PREDICTING POSTPARTUM HEMORRHAGE IN PREGNANT MOTHERS IN LOW-INCOME SETTINGS A MACHINE LEARNING APPROACH

TOM EGANYU

J23MD10/220

A DISSERTATION SUBMITTED TO THE FACULTY OF ENGINEERING, DESIGN, AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF MASTER OF SCIENCE IN DATA SCIENCE AND ANALYTICS OF UGANDA CHRISTIAN UNIVERSITY

October, 2025

UGANDA CHRISTIAN UNIVERSITY

A Centre of Excellence in the Heart of Africa

Abstract

Introduction

Postpartum hemorrhage (PPH) is still one of the leading causes of maternal deaths worldwide. Unfortunately, in many low-income countries, there is still limited information and effort about the severity of Postpartum Hemorrhage (PPH), its risk factors, use of machine learning algorithms to predict mothers at risk of PPH. It is crucial to comprehend the relative contributions of various PPH risk factors beyond the documented 4Ts (Tone, Trauma, Tissue, and Thrombin).

Methodology

A total of 2212 deliveries were recorded, and 2094 were utilized in this project. Data exploration, feature engineering and correlation analysis were done to identify risk factors of PPH. Machine learning models (Extreme gradient boosting, Random forest, Support vector machine, and Artificial neural network) were used to develop a validated machine learning model.

Results

Extreme Gradient Boosting model performed better with AUC of 97.0%, Accuracy of 96.0% precision of 96.0%, recall of 97.0%. The factors associated with increased risk of Postpartum Hemorrhage were: Number of ANC Visits (P-Value: 0.00); Weight of Baby (g) (P-Value: 0.00); Duration of Labour (P-Value: 0.00); Cervical Tear (P-Value: 0.00), Episiotomy (P-Value: 0.00), and Perineal Tears (P-Value: 0.00)

Conclusion

This project successfully identified the risk factors of PPH and developed a validated machine learning model to predict mothers at risk of PPH. Machine learning models can be used to identify the risk factors and predict mothers at risk of Postpartum Hemorrhage.

Declaration

I, Tom Eganyu, declare that the research described in this report is unique in my own words. Every source of data and information utilized in this study has been properly cited and acknowledged. I have given due credit to anybody who helped me with this research. I promise not to use this work for any other degree or certification at any other university. I realize that this study's findings, conclusions, and suggestions are based on my understanding and insights, as well as the analysis and interpretation of data utilized in this project. I accept and take full responsibility for the integrity and correctness of this study. I am solely accountable for the integrity and correctness of the report. Furthermore, I certify that, by the standards of academic integrity and ethical research conduct, I appropriately took ethical considerations into account at every stage.

02 / October / 2025

....

Tom Eganyu Date

Dedication

To my beloved parents and Siblings who have stood with me in prayers and supported me financially, emotionally, and spiritually throughout my academic life. To my better half, who has stood with me throughout this period and encouraged me to be the best version of myself in every aspect, especially academically. Finally, to thousands of women in our communities who struggle time and time again to bring forth new life amidst poverty and dwindling resources.

Approval

This project thesis entitled "Predicting postpartum hemorrhage in pregnant mothers in low-income settings. A machine learning approach" was submitted to the Faculty of Engineering, Design, and Technology, Department of Computing and Technology, Uganda Christian University for partial fulfillment of the award of Masters of Science in Data Science and Analytics and was examined and approved.

Brief quotations from this manuscript are allowable without special permission, provided that accurate acknowledgment of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the Department. In all other instances, however, permission must be obtained from the author

02 / October / 2025

.....

Student: Tom Eganyu

.....

Date

10 / October / 2025

.....

Supervisor: Ian Raymond Osolo

.....

Date

Acknowledgment

First and foremost, I would want to express my gratitude to God, the Almighty, for His many gifts that enabled me to successfully complete my studies.

My sincere appreciation to my research supervisor, Dr. Ian Raymond Osolo, for his relentless guidance and incites throughout the research period. Working and learning under his direction was an enormous honor and privilege. I would also like to appreciate the entire department of computing and technology and the faculty of engineering, design, and Technology of Uganda Christian University.

I am extending my thanks to the Research Ethics Committee of Uganda Christian University and Kawempe National Referral Hospital for the opportunity taking time to review and approve this project. Special appreciation goes to the records staff of Kawempe National Referral Hospital for their support in data collection.

Lastly, I would want to express my gratitude to everyone who helped me, directly or indirectly, to finish the research project.

Mr. Tom Eganyu

Contents

List of Figures

Acronyms

ANC Antenatal Care. 18

ANN Artificial Neural Network. 51, 56, 63

AUC Area Under the Curve. 11, 56, 62

CSL Consortium for Safe Labor. 9

DIC Disseminated Intravascular Coagulation. 1

GAIN Generative Adversarial Nets Framework (. 10

IQR Inter-Quartile Range. 33

MAR Missing Not At Random. 22

MCAR Missing At Random. 22, 32

MCAR Missing Completely At Random. 22, 32

MMR Maternal Mortality Ratio. 3

NICHD National Institute of Child Health and Human Development. 9

PPH Postpartum Hemorrhage. i, viii, 1–5, 7, 8, 10–13, 16–18, 34, 35, 49, 50, 55, 59–65

RFC Random Forest Classifier. 51, 56, 63

ROC Receiver Operating Characteristic. x, 57

SDG Sustainable Development Goals. 3

SHAP SHapley Additive exPlanations. 54, 59

SVM Support Vector Machines. 56, 63

WHO World Health Organization. 1, 3, 4

XGB Extreme Gradient Boosting. 50, 51, 56, 63

Chapter1

Introduction

According to the American College of Obstetricians and Gynecologists (ACOG), obstetric hemorrhage also known as Postpartum Hemorrhageis an estimated cumulative loss of blood $\geq$ 1000mL within 24 hours of delivery [1]. The World Health Organization defines Postpartum Hemorrhage as cumulative loss of blood $\geq$ 500mL within 24 hours of delivery [2]. Both vaginal and cesarean deliveries are covered by this updated definition according to ACOG. ACOG emphasized that blood loss of $\geq$ 500mL following vaginal birth should still be considered abnormal, especially if severe bleeding persists, and should be closely monitored by the medical team [3].

Blood loss of $\geq$ 500mL and $\geq$ 1000mL is associated with both vaginal delivery and cesarean delivery respectively. The FIGO International Federation of Gynecology and Obstetrics) guideline, which was recently updated, defines postpartum hemorrhage as any blood loss that compromises hemodynamic stability, or $\geq$ 500mL within 24 hours after a vaginal birth or $\geq$ 1000mL after a cesarean section [3]. Postpartum Hemorrhage occurs within 24 hours until 12 weeks after delivery. Primary Postpartum Hemorrhage occurs within 24 hours after delivery whereas Secondary Postpartum Hemorrhage occurs from 24 hours to 12 weeks after delivery. Postpartum Hemorrhage causes hypovolemia (Abnormally low extracellular fluid in the body) within 24 hours of delivery, 10% decrease in hemoglobin levels, body shock, Disseminated Intravascular Coagulation, and eventually deaths [4] [5] [6].

1.1 Postpartum Hemorrhage Etiology

PPH's etiology is frequently grouped using the "4 Ts" framework [3]:

- Tone: uterine atony
- Trauma: laceration, rupture;
- Tissue: retained tissue
- Thrombin: coagulopathy

Generally, Postpartum Hemorrhage is caused by genital tract injuries, failure of the blood coagulation system, and uterine atony [7]. Uterine atony is defined as a soft and weak uterus after childbirth and it is responsible for up to 75% of Postpartum Hemorrhage. Additionally, any pregnant mother with a history of PPH, multiple pregnancies, fetal macrosomia, primigravida, grand multi-parity, older age, pre-term births, genital tract injuries, non-use of oxytocics for PPH prophylaxis, labor induction, cesarean birth, and intra-uterine fetal deaths is at risk of PPH [7].

Despite the known factors and risks of PPH, the identification of mothers at risk of PPH is a challenge. Postpartum Hemorrhage still occurs in 1%–5% of maternal deliveries in developed and developing countries and it is still the leading cause of maternal morbidity and mortality worldwide [4]. Maternal mortality associated with Postpartum Hemorrhage is a result of a delay in clinical diagnosis of Postpartum Hemorrhage and its complications. Early-stage diagnoses of Postpartum Hemorrhage could reduce the death ratio for obstetricians [4]. In most developing countries, visual estimation of blood loss by health care providers is done compared to the use of calibrated under-buttock drapes to measure blood loss as a standard in developed countries. Inaccurate estimation of blood during delivery and after delivery delays response to PPH cases. Additionally, it frustrates efforts to predict mothers at risk and reduce maternal morbidity and mortality cases associated with Postpartum Hemorrhage. Therefore, It is very critical to assess patients who are at risk of Postpartum Hemorrhage during antenatal, intrapartum, and postpartum. This will allow for early preparation, improve monitoring and timely prevention, and lower the incidence of Postpartum Hemorrhage and maternal mortality,

Documentation of Postpartum Hemorrhage cases is critical for any future efforts to reduce maternal morbidity and mortality resulting from Postpartum Hemorrhage. Unfortunately, the health system in most developing countries struggles with the digitization of patient data. Currently, data from health facilities is in most cases, aggregated counts of cases,

leaving the most critical individual patient raw data on paper. More effort is required to digitize past data and introduce electronic systems to capture health facility data.

1.2 Background

Globally, there is still high morbidity and mortality among pregnant mothers from preventable causes resulting from pregnancies and childbirth. Postpartum Hemorrhage (PPH) is still one of the leading causes of morbidity and mortality among pregnant mothers. it is credited for 20%-50% of maternal mortality worldwide and 99% of maternal mortality occurs in low-income and medium-income countries (LMICs) [8].

According to the World Health Organization, millions of women are still exposed and suffer from Postpartum Hemorrhage with over 20% maternal fatalities [2]. In addition to likely having had immediate surgical procedures to stop the bleeding, women who survived life-threatening PPH may have long-term physical (such as permanent reproductive impairment, bladder damage, postpartum infection, or anemia) and psychological (such as post-traumatic stress disorder) repercussions [2].

Sustainable Development Goals 2030 target 3.1 aims at reducing maternal mortality to less than 70 maternal deaths per 100,000 live births by 2030 [9]. However, maternal mortality from preventable causes resulting from pregnancies and childbirth seeks to derail the effort in achieving target 3.1 of SDG. The Maternal Mortality Ratio (MMR) by 2020 was 223 per 100,000 live births. This is way above the SDG 3.1 target by 2030 [2].

Although there is significant progress in reducing maternal mortality, progress has stagnated for the past 5 -10 years. Between the years 2000 and 2023, the global maternal mortality ratio (MMR) saw a 40 percent reduction, decreasing from 328 deaths to 197 deaths per 100,000 live births [10]. Globally, Postpartum Hemorrhage prevalence stands at 6% [7]. In the United States alone, there was an increase in Postpartum Hemorrhage cases reported with 2.9% cases in 2010 and 3.2 cases by 2014 [1]. Overall, between 2010 and 2016, there were 11% of Postpartum Hemorrhage cases reported in the United States [11].

Meanwhile, Sub-Saharan Africa is lagging with as high as 10.5% Postpartum Hemorrhage cases. This is worsened by poor documentation and reporting of cases in most developing countries. Maternal death in Kenya is as high as 342 per 100,000 live births and this is

above the current global maternal mortality rate of 197 per 100,000 [8]. In Uganda, maternal mortality stands at 189 per 100,000, and 25% of maternal deaths are as a result of Postpartum Hemorrhage [12].

As a result of the limited progress, World Health Organization developed a roadmap to combat maternal mortality and particularly Postpartum Hemorrhage between 2023 and 2030. The roadmap acknowledges the importance of research in understanding Postpartum Hemorrhage and the development of new prevention, diagnosis, and treatment strategies and interventions for Postpartum Hemorrhage [2].

Several studies have been conducted in various countries to identify the critical risk factors of Postpartum Hemorrhage and attempts have been made to develop models to predict mothers at risk of Postpartum Hemorrhage for example using traditional statistical lasso/logistic regression models. However, there are still no reliable prediction and scoring systems for Postpartum Hemorrhage. There exist several risk factors associated with Postpartum Hemorrhage and identifying the most relevant factors is challenging without a robust clinical prediction model. Currently, uterine atony, soft birth canal laceration, placental factors, and coagulation dysfunction are recognized risk factors of Postpartum Hemorrhage. Unfortunately, Postpartum Hemorrhage patients without those risk factors do exist. Although classic statistical approaches provide a modest yet good discriminate ability utilizing maternal clinical features and medical history the performance of the models is unsatisfactory.

Machine learning algorithms have recently been popularized in healthcare. Machine learning uses statistical approaches to train algorithms to learn from and perform predictions based on data. Machine learning thrives on large datasets by deriving novel insights from large datasets which is most critical for decision-making. Machine learning models utilizing small data have proved to be unreliable. Recently, studies have utilized machine learning approaches to identify Postpartum Hemorrhage risk factors and mothers at risk of Postpartum Hemorrhage. However, these models have used small data sizes, or too many features stemming from the feature selection method among other factors.

In this project, machine learning was used to identify Postpartum Hemorrhage risk factors and develop a model to predict mothers at risk of PPH.

1.3 Problem Statement

Traditionally, a reduction in maternal morbidity and mortality is one of the key indicators in measuring improvement in maternal health. The reduction in maternal and child mortality is one of the key pressing global health priorities as indicated in SDG 3.1. According to the Postpartum Hemorrhage road-map 2023 to 2030, the progress to achieve SDG 3.1 has stalled over the last 5-10 years [2]. Postpartum Hemorrhage still stands at 6% globally, credited for 20% - 50% of maternal deaths of which 99% occur in low-income settings [8].

Several studies have been conducted to assess the risk factors of Postpartum Hemorrhage. In low-income countries, risk factors for Postpartum Hemorrhage include, among others, cesarean section delivery, multiple pregnancies, fetal macrosomia $\geq 4000g$, HIV positive serostatus [7].

Machine learning algorithms have since been utilized in healthcare in diagnostics, improving care, pharmacology, prescription, and data security. Machine learning algorithms can be utilized to identify risk factors and mothers who are at risk of PPH. Unlike traditional approaches that convert input into output according to predetermined rules, Machine learning algorithms learn from data and generate new rules and patterns from both input and output data. Unfortunately, machine learning algorithms have not been adequately evaluated in obstetrics.

Studies have been conducted in developed countries that focus more on cesarean delivery [8]. A study in Kenya developed a machine learning model to identify risk factors and predict women at risk of Postpartum Hemorrhage. Despite very impressive results, the dataset used in the study was small. More research on the application of machine learning in general obstetric and low-income settings, to which most of the countries in sub-Saharan Africa belong, is necessary.

Early-stage identification of risk factors and prediction of mothers at risk Postpartum Hemorrhage is critical in reducing maternal and child mortality, life and permanent long-term physical, reproductive, and psychological repercussions, and eventually achieving SDG 3.1.

1.4 Goal and Objectives

The aim of this project is to decrease postpartum hemorrhage related morbidity and mortality in low-income areas by leveraging machine learning approaches.

The specific objectives were:-

(a) To identify risk factors associated with Postpartum Hemorrhage among pregnant mothers.

(b) To develop a machine learning algorithm to predict mothers at risk of Postpartum Hemorrhage.

(c) To evaluate the machine learning model's performance with a focus on precision.

1.5 Definition of Hypothesis

The purpose of this project was to address the hypothesis questions below:

- What are the significant risk factors of postpartum hemorrhage in low-income settings?
- Based on the risk factors, can a validated machine learning model be developed to predict mothers at risk of postpartum hemorrhage?

1.6 The Research Methodology

This project will utilize quantitative research methodology. Quantitative research methodology is popular in research endeavors and projects involving machine learning approaches. The foundation of quantitative research is the measurement of a certain quantity or amount of a given phenomenon [13]. It focuses on collecting and analyzing numerical data, and it can be used to find averages, patterns, or forecasts. Primary and secondary data collection approaches are often used in quantitative research methodology. Primary data is collected from the source, usually through questionnaires either remotely or through interviewers. However, secondary data can be obtained from reports or data repositories.

A study in Kenya used quantitative research methodology, secondary data was collected from the Kenya Antenatal and Postnatal Care Research Collective cohort for the period

August 2020 to February 2022 [8]. Another study by Yawei Zhang et al used a quantitative approach and collected and maintained a database of 3842 patient records from Beijing Obstetrics and Gynecology Hospital. Another study by Meng Wang et al used a quantitative research methodology. Data was extracted from the medical electronic case system. A wide range of factors, including social, clinical, and diagnostic ones, were used. A Single-Centre Retrospective Analysis study by Wenhuan Wang et al used clinical variables to identify risk factors of Postpartum Hemorrhage and predict women at risk of PPH. In this study, quantitative methodology and secondary data were obtained from Wenzhou People's Hospital.

In this study, a quantitative research methodology approach was considered. Data was obtained from Afro Health Connect from 2017 to 2023. Secondary data was utilized for analysis for identification of risk factors and prediction of Postpartum Hemorrhage.

1.7 Project Process

The project process involved obtaining clearance from the Uganda Christian University Ethics Committee to carry out the project. This was followed by stakeholder engagement in preparation for data collection. Secondary data was obtained online from Mendeley data repository for a study conducted in zimbabwe between 01 March 2016 to 31 May 2016. Data analysis was conducted involving exploratory data analysis, data pre-processing, feature engineering, and prediction model development and evaluation. The project report was written based on the project process and findings from the analysis. This was followed by the dissemination of the report to the stakeholders, including Uganda Christian University for the award of a degree and Afro Health Connect.

Chapter2

Literature review

Postpartum Hemorrhage is a threat to all pregnant mothers. There are known guidelines on identification of mothers who are at risk of Postpartum Hemorrhage, including the 4Ts framework defined as Tone, Tissue, Trauma, and Thrombin. Studies have also shown a combination of social and clinical indicators that are known as risk factors for Postpartum Hemorrhage. Unfortunately, these risk factors are often overlooked in assessing mothers at risk Postpartum Hemorrhage. As a result, a percentage of women have experienced PPH and its adverse events, including death. Machine learning models can be utilized for the early identification of mothers who are at risk of PPH.

Machine learning has revolutionized health by improving efficiency and accuracy in disease diagnosis; as a result, patients do not have to queue for long hours to receive medical attention. Machine learning is increasingly being used to enable personalized medicine and optimized treatment strategies. Machine learning algorithms can assist medical professionals in developing individualized treatment programs that are more appropriate for each patient's particular requirements by examining a patient's genetic composition, lifestyle choices, and environmental variables. It is heavily utilized in predictive analytics for early disease detection. As a result, early interventions and preventive measures are made possible by predictive models' ability to identify people at high risk for particular diseases based on patient data. Machine learning algorithms have enabled effective and efficient analysis of medical images. With the use of machine learning algorithms, radiologists can identify minute anomalies that the human eye might overlook by accurately analyzing X-rays, CT scans, and MRIs. Other areas of machine learning application include accelerated drug discovery and development, streamlined health operations, and public health applications, including disease surveillance

and public health policy. [14]

2.1 Related Literature

Studies have been conducted to identify Postpartum Hemorrhage risk factors and predict mothers at risk of Postpartum Hemorrhage.

Venkatesh KK et al conducted a study to predict a woman's risk of postpartum hemorrhage at labor admission using machine learning and statistical models. They utilized available data at the time of admission for labor. Machine learning and statistical models were developed and measured using the C statistic, or area under the receiver operating characteristic (ROC) curve (machine learning), calibration curves (statistical method), and decision curves. (statistical method). Machine learning models outperformed other models, with Extreme Gradient Boosting and random forest models providing excellent ability to predict postpartum hemorrhage. However, the data utilized for this analysis included variables on time of labor admission and intrapartum variables, for example, length of labor and mode of delivery, which were not included in this analysis [15]. Additionally, other clinical variables like a history of Postpartum Hemorrhage, especially in prior pregnancies, use of thromboprophylaxis drug treatment, uterine fibroids, and placental characteristics, were not part of the analysis. A machine learning model that incorporates the above missing variables is necessary to accurately predict mothers at risk of Postpartum Hemorrhage.

In a study by Ahmadzia H et al, data from the Consortium for Safe Labor created by the Eunice Kennedy Shriver National Institute of Child Health and Human Development was utilized to develop a machine learning model to predict postpartum hemorrhage and transfusion. The study was aimed at ensuring effective risk-based triage and eventually reducing maternal morbidity and mortality resulting from postpartum hemorrhage. In this project, the model predicting the transfusion and PPH composite using antepartum and intrapartum variables produced excellent positive predictive values. Extreme Gradient Boosting (XGB) machine learning model was the best with ROC-AUC=0.833, 95% CI 0.828-0.838; PR-AUC=0.210, 95% CI 0.201-0.220. The variables used for prediction included mode of delivery, intrapartum tocolytic use, oxytocin incremental dose for labor (mU/minute), hospital type, and presence of anesthesia nurse [11]. The results of the study show a very low

precision of 0.111 for XGB, which is the best-performing model in this case. Additionally, the outcome variable is a composite of transfusion and Postpartum Hemorrhage. This means that any mother with blood loss above the threshold and who received a blood transfusion certainly had Postpartum Hemorrhage.

Between 2013 and 2014, a study was carried out in Uganda to evaluate the prevalence of postpartum hemorrhage and its risk factors among Ugandan rural women.. Six health facilities in Uganda were included, and a questionnaire was administered to Women to identify risk factors for postpartum hemorrhage. In this project, PPH was defined as a blood loss of $\geq$500mL and assessed using a calibrated under-buttocks drape at childbirth. The outcome revealed overall, 9.0% (95% CI) of the 1188 women experienced postpartum hemorrhage, and 1.2% (95% CI) experienced severe postpartum hemorrhage ($\geq$1000mL). After giving birth, 97.4% of women received uterotonics to prevent postpartum hemorrhage. [7]. The percentage of women who were administered uterotonic for prophylaxis was high. This complicates the identification of women with Postpartum Hemorrhage. Unfortunately, this project did not utilize a machine-learning approach. There are currently very few, if any, studies that have employed machine learning to predict PPH in Uganda.

A recent study in Kenya used machine learning algorithms to Postpartum Hemorrhage in a Kenyan population. This is one of the studies that explored the application of a machine learning algorithm in sub-Saharan and a generalized setting (virginal and cesarean birth). This dataset contained demographic, clinical, and intrapartum variables. Delivery outcomes were reported by healthcare staff within 24 hours, The study considered both Virginal and Cesarean delivery outcomes. Features from the onset of labor, for example, length of labor, mode of delivery, and Secondary Postpartum Hemorrhage feature were excluded from this study. To prevent multicollinearity, one feature was retained from highly correlated features. Missing values were handled in one of two ways: for categorical variables, a new category "others" was created; for numerical variables, missing values were imputed using Generative Adversarial Nets Framework (imputation methodology. This imputation approach is known for retaining precision in numeric features. The ratio of Postpartum Hemorrhage outcome (PPH and non-PPH) was kept balanced, and data was divided into train and test datasets with train 67% and 33% percentage of data allocated respectively. The data was further subdivided into 67%, and 33% and small datasets were utilized for validation [8]. Logistic

regression, Naïve Bayes, decision tree, and random forest models were evaluated using the dataset. Accuracy, precision, and Area Under Curve (AUC) performance metrics were used. The naive-Bayes model performed best with 0.95% accuracy, 0.97% precision/specificity, and 0.76% Area Under the Curve. Unfortunately, the study utilized a relatively limited dataset. The total sample size was 1,576 records, which were divided into 67% training and 33% testing (1,056/1,576). Studies need to be conducted with larger datasets to assess the effectiveness of machine learning algorithms to predict mothers at risk of Postpartum Hemorrhage.

A Single-Centre Retrospective Analysis study by Wenhuan Wang et al used clinical variables to identify risk factors of Postpartum Hemorrhage and predict women at risk of Postpartum Hemorrhage. This study utilized secondary data from Wenzhou People's Hospital. The data was split into groups based on hemoglobin levels, high and low. Women in the high hemorrhage group lost more than 500mL of blood within 24 hours of giving birth, while those in the low hemorrhage group lost less than 500mL [16]. For machine learning analysis, test sets for the two distinct delivery modalities were chosen from a total of 130 women who gave birth vaginally and 26 women who gave birth via cesarean section. Women with insufficient clinical data, those with coagulation problems, and those using anticoagulant medication were not included. This approach is indeed ideal for predicting Postpartum Hemorrhage outcome in low-resource settings. However, this may have an impact on the sample size available for model development. Recursive Feature Elimination (RFE), Recursive Feature Elimination and Cross-Validation (RFECV), and SelectKBest techniques were used to select features for model development. The most accurate feature was RFE. Data was divided into 80% training, and 20% test. AdaBoostClassifier, GaussianNB, GradientBoostingClasifier, HistGradientBoostingClassifier, and Logistic Regression were the ensemble models used for the predictions. To determine ideal settings, GridSearchCV was utilized to find optimal parameters. Accuracy, Recall, Precision, F1-score, and ROC_AUC were used to evaluate performance [16]. Compared to manual scoring and the machine learning model, Postpartum Hemorrhage prediction model had an AUC of 0.669, and the 95%, and CI was 0.578–0.759, and manual scoring had an AUC of 0.557 and 95% confidence interval of 0.460–0.654. These results confirm the superiority of machine-learning approaches compared to manual approaches.

Another study by Meng Wang et al used machine learning algorithms to predict the amount of PPH during a cesarean section. Data was extracted from the medical electronic case system. A wide range of factors, including social, clinical, and diagnostic ones, were used. The feature engineering done was to rectify the data imbalance. This study used simple imputation techniques using mean and mode for both numeric and categorical variables. The Permutation importance method was used in this study to clarify each indicator's capacity for prediction in these models. Permutation helped identify features that can best be utilized in the prediction model. Other approaches utilized by other studies include Recursive Feature Elimination (RFE), Recursive Feature Elimination and Cross-Validation (RFECV), and SelectKBest techniques [16]. The models Random Forest, Multilayer Perceptron, XGBoost Models were used, Gradient Boosting, Linear Regression, and Logistic Regression. Root Mean Squared Error (RMSE), and Mean Squared Error (MAE) were the only performance metrics used to evaluate the models. The random forest model performed better with an RMSE 33.75 and a prediction error of less than 9.3%. Generally, ensemble models seem to potentially perform better in regression tasks.

Yawei Zhang et al data from Beijing Obstetrics and Gynecology Hospital to predict blood loss and disseminated intravascular coagulation (DIC) using Ensemble learning-based models. Features were classified into present gestation and delivery features. This study used four base learners: SVM, XGBoost (XGB), gradient boosting decision tree (GBDT), and random forest (RF), Grid-search technique to find the ideal value for the primary hyperparameters for each base learner (RF, GBDT, XGB, or SVM) [17]. Compared to Santosh Yogendra Shah et al, this study utilized hyperparameter tuning. Hyperparameters improve the learning rate of the model and indeed the accuracy and precision. The performance of the model was measured using Accuracy, Precision, F-score, Recall, and Matthews correlation coefficient (MCC). The ensemble model accuracy of over 96.7% was recorded for the Postpartum Hemorrhage, and the accuracy of 90.0 for DIC prediction. Generally, the precision value for all prediction tasks was over 70%. This study showed very impressive results while utilizing the ensemble model.

A multi-center study was conducted, involving 203 patients with or without intra / postpartum hemorrhage within the first 24 hours after delivery. The subjects were divided into two categories: individuals with intra/postpartum hemorrhage (PPH) and those in the con-

trol group who did not have Postpartum Hemorrhage. The Postpartum Hemorrhage patients were further grouped into four strata following Advanced Trauma Life Support guidelines [3]. The Naïve Bayes (NB) algorithm demonstrated the best accuracy for predicting PPH, achieving a sensitivity of 96.3% and an overall accuracy of 98.6%, along with a false negative rate of 3.7%. However, the study utilized a very small dataset consisting of 203 patients with very limited features, limiting the generalization of the model, necessitating additional studies utilizing larger datasets.

2.2 Reviewed literature

A comparative assessment of the examined research papers is shown in the following table, which also highlights the studies' methodology, areas of focus, and applicability to this project:

Paper Reviewed	Author	Data Source	Objectives	Method of Analysis	Relevance	Gap
Machine Learning and Statistical Models to Predict Postpartum Hemorrhage	Kartik K Venkatesh et al	Online Article/ Journal	To predict a woman's risk of postpartum hemorrhage at labor admission using machine learning and statistical models.	Statistical Analysis and Machine learning Algorithms	Demonstrates that machine learning models are superior to statistical models in prediction PPH	Focussed on time of labour admission but does not include intrapartum variables like length of labour and mode of delivery.
Machine Learning Models for Prediction of Maternal Hemorrhage and Transfusion: Model Development Study	Homa Khorrami Ahmadzia et al	Online Article/ Journal	To create a validated prediction model using machine learning for postpartum hemorrhage and transfusion to optimize risk-based triage and inform policy makers and stakeholders who aim to further reduce maternal morbidity and mortality associated with hemorrhage	Machine learning algorithms	Demonstrated the power of Ensemble models in prediction of PPH	Limitations of the study include the low reported precision of algorithms. Sensitivity is prioritized for prediction.
Incidence and risk factors for postpartum hemorrhage in Uganda	Sam Ononge et al	Online Article / Journal	Assessed the incidence of, and risk factors for postpartum hemorrhage among rural women in Uganda	Statistical Analysis	One of the few papers that assessed the risk factors of PPH in Uganda	No machine learning concept has been applied in this papera

Paper Reviewed	Author	Data Source	Objectives	Method of Analysis	Relevance	Gap
Machine learning approach for the prediction of postpartum hemorrhage in vaginal birth	Munetoshi Akazawa et al	Online Article / Journal	to construct a deep learning model to predict PPH in vaginal birth	Deep Learning model	Demonstrates the application of deep learning models in prediction of Postpartum Hemorrhage	Small Dataset
Prediction of postpartum hemorrhage using traditional statistical analysis and a machine learning approach	Vahid Mehrnoushet al	Online Article/ Journal	use a traditional analytical approach and a machine learning model to predict postpartum hemorrhage	Statitical analysis and machine learning	Compares traditional statistical analysis and machine learning algorithms.	Missing timeline start of delivery events for all forms of delivery

Figure 2.1: Reviewed Literature

2.3 Conclusion

The above studies show that the machine learning approach, given the availability of data, can be used to accurately predict Postpartum Hemorrhage. Health systems in developing countries like Uganda struggle with the digitization of health data. However, in Uganda, there is hope with the recent piloting of digital tools in regional referral hospitals. Additionally, the private sector has made some progress in digitizing health data. The studies conducted, as seen above, have expressed a need for accurate primary and secondary data collection in healthcare by providers. Postpartum Hemorrhage persists, and it is still the leading cause of maternal morbidity and mortality. PPH can be prevented if models can be designed on available data to identify risk factors and predict mothers at risk of PPH [18].

Chapter3

Methodology

3.1 Philosophical consideration

The foundation of this project is based on positivism philosophical approach, which emphasizes the use of empirical, observable, and quantifiable data to comprehend real-world occurrences.. Positivism, which has its roots in Auguste Comte's writings, holds that knowledge comes from scientific investigation rather than conjecture or personal opinion. It maintains that observation, experimentation, and quantitative reasoning are the methods by which reliable information is acquired.

The study's data-driven technique and positivist research philosophy are highly compatible, emphasizing objective, measurable, and quantifiable facts. The methodical gathering and examination of postpartum hemorrhage leverages the use of available data, exploratory data analysis (EDA), and the use of machine learning algorithms, demonstrating this alignment.

All things considered, the positivist approach strengthens the study's methodological rigor and guarantees that its conclusions, especially those about postpartum hemorrhage, are supported by reproducible, evidence-based findings.

3.2 Research Design

This research incorporates a quantitative method design within a data-focused framework. The positivist approach prioritizes impartiality in the analysis, interpretation of empirical data, and predictive modeling. From a quantitative perspective, the research applies machine

learning techniques to clean and transform data, extract features from data relevant for analysis and select features that are highly correlated with Postpartum Hemorrhage. Being a classification problem, machine learning classification algorithms for example random forest, support vector machines, artificial neural networks and extreme gradient boosting were tested and the results and interpretation were based on data and performance of the machine learning models.

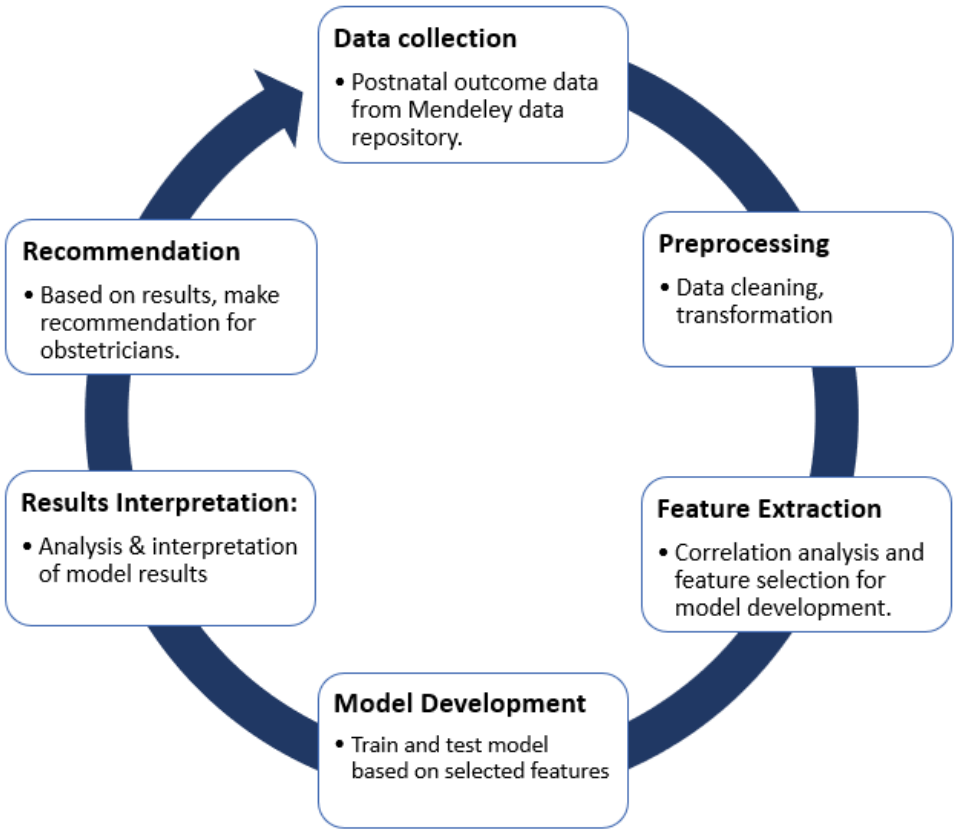


Figure 3.1: Project design

3.3 Project Setting

The project data was to be obtained from Kawempe National Referral Hospital. The Kawempe National Referral Hospital (KNRF) maternity section is the busiest unit in the country, with estimated births of 20,485 between the 2022 and 2023 financial year. Nationally, Uganda reported 1.4 million births in the same period. Kawempe National Referral Hospital doubles as a teaching hospital for medical and nursing schools. The unit has expe-

rienced obstetric and gynecologists, midwives, and residents depending on the universities that use the hospital for clinical placements [19]. Ethical review was obtained from Kawempe National Referral Hospital, and administrative clearance to conduct research in the hospital was obtained. Unfortunately, the hospital health data management system is built to record data and generate aggregated reports, and cannot enable the extraction of secondary data. Nevertheless, the data for model development was sourced online from Mendeley data repositories for published articles and journals [20] while about 10 records were collected from Kawempe National Referral Hospital for model testing.

A study conducted in Zimbabwe between 01 March 2016 to 31 May 2016 evaluated postnatal outcomes in low-income settings [21]. A total of 2212 deliveries were recorded within that period, of which 2094 were selected for the study. Social-demographic, clinical, and perinatal variables were recorded in this dataset including outcome variables stage of labor at delivery, gestational age at delivery, HIV prevalence, Cesarean section rates, hypertensive disorders, preterm delivery, postpartum hemorrhage, birth weight, stillbirth rates, duration of labour, perineal tears, neonatal intensive care unit admissions, special baby care admissions, birth asphyxia, and maternal death.

This dataset was used to develop the model considering the similarities between Zimbabwe's health systems and health systems in other low-resource settings, including Uganda. Both Uganda and Zimbabwe health systems have decentralized, multi-tiered, public and private, and focus on primary health care. Postpartum Hemorrhage incidence in Uganda and Zimbabwe is not well documented. However, Postpartum Hemorrhage incidence in Mpilo Central Hospital, Bulawayo in Zimbabwe was at 1.6% and 9% in Uganda [21]. In Uganda, Postpartum Hemorrhage is credited for over 25% of maternal deaths.

3.4 Data collection approach

This project utilized quantitative research methodology. Studies involving machine learning models are entirely based on measured variables of quantity or a certain amount of quantity or measurement [13][8]. Secondary data were collected online from Mendeley data. The search string "Postpartum Hemorrhage in Low-Resource Settings" listed relevant articles and journals related to the topic. Additionally, articles and journals from 2017 to 2025,

and with a dataset, were included, giving search results of 4,353,692. Articles relevant to the topic were considered, and the dataset was reviewed to ensure that variables relevant to predicting postpartum hemorrhage were included. A study by Solwayo Ngwenya et al assessed Perinatal outcomes in a low-resource setting in Zimbabwe. The study assessed the outcomes between booked and unbooked pregnancies in labour and postnatal wards of Mpilo Central Hospital. A questionnaire was administered to pregnant mothers, and demographic, maternal, and perinatal data were recorded. This study involved 2094 pregnant mothers. The data obtained involved a total of 60 variables for gestational age at delivery, stage of labor at delivery, preterm delivery, hypertensive disorders, HIV prevalence, Mode of Delivery duration of labour, perineal tears, birth asphyxia, stillbirth, postpartum hemorrhage, birth weight, neonatal ICU admissions, maternal death etc [22], The data for this study was used in this project.

3.5 Study population

A total of 2094 pregnant mothers were included in this project. All pregnant mothers having a documented record of antenatal, intrapartum, and perinatal delivery outcome, and with reported Postpartum Hemorrhage outcome were included. Mothers who delivered through vaginal and cesarean sections were included in this project. Estimated Blood Loss was critical in the validation of the prediction variable for postpartum hemorrhage. Estimated Blood Loss 500mL or more and 1000mL or more indicated postpartum hemorrhage for vaginal and cesarean deliveries respectively. Incidentally, all records without estimated blood loss, and less than 100mL according to the dataset, were excluded from the analysis.

3.6 Sample size determination

An optimal amount of data is necessary for a precise and reliable model. Machine learning models utilizing small data samples have suffered from over-fitting of data. Additionally, an increment in the size of the sample produced increased accuracy of the prediction, but may not cause additional effects after a certain sample size [23]. To provide good discriminative ability to the model, all 2094 records were included in this project. Inclusion and exclusion criteria reduced the sample size to 1954 records. All Antenatal Care, maternity

and perinatal features except referral, admission, final maternal outcome, and child delivery outcome features, which were not relevant to the project. Pregnant mothers and study site identification features were also excluded from the project.

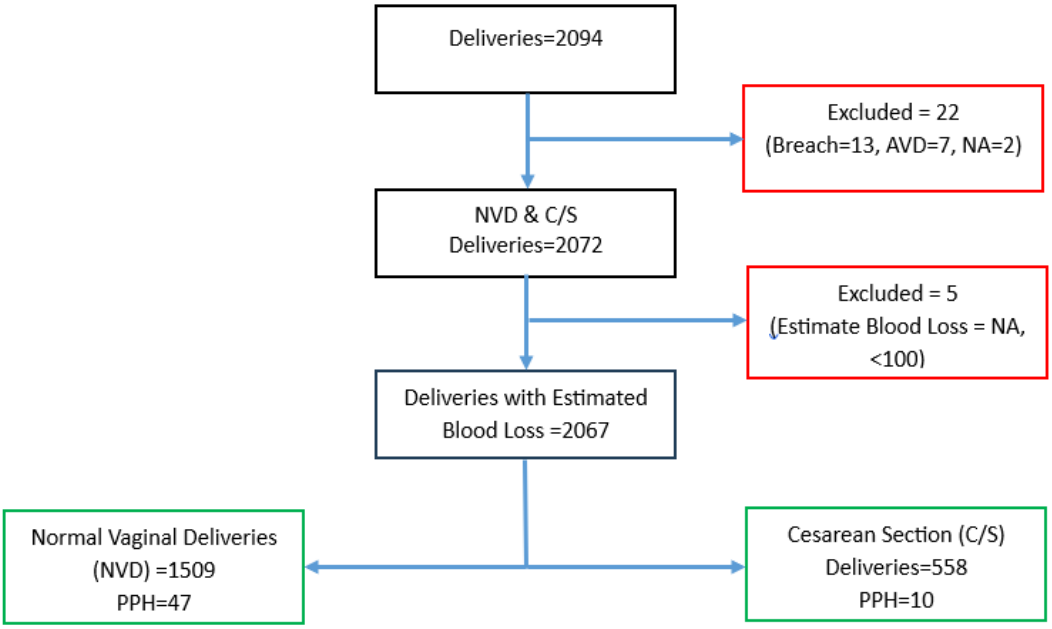


Figure 3.2: Study participants

Chapter4

Data Analysis, and presentation

4.1 Analysis Tools

The analysis was done in Python programming language. Machine learning approaches have transformed the way large data is processed and analyzed, and Python programming language is central to this. Python's powerful libraries like Scikit-learn, TensorFlow, and Keras are fundamental in machine learning. These comprehensive resources are equipped with methods and functions to harness the power of data and create intelligent applications. The scikit-learn library was very central in the development of the model. Data was obtained in the spreadsheet format (.xlsx). The data format is compatible and can be imported and manipulated in a Python environment. These tools were installed on an Acer laptop, Core i7, 1TB, 16GB, and Microsoft Windows 11 en.

4.2 Data exploration

The purpose of data exploration was to develop an understanding of the dataset, especially the different features, data types, missing data, and outliers that can be treated later. It also prompted a need for data pre-processing, for example change of data types, and the creation of new variables.

Data structure

Data structure focuses on the identification of the size, shape, features, and dimensionality. This is the most important first step in data analysis. The Python pandas library provides

functions, for example, df.size, df.shape, df.ndim dt.unique(), etc, that support this process.

The dataset contained 2094 records and 60 features. The features are grouped into 3 different data types: 3 float, 9 integer, and 48 object type features. A quick look at the not null count column shows that several features have missing values, for example, Hypertension Disorder, Other complications, Estimated Blood Loss, Perineal Tears, If Yes, degree, Admission to NICU, Partner Level of Education, etc. Checking for missing values is necessary to confirm this assumption. Certainly, several of these columns identified the interviewers, site, date, and time of recording this data. Several of the features were dropped during data preprocessing because they were not important for this analysis. The data was organized in a two-dimensional array i.e, an array of arrays, with each column called a feature, having its associated rows called records.

Feature unique values

The exploratory data analysis identified features with unique values. This was critical to determine features whose data types can be changed depending on the data they hold. Weight of Baby (g) had the most categories, Other includes Reason for referral to mpilo, Reason for booking at mpilo, Duration of Labour_mins, Estimated Blood Loss, etc. Fortunately, features like Weight of Baby(g) and Estimated Blood Loss are numeric features. Depending on the size of the dataset, a threshold is important for a variable to be considered categorical. In this case, 10 would be more appropriate. It is important to note that variables with descriptive responses had many categories. These variables were dropped during data pre-processing.

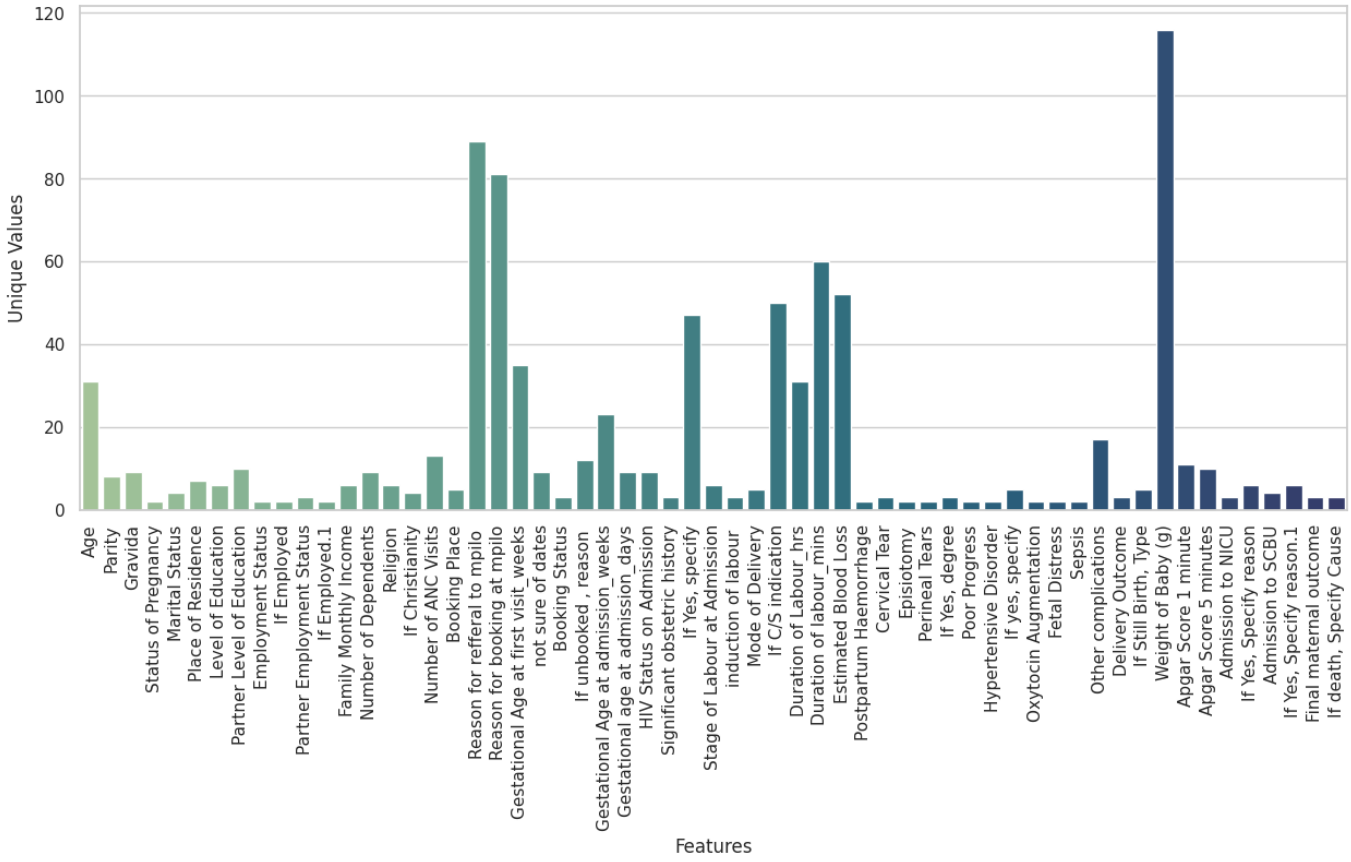


Figure 4.1: Unique values for each feature

Missing values identification

Missing data can be grouped into 3 categories:- Missing Completely At Random (MCAR), Missing At Random (MCAR) and Missing Not At Random (MAR). MCAR occurs when missingness is a result of reasons unrelated to the observed data. This way, the probability of missingness is the same for every observation. In MAR, data missingness can be explained by observed variables. The probability of missingness is the same and related to conditional groups in the observed data. Finally, in MNAR the probability of missingness is unknown. In this case, neither MCAR nor MAR is true [24].

Missing values identification and treatment are a critical step that helps in determining the extent of missingness and whether a feature should be retained or excluded from the data. The existence of missing values in analysis reduces the statistical power of the study, resulting in biased estimates and eventually wrong conclusions. Identification and eventual elimination of missing values are necessary. In this project, dt.isnull().sum() function was

used to identify all missing values.

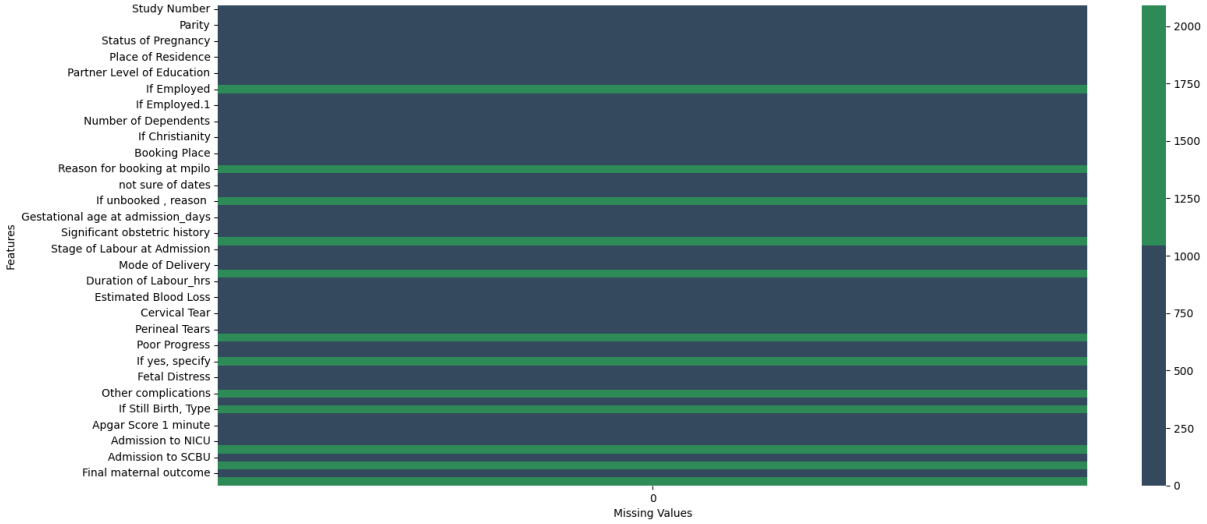


Figure 4.2: Missing data in features

Generally, about 12 features had significant missing data. However, a number of these features, especially those with descriptive responses, for example, If Employed.1, If Christian, If Unbooked, reason, If Yes, specify, If C/S indication, If Yes, degree, etc, are not useful for analysis and shall be excluded during data processing. The percentage of missingness was assessed.

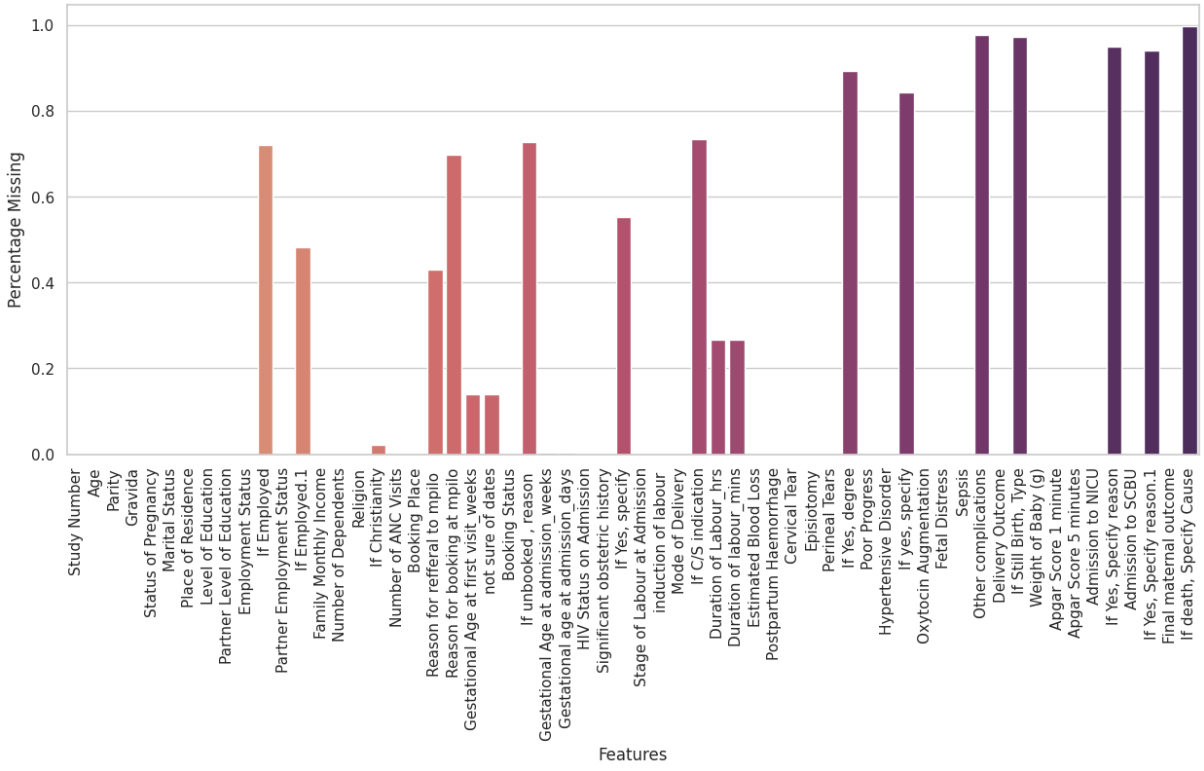


Figure 4.3: Percentage missingness

For this project, any feature with over 50%missingness was dropped from the analysis. The other remaining variables were to establish if they should be retained as part of the dataset.

Duplicate data

Duplicate data results in inaccurate information and a lack of data integrity. It may result in inaccurate computations and conclusions based on inaccurate information. The duplicated().sum() functions were used to identify duplicate records. In this dataset, there were no duplicates.

Outliers identification

Outliers are data points that significantly differ from other observations in a dataset in the context of data analytics. If outliers are not well controlled, they can have a substantial negative influence on data analysis, frequently distorting findings and producing false conclusions. A boxplot was used to identify outliers in each numeric feature in the dataset.

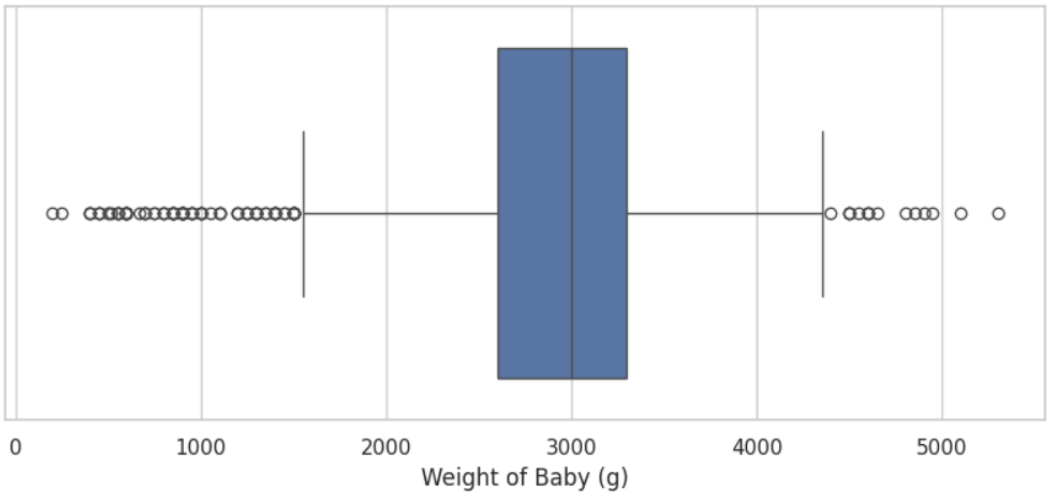


Figure 4.4: Outliers in Weight of Baby (g) feature

Several features, Estimated Blood Loss, Weight of Baby (g), Apgar Score 1 minute, Apgar Score 5 minutes, Duration of Labour_mins, Duration of labour_mins contain outliers because the minimum value is way apart from the mean and maximum values. This also means that the features are skewed. We shall confirm this using box plots.

Garbage data identification

Garbage cleaning is mainly important for categorical variables. It helps identify categories or string characters that may appear on records and need to be collapsed into one category or completely removed from the data. Garbage can also be removed from variables or features, especially if it does not conform to the variable naming standards.

Garbage values associated with categorical type features were seen, especially in critical variables like Mode of Delivery, where certain categories were wrongly spelled and created a new category. For example, in Mode of Delivery, there exist breech and Breech categories that seem similar, and of one category. These categories were collapsed into one category, Breech.

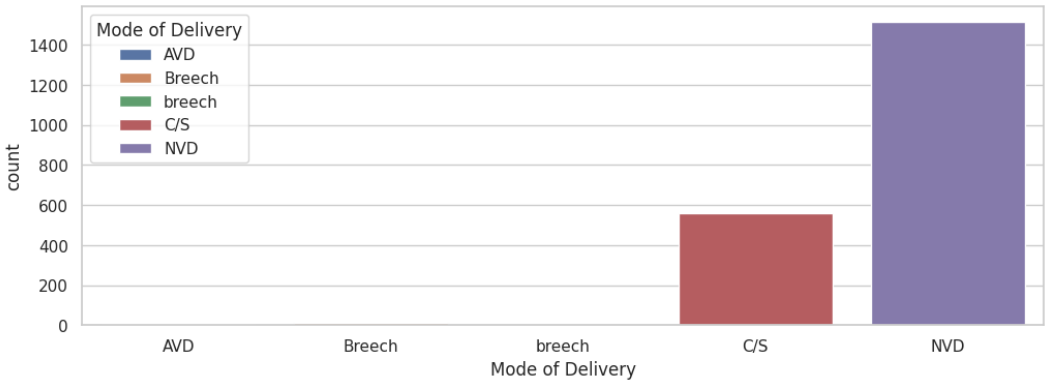


Figure 4.5: Garbage in Mode of Delivery feature

Distribution of features

Normal distribution, also known as the Gaussian distribution, is a continuous probability distribution function with a mean that is symmetrical, and most data falls within one standard deviation of the mean. It is critical to use normally distributed features in machine learning models because they are designed on the assumption that bivariate and multivariate features are normally distributed. Data pre-processing is critical to get features close to a Gaussian distribution.

The features Apgar Score 1 minute and Apgar Score 5 minutes are left-skewed, whereas Estimated Blood Loss, Duration of Labour_mins, Parity, Gravida, Number of Dependents, Number of ANC visits are all right-skewed. Weight of Baby (g) depicted a normal distribution. Data pre-processing was necessary to treat outliers and impute missing values to remove their effects on the distribution.

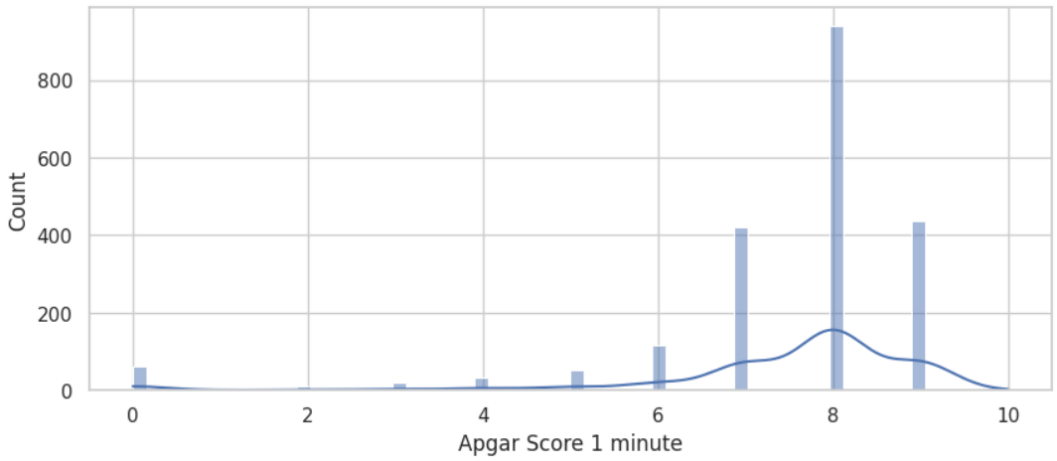


Figure 4.6: Left-skewed

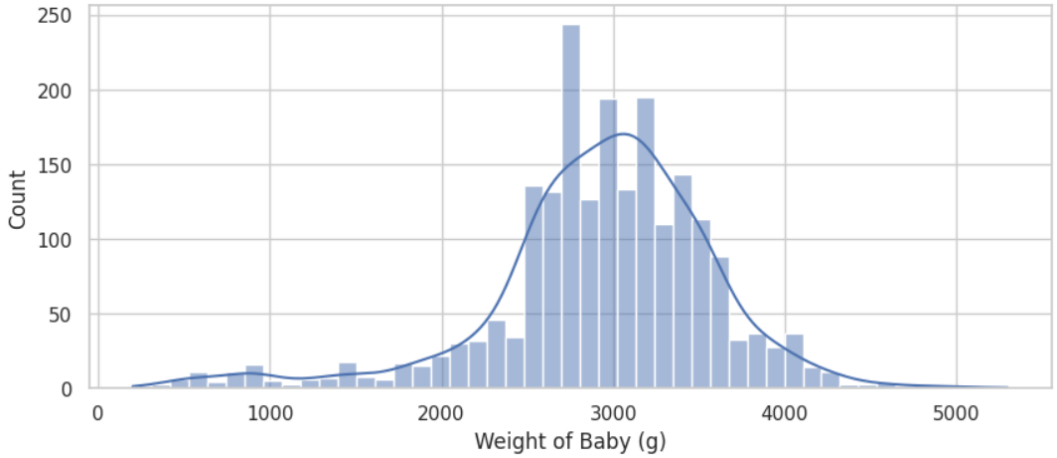


Figure 4.7: Guassian distribution for Weight of Baby (g)

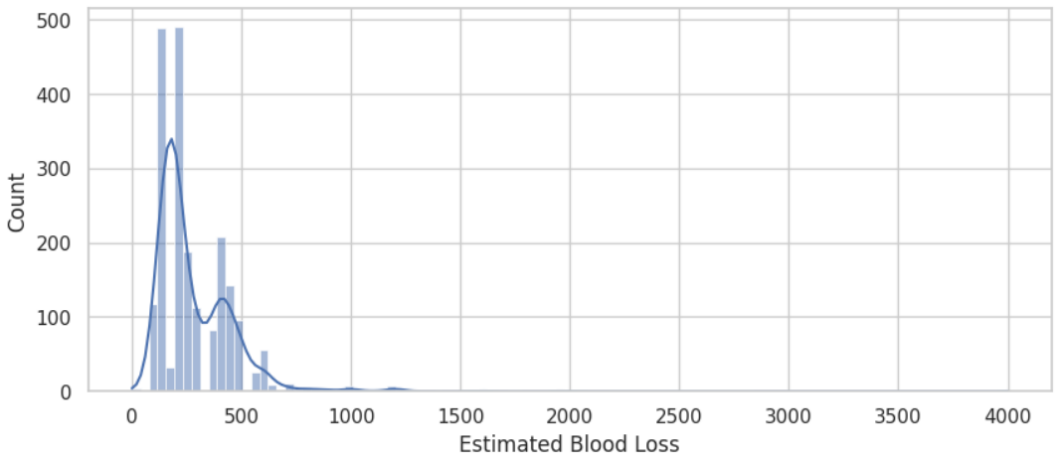


Figure 4.8: Right-skewed distribution for Estimated Blood Loss

Statistical exploration

Statistical exploration focused on descriptive statistics of the data. Descriptive statistics give insights distribution tendency of each variable. dt.describe() was used to obtain descriptive statistics.

Descriptive statistics revealed the following the mean age of women in the study was 26 years. The youngest woman was 14 years old, and the oldest woman was 44 years old. The mean Estimated Blood Loss was 282.0mL, with a minimum value of 0 and a maximum Estimated Blood Loss of 4000mL. The Estimated Blood Loss of 0 is potentially missing data. Descriptive statistics for numeric features are shown below.

	count	mean	std	min	25%	50%	75%	max
Study Number	2094.0	1047.531996	604.641786	1.0	524.25	1047.5	1570.75	2094.0
Age	2094.0	25.936963	6.474681	14.0	21.00	25.0	31.00	44.0
Parity	2094.0	1.059694	1.207861	0.0	0.00	1.0	2.00	7.0
Gravida	2094.0	2.142311	1.274842	1.0	1.00	2.0	3.00	9.0
Number of Dependents	2094.0	1.435530	1.435472	0.0	0.00	1.0	2.00	8.0
Number of ANC Visits	2094.0	3.046800	2.055305	0.0	1.00	3.0	4.00	12.0
Duration of Labour_hrs	1536.0	11.485026	4.792751	2.0	8.00	11.0	14.00	38.0
Duration of labour_mins	1537.0	30.086532	16.790132	0.0	15.00	30.0	45.00	59.0
Estimated Blood Loss	2093.0	282.003822	187.736164	0.0	150.00	200.0	400.00	4000.0
Weight of Baby (g)	2094.0	2934.510506	661.114167	200.0	2600.00	3000.0	3300.00	5300.0
Apgar Score 1 minute	2094.0	7.435530	1.790349	0.0	7.00	8.0	8.00	10.0
Apgar Score 5 minutes	2094.0	8.456543	1.833169	0.0	8.00	9.0	9.00	10.0

Figure 4.9: Right-skewed distribution for Estimated Blood Loss

Descriptive statistics for categorical features show the count (number of rows with responses), unique (number of unique values), top (mode category), and frequency (how often the mode category appears in the dataset).

Descriptive statistics for categorical variables revealed that most women in this dataset were married (1311), most women preferred normal vaginal delivery, most women went to school and attended O level (1482), most partners were employed in formal sector, christianity was the most dominant religion, most women booked for delivery

Additionally, correlation among numerical features was explored. Correlation assesses the relationship between variables using a correlation matrix.

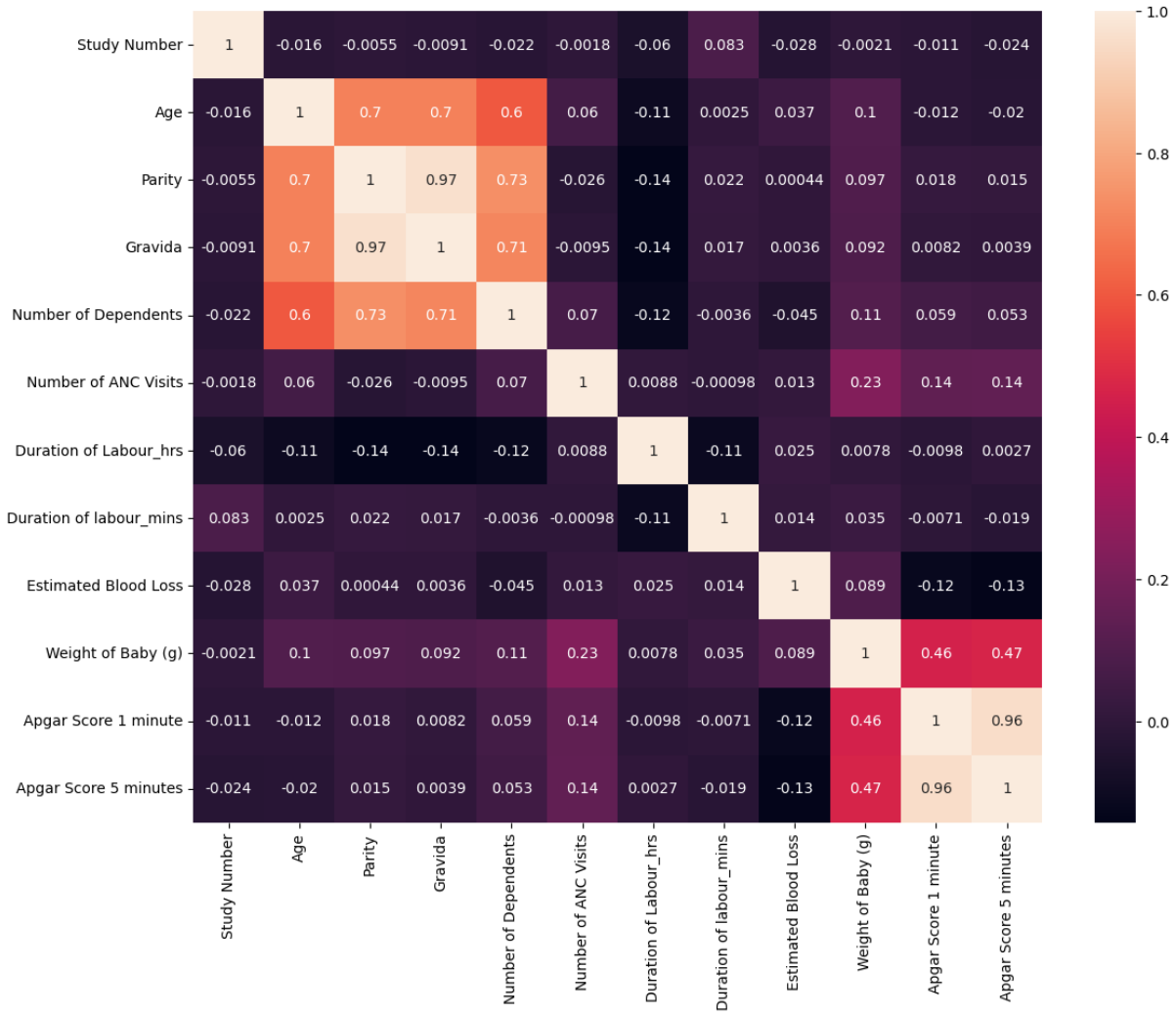


Figure 4.10: Correlation among continuous variables

Inconsistencies in data

Inconsistency was seen during exploratory data analysis, where the postpartum hemorrhage variable observation did not conform to the definition of postpartum hemorrhage. This necessitated the creation of a new prediction variable based on Estimated Blood Loss that conforms to the definition of postpartum hemorrhage, and is to be used as a prediction variable.

4.3 Data pre-processing

This critical step involved addressing the data issues identified during the data exploratory phase. In summary, this involved dropping from the dataset variables that are not useful for this analysis, trimming the excess spaces in values, for example, spaces in Mode of Delivery, making them very easy to manipulate, treatment of missing values, removal of garbage data, treatment of outliers, etc.

Dropped features

The following features were dropped from the dataset:- Reason for referral to mpilo, Apgar Score 1 minute, Study Number, Apgar Score 5 minutes, Admission to NICU, Admission to SCBU, Final maternal outcome, If Employed.1, and Delivery Outcome. The features measured post-delivery outcome that are not relevant for this analysis.

Additionally, features with more than 50% missingness were dropped from the dataset. The features retained consisted of about 25% missingness, which can be imputed and may not create a dominant value or category in the dataset.

For features values, records were dropped where estimated Blood Loss was less than 100mL and where Mode of Delivery was not Normal Vaginal Delivary (NVD) and Cesarean Section (C/S).

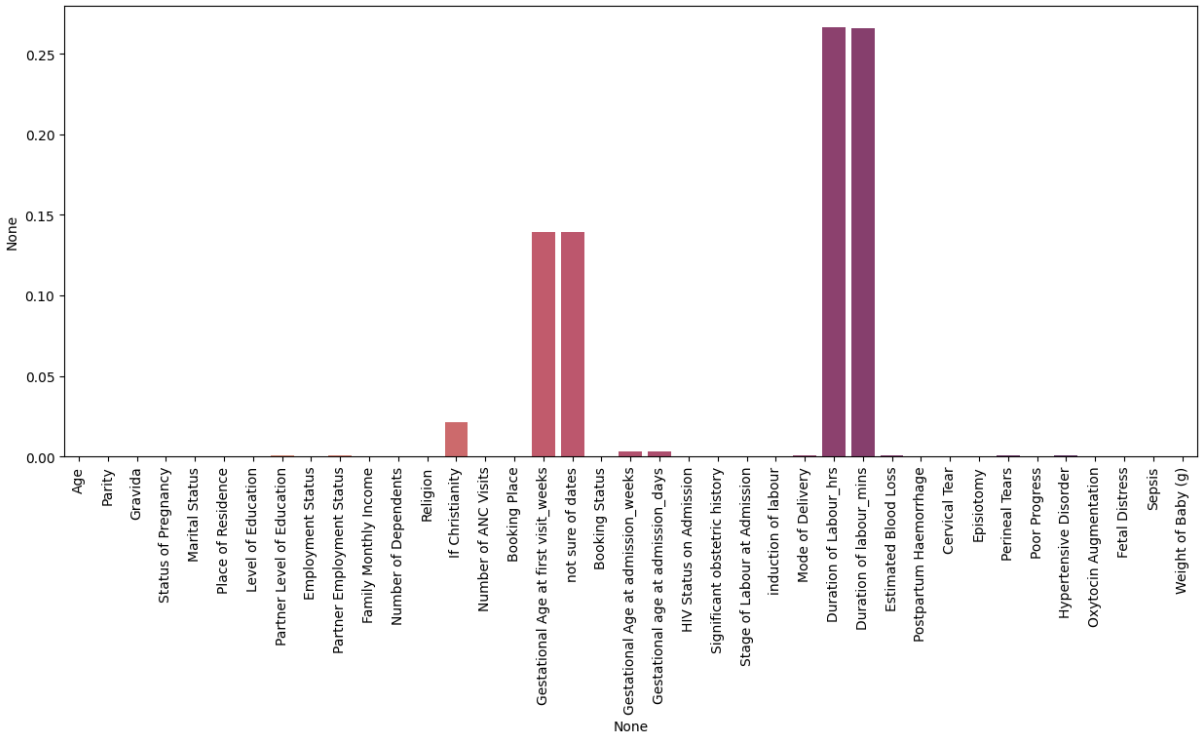


Figure 4.11: Percentage missingness reduced to around 25%

Dropping non-useful data altered the shape of the dataset. The records were reduced from 2094 to 2067 records, and the features were reduced from 60 to 39 features.

Garbage data Treatment

Garbage data addressed the inconsistency in the values of categorical features. For Mode of Delivery, breech and breach were replaced with Breech. For Partner Level of Education, Uknown, and unknown were replaced with Unknown. Features values "not sure of date" in Gestational Age at admission_weeks, Gestational age at admission_days, and Gestational Age at first visit_weeks were replaced with null value, and the datatype was converted from object to numeric.

Missing data Treatment

The dataset consisted of all the above forms of missingness, for example, missing values in Estimated Blood Loss were Missing Completely At Random either because it was not measured or recorded. Other variables depicted Missing At Random.

Missing data in categorical variables was addressed by creating a new category, "Other". This limits the contamination of observed categorical variables. Whereas missing values in continuous variables were replaced with the median. In a measure of central tendency, especially in a dataset with outliers, the median is very robust compared to the mean because it does not shift with the presence of outliers.

Outlier Treatment

Outlier points were identified in several features during the exploratory data analysis phase. Features like Parity, Gravida, Number of Dependents, Number of ANC Visits, Duration of labour_hrs, and Weight of Baby g all contained outliers and required outlier treatment. Inter-Quartile Range was utilized to identify outliers.

For each feature, data points were ordered from the lowest to the highest. The first quartile (25th quantile) and third quartile (75th quantile) were established using the quantile() function. Inter-Quartile Range was calculated by subtracting the first quartile (25th quantile) from the third quartile (75th quantile). The range was calculated to establish the lower bound defined as $First \quad Quartile - 1.5 * IQR$ and upper bound defined as $Third \quad Quartile + 1.5 * IQR$. For each feature, data points below the lower bound and above the upper bound were considered outliers and replaced with the value of the lower bound and upper bound, respectively.

Estimated Blood Loss feature consists for measure for bother vaginal and cesarean mode of delivery and its potentially susceptible to outliers because measure of blood loss varies between the two modes of delivery. Extreme values for Estimated Blood Loss were removed and replaced with logical values.

Despite the removal of outliers in the dataset, several numeric features remain skewed either to the left or right, for example, Parity, Gravida, Number of dependents, and Estimated Blood Loss, among others. Unfortunately, there is never anything life perfect data.

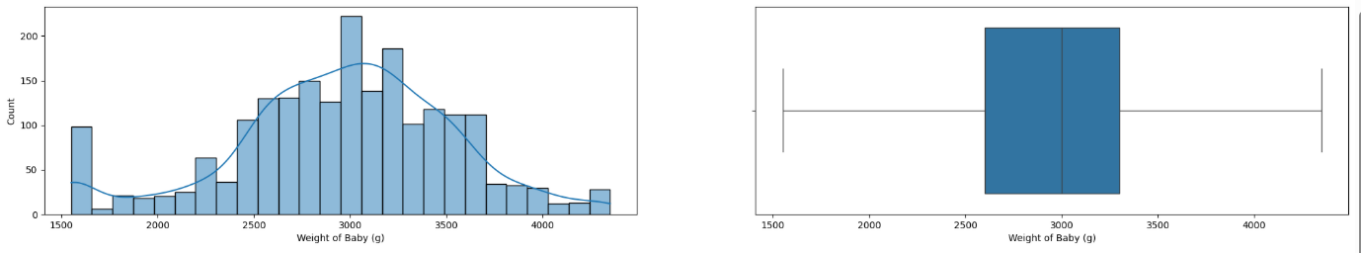


Figure 4.12: Outliers treated in Weight of Baby (g)

4.3.1 Generate outcome variable "pph"

By definition, Postpartum Hemorrhage is estimated loss of blood of $\geq 500mL$ for vaginal delivery and $\geq 1000mL$ for cesarean delivery. The existing Postpartum Haemorrhage feature did not conform to that definition, necessitating the generation of a new feature that conforms to Postpartum Hemorrhage definition. Using the features, Estimated Blood Loss and Mode of Delivery, the pph feature was generated and added to the dataset.

4.4 Exploratory Analysis

The goal of exploratory data analysis (EDA) is to identify the key features of a dataset by examining patterns and correlations through visualizations. Creating questions, visualizing and processing data, and then using the outcomes to answer those questions.

Postpartum Hemorrhage classes

After data pre-processing, the prevalence of pph in the dataset is at 2.8

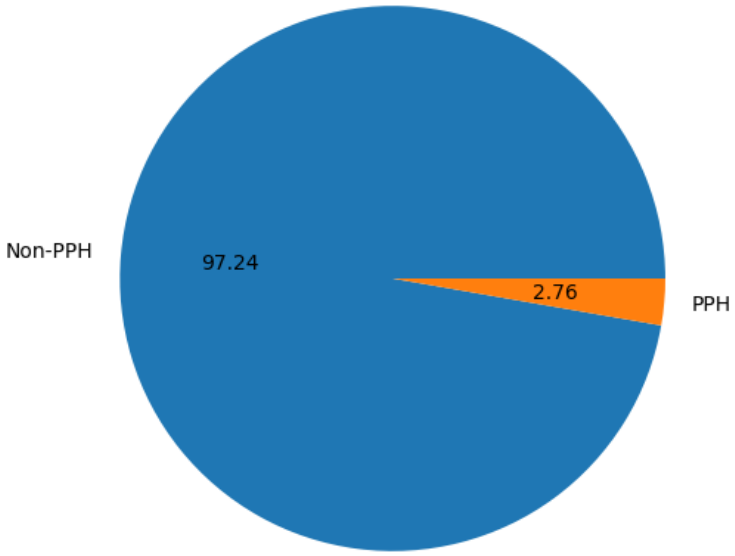


Figure 4.13: Postpartum Hemorrhage classes

Bivariate Analysis

Bivariate analysis focused on the comparison between key features in our dataset. The features considered for bivariate analysis are pph, Mode of Delivery, Estimated Blood Loss, Age_Category (generated from Age), Weight of Baby (g), Duration of Labour (also a composite of Duration of labour_mins and Duration of Labour_hrs), and Number of ANC visits.

Mode of Delivery

The analysis considered two delivery modes; Vaginal and Cesarean modes of deliveries with Vaginal delivery being a more dominate category compared to Cesarean. Similarly, there were more Postpartum Hemorrhage cases in vaginal mode of delivery compared to Cesarean.

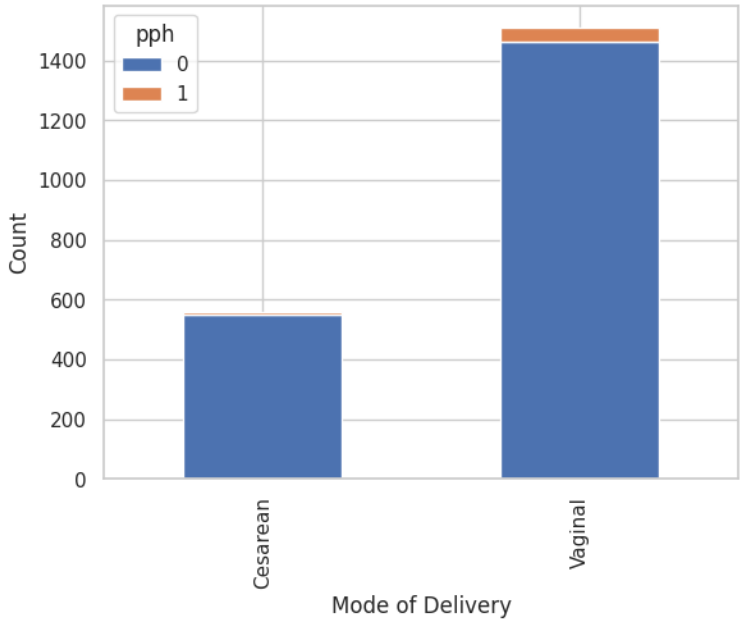


Figure 4.14: Distribution: Mode of Delivery

Comparing Mode of Delivery and Estimated Blood Loss

The analysis showed that more women who had cesarean section experience increased blood loss compared to women who had vaginal deliveries. This is expected because cesarean section involves incision through the mother's abdomen and eventually leads to bleeding.

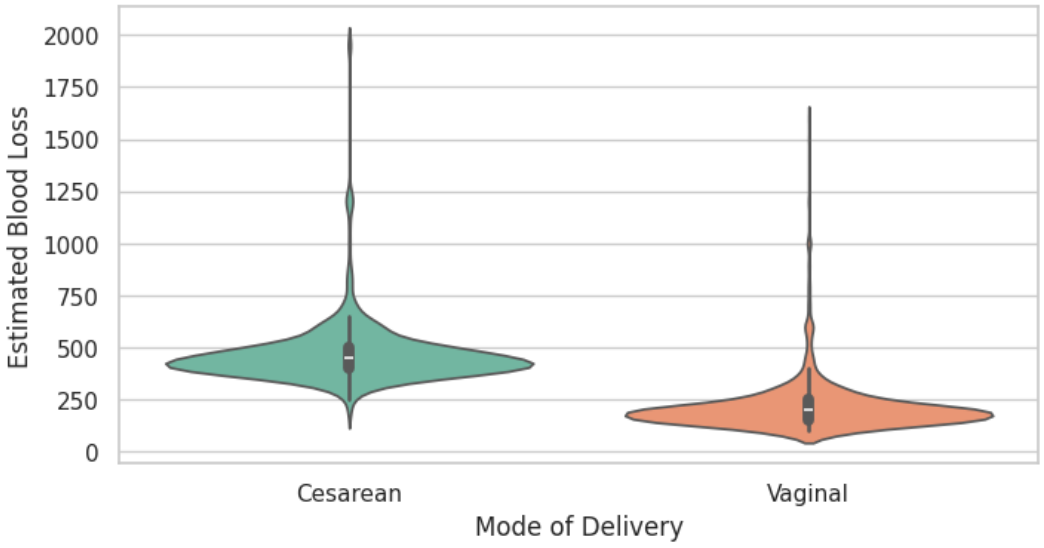


Figure 4.15: Comparison of Mode of Delivery and Estimated Blood Loss

Comparing Mode of Delivery Age Category

The analysis revealed that, generally there were very few women who had cesarean section compared to normal vaginal delivery. Additionally, women between 14 and 34 years of age preferred cesarean section compared to those 35 years and above. This finding is in agreement with Isaac Waniala et al, whose study aimed to to assess prevalence, indications, and community perceptions of cesarean section Delivery in Ngora district found that women preferred vaginal route of delivery compared to cesarean [25].

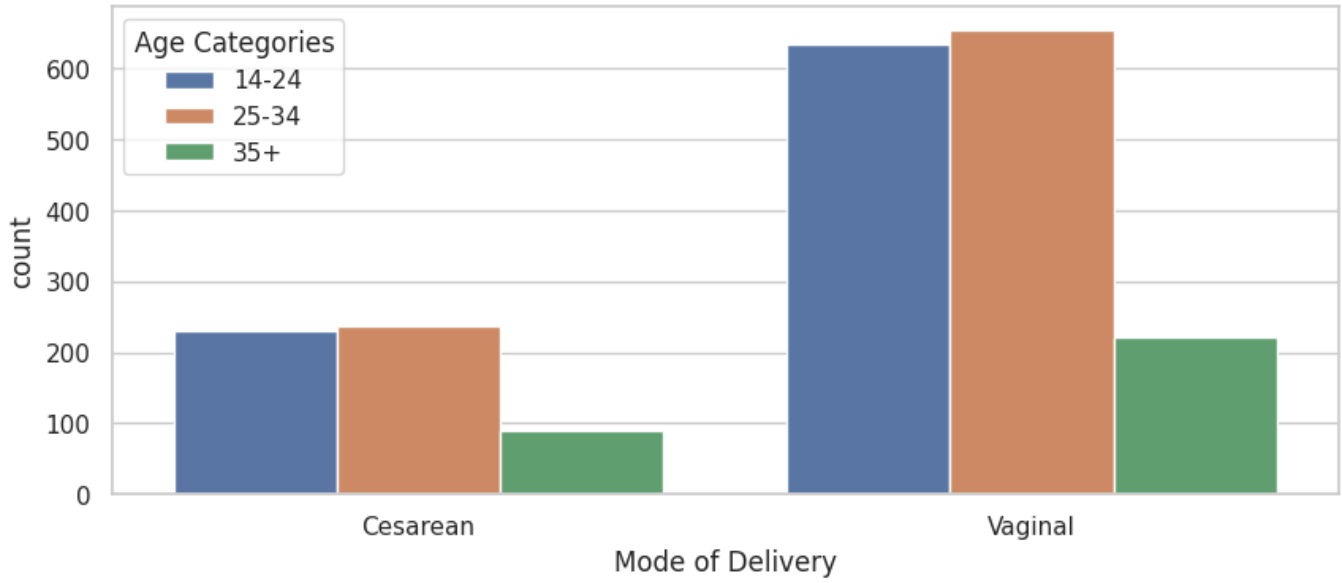


Figure 4.16: Comparison of Mode of Delivery and Estimated Blood Loss

Number of ANC Visits

Antenatal Care (ANC) is critical for pregnant women for the health of the fetus and mother. According to the World Health Organization's (WHO) new guidelines, pregnant women should complete at least 8 visits during pregnancy [26]

However, the analysis revealed very poor ANC attendance with a median attendance of 3 visits, which is below the WHO recommendation. The analysis further reveals that more than half of the women had less than 4 ANC visits at the facility, about 800 women had between 5 and 8 ANC visits, and fewer than 50 women had more than 8 ANC visits.

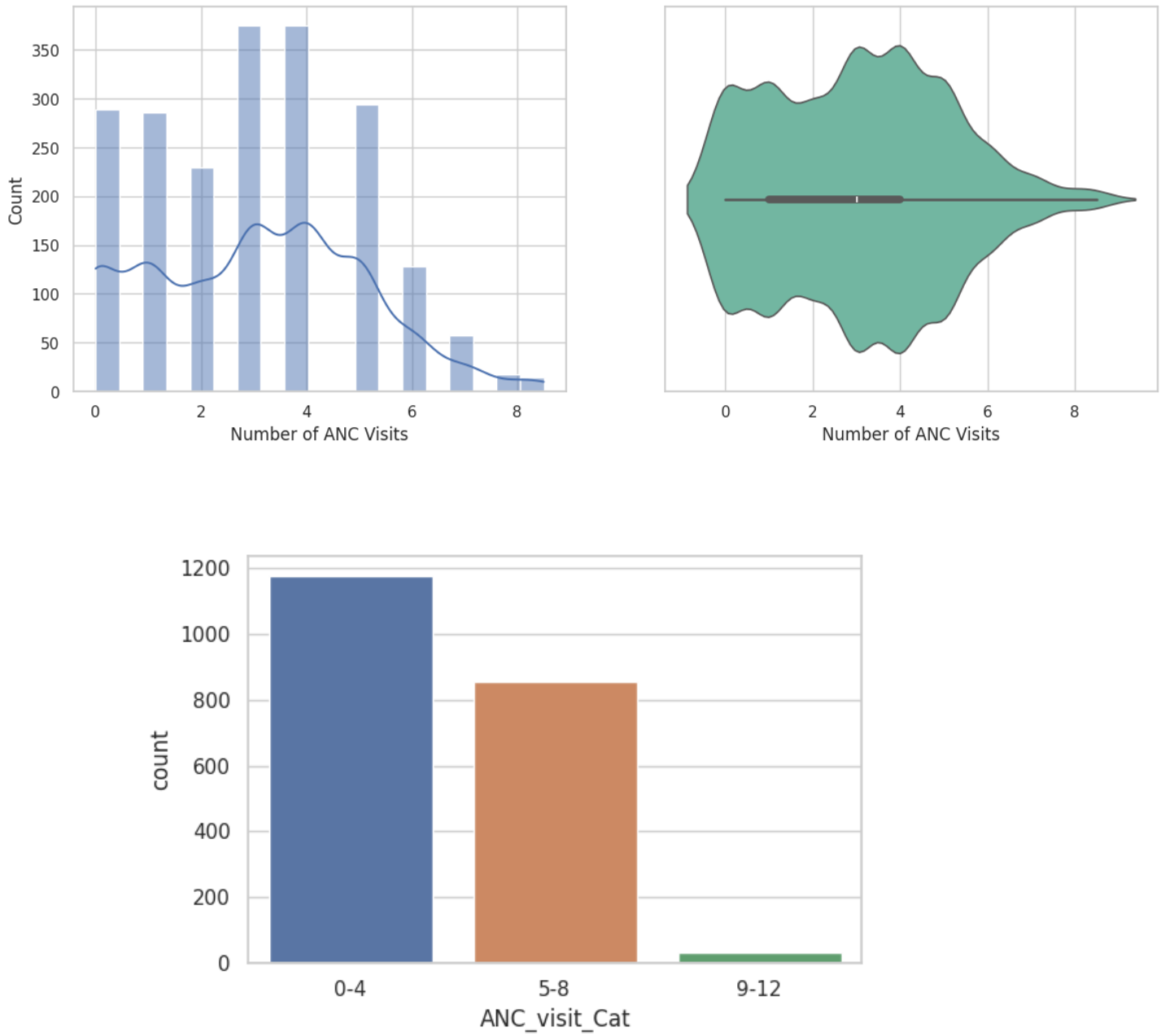


Figure 4.17: Distribution: Number of ANC Visits

Comparing Number of ANC Visits and pph

The visual plots show a varying median between the postpartum hemorrhage categories, indicating a high correlation between the two variables.

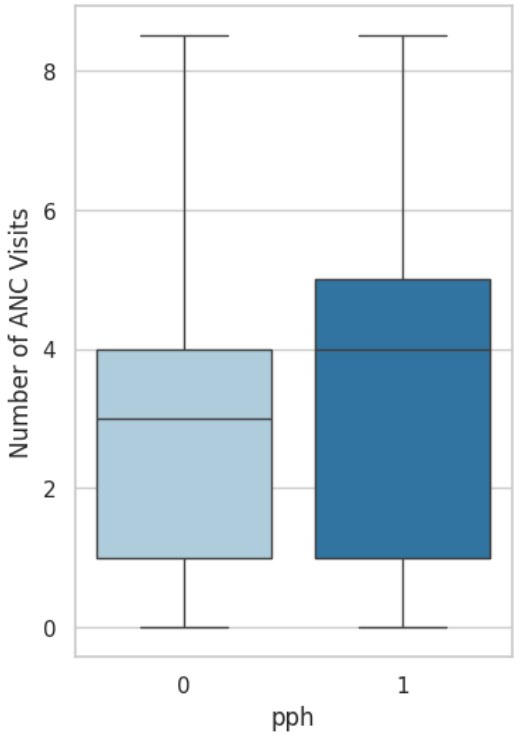


Figure 4.18: Comparison of Number of ANC Visits and pph

Comparing Number of ANC Visits and Age Categories

The visual plots show very similar medians between the Age Categories and Number of ANC Visits. The median is similar across all categories. The ANOVA test shall indicate if there is a correlation or not between the two variables.

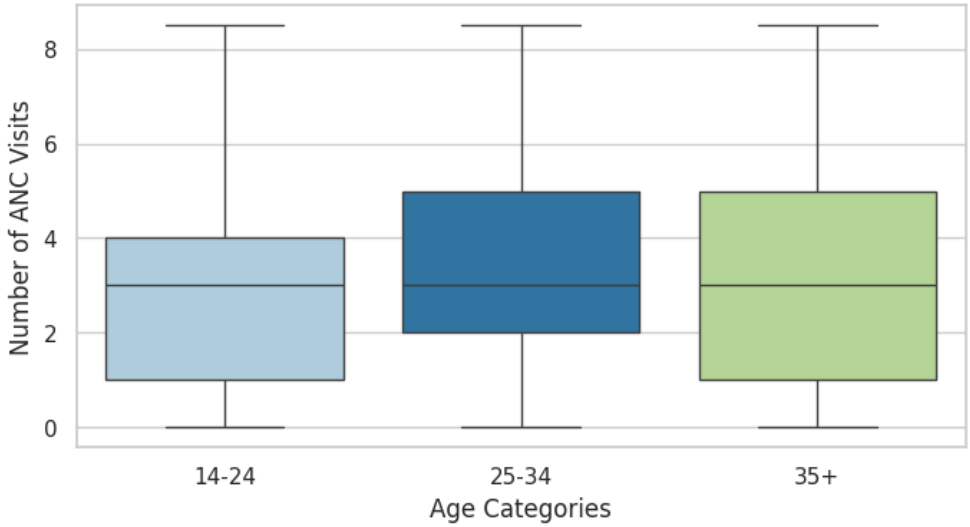
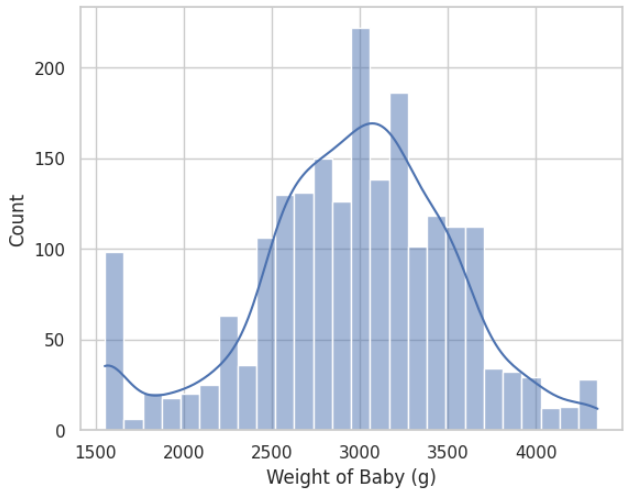


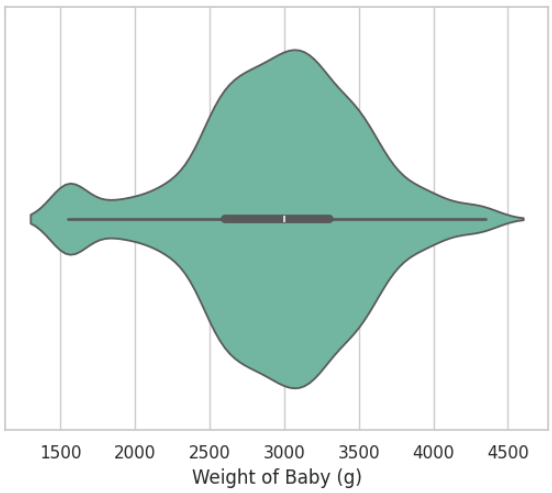
Figure 4.19: Comparison of Number of ANC Visits and Age Categories

Weight of Baby (g)

The weight of the baby is one of the documented risk factors of postpartum hemorrhage. Macrosomic baby is a newly born baby with weight of greater than or equal to 4000 grams (about 8 pounds and 13 ounces).

The data analysis revealed that the median baby weight was 3000 grams and ranged from 200 to 5300 grams. It is important to note that most of the Weights below 1000 grams were miscarried deliveries.





Comparing Weight of Baby (g) and Age

The analysis shows that women above 24 years old delivered slightly overweight babies compared to those between 14-24 years old. Additionally, there is potentially a correlation between Weight of Baby and Age Categories. The median is not similar across all categories. Additional comparison with Age show a slightly significant positive correlation between Weight of Baby and Age. This means that as women grow older they tend to produce larger babies.

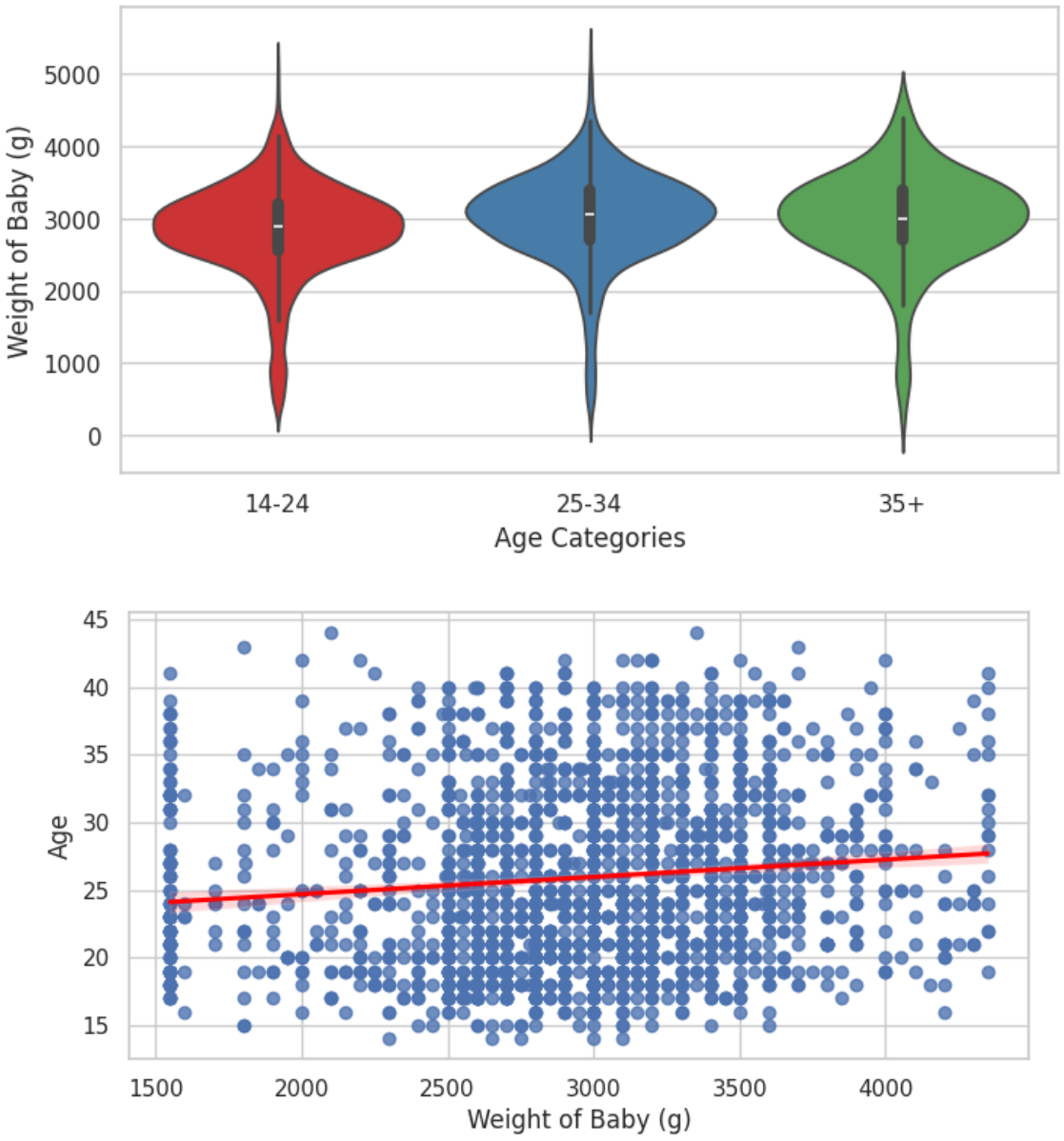


Figure 4.20: Weight of Baby (g) and Age

Comparing Weight of Baby (g) and pph

The analysis showed a positive correlation between the Weight of Baby and pph, with an increase in the weight of Baby increasing pph. Based literature reviewed, women with macrosomic babies are at risk of experiencing postpartum Hemorrhage.

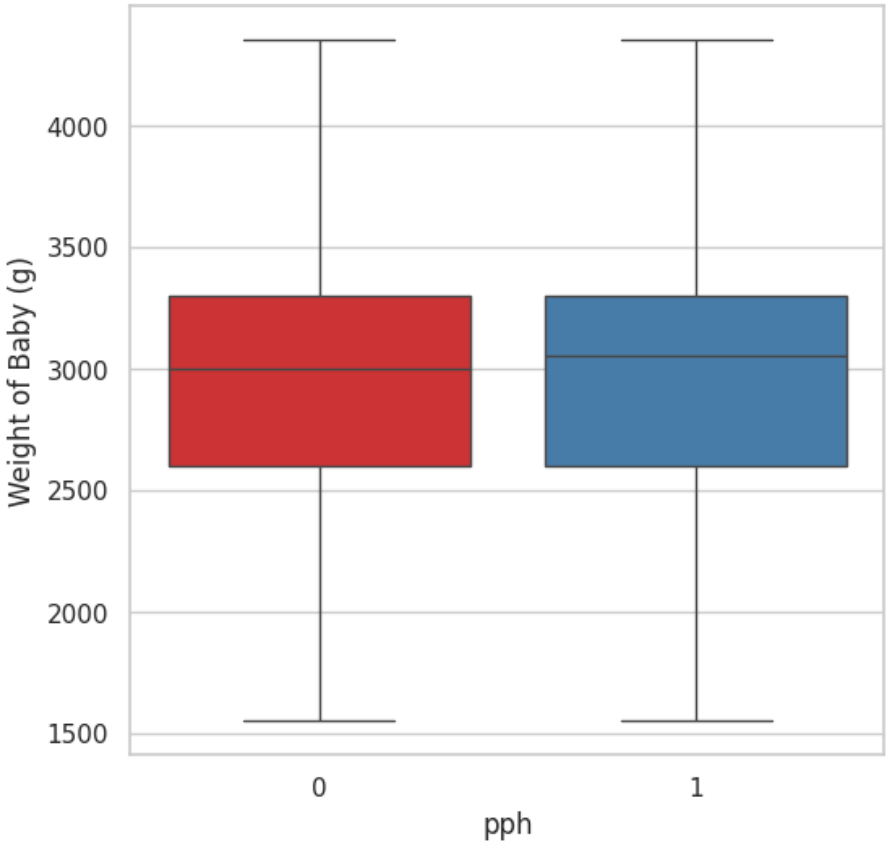
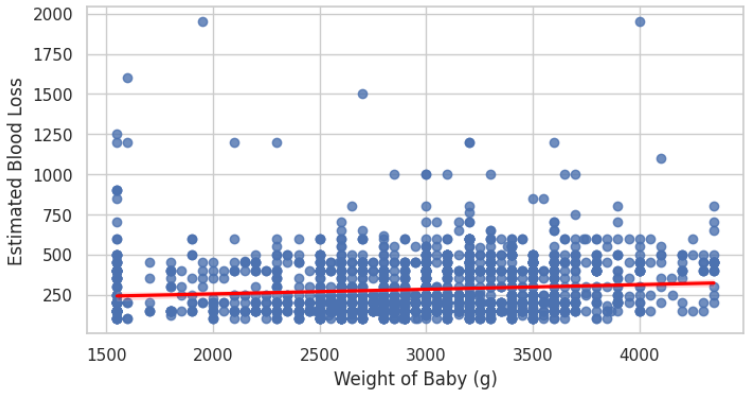


Figure 4.21: Weight of Baby (g) and pph

Comparing Weight of Baby (g) and Estimated Blood Loss

The analysis attempts to establish if there is any association between Estimated Blood Loss and Weight of Baby (g). There is no strong correlation. However, with an increase in baby weight, there seems to be an increase in Estimated Blood Loss. This varied between modes of delivery with slightly significant positive correlation in Vaginal mode of delivery compared to cesarean mode of delivery.



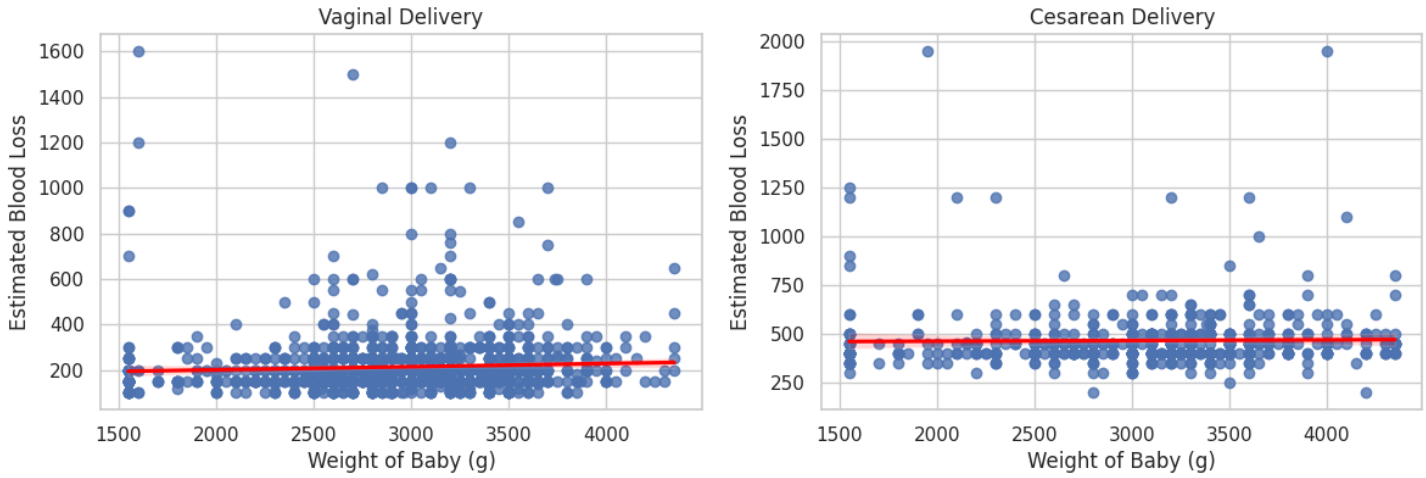
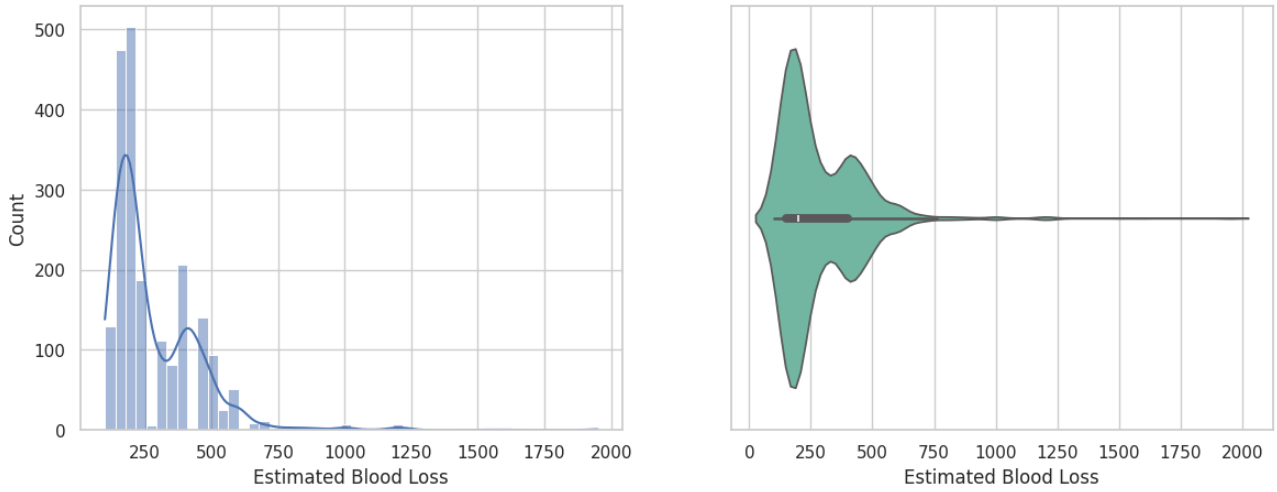


Figure 4.22: Comparison of Weight of Baby (g) and Estimated Blood Loss

Estimated Blood Loss

Estimated blood loss, or simply blood loss, is often used to determine if a particular maternity case is PPH or not, and it is highly dependent on the mode of delivery. Unfortunately, estimation of blood loss is still a challenge especially in low-income settings with visual estimation of blood still common in obstetrics wards.

Generally, median Estimated Blood Loss was about 200mL. However, this may vary depending on the mode of delivery. Considering that there are many records for vaginal delivery, this median value is potentially biased toward vaginal mode of delivery. This can be further assessed by looking at the different modes of delivery.



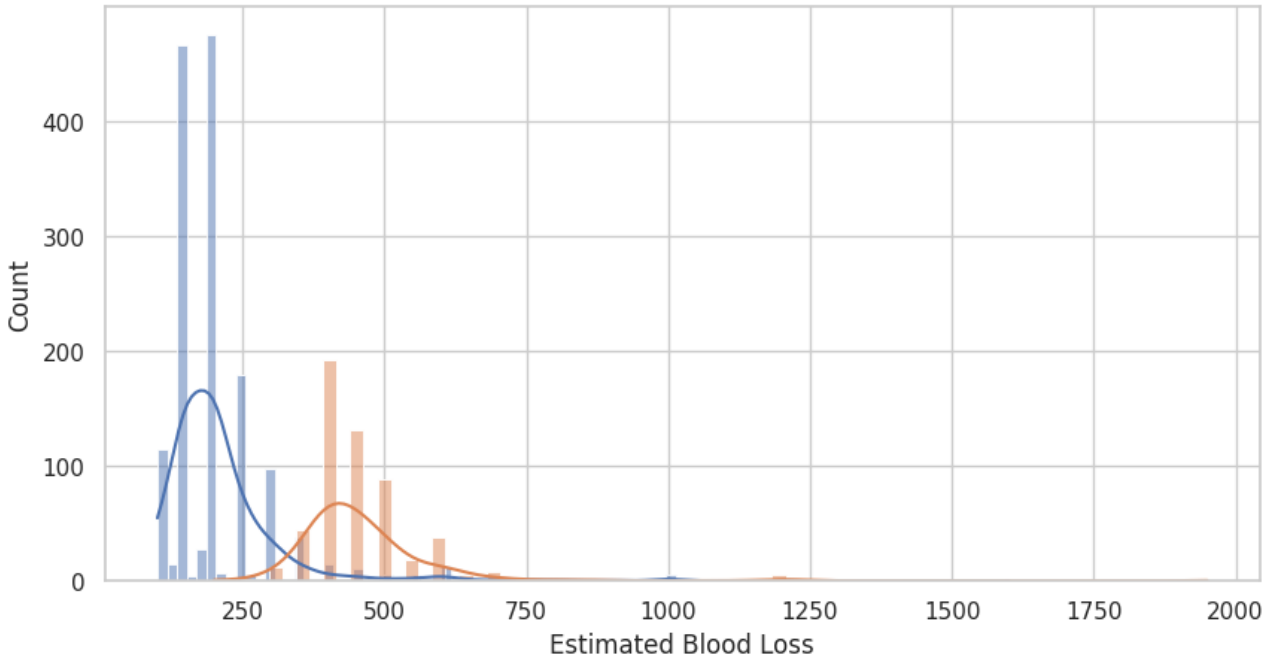


Figure 4.23: Distribution: Estimated Blood Loss

As earlier observe, the median Estimated Blood Loss is different between vaginal and cesarean deliveries with about 200 and 450 respectively.

4.5 Statistical and Visual Correlation

Statistical and Visual Correlation were performed on all features in the dataset to determine the features that are highly associated with PPH.

For visual correlation, a boxplot was used for correlation analysis between continuous independent and categorical dependent variables. For each pair of continuous independent and categorical dependent variables, correlation is true if the median of each category of the dependent variable varies. If the median is similar, then there is no correlation between the continuous independent variable and the categorical dependent variable.

Additionally, the correlation matrix revealed the correlation between the numerical variables. The correlation matrix helped determine if numerical variables exhibited a linear positive or negative correlation. The correlation matrix analysis showed that Parity and Gravida are positively correlated with a correlation measure of 0.97. Number of Dependents and Age, Parity, Gravida are positively correlated with correlations of 0.61, 0.74, and 0.71,

respectively.

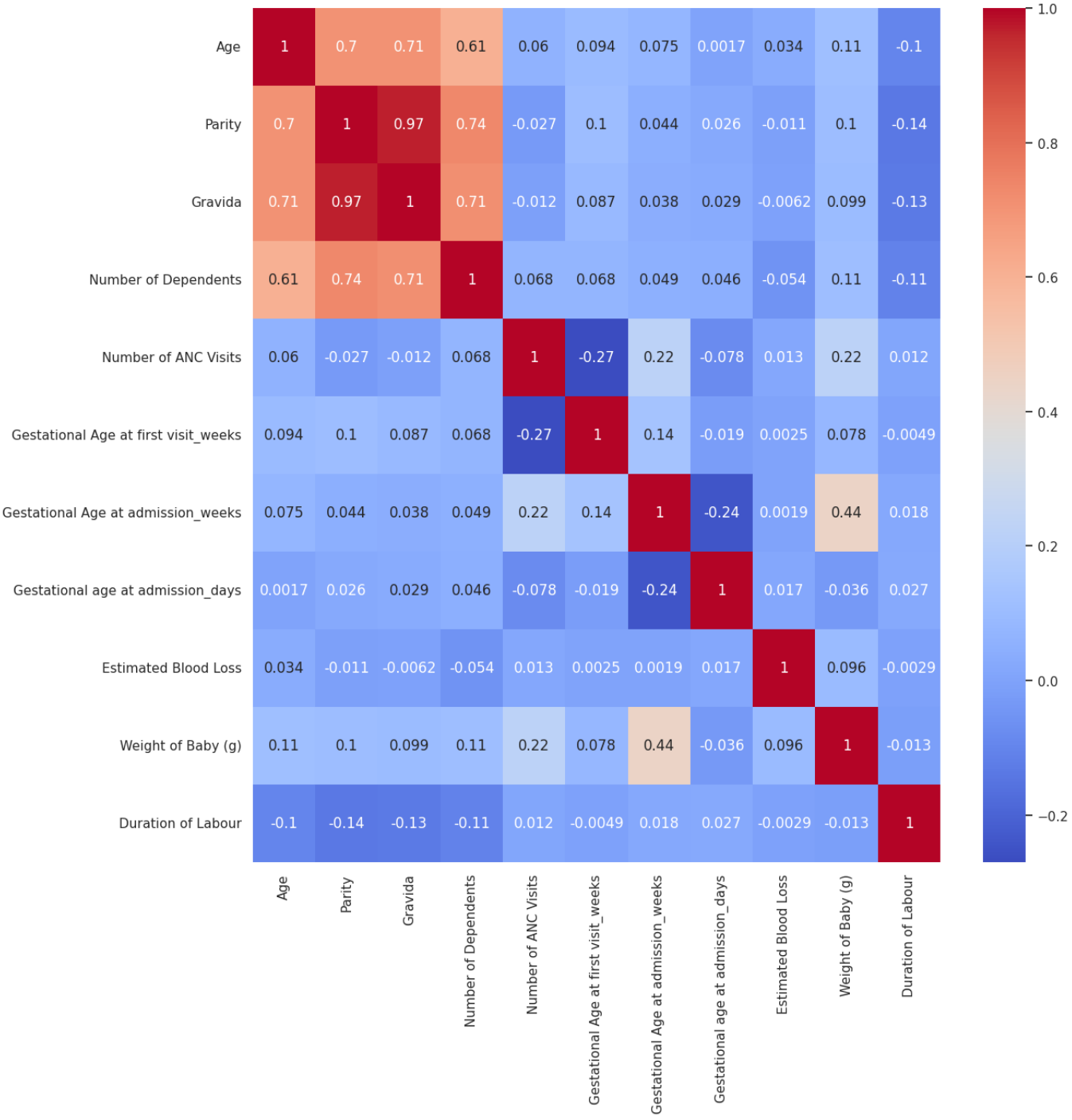


Figure 4.24: Correlation matrix

4.5.1 ANOVA Test

The ANOVA test was used to check for correlation between continuous independent and categorical dependent variables. The null hypothesis and alternative hypothesis were tested.

- H_0: A correlation does not exist between independent continuous and categorical dependent variables. The P-value is less than 0.05.
- H_a: A correlation exists between independent continuous and categorical dependent variables. The P-value is not less than 0.05.

A barplot was used to visualize the correlation between categorical independent variables and categorical dependent variables. For each pair of categorical independent and categorical dependent variables, correlation is true if the distribution across all categories varies. If the distribution across all categories is similar, then there is no correlation between the categorical independent and categorical dependent variable.

4.5.2 Chi-Square Test

The Chi-Square test was used to check for correlation between categorical independent and target variables. The null hypothesis and alternative hypothesis were tested.

- H_0: A correlation does not exist between categorical independent and dependent variables. The P-value is less than 0.05.
- H_a: A correlation exists between categorical independent and dependent variables. The P-value is not less than 0.05.

Feature	P-value
Age	0.58
Parity	0.98
Gravida	0.89
Number of Dependents	0.84
Number of ANC Visits	0.00
Gestational Age at first visit_weeks	0.06
Gestational age at admission_days	0.92
Weight of Baby (g)	0.01
Duration of Labour	0.00
Status of Pregnancy	1.00
Marital Status	0.92
Place of Residence	0.68
Level of Education	0.85
Partner Level of Education	0.94
Employment Status	0.30
Partner Employment Status	0.92
Family Monthly Income	0.84

Feature	P-value
Religion	0.74
If Christianity	0.94
Booking Place	0.68
not sure of dates	1.00
Booking Status	0.85
HIV Status on Admission	0.48
Significant obstetric history	0.31
Stage of Labour at Admission	0.83
induction of labour	1.00
Mode of Delivery	0.14
Cervical Tear	0.00
Episiotomy	0.00
Perineal Tears	0.00
Poor Progress	0.37
Hypertensive Disorder	0.93
Oxytocin Augmentation	0.32
Fetal Distress	0.72
Sepsis	0.32
Microsomia	0.79

Figure 4.25: Summary of correlation analysis

4.5.3 Variance Inflation Factor (VIF)

Regression analysis uses the Variance Inflation Factor (VIF) as a tool to identify and measure multicollinearity, or strongly correlated independent variables in a model. If the correlation between two independent variables is equal to 1 or 1, as in the equation above, we have perfect multicollinearity. If the VIF measure between independent variables is between 1 and 5, then we have medium multicollinearity, and a VIF measure between 5 and 10 is very high multicollinearity, and it is considered very serious. Perfect multicollinearity in data is uncommon in practice. More frequently, when two or more independent variables have an approximate linear correlation, multicollinearity becomes a problem [27].

The purpose of regression analysis is to illustrate the direction and intensity of the relationship between dependent and one or more independent variables, or factors. However, in reality, the only restrictions on the amount of possible features incorporated into a regression model are the creativity and ability to collect the relevant data.

4.5.4 Measures of Variance Inflation Factor

The high risk factors for Postpartum Hemorrhage were assessed for multicollinearity using Variance Inflation Factor (VIF). Cervical Tear and Episiotomy had a VIF measure of 1.0 and 1.4, indicating perfect correlation. Number of ANC Visits indicated a medium correlation with VIF measures of 3.4. Duration of Labour, Perineal Tears, and Weight of Baby (g) VIF measures were 8.7, 10.4, 15.2, respectively, indicating a high correlation between variables.

Variance Inflation Factor showed that only two features, Cervical Tear (1.0) and Episiotomy (1.4), would be perfect for the prediction of Postpartum Hemorrhage. However, using a limited number of features leads to underfitting; the model fails to capture the data patterns and performs poorly. A correlation matrix was used to determine variables highly correlated with each other.

Feature	VIF Measure
Cervical Tear	1.0
Episiotomy	1.4
Number of ANC Visits	3.4
Duration of Labour	8.7
Perineal Tears	10.4
Weight of Baby (g)	15.2

Figure 4.26: Variation Inflation Factor measure

4.6 Model Development

Supervised machine learning model algorithms were used to develop a machine learning model to predict Postpartum Hemorrhage. Supervised machine learning models have been categorized into Ensemble and Non-Ensemble machine learning models.

In Ensemble machine learning models, multiple base learners are combined to improve the accuracy and robustness that can not be obtained from single models. The following ensemble models were considered for this project:

Extreme Boosting Classifier

The Extreme Gradient Boosting algorithm was proposed by Chen and Guestrin and utilized the gradient framework. This algorithm is popular for classification tasks, considering its efficiency and flexibility. It is based on 4 concepts:-

- Ensemble of trees: The model utilizes multiple decision trees for prediction.
- Sequential training: the tree is trained to correct errors of the previous tree.
- Gradient Descent: Minimize the loss function by utilizing the gradient descent approach.
- Regularization: prevents overfitting through regularization.

To avoid over-fitting and smoothen learning weights, Extreme Gradient Boosting loss

function combines a regularization term and objective function. Additionally, Extreme Gradient Boosting supports row and column sampling [28].

$$formula : \Omega(f) = \gamma * |\omega| + \lambda * \omega^2 \tag{4.1}$$

Random Forest Classifier

Random Forests are ensemble models that gain more accurate results by using a large number of decision trees as base learners. The bootstrap sampling technique is used to create individual trees from training data, with random parameters serving as the nodes and roots. Because several decision trees average the results and lessen overfitting, they are more stable than a single tree. Random Forest Classifier uses three main parameters: the number of trees in the forest at each binary node, the number of predictors chosen at random, and the minimum number of observations at the tree nodes [28].

$$formula : Gini\_Importance\_i = \sum_{j=1}^{T} (Gini_{Tree_j} - Gini_{Node_j\_split\_by\_i}) \tag{4.2}$$

Non-ensemble machine learning models, single independent base algorithms to make predictions. Non-ensemble machine learning algorithms are simple and easy to implement and understand. Additionally, they are computationally efficient in training and prediction, especially when used on small datasets. However, the struggle often results in lower accuracy and precision on complex datasets. The following non-ensemble models were considered for this project:

Artificial Neural Network

Artificial Neural Network sometimes referred to as a neural network, is a model capable of handling classification and regression tasks. An artificial neural network (ANN) mimics human neurons to process information. It is a computational model made up of several interconnected neurons. One could think of the neuron as a unit for computation and

storage. Typically, a single neuron receives inputs from several sources. Subsequent neurons will utilize the output of a neuron as their input after some computations (weighted addition). Each link between two neurons in a neural network has a weight that indicates how strong the connection is [29].

ANN basic equation:

$$Z = W \cdot X + b \tag{4.3}$$

Where:

- Z is the output of the neuron.
- W is the weight matrix.
- X is the input vector.
- b is the bias.

Support Vector Machine (SVM)

A supervised machine learning approach for classification and regression applications is called a Support Vector Machine (SVM). SVMs seek to maximize the margin between the closest points of various classes by identifying the optimal hyperplane (or line in 2D) that divides data points into distinct classes. These nearest locations are called "support vectors" and directly influence the position of the separating hyperplane [29].

$$formula : \frac{w^T\left(x_{\mathbf{pos}} - x_{\mathbf{neg}}\right)}{x} = \frac{2}{w}$$

4.6.1 Model tuning

Hyperparameter Tuning

Hyperparameter tuning determines the best hyperparameter configuration for a machine learning model and helps optimize its performance. They help regulate a model's learning process. Experimenting with various hyperparameter values to determine the combination that produces the best results is known as tuning. Using GridSearchCV, each model was

given a set of sample grid parameters, providing the opportunity for the model to choose the best parameter values from a given set of grid parameters.

k-Fold Cross-Validation

10-fold Cross-Validation was used to evaluate the performance of the model on limited data. Unlike training and testing, where a model's performance is evaluated once. K-Fold Cross-Validation allows for the model to be evaluated multiple times (k times), resulting in more confidence in the model design and performance. k-fold validation is embedded in GridSearchCV as a parameter.

4.6.2 Model performance measurement

Accuracy, Confusion Matrix, Precision and Recall, F1-score, and AUC_ROC were used to measure the performance of the machine learning model [30]. Being a healthcare project, precision or sensitivity in our prediction is critical.

Accuracy: Measures how a model correctly predicts the outcome. This is calculated by dividing the number of correct predictions by the total number of predictions [30].

$$Accuracy = \frac{Correct \quad predictions}{All \quad predictions}$$

The confusion matrix measures the performance of the model based on a set of test data. It displays several accurate and inaccurate (True Positives, False Positives, True Negatives, and False Negatives) using the model's prediction [30].

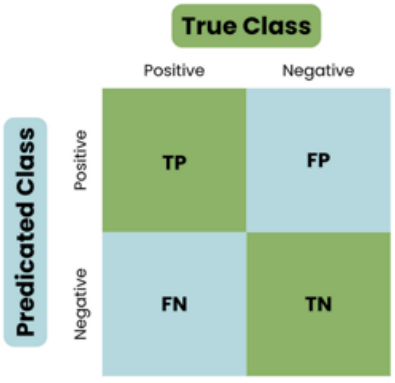


Figure 4.27: Confussion matrix

Precision, also called sensitivity, measures the proportion of positive correctly classified [30].

$$Precision/Sensitivity = \frac{TP}{TP + FP}$$

Recall / Specificity: Evaluates the frequency with which a model accurately distinguishes positive examples (true positives) from the real positive samples. [30].

$$Recall/Specificity = \frac{TP}{TP + FN}$$

F-score: combines precision and recall to measure the performance of the model [30].

$$F - score = \frac{2TP}{2TP + FP + FN}$$

AUC_ROC: ROC is a probability curve, and AUC represents the degree of separability. It shows how much the model can distinguish between classes. The higher the AUC, the better the performance of the model [30].

4.6.3 SHapley Additive exPlanations (SHAP)

SHapley Additive exPlanations (SHAP) framework was used to determine the contributions of each prediction variable to the model's predictions [8]. SHAP assigns an importance value to each feature used in the development of the model. SHAP also aids in fine-tuning the performance of the model by removing features or variables with low importance values.

Chapter5

Results and Discussion

5.1 Descriptive statistics

The data collected from 2067 mothers has a mean Estimate Blood Loss (EBL) of 282.0mL. The Estimated Blood Loss ranged from 100 to 4000. The total number of women with Postpartum Hemorrhage was 57 (3%); 10 (17.5%) and 47 (82.5%) for cesarean section and normal vaginal deliveries respectively. Severe Postpartum Hemorrhage, defined as Estimated Blood loss greater than 2000mL, was experienced in 1 woman. The mean age of women was 26 years, and the age range recorded was between 14 and 44 years. Overall, Postpartum Hemorrhage was prevalent among women aged below 35 years (78.9%). Older women above 34 years had a macrosomic baby (5.3%). The total number of women who received Oxytocin Augmentation was 57 (3%

5.2 Risk factors of Postpartum Hemorrhage

The factors associated with increased risk of Postpartum Hemorrhage were: Number of ANC Visits (P-Value: 0.00); Gestational Age at first visit_weeks (P-Value: 0.02); Weight of Baby (g) (P-Value: 0.01); Duration of Labour (P-Value: 0.00); Cervical Tear (P-Value: 0.00), Episiotomy (P-Value: 0.00), and Perineal Tears (P-Value: 0.00).

Feature	P-value
Duration of Labour	0.00
Cervical Tear	0.00
Episiotomy	0.00
Perineal Tears	0.00
Number of ANC Visits	0.00
Weight of Baby (g)	0.01

Figure 5.1: Summary of correlated variables

5.3 Model Predictions & Performance

A total of 7 features: Gestational Age at first visit (weeks), Number of ANC Visits, Cervical Tear, Episiotomy, Perineal Tears, Weight of Baby (g), and Duration of Labour were selected from 60 features for the development of the model. Extreme Gradient Boosting model performed better with AUC value of 97.0%, followed by Random Forest Classifier with AUC of 96.0%, and Support Vector Machines with AUC of 71.0%,. Artificial Neural Network performed poorly with AUC of 69.0%.

Model	AUC	Accuracy	Recall	Precision	F1-Score
Extreme Gradient Boosting	0.97	0.96	0.97	0.96	0.96
Random Forest	0.96	0.97	0.96	0.95	0.96
Support Vector Machine	0.71	0.71	0.73	0.69	0.71
Artificial Neural Network	0.69	0.69	0.72	0.67	0.69

Figure 5.2: Model performance

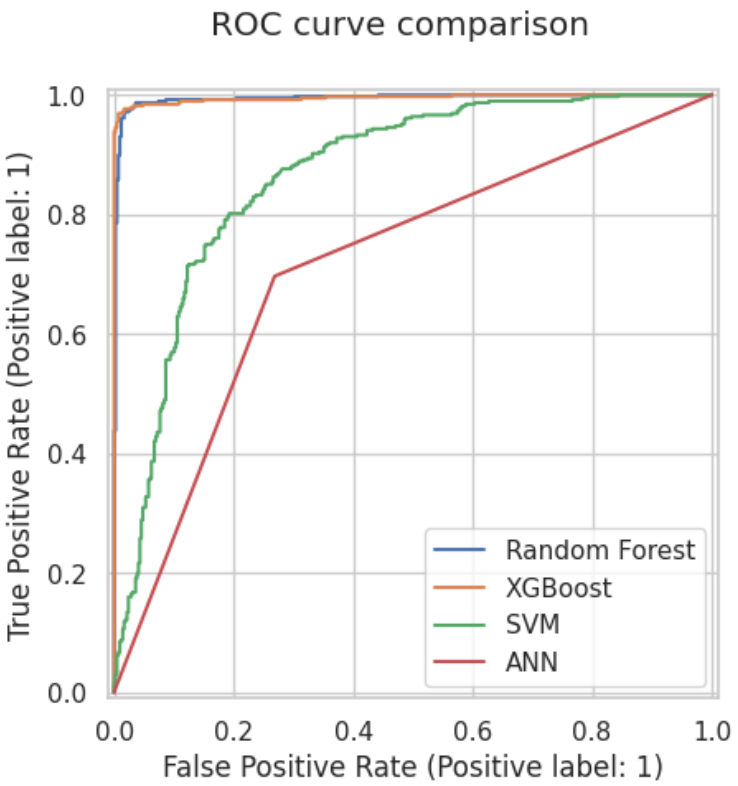


Figure 5.3: Receiver Operating Characteristic Curve

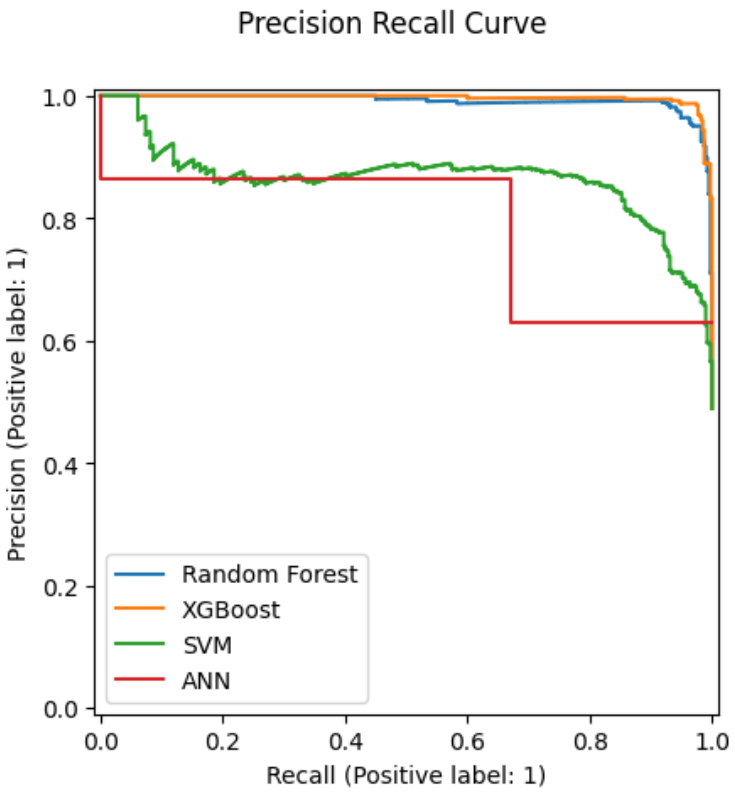


Figure 5.4: Precision curve

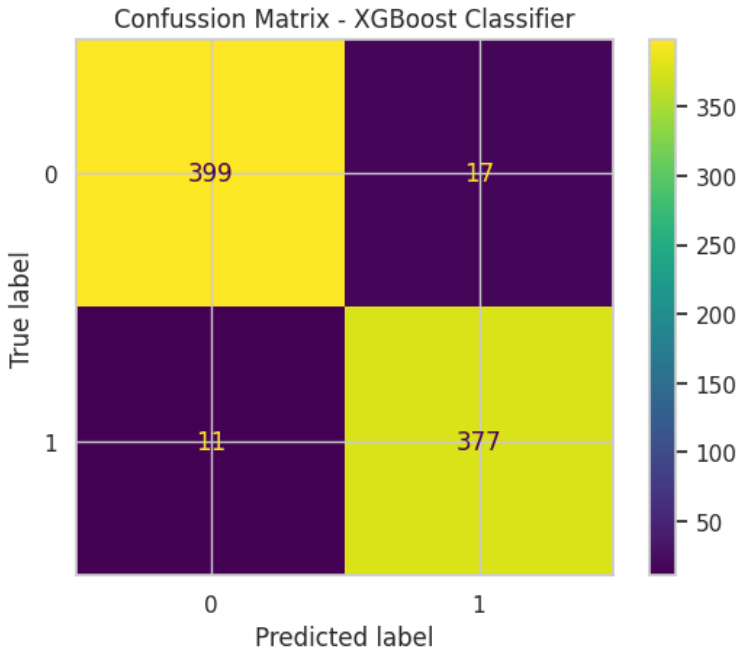


Figure 5.5: Confusion Matrix

5.4 Result of SHapley Additive exPlanations (SHAP) Analysis

SHapley Additive exPlanations analysis assessed how different variables affected the prediction of Postpartum Hemorrhage. The following crucial features were linked to a high risk of Postpartum Hemorrhage based on Extreme Gradient Boosting model: Duration of Labour [1], Number of ANC Visits [2], Weight of Baby (g) [3], Episiotomy [4], and Perineal Tears [5]. In contrast, Cervical Tear is more protective in the prediction of Postpartum Hemorrhage.

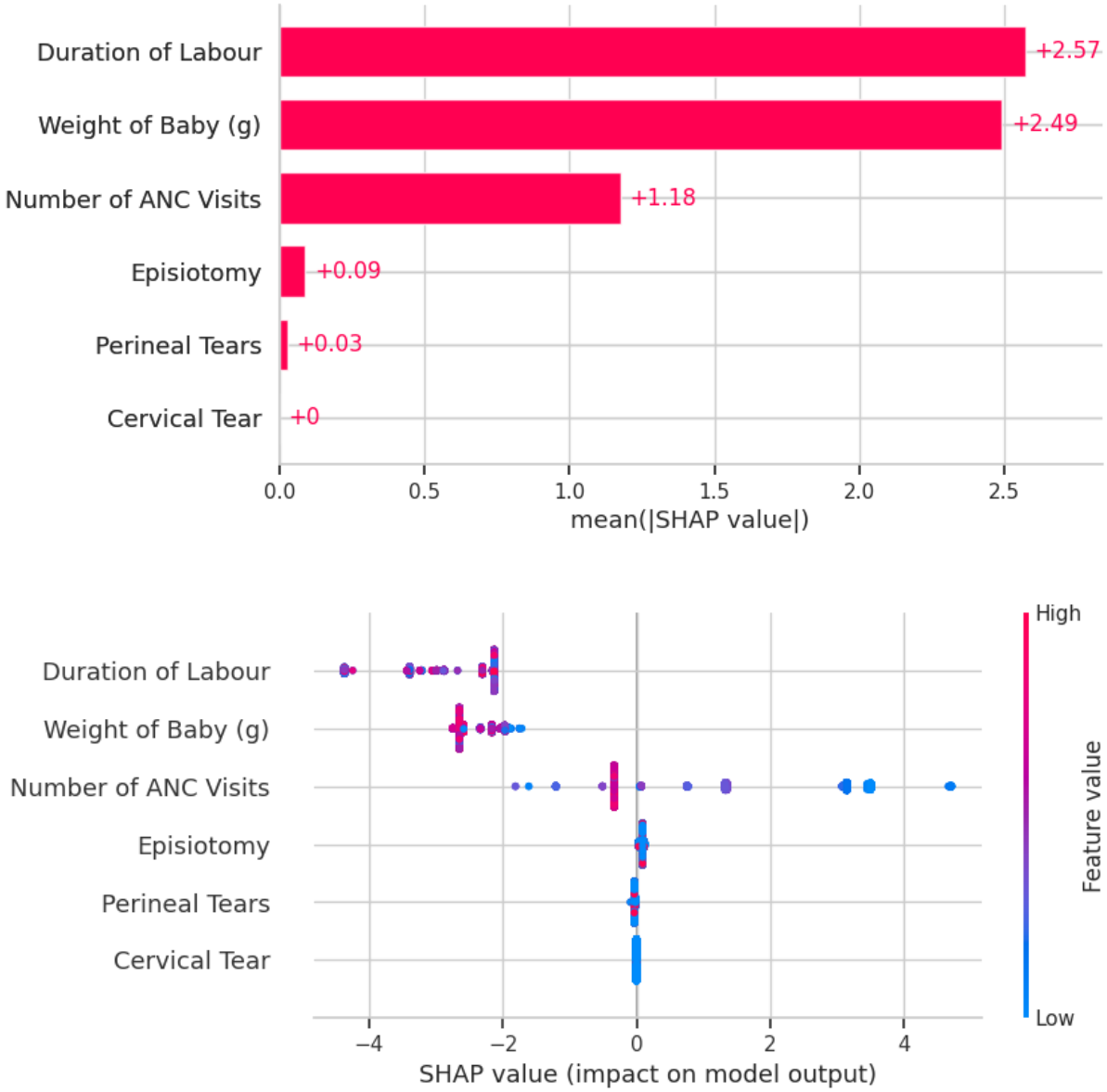


Figure 5.6: SHapley Additive exPlanations analysis

5.5 Discussion

5.5.1 Principal findings

The incidence of Postpartum Hemorrhage (PPH) in Mpilo Central Hospital, Bulawayo is 2.8% between 01 March 2016 to 31 May 2016. This is slightly higher compared to the 1.6% as reported by Solwayo Ngwene et al utilizing the same data [21]. This project excluded records from other forms of deliveries other than Normal Vaginal Delivery and Cesarean

Delivery. Additionally, the Postpartum Hemorrhage outcome variable "pph" is a composite of Mode of Delivery and Blood Loss; True if blood is ≥500mL for normal vaginal delivery and ≥1000mL for cesarean delivery.

In this project, one of the major risk factors for Postpartum Hemorrhage is Duration of labour. Women who experience prolonged labour were most likely to experience Postpartum Hemorrhage [31][32]. This has been documented in previous studies that observed that the median blood loss experienced at the first stage of labour (onset of regular contractions to complete cervical dilation (10 cm)) increased linearly after 10 hours of labour. Postpartum Hemorrhage proportion was highest in women with longer first stage of labour (≥ 12hrs) and longer second stage of labour (≥ 5hrs). Similarly, Postpartum Hemorrhage proportion was minimal in short first stage of labour (< 5hrs) and short second stage of labour (< 3hrs).

Additionally, Women who delivered large babies were highly likely to experience Postpartum Hemorrhage. Weight of the baby and fetus macrosomia are known causes Postpartum Hemorrhage [7]. Large babies lead to extensive distention of the uterus, a condition where the uterus is unable to adapt to an increase in size. Women with twin or multiple births are at risk of extensive distention and eventually Postpartum Hemorrhage. Additionally, women with polyhydramnios (excess buildup of amniotic fluid around the baby in the uterus) are at risk of extensive distension and eventually Postpartum Hemorrhage. Therefore, closer attention should be given to mothers who have delivered overweight babies to avoid the risk of Postpartum Hemorrhage and adverse effects.

The analysis revealed that women with fewer than 4 ANC follow-ups had an increased risk of postpartum hemorrhage. Studies have reported that antenatal care follow-up (ANC) reduces the occurrence of postpartum hemorrhage after birth [32]. Therefore, ANC caregivers should encourage pregnant women to attend ANC visits for early identification of pregnancy and potential delivery-related complications.

This project also found no association between parity, Gravida, Number of Dependents, Age, and risks of Postpartum Hemorrhage. However, studies have shown that these features are risk factors of Postpartum Hemorrhage. A study to assess the incidence and risk factors for postpartum hemorrhage in Uganda also found that grand multi-parity was marginally associated with Postpartum Hemorrhage [7].

Women who experience delivery-related complications like cervical tear, episiotomy, and

perineal tear features were associated were at risk of Postpartum Hemorrhage. These procedures result in excess bleeding, and such women should be given a lot of attention to stop bleeding.

The extreme gradient boosting model performed better than other classification models tested with the AUC value of 97.0%, accuracy, precision, and recall of 96.0%, 96.0%, and 97.0%, respectively. In this project, ensemble models demonstrated excellent discriminative ability, with Extreme gradient boosting and Random forest classifier emerging first and second, respectively in terms of overall performance metrics. A study by Vahid Mehrnoush et al assessed predicting postpartum hemorrhage using traditional statistical analysis and a machine learning approach, and Extreme Gradient Boosting model performed best with AUC of 98%, proving that Extreme gradient boosting Extreme Gradient Boosting model's performance. This model can assist in risk-based triage in the obstetric department in the identification of mothers who are at risk of PPH and providing appropriate care, for example, blood transfusion and oxytocin augmentation, to minimize the risk of PPH and adverse events, especially maternal morbidity and death.

Chapter6

Conclusions and Recommendation

6.1 Summary of Key Findings

The goal of the project was to create a validated machine model that can predict mothers at risk of Postpartum Hemorrhage. This involved the identification of the risk factors associated with Postpartum Hemorrhage. Six features: Cervical Tear, Episiotomy, Number of ANC Visits, Duration of Labour, Perineal Tears, Weight of Baby (g) were associated with Postpartum Hemorrhage. Four machine learning models: Random Forest Classifier, Extreme Gradient Boosting, Support Vector Machines, and Artificial Neural Network were developed. Extreme Gradient Boosting model outperformed other models with the highest AUC of 97.0% and precision of 96.0%.

6.2 Recommendations

Machine learning algorithms' application in public health is gaining ground. Artificial intelligence is being utilized in medical diagnosis, tailored treatment, risk assessment, and decision support. clinical trials, drug discovery, etc. The following are key recommendations:

- Obstetricians in the maternity ward should embrace machine learning models in risk-based triage. Early identification of Postpartum Hemorrhage is critical to prevent maternal morbidity and mortality.
- Documentation of Postpartum Hemorrhage cases is critical to support future research endeavors. Machine learning models thrive on large datasets. The more data available for a model, the better it can discriminate and understand the underlying patterns.

- Accurate measurement of Blood loss is critical. Adoptive accurate techniques like use of under buttock drape is critical for accurate measurement of Blood Loss.

- Obstetricians should closely monitor women who are undergoing long hours of labour. Studies have observed that the median blood loss experienced at the first stage of labour (onset of regular contractions to complete cervical dilation (10 cm)) increased linearly after 10 hours of labour.

- Mothers who have sustained Cervical, Episiotomy, and Perineal tears during delivery should be closely monitored and managed to avoid blood loss.

- Pregnant women should be sensitized and encouraged to attend antenatal care for early risk identification of any risk, including Postpartum Hemorrhage to the mother and fetus. The data revealed very poor antenatal care visits, with a median ANC attendance of up to 3 visits.

6.3 Limitations

One of the study's weaknesses is that Blood Loss was determined using estimation techniques rather than measurements. Estimated Blood Loss was used to generated outcome variable "pph". The estimation of blood loss is extremely challenging. Regretfully, this is the only approach that can be used in low-income settings. Observer variations may also exist since the diagnosis was made by various doctors with varying degrees of training, experience, and grade.

This project utilized a sample size of 2067 records with 60 features. In this dataset, there were about 57 records with Postpartum Hemorrhage. Normal vaginal delivery had 47 PPH cases and 10 PPH cases for ceserean. The small quantity of data and category sizes could have limited the discriminative ability of the model. Future projects could consider using a larger dataset with an increased number of Postpartum Hemorrhage categories.

The project utilized secondary data for the development of the model. There was no control over how the data was collected. This meant missing data from critical features had to be excluded from the analysis, limiting the amount of data available. Therefore, it is critical to utilize a larger dataset.

6.4 Conclusion

The risk factors like the weight of a baby, the Number of ANC attendances, and Parity need to be monitored. Women who have gone through Episiotomy, Cervical, and perineal tears require additional care to prevent blood loss and its associated adverse events. Machine learning models can be used to identify the risk factors and predict mothers at risk of Postpartum Hemorrhage. Using machine learning models to improve PPH prediction with high precision was a legitimate strategy. To find the optimum model, more research is being done to prepare massive data and examine relevant variables.

Bibliography

[1] A. B. Zheutlin, L. Vieira, R. A. Shewcraft, S. Li, Z. Wang, E. Schadt, S. Gross, S. M. Dolan, J. Stone, E. Schadt, and L. Li, "Improving postpartum hemorrhage risk prediction using longitudinal electronic medical records," *Journal of the American Medical Informatics Association*, vol. 29, pp. 296–305, 2 2022.

[2] WHO, "A roadmap to combat postpartum haemorrhage between 2023 and 2030," *https://www.who.int/publications/i/item/9789240081802*, 2023.

[3] C. Susanu, A. Hărăbor, I. A. Vasilache, V. Harabor, and A. M. Călin, "Predicting intra- and postpartum hemorrhage through artificial intelligence," *Medicina (Lithuania)*, vol. 60, 10 2024.

[4] A. Ranjbar, S. R. Ghamsari, B. Boujarzadeh, V. Mehrnoush, and F. Darsareh, "Predicting risk of postpartum hemorrhage using machine learning approach: A systematic review," 9 2023.

[5] V. Mehrnoush, A. Ranjbar, M. V. Farashah, F. Darsareh, M. Shekari, and M. S. Jahromi, "Prediction of postpartum hemorrhage using traditional statistical analysis and a machine learning approach," *AJOG Global Reports*, vol. 3, 5 2023.

[6] M. Akazawa, K. Hashimoto, N. Katsuhiko, and Y. Kaname, "Machine learning approach for the prediction of postpartum hemorrhage in vaginal birth," *Scientific Reports*, vol. 11, 12 2021.

[7] S. Ononge, F. Mirembe, J. Wandabwa, and O. M. Campbell, "Incidence and risk factors for postpartum hemorrhage in uganda," *Reproductive Health*, vol. 13, 4 2016.

[8] S. Y. Shah, S. Saxena, S. P. Rani, N. Nelaturi, S. Gill, B. T. Barr, J. Were, S. Khagayi, G. Ouma, V. Akelo, E. R. Norwitz, R. Ramakrishnan, D. Onyango, and M. Teltumbade, "Prediction of postpartum hemorrhage (pph) using machine learning algorithms in a kenyan population," *Frontiers in Global Women's Health*, vol. 4, 2023.

[9] S. U. Org, "Transforming our world: The 2030 agenda for sustainable development united nations united nations transforming our world: The 2030 agenda for sustainable development."

[10] unicef, "Maternal mortality," *https://data.unicef.org/topic/maternal-health/maternal-mortality/*, 4 2025.

[11] H. K. Ahmadzia, A. C. Dzienny, M. Bopf, J. M. Phillips, J. J. Federspiel, R. Amdur, M. M. Rice, and L. Rodriguez, "Machine learning for prediction of maternal hemorrhage and transfusion (preprint)," *JMIR Bioinformatics and Biotechnology*, 2 2023.

[12] WHO, "The health of mothers and babies is the foundation of healthy families and communities," *https://www.afro.who.int/countries/uganda/news/health-mothers-and-babies-foundation-healthy-families-and-communities*, 4 2025.

[13] S. Goundar, "Chapter 3-research methodology and research method."

[14] N. I. of Technology, "The role of data science in healthcare," *https://online.nyit.edu/blog/data-science-in-healthcare*, 8 2024.

[15] K. K. Venkatesh, R. A. Strauss, C. A. Grotegut, R. P. Heine, N. C. Chescheir, J. S. Stringer, D. M. Stamilio, K. M. Menard, and J. E. Jelovsek, "Machine learning and statistical models to predict postpartum hemorrhage," *Obstetrics and Gynecology*, vol. 135, pp. 935–944, 4 2020.

[16] M. Wang, G. Yi, Y. Zhang, M. Li, and J. Zhang, "Quantitative prediction of postpartum hemorrhage in cesarean section on machine learning," *BMC Medical Informatics and Decision Making*, vol. 24, 12 2024.

[17] Y. Zhang, X. Wang, N. Han, and R. Zhao, "Ensemble learning based postpartum hemorrhage diagnosis for 5g remote healthcare," *IEEE Access*, vol. 9, pp. 18538–18548, 2021.

[18] M. University, "How positivism shaped our understanding of reality," *https://meridianuniversity.edu/content/how-positivism-shaped-our-understanding-of-reality*, 3 2023.

[19] H. Ainebyona, E. Ayebare, A. Nabisere, and M. A. Saftner, "Prevalence of maternal fever and associated factors among postnatal women at kawempe national referral hospital, uganda: A preliminary study," *International Journal of Environmental Research and Public Health*, vol. 21, 3 2024.

[20] M. data, "Mendeley data."

[21] H. D. S. Ngwenya, "Perinatal outcomes in a low-resource setting," *http://data.mendeley.com/datasets/rrxv5twyd3/1*, 10 2019.

[22] D. H. Ngwenya Solwayo, "Perinatal outcomes in a low-resource setting."

[23] D. Rajput, W. J. Wang, and C. C. Chen, "Evaluation of a decided sample size in machine learning applications," *BMC Bioinformatics*, vol. 24, 12 2023.

[24] A. W. Lo, K. W. Siah, and C. H. Wong, "Machine Learning with Statistical Imputation for Predicting Drug Approvals: Supplementary Materials," *Harvard Data Science Review*, vol. 1, jul 1 2019. https://hdsr.mitpress.mit.edu/pub/4tx7h11w.

[25] I. Waniala, S. Nakiseka, W. Nambi, I. Naminya, M. O. Ajeni, J. Iramiot, R. Nekaka, and J. Nteziyaremye, "Prevalence, indications, and community perceptions of caesarean section delivery in ngora district, eastern uganda: Mixed method study," *Obstetrics and Gynecology International*, vol. 2020, 2020.

[26] E. Noble, "New guidelines on antenatal care for a positive pregnancy experience," *https://www.who.int/news/item/07-11-2016-new-guidelines-on-antenatal-care-for-a-positive-pregnancy-experience*, 11 2016.

[27] M. O. Akinwande, H. G. Dikko, and A. Samson, "Variance inflation factor: As a condition for the inclusion of suppressor variable(s) in regression analysis," *Open Journal of Statistics*, vol. 05, pp. 754–767, 2015.

[28] M. Ahmad, R. A. Al-Mansob, K. R. Kashyzadeh, S. Keawsawasvong, M. M. S. Sabri, I. Jamil, and A. C. Alguno, "Extreme gradient boosting algorithm for predicting shear strengths of rockfill materials," *Complexity*, vol. 2022, 2022.

[29] Y. Tian, M. Shu, and Q. Jia, *Artificial Neural Network*, pp. 1–4. 2021.

[30] B. J. Erickson and F. Kitamura, "Magician's corner: 9. performance metrics for machine learning models," *Radiology: Artificial Intelligence*, vol. 3, 5 2021.

[31] L. V. Ladfors, X. Liu, A. Sandström, L. Lundborg, J. M. Snowden, M. Ahlberg, and O. Stephansson, "First stage labor duration and the risk of postpartum hemorrhage - a population-based cohort study," *American Journal of Obstetrics and Gynecology*, vol. 228, p. S175, 1 2023.

[32] J. Nigussie, B. Girma, A. Molla, T. Tamir, and R. Tilahun, "Magnitude of postpartum hemorrhage and its associated factors in ethiopia: Systematic review and meta-analysis.," 11 2021.

UGANDA CHRISTIAN UNIVERSITY
A Centre of Excellence in the Heart of Africa

SCHOOL OF RESEARCH & POSTGRADUATE STUDIES
DISSERTATION CORRECTION COMPLIANCE FORM (POST VIVA FORM)

Date:
Name of Candidate: Tom Eganyu Reg. No: J23MD10/220

Title of Dissertation:
Predicting Postpartum Haemorrhage in Pregnant Mothers in low-income settings. A machine learning approach

S/N	COMMENTS BY EXTERNAL EXAMINER	ACTION TAKEN	INDICATOR
1	The objectives need to be revised. There is no difference between the main objective and specific objective number 2. Avoid using bullets. It is very hard to refer to a bullet.	Change the main objective to: The aim of this project is to decrease maternal morbidity and mortality caused by postpartum haemorrhage in low-income areas by utilizing machine learning algorithms. Replaced the bullets with enumerated alphabetical list.	Goal and Objectives in Page 6
2	There is no relationship between objectives and the methods used to achieve the objectives. The exact methodology and research design are not described.	Added the research design and linked objectives with the methodology.	Page 1, etc corrected
3	Section 3.6 Implementation and analysis. Exactly what have you implemented? Still, under methodology, the candidate seems to discuss results meant to be under chapter four.	Moved implementation and analysis under entomology.	Whole section 3.6 implementation and analysis.
4	Clearly differentiate your chapters, and let have the right content under the right chapters.	Clearly, differentiated the chapters and content.	

5			

S/N	COMMENTS BY INTERNAL EXAMINER	ACTION TAKEN	INDICATOR
1			e.g. Cover page
2			Page 1, etc corrected
3			

S/N	COMMENTS BY VIVA VOCE PANNEL	ACTION TAKEN	INDICATOR
1			e.g. Cover page
2			e.g Page 1, etc corrected
3			

Candidate's Name Signature

Supervisor's Name/ Signature

NB: Post Viva compliance form is designed to capture all the corrections recommended by internal examiner (supervisor), external examiner and viva panel.