org/10.1093/fampra/cmaa053>
- [17] Jebb, A. T., Parrigon, S., & Woo, S. E. (2017). Exploratory data analysis as a foundation of inductive research. *Human Resource Management Review*, 27, 265-276. <https://doi.org/10.1016/j.hrmr.2016.08.003>
- [18] Winson-Geideman, K., Krause, A., Lipscomb, C. A., & Evangelopoulos, N. (2017). Exploratory data analysis. *Real Estate Analysis in the Information Age*, 69-85. <https://doi.org/10.4324/9781315311135-8>
- [19] <https://ai.stanford.edu/~ang/slides/DeepLearning-Mar2013.pptx>
- [20] Nargesian, F., Samulowitz, H., Khurana, U., Khalil, E. B., & Turaga, D. (2017). Learning feature engineering for classification. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. <https://doi.org/10.24963/ijcai.2017/352>
- [21] Tran, L., Liu, X., Zhou, J., & Jin, R. (2017). Missing modalities imputation via cascaded residual Autoencoder. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. <https://doi.org/10.1109/cvpr.2017.528>

-
- [22] Kwak, S. K., & Kim, J. H. (2017). Statistical data preparation: Management of missing values and outliers. *Korean Journal of Anesthesiology*, 70, 407. <https://doi.org/10.4097/kjae.2017.70.4.407>
- [23] Salgado, C. M., Azevedo, C., Proença, H., & Vieira, S. M. (2016). Missing data. *Secondary Analysis of Electronic Health Records*, 143-162. https://doi.org/10.1007/978-3-319-43742-2_13
- [24] Yadav, M. L., & Roychoudhury, B. (2018). Handling missing values: A study of popular imputation packages in R. *Knowledge-Based Systems*, 160, 104-118. <https://doi.org/10.1016/j.knosys.2018.06.012>
- [25] Yang, P., Wang, D., Wei, Z., Du, X., & Li, T. (2019). An outlier detection approach based on improved self-organizing feature map clustering algorithm. *IEEE Access*, 7, 115914-115925. <https://doi.org/10.1109/access.2019.2922004>
- [26] . Liquet, B., & Riou, J. (2019). Cpmcglm: An R package for P-value adjustment when looking for an optimal transformation of a single explanatory variable in generalized linear models. *BMC Medical Research Methodology*, 19. <https://doi.org/10.1186/s12874-019-0711-2>
- [27] Jackson, E., & Agrawal, R. (2019). Performance evaluation of different feature encoding schemes on cybersecurity logs. 2019 SoutheastCon. <https://doi.org/10.1109/southeastcon42311.2019.9020560>
- [28] Cerda, P., Varoquaux, G., & Kégl, B. (2018). Similarity encoding for learning with dirty categorical variables. *Machine Learning*, 107(8-10), 1477-1494. <https://doi.org/10.1007/s10994-018-5724-2>
- [29] Cerda, P., Varoquaux, G., & Kégl, B. (2018). Similarity encoding for learning with dirty categorical variables. *Machine Learning*, 107(8-10), 1477-1494. <https://doi.org/10.1007/s10994-018-5724-2>
- [30] Tan, J., Yang, J., Wu, S., Chen, G., & Zhao, J. (2021). A critical look at the current train/test split in machine learning. arXiv preprint arXiv:2106.04525
- [31] Behnamian, A., Millard, K., Banks, S. N., White, L., Richardson, M., & Pasher, J. (2017). A systematic approach for variable selection with random forests: Achieving stable variable importance values. *IEEE Geoscience and Remote Sensing Letters*, 14, 1988-1992. <https://doi.org/10.1109/lgrs.2017.2745049>
- [32] Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients. *Anesthesia & Analgesia*, 126, 1763-1768. <https://doi.org/10.1213/ane.0000000000002864>
- [33] Bertsimas, D., & King, A. (2017). Logistic regression: From art to science. *Statistical Science*, 32. <https://doi.org/10.1214/16-sts602>

-
- [34] Ding, S., Shi, Z., & Azar, A. T. (2015). Research and development of advanced computing technologies. *The Scientific World Journal*, 2015, 1-2. <https://doi.org/10.1155/2015/239723>
- [35] Ambikavathi, C., & Srivatsa, S. K. (2020). Predictor selection and attack classification using random forest for intrusion detection. *Journal of Scientific and Industrial Research (JSIR)*, 79(??) 365-368.
- [36] Bates, S., Hastie, T., & Tibshirani, R. (2021). Cross-validation: what does it estimate and how well does it do it?. arXiv preprint arXiv:2104.00673
- [37] Nordhausen, K. (2009). The elements of statistical learning: Data mining, inference, and prediction, second edition by Trevor Hastie, Robert Tibshirani, Jerome Friedman. *International Statistical Review*, 77, 482-482. https://doi.org/10.1111/j.1751-5823.2009.00095_18.x
- [38] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *Statistical learning*. Springer Texts in Statistics, 15-57. https://doi.org/10.1007/978-1-4614-7138-7_2
- [39] Novaković, J. D., Veljović, A., Ilić, S. S., Papić, Ž., & Milica, T. (2017). Evaluation of classification models in machine learning. *Theory and Applications of Mathematics & Computer Science*, 7, 39-46.
- [40] Oshinyemi, T. E., Aluko, J. O., & Oluwatosin, O. A. (2018). Focused antenatal care: Re-appraisal of current practices. *International journal of nursing and midwifery*, 10, 90-98.
- [41] Ministry of Health (MOH), Kenya. Kenya Health Policy, 2012-2030. 2012. Accessed March 29, 2019.
- [42] Micronutrient Initiative Kenya (2016). Improving the demand and health services for pregnant women and newborns in underserved communities. 2012. Accessed March 29, 2019.
- [43] Ng, A. (2013). Machine Learning and AI via Brain simulations. Accessed: May, 3, 2018.