



ADDIS ABABA UNIVERSITY

SCHOOL OF INFORMATION SCIENCE

**Predicting Adverse Perinatal Outcomes among Women with Hypertensive Disorders of
Pregnancy in Addis Ababa**

**A thesis proposal submitted to the School of Information Science at Addis Ababa
University in partial fulfillment of Masters of Science Degree in Public Health Data Science**

By; Rebecca Ashagire

GSR/4603/16

June 2025

Addis Ababa, Ethiopia

DECLARATION

I declare that this thesis work is my original work and has not been presented for a degree in this or any other universities, and all sources of materials used for the thesis work have been duly acknowledged.

Rebecca Ashagire Demissie

Full Name of Student

Signature

This thesis has been submitted for examination with my approval as advisor.

Dr. Wondwossen Mulugeta (PHD)

Full Name of Advisor

Signature

Dr. Meselech Assegid (PHD)

Full Name of Advisor

Signature

Addis Ababa Ethiopia

June, 2025

ADDIS ABABA UNIVERSITY
SCHOOL OF INFORMATION SCIENCE

**Predicting Adverse Perinatal Outcomes among Women with Hypertensive Disorders of
Pregnancy in Addis Ababa**

By; Rebecca Ashagire

Name and signature of Members of the Examining Board

Name	Signature	Date
Dr. Martha Yifiru	_____	_____
Mr. Teshome Kabeta	_____	_____

Acknowledgement

I would like to express my deepest gratitude to my family for their love, encouragement, and support throughout this journey. I am equally thankful to my advisors Dr. Wondwossen and Dr. Meselech for their invaluable guidance and their constructive comments. I also want to give sincere thanks to my friends, whose belief and presence made even the hard times enjoyable. This accomplishment is as much yours as it is mine. I would also like to thank APHREA-DST for supporting this research work to be done. This would not have been possible without each and every one of you, Thank you very much.

Table of Contents

Table of Contents.....	v
List of figures	vii
List of tables	vii
Acronyms and Abbreviations	viii
CHAPTER-ONE	1
1. Introduction	1
1.1 Background	1
1.2 Statement of the problem	3
1.3 Research questions.....	5
1.4 Objectives	5
1.4.1 General objective.....	5
1.4.2 Specific objectives.....	5
1.5 Significance of the study	6
1.6 Scope of the study	6
1.7 Ethical considerations	6
1.8 Application of results (Result Dissemination)	6
1.9 Organization of the research	7
CHAPTER-TWO	8
2. Literature review	8
2.1 Pathogenesis of HDPs.....	8
2.2 Impact of HDPs on neonatal health.....	8
2.3 Strategies for managing and preventing APOs in HDP	9
2.4 Factors associated with APOs among women with HDP	9
2.5 Machine Learning methodology: Preprocessing, Models and Tuning	10
2.5.1 Data Preprocessing	10
2.5.2 Classification Machine Learning Models	15
2.5.3 Hyperparameter Tuning	17
2.6 Model Evaluation.....	18
2.7 The role of ML in predicting APOs.....	20
2.8 Related Works	21
CHAPTER-THREE.....	25
3. Data and Methodology.....	25

3.1 Data Source	25
3.2 Variables	26
3.3 Exploratory Data Analysis	29
3.4 Data Preprocessing.....	30
3.4.1 Data Cleaning.....	30
3.5 Selection of ML models	36
3.6 Software and Tools	37
CHAPTER FOUR	38
4. Experimentation and Analysis of Results.....	38
4.1 Overview	38
4.2 Experimentation	38
Model 1: LR	38
Model 2 : Naïve Bayes.....	42
Model 3: SVM	45
Model 4: DT	49
Model 5: XGBoost	52
Model 6: RF	55
4.3 Identifying Key Features for Prediction	58
4.4 Comparison of ML models	59
4.5 Discussion of the result	62
CHAPTER FIVE.....	64
5. Conclusion & Recommendation	64
5.1 Conclusion.....	64
5.2 Recommendation	65
References	67
Annex.....	72
A: Table One: List of final attributes selected for modelling.....	72
B: Table Two: Performance of models using different imputation methods	73
C: Table Three: Performance of models using different tuning methods.....	73

List of figures

Figure 1 Decision Tree Structure-----	16
Figure 2 A structure of RF made up of multiple DTs-----	16
Figure 3 Confusion matrix-----	18
Figure 4 Example of inconsistencies in attributes -----	30
Figure 5 Pie chart showing distribution of APOs-----	32
Figure 6 Label encoded target class & distribution-----	35
Figure 7 Architecture of adverse outcome risk prediction model-----	37
Figure 8 Feature importance plot using XGBoost model-----	59
Figure 9 Comparison of the models by Precision-----	60
Figure 10 Comparison of the models by Recall-----	60
Figure 11 Comparison of the models by AUC-ROC-----	61
Figure 12 Comparison of the models by PR-AUC-----	61

List of tables

Table 1 Name of hospital and amount of data from each hospital-----	26
Table 2 Summary of categories & attributes in the data-----	26
Table 3 Summary of datatypes in the data-----	26
Table 4 List of all variables included initially-----	28
Table 5 List of variables removed due to high amount of missingness-----	33
Table 6 Outcome variable engineering-----	34
Table 7 Number of instances in training & testing set-----	35
Table 8 Selected ML models for experiment-----	36
Table 9 Test set classification report of LR model-----	40
Table 10 Test set classification report of NB model-----	43
Table 11 Test set classification report of SVM model-----	46
Table 12 Test set classification report of DT model-----	49
Table 13 Test set classification report of XGBoost model-----	52
Table 14 Test set classification report of RF model-----	55
Table 15 Summary of the final tuned model's performance on test set-----	58

Acronyms and Abbreviations

APO - Adverse Perinatal Outcome
AUC- ROC – Area under the curve– Receiver operating characteristics
CSV – Comma separated values
DT – Decision tree
EMR – Electronic medical records
fm – Free molecule
HDP – Hypertensive disorders of pregnancy
KNN – K-nearest neighbors
LMIC – Low- and middle-income countries
LR – Logistic regression
MAR – Missing at random
MCAR – Missing completely at random
MICE – Multiple imputation by chained equations
MNAR – Missing not at random
mmHg - Millimeters of mercury
ML – Machine Learning
MLP – Multilayer perceptron
NB – Naïve Bayes
NPV – Negative predictive value
PlGF - Placental Growth Factor
PPI – Personally identifiable information
PPV – Positive predictive value
RF – Random Forest
RR – Risk Ratio
SFlt-1 - Soluble fms-like tyrosine kinase-1
SVM – Support Vector Machine

Abstract

Background - Hypertensive disorders of pregnancy (HDP) are among the leading causes of adverse perinatal outcomes (APOs) worldwide, with a particularly high burden in low- and middle-income countries (LMICs) like Ethiopia. Despite advancements in maternal healthcare, traditional risk assessment methods often fail to account for the complex nature of HDP, resulting in delayed interventions and preventable maternal and neonatal complications. The lack of efficient and accurate risk prediction tools in resource-limited settings contributes to high morbidity and mortality rates among pregnant women and newborns.

Objective - This study aimed to develop a machine learning (ML) predictive model using electronic medical record (EMR) data to identify women with HDP who are at high risk for APOs.

Methodology - This study applied a retrospective analysis of EMRs collected from selected public hospitals in Addis Ababa (AA), focusing on predicting APOs among pregnant women diagnosed with HDP. A total of 1042 instances was collected from four hospitals in AA. After performing data cleaning and preprocessing, including handling missing values using three imputation techniques (Median, KNN, and MICE), six supervised ML algorithms—Logistic Regression, Naïve Bayes, Support Vector Machine, Decision Tree, Random Forest, and XGBoost—were developed. Model performance was evaluated using, internal cross-validation, two hyperparameter optimization strategies (Grid Search and Bayesian Optimization), and with metrics including accuracy, sensitivity, precision, F1-score, AUC-PR and ROC-AUC.

Result - Random Forest and XGBoost models had the highest predictive performance, with XGBoost achieving the best overall results. However, even the best-performing model reached a maximum recall of only 69% for predicting the positive class, which indicates limitations in identifying APOs.

Conclusion - The findings indicate that ML models show promise in predicting APOs among women with HDP using routinely collected clinical data. Most models achieved high specificity and ROC-AUC value, showing their potential for use in risk stratification. However, the recall performances were suboptimal for clinical application, which suggests the current models are not yet suitable for deployment. Further research with larger datasets, more diverse features, and broader validation is necessary.

CHAPTER-ONE

1. Introduction

1.1 Background

Hypertensive disorders of pregnancy (HDP), including chronic hypertension, gestational hypertension, preeclampsia, and eclampsia, are significant contributors to adverse maternal and perinatal outcomes worldwide. Chronic hypertension, which predates pregnancy or is diagnosed before 20 weeks of gestation, is associated with complications such as preeclampsia, preterm delivery, fetal growth restriction, and acute maternal organ failure (Roberts et al. 2003). Effective management involves tight blood pressure control (110–140/85 mm Hg), fetal growth monitoring, and frequent assessments for preeclampsia development (Brown et al. 2018).

Gestational hypertension, a de novo hypertension occurring after mid-pregnancy, often results in favorable maternal and fetal outcomes. However, approximately 25% of cases progress to preeclampsia, especially when diagnosed before 34 weeks of gestation. The development of severe hypertension in gestational hypertension increases the likelihood of adverse maternal and neonatal outcomes, including low birth weight, preterm delivery, neonatal morbidity, and maternal complications such as thrombocytopenia and abnormal liver enzymes (Brown et al. 2018).

Preeclampsia, a pregnancy-specific syndrome defined by hypertension ($\geq 140/90$ mm Hg) and proteinuria (≥ 300 mg per 24 hours), remains a major global health challenge (Roberts et al. 2003). It is responsible for over 500,000 fetal and neonatal deaths and 70,000 maternal deaths annually (Brown et al. 2018). This condition poses acute risks, including stillbirth, small-for-gestational-age births, and maternal mortality, particularly in cases of eclampsia, the most severe form of HDP (Bridwell et al. 2019).

HDP remains one of the leading causes of perinatal mortality, alongside obstructed labor, fetal mal-presentation, and uterine rupture (Asseffaid and Demissie 2019). Early identification of women at heightened risk of adverse outcomes is critical for improving maternal and neonatal health. Prediction models using routinely accessible clinical features, such as blood pressure, fetal weight, and maternal health indicators, can guide targeted screening and preventative measures. These models can differentiate between women at high risk requiring closer surveillance and those with mild disease, reducing unnecessary interventions and associated risks (Muche et al. 2024; von Dadelszen et al. 2009). In Ethiopia, the integration of electronic medical records (EMRs) in

health facilities offers a unique opportunity to apply ML algorithms for developing different prediction models.

In this context, exploring predictive tools such as ML algorithms to anticipate APOs among women with HDP offers a promising path for improving healthcare delivery. By leveraging data-driven insights, these tools can improve risk stratification, enable timely interventions, and ultimately improve perinatal outcomes in resource-limited settings (von Dadelszen et al. 2009).

1.2 Statement of the problem

HDP are among the most common medical complications of pregnancy (Hutcheon, Lisonkova, and Joseph 2011). The burden of HDP is disproportionately higher in low- and middle-income countries (LMICs), where most deaths associated with these disorders occur (Bridwell et al. 2019; Payne et al. 2014). Africa has the highest regional prevalence of HDP globally, with 334.9 cases per 100,000 women of childbearing age, highlighting the urgency of addressing this issue (Jiang et al. 2022).

In Ethiopia, the prevalence of HDP is estimated to be 6.07% - 6.82% (Tesfa et al. 2020; Berhe et al. 2018), with the prevalence of preeclampsia/eclampsia reaching 5.47%. Despite efforts to reduce HDP rates in Ethiopia, the reduction has been minimal, and HDP remains a prevalent and serious pregnancy-related disorder in the country (Berhe et al. 2018). HDP is associated with significant APOs, including low birth weight, preterm delivery, and perinatal mortality. In Ethiopia, the prevalence of low birth weight among women with HDP is as high as 37%, and perinatal mortality among this group is estimated to be 25% (Mersha, Abegaz, and Seid 2019).

In LMICs, the high burden of illness and death from HDP is often attributed to delays in triage (identifying women in need of specialized care who may become seriously ill) and transportation to facilities capable of providing appropriate care, compounded by a shortage of trained healthcare workers (Payne et al. 2014). Changes in maternal factors (e.g., maternal age and pre-pregnancy weight) have also led to secular increases in pre-eclampsia, gestational hypertension, and chronic hypertension (Hutcheon, Lisonkova, and Joseph 2011). Early detection and referral of pregnant women with hypertensive disorder could improve the situation by designing a simple tool to identify those at increased risk of complications or death (Mersha, Abegaz, and Seid 2019; Payne et al. 2014).

A study done by (Chen et al. 2022) developed a risk assessment tool using ML, with the goal of identifying patients at risk of adverse maternal outcomes or severe neonatal complications associated with de novo HDP in suburban population in northern China (Beijing). Even though the model had strong predictive abilities in identifying women at higher risk for adverse outcomes, the generalizability of the model needs further study.

A study by (Jaleta, Gebremedhin, and Jebena 2021) tried to determine the risk of APOs among women with HDP. Different researchers also assessed the existing evidence on the effect of HDP on various APOs and its associated factors in Ethiopia (eg., Tadesse et al. 2024; Yahaya, Adeyemo, and Kumah 2024; Jaleta and Abdisa 2022). These studies were valuable in identifying different risk factors related to the different types of HDP. The studies compared clinical, socio-demographic and obstetric data of patients who developed different adverse maternal and perinatal outcomes against those who didn't by using statistical methods.

Therefore, the gap lies in the fact that current clinical practices often rely on traditional risk assessment methods that may not adequately capture the complexities of HDP and its associated risks. This limitation highlights the need for more advanced predictive models that can make use of the great amount of data available in health sectors using ML algorithms which are able to increase predictive accuracy by analyzing large datasets to identify patterns and risk factors that traditional methods might overlook.

There is also a lack of localized predictive research focusing on the Ethiopian context, particularly in urban settings like Addis Ababa. Existing studies have primarily been conducted in high-income countries and different demographic settings (eg., Bucher et al. 2024; Magee et al. 2016; Rana et al. 2012), which may not be directly applicable to Ethiopian women with HDP.

This thesis focuses on the critical need to identify women with HDP in Addis Ababa who are at risk of APOs by creating a ML-based predictive model that is tailored to the local context. It will focus on women with all types of HDP since including all types of HDP improve the clinical applicability because it includes a broader population of women at risk (Ukah et al. 2018). By utilizing easily accessible maternal characteristics that are routinely collected during prenatal visits, such as demographic and clinical data, prediction models can improve early identification of at-risk pregnancies. ML techniques can enhance these models' accuracy and applicability, enabling evidence-based interventions and reducing the reliance on advanced diagnostics that are unavailable in resource-limited settings. These algorithms have the potential to bridge critical gaps in maternal healthcare, reduce unnecessary treatments, and improve outcomes for mothers and their newborns (Hassen et al. 2020; Feleke et al. 2022; Bogale, Abuhay, and Dejene 2022).

1.3 Research questions

- What are the key predictors of APOs in women with HDP?
- Which ML model has the best performance to develop predictive models for adverse outcomes among women with HDP?

1.4 Objectives

1.4.1 General objective

- To develop and evaluate ML predictive models for early identification APOs among women with HDP in AA using routinely collected maternal and clinical data.

1.4.2 Specific objectives

- To develop a ML model for identifying women with HDP at risk for APOs.
- To identify key clinical predictors of APOs among women with HDP
- To compare the performance of different ML classifiers and select the best one

1.5 Significance of the study

By using ML algorithms to predict APOs, this study helps in early identification of high-risk pregnancies and reducing maternal and fetal morbidity and mortality. In addition, it identified risk factors that can help medical practitioners to control and treat the condition early and improve quality of care offered in health centers in low-resource settings.

In environments with limited resources, concentrating on high-risk pregnancies can improve the overall effectiveness of the healthcare system. The creation and improvement of predictive models for pregnancy outcomes promotes advancements in ML and medical informatics which then can be applied to other medical conditions. Furthermore, this study could be used as a reference or a bench mark study for related studies and adds to the existing body of knowledge in using predictive tools in the health care field.

1.6 Scope of the study

The research focuses on developing a ML predictive model by using data from routinely collected health metrics from EMRs used in selected hospitals within Addis Ababa, Ethiopia and addressing different factors influencing APOs in women with all subtypes of HDP.

1.7 Ethical considerations

Ethical approval was obtained from the College of Natural Science Institutional Review Board (CNS-IRB) and the selected hospitals before using the data. Ethical clearance was also secured from Addis Ababa Public Health and Emergency Management Directorate prior to data extraction. The data was anonymized to eliminate personally identifiable information (PII) and prevent traceability to individual patients. The data was also used only for academic purposes.

1.8 Application of results (Result Dissemination)

The findings of this research will be disseminated through multiple channels to ensure they reach key stakeholders, policymakers, and healthcare professionals. Results of the study will also be submitted to Addis Ababa Public Health and Emergency Management Directorate, where insights from the study will be shared to inform public health strategies and improve maternal care policies.

Additionally, results will be published in peer-reviewed journals and shared with relevant health institutions to increase impact and help evidence-based decision-making.

1.9 Organization of the research

This thesis is divided into five chapters, each describing the process of building a predictive model. Chapter One provides the background, significance, and objectives of the study, highlighting the clinical and public health relevance of early prediction in maternal care. The second chapter reviews existing research on hypertensive disorders in pregnancy, factors associated with neonatal complications and current applications of ML in obstetrics and identifies the gap that this study aims to address. Chapter Three outlines the study design, data sources, variables, the preprocessing steps to be performed and about the selected models. Chapter Four explains about the data preprocessing conducted, implementation of ML algorithms, model training and validation, performance evaluation metrics and discussion of the results. The fifth chapter summarizes key findings, discusses limitations, and offers recommendations for future research.

CHAPTER-TWO

2. Literature review

2.1 Pathogenesis of HDPs

Pregnant women who suffer from hypertensive disorders are at high risk for maternal and fetal morbidity and mortality which makes it a major medical concern. These disorders include chronic hypertension, gestational hypertension, preeclampsia, and eclampsia. All have different pathophysiological features but share a common underlying mechanism that complicate pregnancy outcomes (Khedagi and Bello 2020). A crucial area of study and clinical focus is HDP pathogenesis, which is complex and involves intricate interactions between genetic, immunological, and environmental variables. One of the most prominent types of HDPs, preeclampsia, usually manifests after 20 weeks of pregnancy and is characterized by proteinuria and hypertension. The disorder arises from a combination of abnormal placentation, systemic inflammation, and endothelial dysfunction with impaired angiogenesis being central to its development (Torres-Torres et al. 2024).

Gestational hypertension, occurring without proteinuria, may act as a precursor to preeclampsia and is similarly associated with endothelial dysfunction. Pregnancy-related or pre-existing chronic hypertension can exacerbate these illnesses and have negative effects on the mother and fetus, including higher cardiovascular risks and intrauterine growth restriction (Khedagi and Bello 2020; Mammaro et al. 2009).

Understanding the pathophysiology of HDP is crucial for efficient care and preventative methods, given the potential for serious consequences, such as HELLP syndrome and premature birth. In order to reduce the dangers associated with these disorders, clinicians highlight the significance of early detection, monitoring, and appropriate therapeutic approaches (Bisson et al. 2023; Gao et al. 2023).

2.2 Impact of HDPs on neonatal health

Preeclampsia and gestational hypertension are examples of HDP, which have serious clinical implications for both the mother and the fetus. Intrauterine growth restriction is one of the main issues, as high blood pressure can cause fetal growth to be slowed or lowered, which could have negative effects on the baby (Bisson et al. 2023; Gao et al. 2023). Furthermore, if left unchecked,

hypertension can harm the mother's brain, eyes, heart, lungs, kidneys, and liver, among other organs, potentially resulting in death. Preeclampsia can induce premature delivery as a preventive measure against further complications from high blood pressure during pregnancy. Although occasionally necessary, this early birth carries its own risks and challenges for the health of the neonate (Bisson et al. 2023; Gao et al. 2023).

Detection of high blood pressure during pregnancy is important, as the condition can escalate to HELLP syndrome, a medical emergency characterized by hemolysis, elevated liver enzymes, and low platelet counts (Bisson et al. 2023; Umesawa and Kobashi 2016). In cases of severe preeclampsia, placental abruption may occur, leading to stillbirth, which underscores the need for close monitoring and management (Bisson et al. 2023; Gao et al. 2023).

2.3 Strategies for managing and preventing APOs in HDP

Controlling blood pressure and protecting the mother and fetus are the main goals of managing hypertensive diseases during pregnancy. Pharmacological therapies, ongoing monitoring, and lifestyle changes based on the particular kind of hypertension are examples of therapeutic approaches. Initial antihypertensive therapy for managing chronic hypertension or gestational hypertension should typically involve monotherapy from a selection of first-line medications. These include long-acting oral nifedipine, oral labetalol, oral methyldopa, and other oral β -blockers such as propranolol, acebutolol, metoprolol, and pindolol (LEEMAN, DRESANG, and FONTAINE 2016; ACOG 2022). Urine and blood should be continuously tested in the lab to look for abnormalities that could indicate preeclampsia or a worsening of prenatal hypertension (Poye Laney 2023; Radparvar et al. 2024). Although magnesium sulfate is not used as an antihypertensive medication, it is the preferred medication for seizure prophylaxis in situations of severe preeclampsia and for seizure management in eclampsia (Brown et al. 2018; Poye Laney 2023). Effective management of hypertension diseases during pregnancy requires proper therapeutic approaches in addition to close monitoring of the mother's clinical status and the health of the fetus.

2.4 Factors associated with APOs among women with HDP

APOs in women with HDP are associated with multiple interrelated factors. Disease severity plays the main role, with early diagnosis of HDP, severe hypertension and symptoms like headache, epigastric pain, and blurred vision substantially increasing risks, particularly for eclampsia and

HELLP syndrome (Chen et al. 2022; Thi Huyen Anh, Manh Thang, and Thanh Huong 2024). Fetal and neonatal factors, including low birth weight (<2500 g), impaired fetal growth (reduced abdominal circumference, abnormal femur length), and abnormal umbilical artery Doppler flow are strongly predictive of poor outcomes (Thi Huyen Anh, Manh Thang, and Thanh Huong 2024; Chen et al. 2022). Biochemical markers such as elevated liver enzymes (AST, ALT), high creatinine and uric acid levels, thrombocytopenia, and significant proteinuria serve as important indicators of disease progression and adverse outcomes (Alemie et al. 2023; Chen et al. 2022; Sarno et al. 2022). Socioeconomic determinants and healthcare access issues, including low maternal education, inadequate antenatal care, rural residence, and delayed administration of antihypertensive or anticonvulsant medications, further exacerbate risks (Amare et al. 2021; Tadese et al. 2024; Syoum et al. 2022). Pregnancy-related characteristics such as primiparity, early-onset preeclampsia (<34 weeks gestation), and need for cesarean delivery or labor induction also contribute to complications (Sarno et al. 2022; Amare et al. 2021; Alemie et al. 2023). Additionally, maternal factors including family history of hypertension and development of severe complications like HELLP syndrome or acute kidney injury significantly worsen prognosis (Tadese et al. 2024; Alemie et al. 2023).

2.5 Machine Learning methodology: Preprocessing, Models and Tuning

2.5.1 Data Preprocessing

Data preprocessing is a critical phase in the development of reliable and efficient ML models, particularly when working with secondary data that contains noisy, inconsistent, erroneous, irrelevant, or incomplete values. It includes data cleaning, which entails the removal of missing values, outliers, and duplicate entries, as well as applying appropriate imputation strategies for any remaining missing data. Additionally, data discretization, feature selection, feature transformation techniques and variable encoding are part of the data preparation process.

I. Handling missing values

Missing data is a common issue in clinical research, including studies involving EMRs. Even in well-controlled environments, missing values can occur due to technical errors, incomplete patient documentation, or procedural inconsistencies. The implications of missing data are broad: they

can reduce statistical power, introduce bias in parameter estimation, and impair the representativeness of the sample, potentially leading to invalid conclusions (Kang 2013). Missing data may also complicate the analysis process, especially in ML workflows where most algorithms require complete data.

Mechanisms of Missingness

Understanding the mechanism behind the missing data is critical before selecting appropriate imputation techniques. There are three main mechanisms:

- Missing Completely at Random (MCAR): Data is MCAR when the likelihood of missingness is unrelated to any observed or unobserved values. Under MCAR, analyses remain unbiased, although power may be reduced. (Kang 2013).
- Missing at Random (MAR): A more realistic assumption in real-world datasets, MAR implies that missingness depends only on observed data, not on the missing values themselves. This allows for valid imputation using other known variables (Kang 2013; Li et al. 2024).
- Missing Not at Random (MNAR): When the probability of missingness is related to the unobserved values themselves, the data is MNAR. These cases are especially difficult to handle and often require explicit modeling of the missingness process (Li et al. 2024; Kang 2013).

Various imputation methods exist for handling MCAR-type missing data.

a. Listwise Deletion

One of the most commonly used approaches is listwise deletion (also known as complete case analysis), where any case with missing data is excluded from analysis entirely. While simple to implement, listwise deletion can significantly reduce the sample size and statistical power, especially in small datasets (Kang 2013).

b. Simple Imputation

Simple imputation replaces each missing value with a single statistic calculated from the observed data. Common techniques include substituting the mean for continuous variables and the mode for categorical ones. This method is straightforward, computationally efficient, and widely used in applied studies. However, it assumes that the data are MCAR and can result in underestimated variance, distorted distributions, and biased results especially in datasets with complex inter-variable relationships (Li et al. 2024).

Median imputation is a simple yet effective technique for numeric variables, particularly when the data are skewed or contain outliers. This method substitutes missing values with the median of the observed values within the same column. It is computationally efficient and non-parametric, making no distributional assumptions. Although basic, it is a robust approach often used in initial stages of data preprocessing (Hang et al. 2021).

c. K-Nearest Neighbors (KNN) Imputation

KNN imputation estimates missing values by identifying k most similar records (neighbors) using a defined distance metric and imputing based on their values (Hang et al. 2021; Ngusie et al. 2024). This technique is highly adaptable and can handle both numerical and categorical variables. Its key advantage is in preserving local data structures and relationships, which is particularly important in heterogeneous datasets like EMRs. KNN has also been widely adopted in recent literature due to its consistency and ease of implementation (Ngusie et al. 2024).

d. Multiple Imputation by Chained Equations (MICE)

MICE is a more sophisticated method involving the creation of multiple imputed datasets. Each variable with missing data is modeled conditionally based on other variables, and missing values are iteratively imputed in a chained sequence (Li et al. 2024; Hang et al. 2021). The process produces multiple complete datasets that reflect both the estimated values and the uncertainty surrounding them. These datasets are then analyzed separately, and results are pooled to produce final estimates. MICE is particularly advantageous when dealing with high-dimensional data or when missingness is extensive. It also performs well even when the assumptions of normality are violated or when sample sizes are small (Kang 2013; Li et al. 2024).

II. Variable Encoding

In data preprocessing, handling categorical variables is a major step. Unlike numerical variables, the categories in a categorical variable do not have mathematical meaning. Therefore, these values must be converted into a numerical form before they can be used in ML models. Simply removing these non-numeric fields may lead to a loss of information, which will also negatively impact the model's performance. To overcome this, categorical data must be encoded into a format that maintains its original context while making it usable by algorithms (Bolikulov et al. 2024). This transformation process is known as categorical variable encoding.

A. Label Encoding

Label encoding is one of the simplest and most straightforward techniques for encoding categorical data. In this method, each unique category is assigned a sequential integer value. However, this encoding method introduces a limitation where by the numeric values assigned to categories may be interpreted by ML models as having ordinal or hierarchical relationships, even when none exists. This can lead to misinterpretation and decreased model accuracy, especially for nominal variables where no inherent order exists (Bolikulov et al. 2024).

Label encoding is therefore best suited for binary categorical variables. In such cases, the risk of incorrect magnitude interpretation is minimal (Bolikulov et al. 2024).

B. One-Hot Encoding

To overcome the shortcomings of label encoding for nominal variables, one-hot encoding is used. For a categorical variable with k unique values, one-hot encoding creates k new columns, each corresponding to one category. Only one of these columns will contain a 1 for any given observation, while the rest will be 0, effectively indicating mutual exclusivity (Zheng and Casari 2018).

However, it comes with a trade-off: when the number of unique categories is large, the dimensionality of the dataset increases significantly, which can lead to sparsity and computational inefficiency (Bolikulov et al. 2024).

III. Feature Engineering

The effectiveness of classification models depends on the dataset and its features. A well-designed feature should be informative, never change for the set of transformations, and computationally efficient. Feature engineering is foundation to the learning algorithms by applying domain knowledge to improve model effectiveness, with variable creation being a key step in generating new features from existing data attributes (Zheng and Casari 2018).

IV. Data Splitting

The data is split into different sections. The training set is used to train a learning algorithm and consists of a set of instances along with their associated labels. Validation set are examples of instances used to tune the parameters of a learning algorithm when working with labeled data. The test sample is used to evaluate the performance of a learning algorithm. It is separate from the training and validation data and is not made available in the learning stage. It consists of a collection of instances for which the learning algorithm must predict labels based on features (Mohri, Rostamizadeh, and Talwalkar 2012).

V. Feature Scaling

Feature scaling, often called feature normalization, is essential for models sensitive to input scale such as linear and logistic regression while tree-based models remain unaffected. It adjusts feature values individually for each feature, to ensure consistency. Standardization, also known as variance scaling, is one way of feature scaling where by it transforms a feature by subtracting its mean and dividing by its variance, resulting in a feature with a mean of 0 and variance of 1. This ensures that all features contribute equally to model learning without being dominated by large-scale variables (Zheng and Casari 2018).

V. Feature Selection

Feature selection is an important technique in ML that removes irrelevant or redundant features that aims to simplify the model without sacrificing predictive accuracy. The objective is to create a model that performs well while being computationally efficient. Filtering techniques, which are more cost-effective, help remove features that are unlikely to be useful for a model by applying preprocessing steps, such as computing feature correlation and discarding those below a threshold (Rawat T. 2017).

2.5.2 Classification Machine Learning Models

Classification is a fundamental task in ML where the goal is to assign a label or category to a given input based on previously labeled training data. It is a form of supervised learning where the model learns a mapping from input features to predefined classes using known examples. This process is called inductive inference, where general models are developed from specific examples. Typically, each instance is represented as an attribute-value vector, and the model learns a function that can generalize to both known and unseen data (Shah et al. 2018).

2.5.2.1 Logistic Regression

One of the most widely used models for binary classification is Logistic Regression (LR). It models the probability of a binary outcome (0 or 1) as a function of independent variables, using the sigmoid or logistic function. It is especially suited for datasets where the dependent variable is categorical and the independent variables can be continuous, discrete, or a mix of both. LR is simple, fast, and effective, and regularization techniques (L1 and L2) can be used to address overfitting, particularly in high-dimensional datasets (Shah et al. 2018; Sarker 2021; Sarno et al. 2022).

2.5.2.2 Naïve Bayes

Another probabilistic model is Naive Bayes (NB), which is based in Bayes' theorem and assumes strong independence between features. Despite its simplicity, NB performs well in many real-world classification problems such as spam detection and document classification. It requires a small amount of training data and can handle both binary and multi-class problems efficiently (Shah et al. 2018; Sarker 2021).

2.5.2.3 Decision Trees

Decision Trees (DT) are non-parametric models used for both classification and regression tasks. They operate by recursively splitting the dataset based on the value of input features, resulting in a tree-like structure of decisions. Decision trees are highly interpretable and commonly used in medical diagnostics due to their clear logic and ease of understanding (Sarker 2021; Shah et al. 2018; Teshome et al., n.d.).

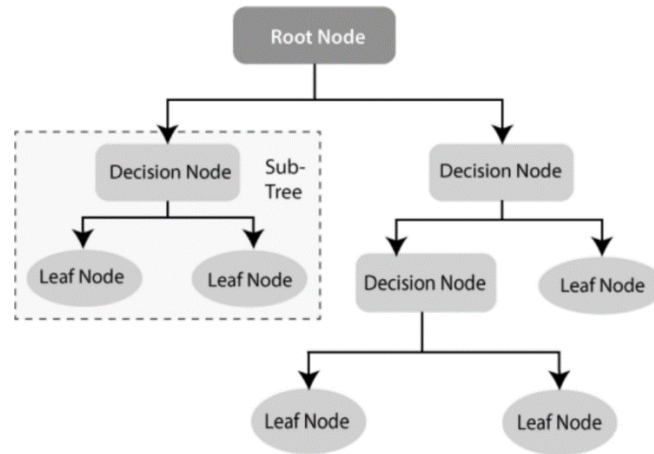


Figure 1 Decision Tree Structure (Sarker 2021)

2.5.2.4 Random Forests

To increase the predictive power of decision trees and reduce overfitting, Random Forests (RF) were developed as an ensemble learning technique. RF builds multiple decision trees on bootstrapped subsets of the data and aggregates their predictions using majority voting. This model improves accuracy, handles both continuous and categorical variables, and requires minimal data preprocessing. It is known for being robust and efficient, even with large feature sets (Sarker 2021; Bekele 2022).

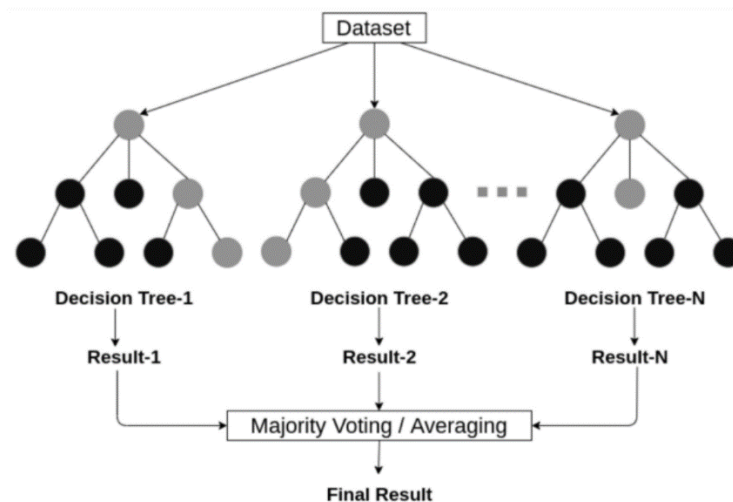


Figure 2 A structure of RF made up of multiple DTs (Sarker 2021)

2.5.2.5 Support Vector Machines

Support Vector Machines (SVMs) are another classifier that try to find the optimal hyperplane that separates classes with the maximum margin. They are well-suited for high-dimensional spaces and

can effectively model complex non-linear relationships. SVMs are particularly useful for smaller datasets and have been widely applied in the health domain due to their robustness and generalization capabilities (Sarker 2021; Shah et al. 2018; Sarno et al. 2022; Hang et al. 2021).

2.5.2.6 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is an advanced implementation of gradient boosting algorithms. Like Random Forests, it is an ensemble of decision trees, but it builds trees sequentially and improves model accuracy by minimizing a loss function using gradient descent. XGBoost incorporates second-order gradient information and regularization techniques (L1 and L2), making it resistant to overfitting while maintaining fast performance. It is considered one of the most accurate and efficient models for structured data problems (Sarker 2021).

2.5.3 Hyperparameter Tuning

Hyperparameter tuning plays a major role in ML, as hyperparameters control the learning process and significantly influence the behavior and complexity of the model. Unlike model parameters, which are learned from data during training, hyperparameters must be carefully selected beforehand. By properly adjusting the hyperparameters, the model may better capture patterns and prevent overfitting while also improving performance and generalization (A Ilemobayo et al. 2024).

In different ML models, various hyperparameters contribute to model performance. For decision trees and random forests, parameters such as tree depth, minimum samples per leaf, and the number of trees directly affect model complexity. Similarly, in gradient boosting machines like XGBoost, hyperparameters like learning rate, number of estimators, and maximum tree depth are crucial for capturing complex patterns. Support vector machines (SVM) rely on kernel parameters, such as gamma in the radial basis function (RBF) kernel, to manage non-linearly separable data effectively (A Ilemobayo et al. 2024).

Several approaches are used to optimize hyperparameters. Grid search is a brute-force method that systematically examines all possible combinations within a predefined set of hyperparameters. Although simple to implement, grid search can be computationally expensive when the number of hyperparameters is high (Hossain R, et al. 2021).

Bayesian optimization takes a more probabilistic approach, building a surrogate model to approximate the objective function (A Ilemobayo et al. 2024). It is more efficient than grid and random search because it can detect optimal combinations of hyperparameters by analyzing previously tested values (Hossain Riyad, Timmer Douglas, and Moya Hiram 2021). Techniques such as cross-validation and regularization further help mitigate overfitting risks during the tuning process, improving model robustness (A Ilemobayo et al. 2024).

Hyperparameter tuning is often challenging due to computational costs. Many ML software packages offer default hyperparameter settings which are refined to increase model accuracy and efficiency. The right selection of hyperparameters ensures optimal performance and better generalization which makes hyperparameter tuning an essential step in developing high-performing ML models (A Ilemobayo et al. 2024).

2.6 Model Evaluation

A ML model is validated by evaluating its prediction performance. Ideally, this evaluation should be representative of how the model would perform when used in a real-life setting. Metrics allow to quantify the performance of an ML model and for classification tasks, the metrics results can be summarized in a matrix called the confusion matrix. For binary classification, the confusion matrix divides the test samples into four categories, depending on their true and predicted labels (Varoquaux and Colliot 2023):

		True label	
		Positive $D+$	Negative $D-$
Predicted label	Positive $T+$	TP	FP
	Negative $T-$	FN	TN

Figure 3 Confusion Matrix (Varoquaux and Colliot 2023)

True Positives (TP): are samples for which the true and predicted labels are both 1.

True Negatives (TN): are samples for which the true and predicted labels are both 0.

False Positives (FP): are samples for which the true label is 0 and the predicted label is 1.

False Negatives (FN): are samples for which the true label is 1 and the predicted label is 0 (Varoquaux and Colliot 2023).

Accuracy: is defined as the proportion of true instances retrieved, both positive and negative, among all instances retrieved and is calculated by:

$$Accuracy : A = \frac{tp + tn}{tp + tn + fp + fn} \text{ (Dalianis 2018)}$$

Precision: measures the number of correct instances retrieved divided by all retrieved instances.

Recall: measures the number of correct instances retrieved divided by all correct instances.

The F1-score: means the harmonic mean between precision and recall (Dalianis 2018).

$$Precision : P = \frac{tp}{tp + fp}$$

$$Recall : R = \frac{tp}{tp + fn}$$

$$F\text{-score} : F_{\beta} = (1 + \beta^2) * \frac{P * R}{\beta^2 * P + R} \text{ (Dalianis 2018)}$$

Specificity: measures the proportion of negatives that are correctly identified as negative, or not having the condition.

Sensitivity (the same as recall): measures the proportion of positives that are correctly identified (Dalianis 2018).

Negative predictive value (NPV): A fraction of the negatively classified samples that are indeed negative (Varoquaux and Colliot 2023).

Positive predictive value (PPV): corresponds to precision (Dalianis 2018)

$$PPV = \frac{TP}{TP + FP}$$

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}} \text{ (Varoquaux and Colliot 2023)}$$

The ROC curve plots the Sensitivity as a function of $1 - \text{Specificity}$. It is the probability that a positive sample has a higher classification score (as positive) than a negative sample. A perfect classification corresponds to an ROC AUC of 1 and a random classification to an ROC AUC of 0.5. Chance stays at 0.5 regardless of the class imbalance, but for severely imbalanced classes, the ROC curve is less useful because a small difference on specificity or sensitivity might have a significant impact on the application without significantly altering the ROC curve. Therefore, it is often used together with the precision–recall (PR) curve that focuses on the minority class (Varoquaux and Colliot 2023).

The PR curve plots the precision as a function of recall and can be summarized using the PR AUC, also called average precision. Just like the ROC AUC, a perfect classification corresponds to a value of 1. However, unlike for ROC AUC, a chance classification does not necessarily lead to a value of 0.5. It depends on the prevalence (Varoquaux and Colliot 2023). Model evaluation can have many different purposes. In some cases, high recall is considered a priority over high precision, and in some cases, it is the opposite (Dalianis 2018).

n-Fold Cross-Validation is also a technique used in ML to evaluate model performance. It involves randomly partitioning a dataset of m labeled examples into n equal subsets (folds). The process follows these steps; initially, the model is trained on $n-1$ folds and tested on the remaining fold. Then this process repeats n times, each time using a different fold for testing (Mohri, Rostamizadeh, and Talwalkar 2012).

The average accuracy across all folds is computed as the cross-validation accuracy. This helps in model selection, where the best parameter is chosen based on the highest accuracy. Typically, n is set to 5 or 10 in ML applications to balance bias and variance while ensuring reliable model evaluation (Mohri, Rostamizadeh, and Talwalkar 2012).

2.7 The role of ML in predicting APOs

ML models have become increasingly valuable in the healthcare domain, particularly for predicting APOs and complications associated with pregnancy. These models offer advanced

abilities to process and analyze complex, high-dimensional data from diverse sources such as EMRs, self-reported questionnaires, and national demographic health surveys. For example, predictive models for HDP built using self-reported early pregnancy lifestyle data have shown a remarkable performance, achieving an area under the curve (AUC) of 0.93 with SVM classifiers, significantly outperforming traditional statistical methods (MIZUNO et al. 2024). Similarly, EMR-based models using algorithms like logistic regression, random forests, and XGBoost have shown high accuracy and recall in identifying patients at risk of maternal and neonatal complications (Hang et al. 2021). ML algorithms such as ensemble methods and neural networks have also shown a near-perfect performance (AUC = 0.98) in predicting neonatal mortality based on clinical data (Sheikhtaheri et al. 2021).

These models are capable of identifying and interpreting critical risk factors, increasing clinical decision-making. Techniques such as normalized regression coefficients, Gini importance, and SHAP values have been applied to interpret and rank the influence of different features (MIZUNO et al. 2024; Ngusie et al. 2024; Bogale, Abuhay, and Dejene 2022). Across these studies, ML has proven to be effective in prediction tasks particularly in handling non-linear relationships and high-dimensional datasets which shows their advantage over conventional statistical methods.

2.8 Related Works

Predicting poor perinatal outcomes in women with HDP has attracted a lot of attention from researchers, who have looked at biomarkers, clinical traits, and advanced computational models. These studies aim to improve risk stratification, enhance clinical decision-making, and optimize resource allocation.

The systematic review and meta-analysis conducted by (Bucher et al. 2024) provides a comprehensive evaluation of predictive tests for APOs among women with HDP. The authors emphasize that early identification of women at high risk for complications could facilitate appropriate triage and resource allocation, ultimately leading to targeted interventions. For perinatal outcomes, biomarkers such as the soluble FMS-like tyrosine kinase-1 (sFlt-1)/placental growth factor (PlGF) ratio showed promising predictive performance; however, they lacked external validation.

The study conducted by (Rana et al. 2012) adds to this discourse by investigating the predictive value of circulating angiogenic factors, specifically the soluble fms-like tyrosine kinase 1 (sFlt1) and placental growth factor (PlGF), in assessing APOs among women with suspected preeclampsia. The researchers analyzed a cohort of 616 women who presented for evaluation of suspected preeclampsia, measuring plasma levels of sFlt1 and PlGF at presentation. The regression results indicated that an elevated sFlt1/PlGF ratio was significantly associated with adverse maternal and perinatal outcomes within two weeks of presentation. The study found that the highest tertile of the sFlt1/PlGF ratio was linked to a staggering odds ratio of 47.8 for adverse outcomes, indicating its strong predictive capability. Furthermore, the addition of the sFlt1/PlGF ratio to traditional clinical parameters such as hypertension and proteinuria significantly improved prediction accuracy (AUC = 0.93 vs. 0.84), highlighting its potential use in risk stratification and management decisions for women with suspected preeclampsia.

The article by (Magee et al. 2016) investigates the predictive ability of clinical characteristics for APOs among women with HDP. The study analyzed data from 987 women with chronic or gestational hypertension, using logistic regression to evaluate 19 candidate predictors, including maternal age, body mass index, and blood pressure measurements. Despite the comprehensive approach, the findings revealed that none of the predictive models achieved an area under the receiver operating characteristic curve (AUC ROC) of 0.70 or higher for most adverse outcomes, indicating insufficient predictive accuracy. Notably, only severe hypertension and delivery before 34 weeks demonstrated borderline predictive value (AUC ROC = 0.70 and 0.71, respectively). The authors conclude that when women with chronic hypertension develop elevated blood pressure during pregnancy, existing clinical predictors are inadequate for reliably predicting APOs. But this study did not use any laboratory serum or urinary biomarkers which have been regarded as predictors in the other studies.

Two studies conducted in China focused on developing and validating predictive models for adverse maternal and neonatal outcomes among women with HDP, using advanced methods and ML algorithms. The first study (Chen et al. 2022) used EMRs from a birth cohort in Beijing, China, to create prognostic models for adverse maternal and neonatal outcomes in women with de novo HDP. Adverse outcomes included various maternal and neonatal complications. Multiple ML algorithms, including random forest and boosted tree methods, were employed. Predictor

variables, such as gestational age at diagnosis, fetal measurements, and laboratory results, were selected based on their association with outcomes and validated using 10-fold cross-validation. The maternal model achieved an AUC of 0.972, indicating high predictive accuracy, while the neonatal model had an AUC of 0.925. For adverse maternal outcomes, the random forest algorithm performed best, whereas for adverse neonatal outcomes, the boosted tree algorithm was the top performer, followed by random forest. This study used a wide range of clinical variables readily available in EMRs and utilized relatively large datasets to have a robust model performance.

The second study (Sarno et al., 2022.), a retrospective cohort analysis, compared ML models and LR in predicting adverse maternal and neonatal outcomes among 867 women diagnosed with preeclampsia. Two separate models were developed, one for maternal outcomes and another for neonatal outcomes, and AUC values were calculated to evaluate the discrimination efficiency on the testing datasets. For maternal outcomes, the SVM model achieved the highest performance, with an AUC of 0.976 and sensitivity, specificity, PPV, and NPV of 92.3%, 92.3%, 83.3%, and 96.6%, respectively. For neonatal outcomes, RF (sensitivity: 93.8%, specificity: 87.9%, PPV: 91.7%, NPV: 90.9%) and multilayer perceptron (MLP) were the top-performing models, with AUC values of 0.967. The study also noted that predictions spanning longer time intervals tend to have lower accuracy, suggesting that short-term predictions might be more effective when integrated into a customized predictive set for individual risk assessment.

(Payne et al. 2014) and a study conducted in India (Kumar et al. 2021) both used LR to develop predictive models for adverse outcomes associated with HDP in LMICs. (Payne et al. 2014) developed the miniPIERS (Pre-eclampsia Integrated Estimate of RiSk) model to predict adverse maternal outcomes among women with HDP in LMICs. Using data from 2,081 women across five LMICs, the researchers identified key predictors, including parity, gestational age, symptoms (e.g., headache, chest pain), systolic blood pressure, and proteinuria. LR analysis yielded a model with an area under the receiver operating characteristic curve (AUC ROC) of 0.768 for internal validation and 0.713 for external validation, showing moderate predictive performance. However, the authors also acknowledged limitations, such as modest sensitivity (41.4%) at the recommended threshold for high-risk classification, and emphasized that the model should be externally validated in community-based settings before routine implementation.

Similarly, the study conducted in India aimed to develop a predictive model for late stillbirth among women with HDP in LMICs. This observational, analytical case-control study was conducted from November 2015 to December 2020 at a single tertiary care center. LR was employed to analyze all stillbirths occurring in the facility, with controls being age- and parity-matched women with HDP who had live births. The predictive model achieved a sensitivity of 80.3% and specificity of 70.03%, with an area under the curve (AUC) of 0.846, indicating good predictive capability.

Research using multivariable regression analysis on the perinatal outcomes of women with HDP at Jimma Medical Center found that these women experienced a significantly higher incidence of adverse outcomes compared to normotensive women (64.1% vs. 32.8%). The study reported that women with HDP had increased risks for low birth weight (adjusted RR = 2.88), preterm birth (adjusted RR = 2.31), and perinatal mortality (adjusted RR = 3.88). These findings further highlight the need for healthcare providers to enhance prevention and management efforts for HDP to improve perinatal outcomes (Jaleta, Gebremedhin, and Jebena 2021).

Overall, this limited review demonstrates a range of research aimed at predicting distinct adverse conditions in pregnancies. It emphasizes how important it is to create highly sensitive predictive models in order to identify women who may experience unfavorable outcomes. A variety of approaches are used in the reviewed studies, ranging from ML models that employ various predictors to statistical methods. Conclusions from these studies show that various predictive models are effective; some of them achieve high area under the curve (AUC) values, which indicate strong predictive performance. Nonetheless, the above-mentioned researches use predictors that require advanced resources for testing (Rana et al. 2012), used statistical methods, didn't incorporate relevant predictors and few are focused on a subtype of HDP. So, this study tried to generate a prediction model using ML algorithms that can effectively predict APOs associated with HDP using routinely collected maternal and fetal data found in hospitals that can be utilized in low-resource settings.

CHAPTER-THREE

3. Data and Methodology

ML has revolutionized health science research by its ability to handle complex data and perform various tasks. ML models excel at finding intricate and unexpected patterns within datasets that traditional methods might miss. The health science field, rich with data from genomics sequencing and electronic health records (EHR), is particularly well-suited for these advanced models (Cho et al. 2024). An experimental research design was used to achieve the objectives of this research and have reliable results.

3.1 Data Source

This study was conducted based on selected hospitals within Addis Ababa, the capital city of Ethiopia. The city's healthcare facilities include tertiary-level hospitals equipped with EMR systems. The EMR is widely regarded as a critical health information technology tool for improving the quality of medical care (Oumer et al. 2021). The data for this study was sourced from EMRs of selected hospitals, which document different clinical, demographic, and obstetric information. Hospitals were selected based on accessibility of high-quality EMR data and volume of maternal healthcare services provided. The selected hospitals were Gandhi Memorial General Hospital, Zewditu Memorial Hospital, Yekatit 12 General Hospital Medical College and Tirunesh Beijing hospital. These hospitals share a common EMR system which is a web application designed by Pulse Techs. This ensures consistency in attribute structure, naming, and data formatting across facilities. Despite having this uniform system, there are differences in data completeness and the number of records contributed by each hospital, which can be attributed to varying levels of resource availability, documentation practices, and patient volume. Secondary data was extracted from EMRs and exported into a standardized format (CSV).

The dataset exclusively included pregnant women diagnosed with the different types of HDP. This selection criteria reflects the objective of the study which is to build a ML model for prediction of adverse neonatal outcomes among women with HDP. By restricting visit time to early gestational ages, the model created will support early identification and intervention, potentially improving maternal and neonatal care.

Name of Hospital	Number
Gandhi Memorial General Hospital	314
Zewditu Memorial Hospital	226
Yekatit 12 General Hospital Medical College	423
Tirunesh Beijing hospital.	79

Table 1 Name of hospital and amount of data from each hospital

Each record in the dataset contained the earliest recorded antenatal visit for the mother, along with the corresponding delivery summary. This structure was important because it enabled the model to make predictions based only on information available at or shortly after the initial ANC visit. This represents situations where decisions must be made early, before full pregnancy progression. The delivery outcomes are used as the target variable, identifying whether or not an adverse neonatal outcome occurred (e.g., stillbirth, low birth weight or other complications).

Across the four hospitals, the dataset comprised a total of 1042 patient records and 71 attributes. These attributes covered a wide range of data types and domains:

Category	Attributes
Demographic Data	Age, Marital status, Education, Occupation
Clinical Measurements	BP, BMI, SPO2, MUAC, PR, RR
Lab tests	CBC, LFTs, Urine analysis, Kidney function tests
Pregnancy related data	Gestational age, ultrasound result (lie, presentation)

Table 2 Summary of Categories and Attributes in the Data

The data was a mix of categorical and continuous variables, with float-type values representing clinical measurements and lab results, object-type values used for patient history and qualitative assessments and integer data types representing ID of patients and age.

Data Type	Number of Attributes
Float	36 Columns
Object & Categorical	33 columns
Integer	2 columns

Table 3 Summary of data types in the data

3.2 Variables

The outcome variable in this study was APOs; a binary variable identifying participants as having adverse outcome and not having adverse outcome. The most prevalent APOs selected as outcomes of interest were preterm birth, low birth weight, still births (Gedefaw, Alemnew, and Demis 2020). Predictor selection was guided by the literature and their availability in the respective EMRs. The predictors initially included in this study are shown in the table below.

No	Attribute Name	Description	Data type
1	MRN	ID of the patient	Integer
2	Age	Age of the patient	Integer
3	Marital Status	Marital Status of the mother	Object
4	Educational Level	Educational Level of the mother	Object
5	History	Past Medical or Surgical history of the mother	Object
6	Occupation	Occupation of the mother	Object
7	Birth type	Type of birth (single, twin, or multiple gestation)	Object
8	Gestational Age	The first recorded gestational age of the mother	Object
9	SBP/DBP	Blood pressure of the mother	Object
10	Height	Height of the mother	Object
11	Weight	Weight of the mother	Object
12	BMI	Body mass index of the mother	Object
13	PR	Pulse rate of the mother	Float
14	RR	Respiratory rate of the mother	Object
15	Temperature	Temperature of the mother	Object
16	SPO2	Oxygen saturation level	Float
17	Blood Group	Blood Group of the mother	Object
18	WBC	White blood cell count	Float
19	Neu	Neutrophile count	Float
20	Lym	Lymphocyte count	Float
21	Mono	Monocyte count	Float
22	eos	Eosinophile count	Float
23	baso	Basophile count	Float
24	RBC	Red blood cell count	Float
25	hemo	Hemoglobin level	Float
26	hema	Hematocrite level	Float
27	mcv	Mean cellular volume	Float
28	mch	Mean corpuscular hemoglobin	Float
29	mchc	Mean corpuscular hemoglobin concentration	Float
30	RDW-CV	Red Cell Distribution Width	Float
31	Reticulocyte	Reticulocyte count	Float
32	PDW	Platelet Distribution Width	Float
33	MID	Mid-sized Cell Count	Float
34	MPV	Mean Platelet Volume	Float
35	Platelet	Platelet count	Float
36	Appearance	Appearance of the urine sample	Object

37	PH	PH level of the urine	Float
38	SG	Specific gravity of the urine	Float
39	Protein	Amount of protein found in the urine	Object
40	Glucose	Amount of glucose found in the urine	Object
41	Ketone	Amount of ketone in the urine	Object
42	Ester	Leukocyte esterase amount in the urine	Object
43	Blood	Amount of RBCs found in the urine	Object
44	Bilirubin	Amount of bilirubin found in the urine	Object
45	Urobilin	Amount of urobilinogen found in the urine	Object
46	Nitrite	Amount of nitrite in the urine	Object
47	Urea	Amount of urea in the urine	Float
48	BUN	Blood urea nitrogen in the blood	Float
49	Creatinine	Amount of creatinine in the blood	Float
50	Cr-Clearance	Creatinine clearance test	Float
51	AST	Amount of aspartate transaminase found in the serum	Float
52	ALT	Amount of alanine transaminase found in the serum	Float
53	ALP	Amount of alkaline phosphatase found in the serum	Float
54	Total bilirubin	Amount of total bilirubin found in the serum	Float
55	Direct Bilirubin	Amount of direct bilirubin found in the serum	Float
56	Total protein	Amount of protein found in the serum	Float
57	Albumin	Amount of albumin found in the serum	Float
58	Globulin	Amount of globulin found in the serum	Float
59	GA	Gestational age of the newborn	Object
60	SDP	Systolic blood pressure	Float
61	DBP	Diastolic blood pressure	Float
62	MUAC	Middle upper arm circumference	Object
63	Malnutrition	Whether the mother is malnourished or not	Object
64	Anemia	Presence or absence of anemia	Object
65	Presentation	Part of the baby that is leading the way through the birth canal	Object
66	Lie	Position of the fetus in relation to the mother's uterus	Object
67	Physician Diagnosis	Diagnosis of the mother	Object
68	HC	Head Circumference	Object
69	Neonatal Status	Outcome of a newborn	Object
70	Stillbirth	Whether stillbirth has occurred or not	Object
71	Assessment	Diagnosis of the newborn	Object

Table 4 List of all variables included initially

3.3 Exploratory Data Analysis

The first step of this stage involved importing the datasets into a Python environment using pandas. Four Excel files which were obtained from four hospitals were loaded, each containing clinical and demographic records.

An initial inspection of the data revealed the composition of the data, which had different datatypes including numerical variables (e.g., age, hemoglobin, blood cell counts) and categorical variables (e.g., marital status, blood group, physician diagnosis). There were also some misclassified numerical data type attributes such as 'BMI', 'height', 'weight'... into object data type which was processed in the later stages.

There was also significant missing data across some columns. Variables, such as different lab measurements like RDW_CV and Globulin, had high amount of missingness possibly due to these tests being unavailable or rarely ordered. Others, such as height, BMI, and blood_group_rh, had moderate missingness, while key clinical indicators like blood pressure (SBP and DBP) were fully recorded.

Furthermore, there is evidence of inconsistent data formatting (e.g., blood pressure recorded as strings like "120/80") and variability in categorical value casing and spelling (e.g., "Married", "married", "MARRIED"), since the EMRs rely heavily on manual data entry. Thorough preprocessing, cleaning, and imputation strategies was applied to increase data quality and make it usable for ML models in the following stages.

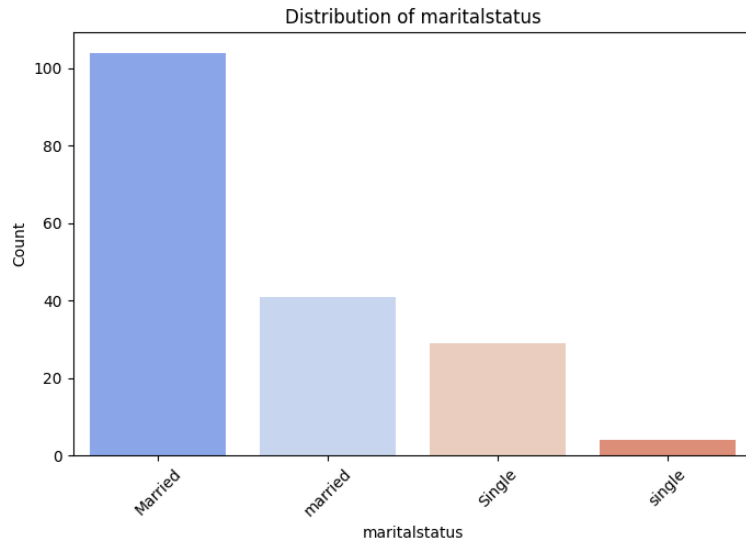


Figure 4 Example of inconsistency in categorical attribute

3.4 Data Preprocessing

3.4.1 Data Cleaning

I. Data Integration

As part of the initial preprocessing steps, column name standardization was performed across the datasets to ensure consistency. The column headers in the datasets were processed to:

- A. Convert all characters to lowercase (e.g., "AGE" → "age") which prevents case-sensitivity issues during column referencing.
- B. Remove leading/trailing whitespace which eliminates hidden formatting errors if there are any.
- C. Replace spaces with underscores for better readability and compatibility with Python variable naming conventions (e.g., ['marital status'] -> ['marital_status'])

Before merging, the column names of the datasets were also renamed and standardized to match and have consistent naming across all the data sets. Then, all four data sets were integrated while retaining all instances. Datasets that contained different number of attributes were merged based on their common attributes.

II. Data Standardization

During the initial data exploration, significant inconsistencies were identified in the *'bmi'*, *'height'*, *'weight'* and *'rr'* variables, in which the attributes contained mixed data types and formats:

- Numeric values (e.g. 59, 64, 57)
- String values with units (e.g. '63.00kg', '158 Cm', "20'")
- Inconsistent formatting (e.g. '63.00KG' & '65.00 kg', '159 CM' & '158 Cm') and
- Invalid entries (e.g. ' ', '-')

To ensure numerical consistency for analysis, data transformation was applied which converted all string values to numeric by extracting numbers from strings with units and removing whitespace and case sensitivity issues

Next, as part of the data preprocessing step, the unique identifier column *'mrn'* was removed since the presence of the identifier served no statistical or ML purpose. Removal of this attribute prevents the potential bias in models that might incorrectly interpret sequential IDs as meaningful numerical data.

Then categorical variables were standardized. Different invalid entries ('-', '=') were converted to missing values (NaN) which ensures only valid categorical values remain for analysis. Whitespace Removal, Case Normalization and Value Consolidation (e.g. "term," to "term", "neg" to "negative") were applied across the dataset to have consistent value entries. Other blank space entries were also properly converted to missing values.

The gestational age variable (*'ga'*) had significant formatting inconsistencies:

- Week+day notation ("25+4", "20+3")
- Free-text annotations ("19+2 weeks", "26 wks")
- Inconsistent spacing (" 26+5", "22 wk")
- Standalone numbers (23, 17)
- Irregular values ("23+", "20+ week")

From these entries, the week component from week+day notation was initially selected. Then any textual annotations were removed. All values were reduced to whole weeks only which maintained

essential gestational age information and enables proper numerical analysis. It resolved 100+ variations to 14 discrete values.

Functions were also created that standardize clinical status classifications from text entries. The first function was applied on the ‘assessment’ attribute which contained information about the new born neonate. On this attribute;

- 15+ observed spelling variations were handled and distinctions between low-birth-weight classifications, preterm conditions and normal cases were made following the following logic:

- If “VLBW”/ ”LBW” with preterm -> Preterm regardless of weight
- If "LBW" only -> Low Birth Weight and
- If “other” -> Normal (‘no’) classification.

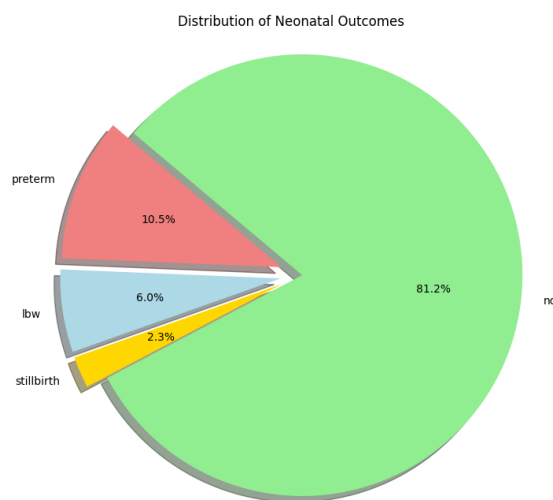


Figure 5 Pie chart showing distribution of APOs

The second function was applied on the ‘physician diagnosis’ attribute which contained information about the type of HDP the mothers were diagnosed with. Based on the inputs, the HDP were classified into four namely;

- GHTN: Gestational hypertension
- PE: Preeclampsia without severe features
- Preeclampsia w SF: Preeclampsia with severe features and
- CHTN: Chronic hypertension

Overall, these standardization processes transformed a highly inconsistent data values into clean, analysis-ready numerical and categorical data types. These steps improved the quality of the variables by eliminating inconsistent formatting, reducing arbitrary variation, properly handling invalid entries and ensuring consistency across all datasets.

III. Missing Value Handling

Initially, attributes with high amount of missing data were removed since imputing them may not be representative and become potentially misleading, especially in a clinical setting.

No	Attribute Name
1	RDW-CV
2	PDW
3	MID
4	Reticulocyte
5	Occupation
6	HC
7	Educational Level
8	Marital Status
9	Total protein
10	Height
11	Albumin
12	Globulin
13	History
14	Ester
15	Bilirubin
16	BUN
17	Creatinine
18	Cr-Clearance
19	Direct Bilirubin
20	SPO2
21	Temperature
22	SBP/DBP
23	Weight

Table 5 List of variables removed due to relatively high amount of missingness

Before imputation, Little's MCAR test, done using the 'naniar' library in R Studio, confirmed that the missing data followed a Missing Completely at Random (MCAR) pattern. After that, a combination of methods to handle missing data were used. Median imputation, KNN, MICE and mode were selected based on the characteristics of the data.

Median imputation was used for continuous variables, ensuring central tendency preservation without being distorted by outliers (Hang et al. 2021). It is also fast and simple to apply. KNN imputation was chosen for mixed-type data, benefiting from its ability to utilize structural and contextual information (Hang et al. 2021; Ngusie et al. 2024). MICE was selected for continuous variables since it can deal with complex inter-variable relationships, particularly where the preservation of multivariate patterns and accurate reflection of uncertainty are important (Li et al. 2024; Hang et al. 2021). For the categorical variables, mode (most frequent value) was used for imputation.

After applying the above four imputation techniques, the resulting datasets were three each with different imputation methods. Number of instances was the same across all datasets and consistency of data types for each variable in each dataset was also the same. This imputation helped in analyzing the data without reducing sample size.

IV. Feature Engineering

In this study, one created attribute was 'outcome', which was the target variable for predicting presence or absence of adverse outcome based on existing attributes in the data. The attribute was derived from indicators 'stillbirth' (occurrence of still birth), gestational age (gestational age when the child was born), 'Neonatal status' and 'assessment' (neonatal assessment given by the physician), using a conditional function that categorizes observations into 'yes' or 'no' based on predefined criteria. If a patient was not classified as stillbirth and had not been medically assessed, the outcome was assigned as 'no' (no adverse outcome), while cases involving preterm birth, low birth weight, or confirmed stillbirth were categorized as 'yes' (having adverse outcome).

Stillbirth	Assessment	Gestational age	Neonatal status	Outcome
No	No	Term	Alive	No
No	Preterm	Preterm	Alive	Yes
No	Low birthweight	Term	Alive	Yes
Yes	Any	Any	Not-Alive	Yes

Table 6 Outcome variable engineering

These four attributes were also dropped after creating the outcome attribute. By including this new feature, the dataset can further be used for binary classification models.

V. Variable encoding

One hot encoding was applied on the categorical attributes since none of the attributes had order in the classes. Label encoding was applied on the outcome variable to convert it into a model friendly format.

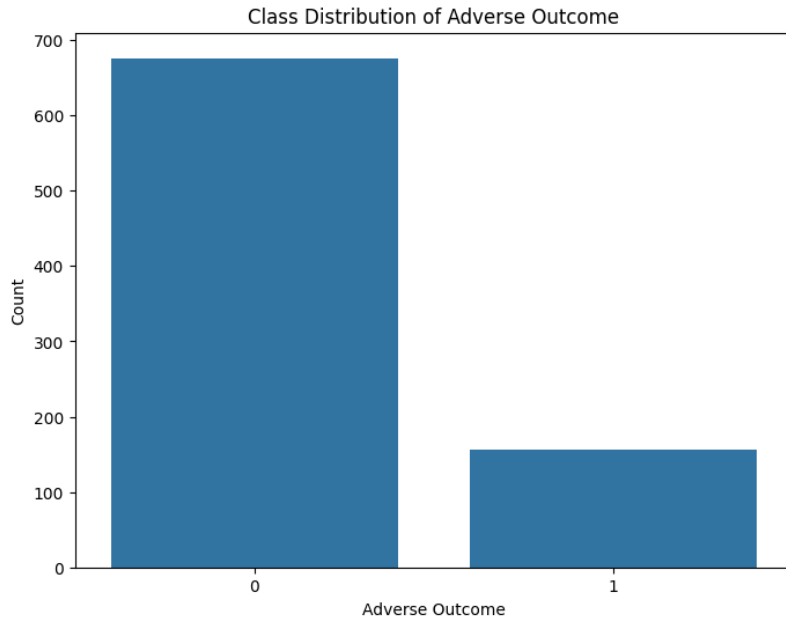


Figure 6 Label encoded target class and distribution

VI. Data Splitting into Train-Test sets

Next, the data was split into an 80-20 ratio, where 80% of the dataset was chosen as the training sample and 20% as the test sample. The 80-20 ratio was used since the amount of positive class labels is relatively small. When the labeled sample is relatively small, the amount of training data is often chosen to be larger than the test data since the learning performance of the model directly depends on the training sample (Mohri, Rostamizadeh, and Talwalkar 2012).

Class	Number of Training sample	Number of Testing sample	Total
0 (non-risk women)	676	170	846
1 (at-risk women)	157	39	196

Table 7 Number of instances in training and testing set

VII. Data Scaling/Normalization

In this study, data normalization was applied on all the numeric features found in the training dataset after splitting the data into training and test set. This made sure that no variable dominates because of differences in units or scales and avoid data leakage.

VIII. Feature Selection

The attribute ‘muac’ (middle upper arm circumference), a binary attribute with values > 23 or < 23 which is also an indicator of malnutrition, was removed because it was highly correlated with the attribute ‘malnutrition’ (0.96).

Other variable dropped in the feature selection process was the variable ‘lie’, an attribute that refers to the orientation of the fetus in the womb relative to the mother's spine, which offers limited clinical relevance to outcome since it had the same value for all instances (‘longitudinal’).

3.5 Selection of ML models

The choice of ML models largely depends on the type of task and data type (Cho et al. 2024). Different supervised ML models were selected for use in this study since they are used for classification tasks and based on their wide application in similar public health setting.

Models	Reason for Selection
LR	Baseline model Simple, fast and Interpretable
NB	Requires small amount of data Fast and Simple
SVM	Good for high dimensional & small data Handles complex boundaries well
DT	Handles non-linear relationship Interpretable
XGBoost	Good for small and structured data High Accuracy Resistant to overfitting & Fast
RF	Good for high dimensional & small data Handles complex interaction Reduces Overfitting

Table 8 Selected ML models for experiment

3.6 Software and Tools

Python in google colab and R-studio version 2024.12.1 version tools were used for the various data handling, cleaning, application of ML models and visualization of the study.

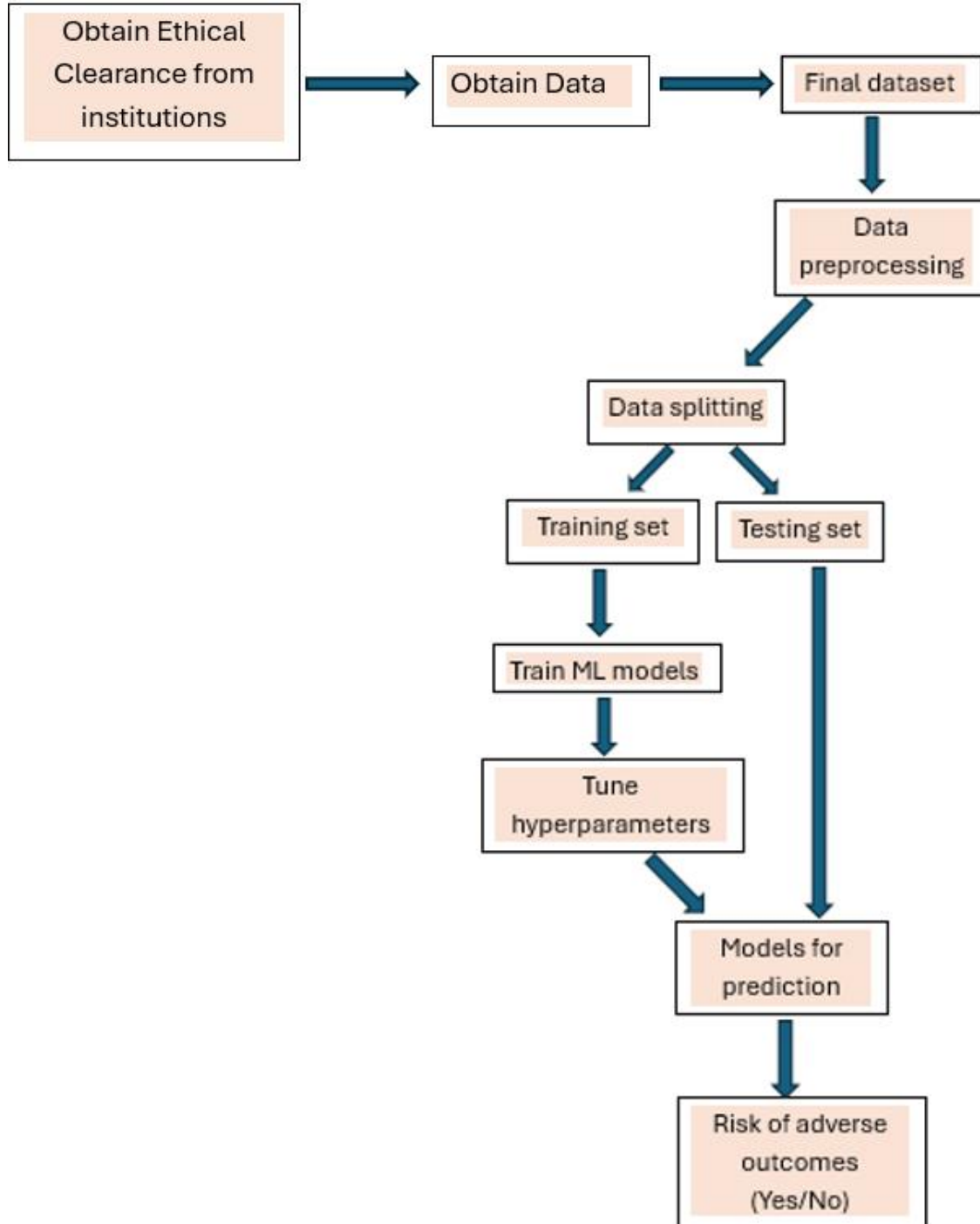


Figure 7 Architecture of adverse outcome risk prediction model

CHAPTER FOUR

4. Experimentation and Analysis of Results

4.1 Overview

After the data was prepared following different interrelated preprocessing steps which resulted with three final data sets that used different imputation techniques (median, MICE, KNN) to handle missing data of the original dataset. Each dataset contained 1042 instances with 42 attributes. Then, using each imputation method, six different ML models were trained, ensuring each model was evaluated across all three techniques using cross-validation. From each combination, the best-performing version of each model was selected based on evaluation of classification report of the cross-validation. Next, Bayesian optimization and Grid Search were used to fine-tune hyperparameters, comparing their effectiveness. After selecting the best parameter configurations between the two approaches, the final six models were also evaluated using cross-validation on the training set. Performance metrics, like the classification report, was used to compare models. Finally, the optimized models were applied to the test set for final evaluation, measuring their generalization and classification abilities of instances into risk and no risk.

4.2 Experimentation

Model 1: LR

In this section, the first prediction model was built using logistic regression. The type of logistic regression used was binary logistic regression where the dependent variable can only take two values. The ‘sklearn’ library in Python was used to import and build the LR model.

➤ Model Training

The training set was fitted into a created function that trains a LR model and evaluates it using stratified K-Fold cross-validation (5 splits). The model was initially created with default hyperparameters. The function then performs predictions on the cross-validation, ensuring that the model is tested on different subsets of the training data while preserving class distribution with stratified sampling. Finally, it prints a classification report showing accuracy, precision, recall, and

F1-score for performance testing of the model which was useful for evaluating the LR model before hyperparameter tuning.

So, initially, the three datasets with different imputation strategies were fitted using the created function. Of the three, the one that performed well was selected for further hyperparameter tuning and prediction using the LR model.

After training the LR model using three different imputation methods, the version trained on the dataset imputed with the MICE technique achieved the best performance on the training set.

➤ Hyperparameter Tuning

After selecting the best model from the three initially created, hyperparameter tuning was performed.

- Grid search: A function was created that trains the selected LR model while performing hyperparameter tuning using ‘GridSearchCV()’ with 5-fold cross-validation by defining a dictionary of hyperparameters. The function searches for the best hyperparameter combination by evaluating different combinations and after identifying the best parameters, the model is trained on the best parameters and further evaluated using stratified K-Fold cross-validation. It then prints a classification report of the cross validation to assess model performance.
- Bayesian Optimization: A function was created that trains the selected LR model while performing hyperparameter tuning using ‘BayesSearchCV’. It first defines a search space for hyperparameters with 5-fold cross-validation. After finding the best hyperparameters, the optimized model is retrained using stratified K-Fold cross-validation. Then, it evaluates the model by generating a classification for performance assessment.

From the two methods experimented with, for LR model, hyperparameters tuned with Bayesian optimization yielded a better result.

➤ Model Prediction

After hyperparameter tuning, the final model was trained using the best hyperparameters on the entire training set and used to predict on the test set.

➤ Model Evaluation

After training the tuned LR model, the model was then evaluated on the test set using Accuracy, Precision, Recall, f1-Score, Confusion Matrix, AUC-ROC curve and Precision-Recall curve.

Classification Report

	Precision	Recall	F1-score	Support
0	0.90	0.95	0.92	170
1	0.70	0.54	0.61	39
Accuracy			0.87	209

Table 9 Test set classification report

This classification report shows the LR model's ability to generalize to the unseen test set. It can be seen that the LR model has an accuracy of 0.87. A recall value of 95% & 54% indicates that the model has a strong performance in detecting “no risk” class but struggles with detecting “at-risk” women. This means that the model only captures 54% of actual "at-risk" women and misses 46%. Of the women predicted as at-risk by the model, only 70%, are correctly classified as shown by the precision. The f1-score value of 61% indicates moderate effectiveness for detecting the at-risk class.

Confusion Matrix

After predicting on the test set, a confusion matrix was created to compare the correct and incorrect predictions by the model.

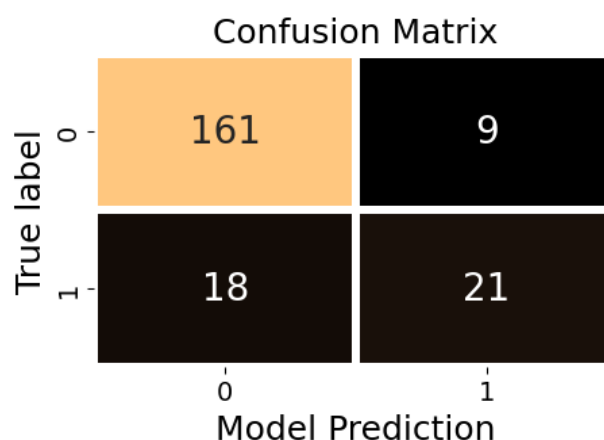


Figure Confusion Matrix of LR model

This confusion matrix shows that of the total 209 test data instances, 182 instances are correctly classified in their respective classes. 161 women from a total of 170 women are correctly classified as no-risk (0) and 21 women from a total of 39 women are correctly classified as risk (1). However, 9 women from a total of 170 women are incorrectly classified as at-risk and 18 women from a total of 39 are incorrectly classified as no-risk.

AUC-ROC curve

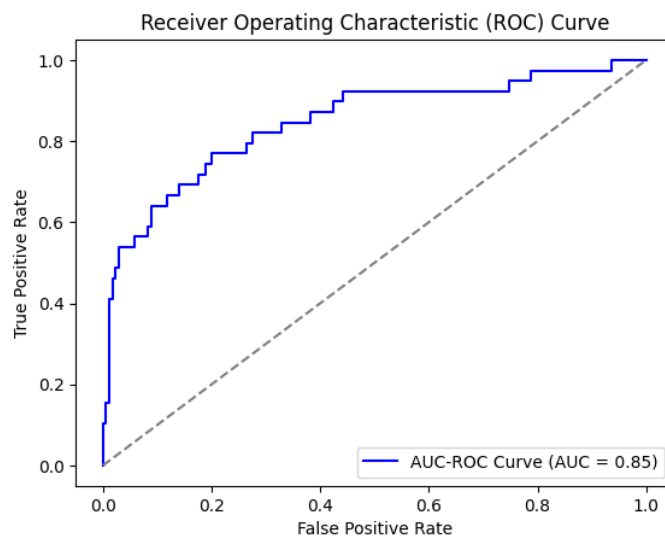


Figure AUC-ROC curve of LR model

The AUC-ROC curve value of 0.85 for the LR model indicates that the model has good discriminatory ability to differentiate between the at-risk women for unfavorable outcome and non-risk women. The plot indicates that the model has a relatively high sensitivity without many false positives.

Precision-Recall curve

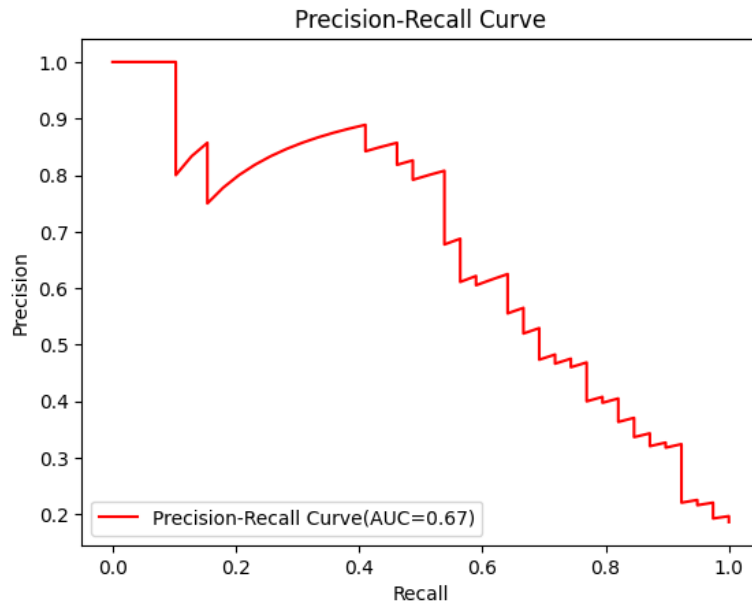


Figure PR-AUC curve of LR model

The LR model achieved a Precision-Recall (PR) score of 67%, which shows a moderate balance between precision and recall in detecting positive instances. The high precision values at low recall values shows that the model predicts few false positives but misses more true positives.

Model 2 : Naïve Bayes

In this section, the second prediction model built was naïve bayes. The ‘sklearn’ library in Python was used to import and build the model.

➤ Model Training

For model training, a function was created with default hyperparameters that trains a NB model and evaluates it using stratified K-Fold cross-validation (5 splits) on a given training dataset. It then performs cross-validation predictions and prints a classification report.

Similarly, as the previous LR model, the three datasets with different imputation strategies were fitted using the created function. Of the three, the one that performed well was selected for further hyperparameter tuning and prediction using the NB model.

For the NB model, the data set imputed with KNN method performed better than the rest.

➤ Hyperparameter Tuning

After selecting the best model from the three initially created, hyperparameter tuning was performed. For NB model, both of the two tuning methods experimented with gave the same result.

➤ Model Prediction

Then, the tuned NB model was used to train on the entire training data and used to predict on the test set.

➤ Model Evaluation

After training the NB model using the best hyperparameters, the model was then evaluated on the test set.

Classification Report

	Precision	Recall	F1-score	Support
0	0.91	0.88	0.89	170
1	0.54	0.64	0.59	39
Accuracy			0.83	209

Table 10 Test set classification report

This classification report shows the model's ability to generalize to the unseen test set. It can be seen that the NB model has an accuracy of 0.83. A recall value of 88% & 64% indicates that the model has a strong performance in detecting "no risk" class but struggles with detecting "at-risk" women. This shows that the model only correctly predicts 64% of actual "at-risk" women and misses 36% of the women. Of the women predicted as at-risk, only 54%, are correctly classified. The f1-score value of 59% indicates slightly poor effectiveness for detecting the at-risk class.

Confusion Matrix

After predicting on the test set, a confusion matrix was created to compare the correct and incorrect predictions by the model.

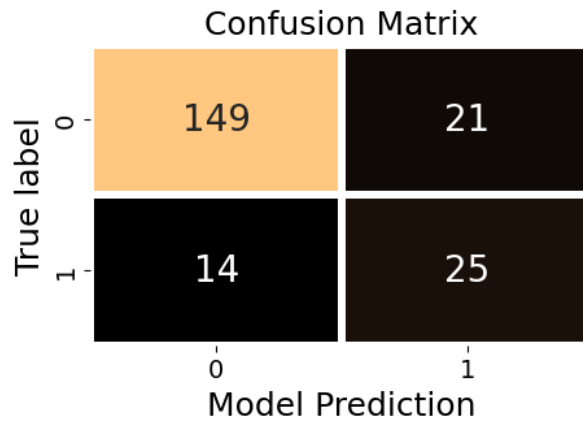


Figure Confusion Matrix of NB model

This confusion matrix shows that of the total 209 test data instances, 174 instances are correctly classified in their respective classes. 149 women from a total of 170 women are correctly classified as no-risk (0) and 25 women from a total of 39 women are correctly classified as risk (1). However, 21 women from a total of 170 women are incorrectly classified as at-risk and 14 women from a total of 39 are incorrectly classified as no-risk.

AUC-ROC curve

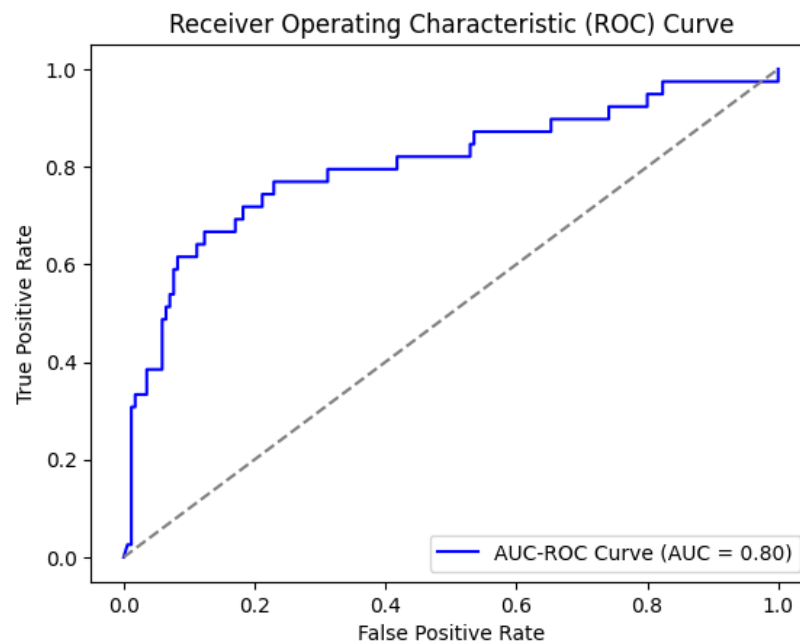


Figure AUC-ROC curve of NB model

The AUC-ROC curve value of 0.80 for the NB model indicates that the model has good discriminatory ability to differentiate between the at-risk women for unfavorable outcome and non-risk women. The curve's blue line, above the dashed diagonal, shows the model performs much better than random guessing (AUC=0.5).

Precision-Recall Curve

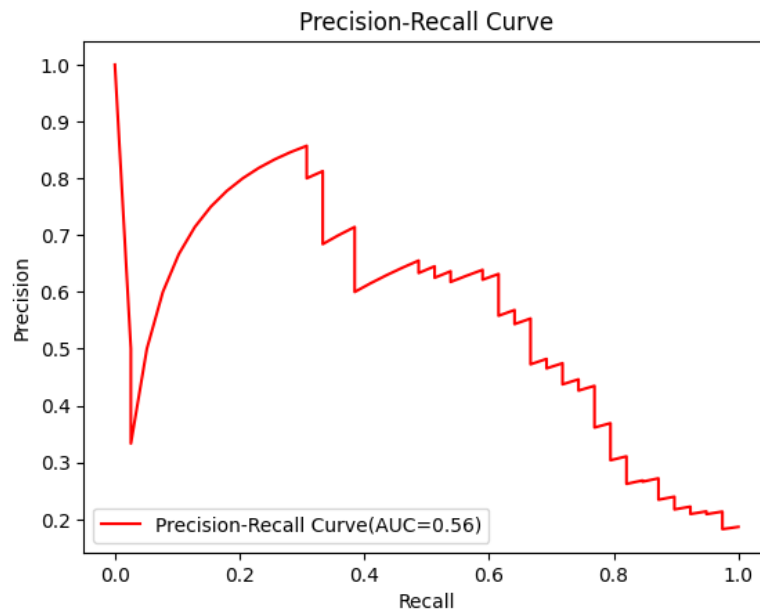


Figure PR-AUC curve of NB model

The NB model achieved a Precision-Recall (PR) score of 56%, which shows a poor balance between precision and recall in detecting positive instances. The curve shows a fluctuating pattern, with precision very initially high at low recall then sharply dropping, and then increasing gradually.

Model 3: SVM

Here, the general process follows the same procedure as the previous two models.

➤ Model Training

So first, similar as the previous models, the three datasets with different imputation strategies were fitted using the created function. Of the three, the one that performed well was selected for further hyperparameter tuning and prediction using the SVM model.

For the SVM model, all of the data set with different imputations performed the same when fitted in the baseline model.

➤ Hyperparameter Tuning

In this model experiment, since all the imputation methods had similar performance, all variants of the data set were further tuned using grid search. After tuning, the dataset that was imputed using MICE method performed slightly better than the rest and was selected.

For this model, the grid search hyperparameter tuning performed better than the Bayesian optimization on the training data.

➤ Model Prediction

Then, the SVM model was used to train on the entire training data and used to predict on the test set.

➤ Model Evaluation

After training the SVM model on the whole training set using the best hyperparameters, the model was then evaluated on the test set.

Classification Report

	Precision	Recall	F1-score	Support
0	0.89	0.98	0.93	170
1	0.83	0.49	0.61	39
Accuracy			0.89	209

Table 11 Test set classification report

From this classification report it can be seen that the SVM model has an accuracy of 0.89 which shows a good overall predictive ability. A recall value of 98% & 49% indicates that the model has a strong performance in detecting “no risk” class but struggles with detecting “at-risk” women. This means the model only captures 49% of actual "at-risk" women and misses more than half. Of the women predicted as at-risk, 83%, are correctly identified. The f1-score value of 61% indicates moderate effectiveness for detecting the at-risk class.

Confusion Matrix

After predicting on the test set, a confusion matrix was created to compare the correct and incorrect predictions by the model.

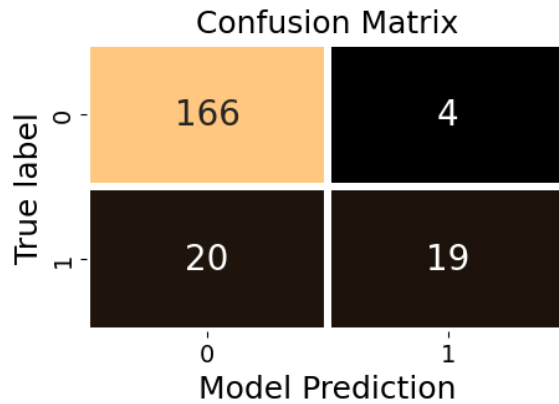


Figure Confusion Matrix of SVM model

This confusion matrix shows that of the total 209 test data instances, 185 instances are correctly classified in their respective classes. 166 women from a total of 170 women are correctly classified as no-risk (0) and 19 women from a total of 39 women are correctly classified as risk (1). However, only 4 women from a total of 170 women are incorrectly classified as at-risk and while 20 women from a total of 39 are incorrectly classified as no-risk.

AUC-ROC curve

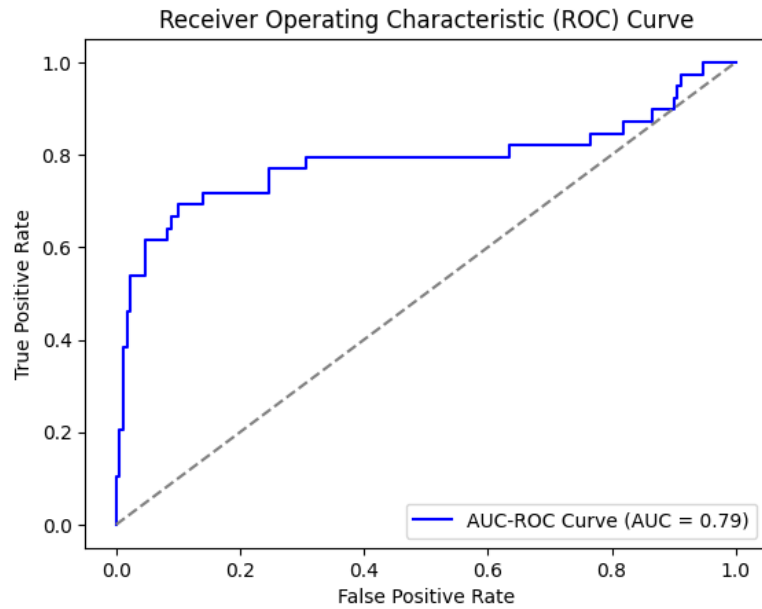


Figure AUC-ROC curve of SVM model

The AUC-ROC curve value of 0.79 for the SVM model indicates that the model has good predictive ability to differentiate between the two classes (at-risk women for APOs and non-risk women).

Precision-Recall curve

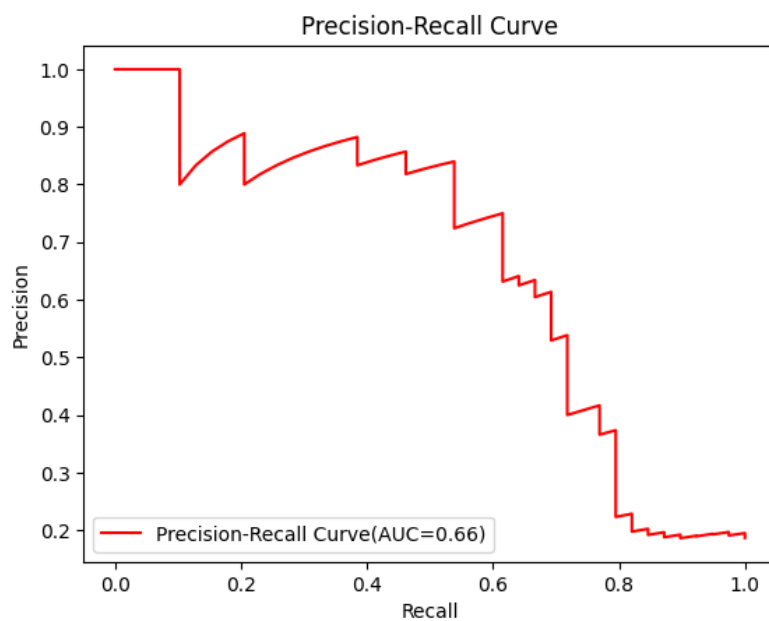


Figure PR-AUC curve of SVM model

The SVM model achieved a Precision-Recall (PR) score of 66%, which shows a moderate balance between precision and recall in detecting positive instances.

Model 4: DT

The fourth model experimented with was the DT classifier model.

➤ Model Training

After fitting the three datasets with different imputation strategies in the baseline model, for DT classifier, the one that performed well was selected for further hyperparameter tuning and prediction.

For the DT model, the dataset variant imputed with KNN performed better than the others.

➤ Hyperparameter Tuning

After experimenting with both hyperparameter tuning methods, for the DT model, Bayesian Optimization gave better performance result on the training set.

➤ Model Prediction

Then, the DT model with the selected best hyperparameters was used to train on the entire training data and used to predict on the test set.

➤ Model Evaluation

After training the DT model using the best hyperparameters, the model was then evaluated on the test set.

Classification Report of the test set

	Precision	Recall	F1-score	Support
0	0.89	0.96	0.93	170
1	0.76	0.49	0.59	39
Accuracy			0.88	209

Table 12 Test set classification report

The above classification report shows that the DT model has an accuracy of 0.88. A recall value of 96% & 49% indicates that the model has a strong performance in detecting “no risk” class but has problems with detecting “at-risk” women meaning that the model only captures 49% of actual "at-risk" women and misclassifies 51% of the women. Of the women predicted as at-risk, only 76%, are correctly identified. The f1-score value of 59% indicates poor effectiveness for detecting the at-risk class.

Confusion Matrix

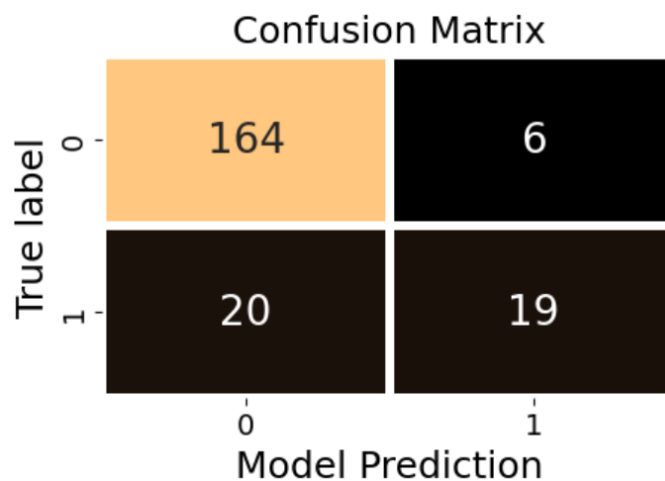


Figure Confusion Matrix of DT model

This confusion matrix shows that of the total 209 test data instances, 183 instances were correctly classified in their respective classes. 164 women from a total of 170 women were correctly classified as no-risk (0) and 19 women from a total of 39 women were correctly classified as risk (1). However, only 6 women from a total of 170 women were incorrectly classified as at-risk and while 20 women from a total of 39 were incorrectly classified as no-risk.

AUC-ROC curve

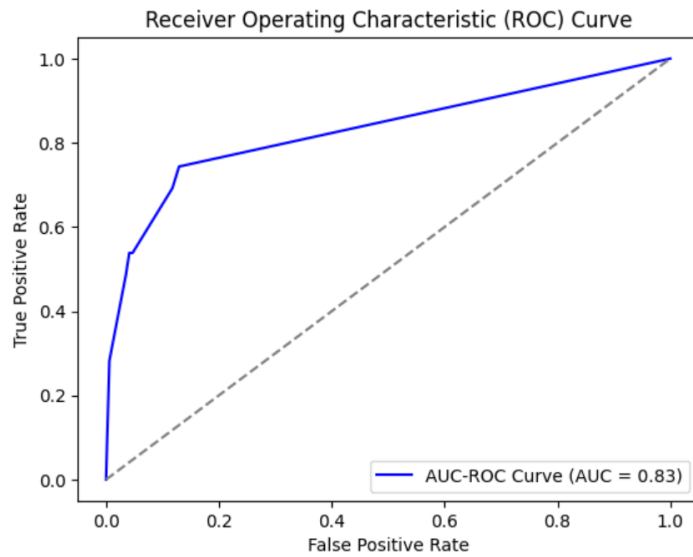


Figure AUC-ROC curve of DT model

The AUC-ROC curve value of 0.83 for the DT model indicates that the model has moderate predictive ability to differentiate between the positive and negative classes.

Precision-Recall curve

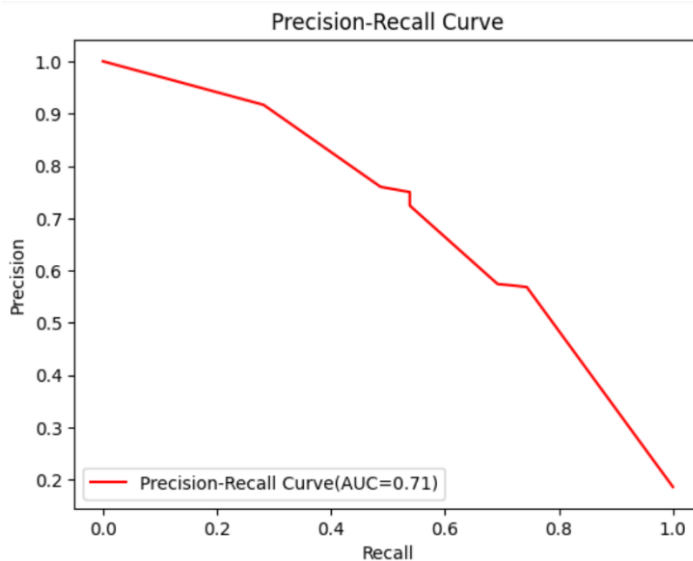


Figure PR-AUC curve of DT model

The DT model achieved a Precision-Recall (PR) score of 71%, which shows a fair ability to correctly distinguish the at-risk group.

Model 5: XGBoost

The fifth model used was the XGBoost classifier model.

Model Training

For this ML model, the dataset variant imputed with simple imputation (median) performed slightly better than the others.

➤ Hyperparameter Tuning

After experimenting with both hyperparameter tuning methods, for the XGBoost classification model, Bayesian Optimization gave better performance result on the training set.

➤ Model Prediction

Then, the XGBoost model with the selected best hyperparameters was used to train on the entire training data and used to predict on the test set.

➤ Model Evaluation

After training the XGBoost model using the best hyperparameters, the model was then evaluated on the test set.

Classification Report of the test set

	Precision	Recall	F1-score	Support
0	0.93	0.99	0.96	170
1	0.93	0.69	0.79	39
Accuracy			0.93	209

Table 13 Test set classification report

From this classification report, it can be seen that the XGBoost model has an accuracy of 0.93 which indicates relatively good overall predictive ability of the model. A recall value of 99% & 69% indicates that the model has a near perfect performance in detecting “no risk” class but slightly lower performance with detecting “at-risk” women. The recall value shows that the model captures 69% of actual "at-risk" women and misses 31%. Of the women predicted as at-risk by the model,

93% are correctly identified. The f1-score value of 79% indicates good effectiveness for detecting the at-risk class.

Confusion Matrix

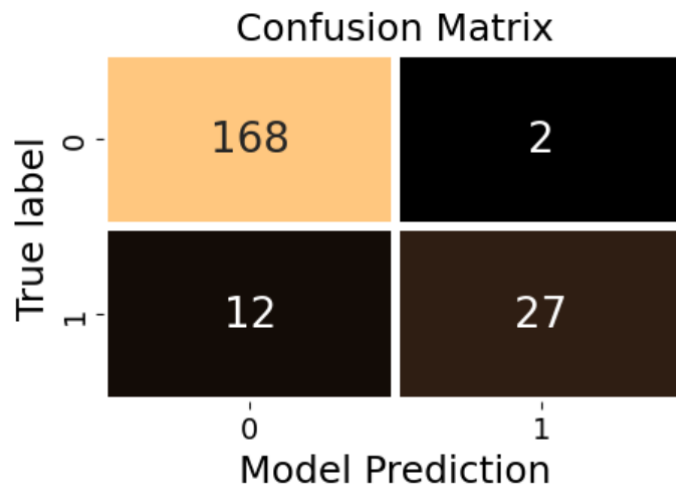


Figure Confusion Matrix of XGBoost model

This confusion matrix shows that of the total 209 test data instances, 195 instances were correctly classified in their respective classes. 168 women from a total of 170 women were correctly classified as no-risk (0) and 27 women from a total of 39 women were correctly classified as risk (1). However, only 2 women from a total of 170 women were incorrectly classified as at-risk and while 12 women from a total of 39 were incorrectly classified as no-risk.

AUC-ROC curve

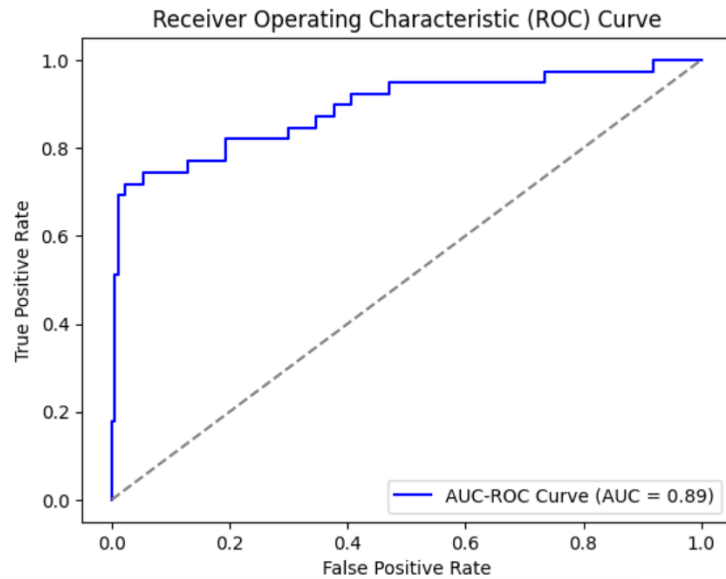


Figure AUC-ROC curve of XGBoost model

The AUC-ROC curve value of 0.89 for the XGBoost model indicates that the model has a strong classification ability to differentiate between the two classes (at-risk women for APOs and non-risk women).

Precision-Recall curve

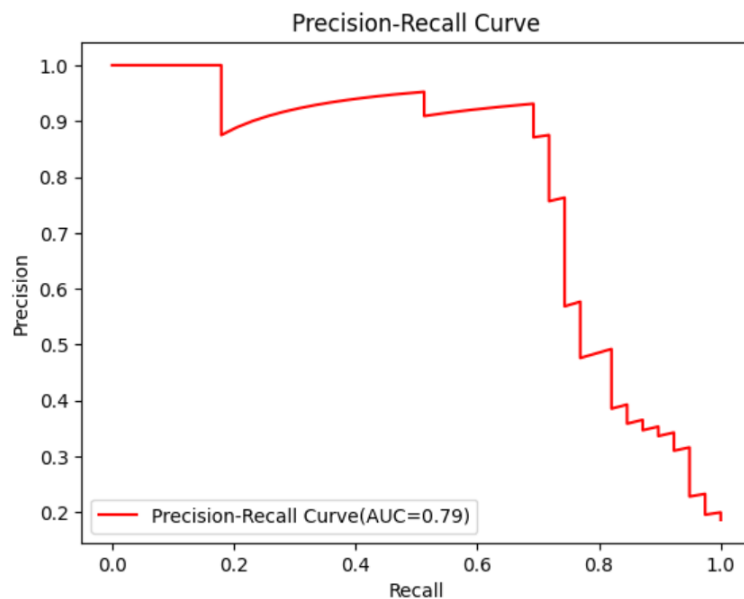


Figure PR-AUC curve of XGBoost model

The XGBoost model achieved a Precision-Recall (PR) score of 79%, which shows a really good ability to distinguish the at-risk group and balance between precision and recall in detecting positive instances. Even as recall increases (>0.7), precision is relatively high (>0.7) before falling off, indicating the model initially captures most positive cases without being much affected by noise.

Model 6: RF

The final model selected and experimented with was the RF tree-based model.

Model Training

For this ML model, the dataset variant imputed with KNN strategy performed slightly better than the others.

➤ Hyperparameter Tuning

After experimenting with both hyperparameter tuning methods, for the RF classification model, grid search optimization gave better performance result on the training set.

➤ Model Prediction

Then, the RF model with the tuned best hyperparameters was used to train on the entire training data and used to predict on the test set.

➤ Model Evaluation

After training the RF model using the best hyperparameters, the model was then evaluated on the test set.

Classification Report of the test set

	Precision	Recall	F1-score	Support
0	0.91	0.98	0.95	170
1	0.88	0.59	0.71	39
Accuracy			0.91	209

Table 14 Test set classification report

From this classification report, it can be seen that the RF model has an accuracy of 0.91. A recall value of 98% & 59% indicates that the model has a strong performance in detecting “no risk” class but lower performance with detecting “at-risk” women. This shows that the model captures 59% of actual "at-risk" women and misclassifies 41%. It also means that it correctly classifies 98% of non-risk women as non-risk and misclassifies 2%. Of the 59% of women predicted as at-risk, 88%, are correctly identified. The f1-score value of 71% indicates good effectiveness for detecting the at-risk class.

Confusion Matrix

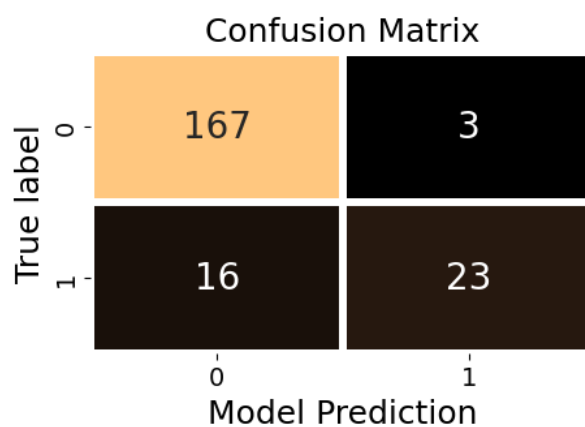


Figure Confusion Matrix of RF model

This confusion matrix shows that of the total 209 test data instances, 190 instances were correctly classified in their respective classes. 167 women from a total of 170 women were correctly classified as no-risk (0) and 23 women from a total of 39 women were correctly classified as risk (1). However, only 3 women from a total of 170 women were incorrectly classified as at-risk and while 16 women from a total of 39 were incorrectly classified as no-risk.

AUC-ROC curve

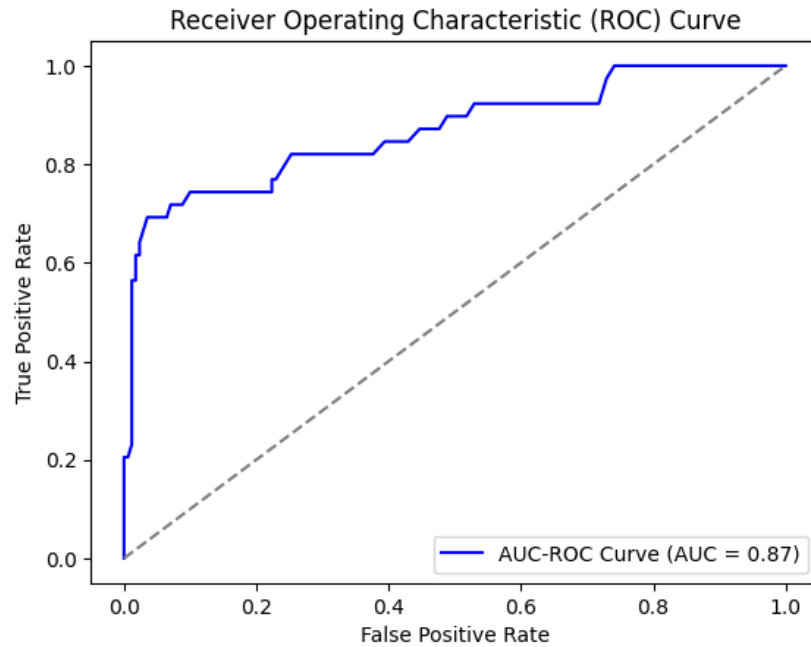


Figure AUC-ROC curve of RF model

The AUC-ROC curve value of 0.87 for the RF model indicates that the model has a strong classification ability to differentiate between the two classes (at-risk women for APOs and non-risk women).

Precision-Recall curve

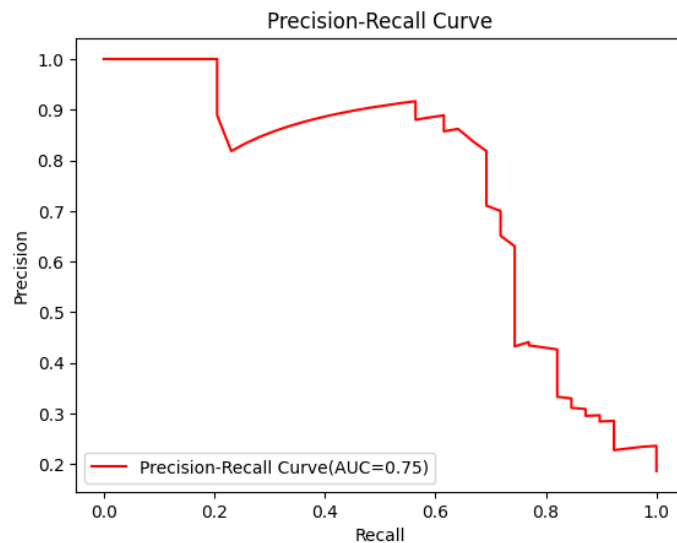


Figure PR-AUC curve of RF model

The RF model achieved a Precision-Recall (PR) score of 75%, which shows a good ability in identifying the positive class and balance between precision and recall. It classifies at-risk class with higher values of precision (>0.7) for higher values of recall (0.6-0.7).

Model	Imputation	Tuning	Class	Precision	Recall	F1-score	Accuracy	AUC-ROC	AUC-PR
LR	MICE	Bayesian Optimization	0	90	95	92	87	85	67
			1	70	54	61			
NB	KNN	Grid Search	0	91	88	89	83	80	56
			1	54	64	59			
SVM	MICE	Grid Search	0	89	98	93	89	79	66
			1	83	49	61			
DT	KNN	Bayesian Optimization	0	89	96	93	88	83	71
			1	76	49	59			
XGB	Median	Bayesian Optimization	0	93	99	96	93	89	79
			1	93	69	79			
RF	KNN	Grid Search	0	91	98	95	91	87	75
			1	88	59	71			

Table 15 Summary of the final tuned models' performance on the test set

4.3 Identifying Key Features for Prediction

To determine the most influential factors affecting the prediction model, the XGBoost's default importance scoring method was used to extract the top 15 features. The more an attribute is used to make key decisions, the higher its relative importance. This importance is calculated explicitly for each attribute in the dataset, which allows for the attributes to be ranked and compared to each other (Brownlee Jason 2020). After preprocessing steps were done, the final obtained 42 attributes were all used to train the XGBoost model. Identifying which indicators are the most influential on the model can help identify possible risk factors that are associated with APOs among women with HDP.

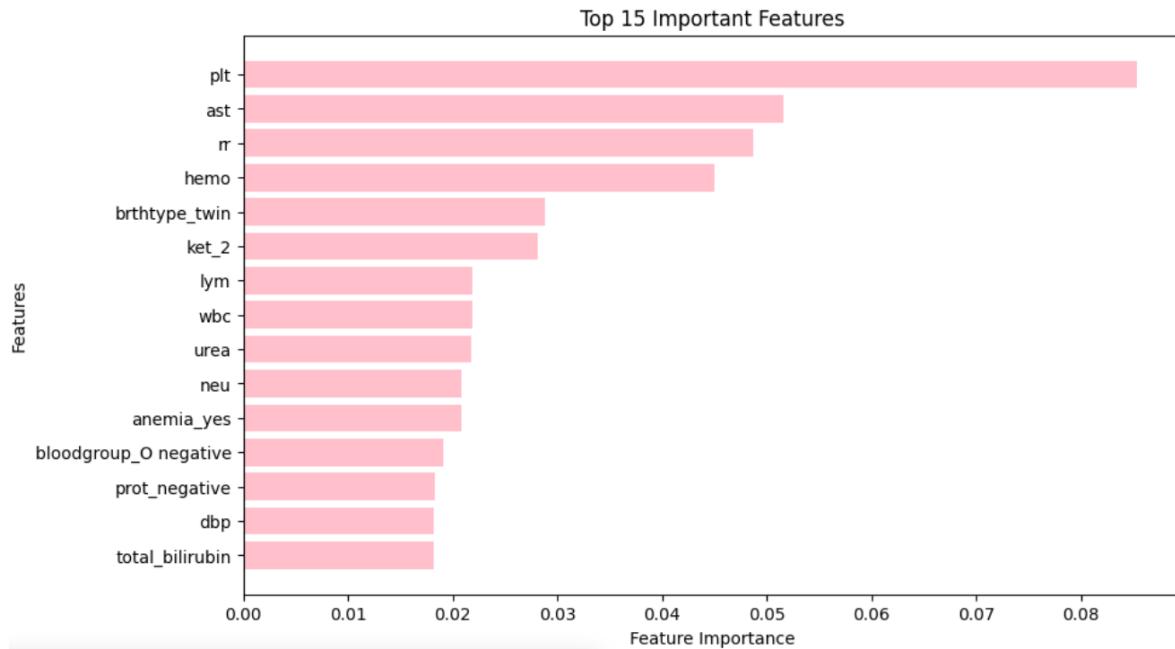


Figure 8 Feature importance plot using XGBoost model

From this feature importance plot, we can observe that among the total 42 evaluated factors, these were the top 15 attributes that had more contribution to the performance of the model. Platelets (PLT) play the most significant role in the model's prediction with aspartate aminotransferase (AST), which is a liver function test, together making up almost 14% of the importance.

Other highly ranked features by the model include rr (respiratory rate), hemo (hemoglobin), birthtype_twin (giving birth to twins), ket_2 (having urinary ketone value of +2) and lym (lymphocyte blood cell count) which are mostly laboratory indicators. Other features, such as white blood cell count, urea, neutrophile count and presence of anemia, have relatively lower importance scores, which shows that their role in prediction is low.

4.4 Comparison of ML models

After performing model training, hyperparameter tuning and modeling steps, the final six models were compared with each other using key metrics such as accuracy, precision, recall, AUC-PR and AUC-ROC curves. The following figures are used to show the performances of the models on the key evaluation metrics.

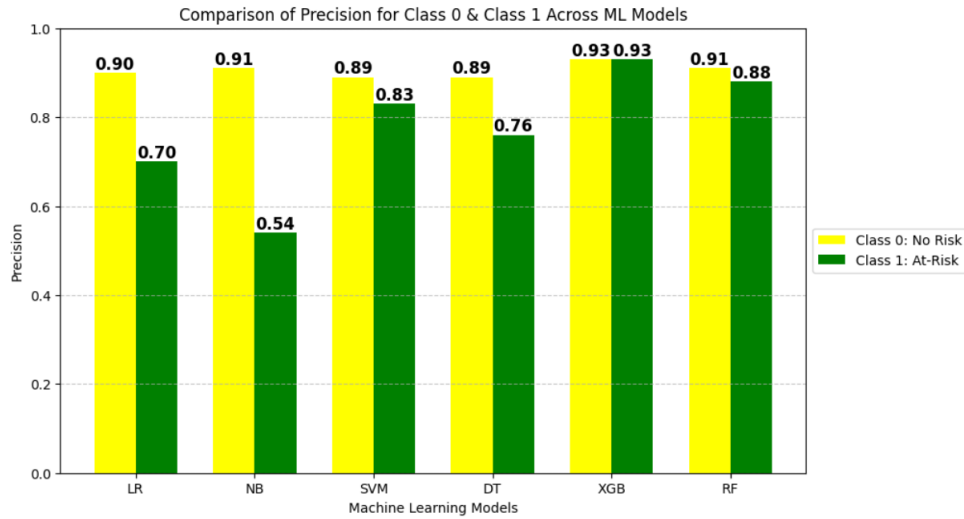


Figure 9 Comparison of the models by Precision

When comparing on precision, XGBoost model had the best performance in classifying the non-risk group followed by RF, NB and Logistic regression. For the minority class, at-risk group, the XGBoost model still had the best performance followed by RF and SVM respectively. The LR and NB models had the lowest precisions.



Figure 10 Comparison of the models by Recall

When comparing based on recall, all models had a really good performance in predicting mothers not at risk of developing adverse outcomes and XGBoost model had a near perfect value followed by SVM and RF. For predicting the at-risk mothers, the XGBoost still out performed better than the rest of the algorithms followed by NB model. The DT and SVM models had the lowest

sensitivity in predicting women who are at-risk for developing adverse outcomes. The figure also shows that all models had better performance in predicting the majority class than the minority class.

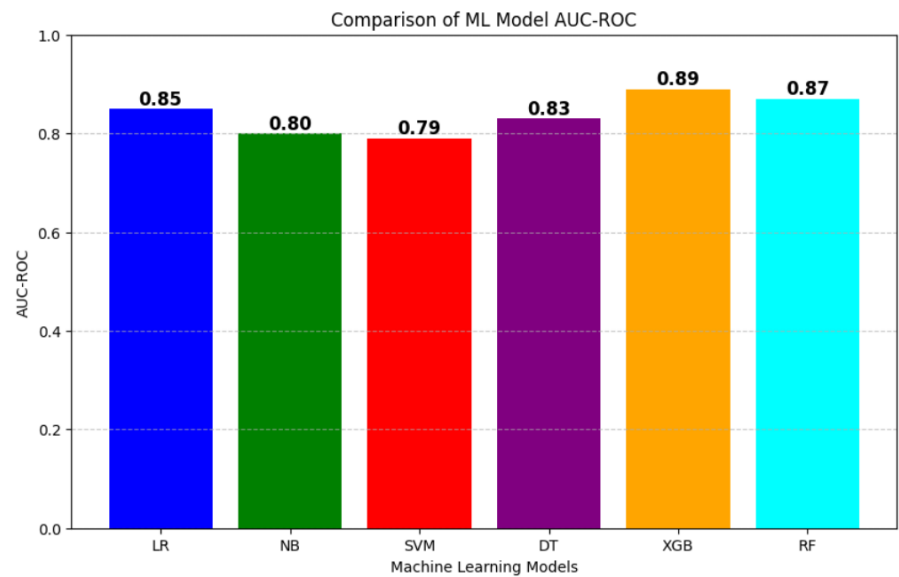


Figure 11 Comparison of the models by AUC-ROC

The AUC-ROC plot shows that XGBoost algorithm has the highest performance showing a better ability to discriminate between classes. RF follows closely together with LR outperforming the other models.

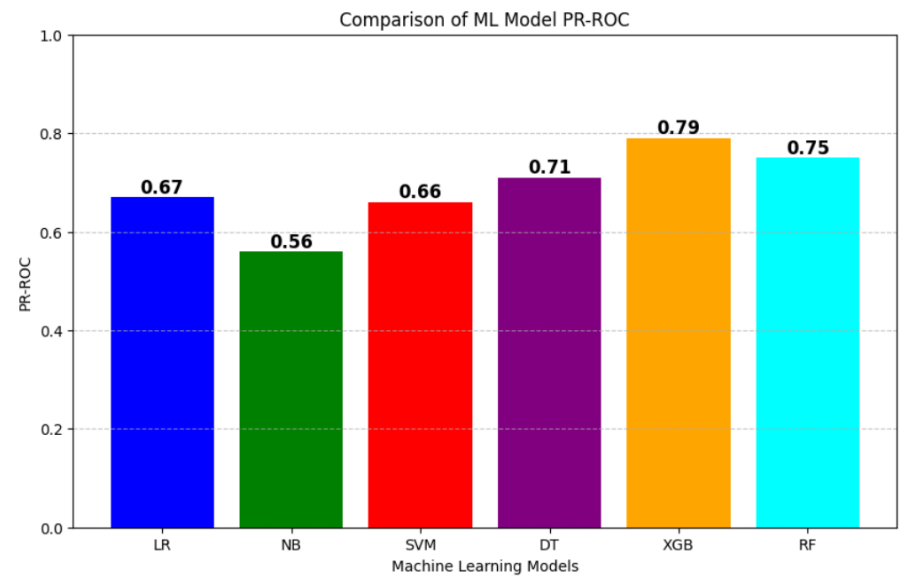


Figure 12 Comparison of the models by PR-AUC

The AUC-PR comparison plot evaluates the performance of the six models on their ability to balance high precision and recall. From the above plot, we can see that XGBoost performs best with followed by the RF model.

4.5 Discussion of the result

The main objective of this research was to build a ML model that can identify women with HDP at-risk of developing APO. All the experiments were done on the same data and the same training and testing dataset split so that it wouldn't influence the performance of the models.

This study experimented with six ML models used for developing a predictive model. All models achieved overall accuracies greater than 80% which means the models have good overall predictive ability. But out of these models, the results show that XGBoost outperformed other models which was similar with other studies (Hang et al. 2021; Chen et al. 2022; Thi Huyen Anh, Manh Thang, and Thanh Huong 2024). It had achieved a 69% recall (sensitivity) and 99% specificity and also had the highest AUC-PR (0.79), which shows the model's good performance on imbalanced data. The RF model had the next best performance with 59% recall (sensitivity) and 98% specificity. In contrast, SVM and DT showed very poor performances, with recall rates of only 49%, which makes them completely unsuitable for clinical decision-making. This contradicts findings from (Sarno et al. 2022) in which DT had better sensitivity than other models that may be attributed to the use of feature selection methods in the pre-processing steps in their study.

For this study, sensitivity (ability to correctly identify individuals who have the disease) was primarily prioritized since false negative (classifying a woman who is at-risk as non-risk) has more consequence. It could lead to delays in initializing treatment which puts the mother and the fetus at more risk for worse health conditions. Specificity (ability to correctly identify individuals who do not have the disease) is also important because a false positive (classifying a mother who is non-risk as at-risk) can lead to unnecessary treatments and further tests. It may also cause undue stress and psychological harm on the mother. So even though most of the models have good accuracy, it can't be used as the only metric for evaluating the performance of the models since it takes into account all the correctly predicted instances in both classes. So, a model having high specificity and low sensitivity (as most of the models in this study) might have good accuracy.

Based on this, XGBoost is chosen as the best predictive model for the problem presented in the study since it has a better sensitivity and specificity balance.

The rest of the models (LR, NB, SVM, DT, RF) had good AUC values and good specificity values but lower sensitivity values which were similar with other studies (Chen et al. 2022; Hang et al. 2021). However, these results contrast with the work of (Sarno et al. 2022), who had higher sensitivity results for all models used including LR, SVM, DT and RF models which can be due to higher proportion of APOs in their data which enabled the models to learn more from the underlying patterns.

Among the features that were identified as important by the XGBoost model, platelet count was identified as the most important feature for the model's performance. This variable was also identified as an important predictor by other studies (Thi Huyen Anh, Manh Thang, and Thanh Huong 2024; Jaleta and Abdisa 2022). Other attributes identified as important by the model is AST (aspartate aminotransferase) which was also identified as another important predictor for APO among women with HDP (Chen et al. 2022) alongside neutrophile count (Sarno et al. 2022). There were also notable differences in the important features selected by this study and other studies which may be because of differences in the predictors used. The study by (Chen et al. 2022) identified fetal femur length, fetal head-to-abdominal ratio and fetal biparietal diameter among other predictors as some of the important features which were not used in this study.

Based on this study, the choice of imputation method (MICE, KNN, or Median) had low impact on determining the best model for prediction. This was shown by the lack of performance difference of the baseline models trained with different imputation techniques (see annex). From the hyperparameter tuning methods experimented with, Bayesian optimization was preferred since it had lower computation time and gave the XGBoost model which had the best performance. The models tuned with Bayesian optimization also gave better average AUC-PR values (see annex).

CHAPTER FIVE

5. Conclusion & Recommendation

5.1 Conclusion

The main objective of this study was to develop a predictive model for APOs among women with HDP using ML techniques. The study was conducted by following a structured data science methodology, involving data understanding, data preprocessing, model building, hyperparameter tuning, and performance evaluation. The dataset used for this research was obtained from electronic medical records of four hospitals under the Addis Ababa Health Bureau, including antenatal and delivery information of pregnant women. For model development and experimentation, six ML algorithms were used; Logistic Regression, Naïve Bayes, Support Vector Machines, Decision Tree, Random Forest, and XGBoost combined with three missing value imputation techniques: Median, KNN, and MICE. Each model was evaluated based on multiple performance metrics using cross-validation and the unseen test datasets.

This study has several strengths. First, it utilized routinely collected clinical and laboratory variables already part of standard antenatal care in Ethiopian health facilities. This increases the usability and applicability of model implementation in low-resource settings. Additionally, the features identified as important for prediction are non-invasive and inexpensive tests that can be done easily. However, the study also had limitations. The fact that the dataset was obtained from selected hospitals in Addis Ababa limits the generalizability of the findings to broader populations or other regions. Additionally, the sample size was relatively small, and rare outcomes (adverse outcomes) were underrepresented, which affects model reliability.

Finally, while the models showed promise, the best-performing model achieved only 69% recall for identifying women at risk of adverse outcomes. Since there is clinical significance of missing high-risk cases, this level of sensitivity is suboptimal for real-world application. Therefore, the current models should not be deployed and used in clinical settings without further development and validation.

The results of this study achieved the research questions previously stated which were to identify key clinical predictors and to identify the best ML model for prediction of APOs among women with HDP. In regards to the first research question, this study has identified the main clinical factor to have an effect on APOs among women with HDP to be the number of platelets count of the

mother followed by AST levels of the mother. Other important predictors include respiratory rate and hemoglobin levels of the mother, having multiple pregnancy and ketone levels of +2 found in the urine.

With respect to the second research question, after performing different experimentations, the study revealed that XGBoost and Random Forest classifiers demonstrated the best overall predictive performance, particularly in terms of sensitivity and AUC. So, based on these findings, XGBoost model when paired with Bayesian optimization, has the potential to be used for predicting APOs in early pregnancy.

5.2 Recommendation

This study has shown the potential application of ML in predicting APOs among women with HDP. Using routinely collected medical data showed that meaningful insights can be derived from existing EMRs, which can be used to support timely risk stratification and targeted care. Based on this research's findings, there are some areas that need further investigation to have a better risk prediction model.

Future researchers should focus on:

Explore more advanced ML models: In this study, only six commonly used models were used. So, future work should consider testing advanced models such as deep learning and hybrid ensembles that may give an improved predictive performance, especially in imbalanced datasets.

Include other adverse outcomes: Including other APOs like IUGR, Birth asphyxia, Low APGAR scores can address the class imbalance. Since rare outcomes may suffer from having low sensitivity, adding related but more frequent outcomes can improve model training.

Incorporate other relevant features: Features that were selected as important in other similar studies like fetal femur length, fetal head-to-abdominal ratio and fetal biparietal diameter, 24- hour urine protein among others were not used in this research since they were not available or had high missingness. Including such variables could possibly increase the models' performances and clinical value.

Clinicians and health institutions should:

Ensure complete registration of patient data: Some variables in this study were not used since they had high proportion of missingness. One reason for missingness could be due to limitations in

routine EMR data documentation. So, healthcare workers should try to ensure complete patient data registration in the EMRs.

Provide awareness and education about the use of ML in maternal healthcare: As ML based tools continue to be developed, it is important to educate the users and the public about how these kinds of tools work and how they can support rather than replace clinical judgment. Having this awareness among the public will be useful when models with sufficient performance are ready for clinical use.

Public Health Stakeholders and Policymakers should:

Establish ethical and regulatory frameworks: Policies must be developed that guarantee that ML models used in maternal health are ethical, transparent and secure.

References

- A Ilemobayo, Justus, Olamide Durodola, Oreoluwa Alade, Opeyemi J Awotunde, Adewumi T Olanrewaju, Olumide Falana, Adedolapo Ogungbire, et al. 2024. "Hyperparameter Tuning in Machine Learning: A Comprehensive Review." *Journal of Engineering Research and Reports* 26, no. 6 (June): 388–95. <https://doi.org/10.9734/JERR/2024/V26I61188>
- ACOG. 2022. "Preeclampsia and High Blood Pressure During Pregnancy | ACOG." American College of Obstetricians and Gynecologists. 2022. <https://www.acog.org/womens-health/faqs/preeclampsia-and-high-blood-pressure-during-pregnancy>.
- Alemie, Tamalew, Asmamaw Abebe, Ousman Adal, Aklilu Azazh, and Destaw Endeshaw. 2023. "Clinical Features and Outcomes of Patients with Preeclampsia and Eclampsia at Gondar University Hospital, Amhara, Ethiopia 2021." *European Journal of Obstetrics and Gynecology and Reproductive Biology: X* 20, no. December (December). <https://doi.org/10.1016/j.eurox.2023.100254>.
- Amare, Mesganew, Adugna Olani, Habtamu Hassen, Bikila Jiregna, Nigusu Getachew, and Sena Belina. 2021. "Perinatal Outcomes and Associated Factors among Women with Hypertensive Disorders of Pregnancy Delivered in Jimma Zone Hospitals, Southwest Ethiopia." *Ethiopian Journal of Health Sciences* 31, no. 6 (November): 1145–54. <https://doi.org/10.4314/ejhs.v31i6.9>.
- Asseffaid, Netsanet Abera, and Birhanu Wondimeneh Demissie. 2019. "Perinatal Outcomes of Hypertensive Disorders in Pregnancy at a Referral Hospital, Southern Ethiopia." *PLOS ONE* 14, no. 2 (February): e0213240. <https://doi.org/10.1371/JOURNAL.PONE.0213240>.
- Bekele, Wondesen Teshome. 2022. "Machine Learning Algorithms for Predicting Low Birth Weight in Ethiopia." *BMC Medical Informatics and Decision Making* 22, no. 1 (December). <https://doi.org/10.1186/s12911-022-01981-9>.
- Berhe, Abadi Kidanemariam, Getachew Mullu Kassa, Gedefaw Abeje Fekadu, and Acheneff Asmamaw Muche. 2018. "Prevalence of Hypertensive Disorders of Pregnancy in Ethiopia: A Systemic Review and Meta-Analysis." *BMC Pregnancy and Childbirth* 18, no. 1 (January): 1–11. <https://doi.org/10.1186/S12884-018-1667-7/FIGURES/5>.
- Bisson, Courtney, Sydney Dautel, Easha Patel, Sunitha Suresh, Patricia Dauer, and Sarosh Rana. 2023. "Preeclampsia Pathophysiology and Adverse Outcomes during Pregnancy and Postpartum." *Frontiers in Medicine* 10, no. March (March): 1144170. <https://doi.org/10.3389/FMED.2023.1144170/BIBTEX>.
- Bogale, Dawit S., Tesfamariam M. Abuhay, and Belayneh E. Dejene. 2022. "Predicting Perinatal Mortality Based on Maternal Health Status and Health Insurance Service Using Homogeneous Ensemble Machine Learning Methods." *BMC Medical Informatics and Decision Making* 22, no. 1 (December). <https://doi.org/10.1186/s12911-022-02084-1>.
- Bolikulov, Furkat, Rashid Nasimov, Akbar Rashidov, Farkhod Akhmedov, and Young Im Cho. 2024. "Effective Methods of Categorical Data Encoding for Artificial Intelligence Algorithms." *Mathematics* 2024, Vol. 12, Page 2553 12, no. 16 (August): 2553. <https://doi.org/10.3390/MATH12162553>.
- Bridwell, Matthew, Endang Handzel, Michelle Hynes, Reginald Jean-Louis, David Fitter, Carol Hogue, Reynold Grand-Pierre, Hedwige Pierre, and Bradley Pearce. 2019. "Hypertensive Disorders in Pregnancy and Maternal and Neonatal Outcomes in Haiti: The Importance of Surveillance and Data Collection." *BMC Pregnancy and Childbirth* 19, no. 1 (June): 1–11. <https://doi.org/10.1186/S12884-019-2361-0/TABLES/5>.
- Brown, Mark A., Laura A. Magee, Louise C. Kenny, S. Ananth Karumanchi, Fergus P. McCarthy, Shigeru Saito, David R. Hall, Charlotte E. Warren, Gloria Adoyi, and Salisu Ishaku. 2018. "Hypertensive Disorders of Pregnancy: ISSHP Classification, Diagnosis, and Management Recommendations for International Practice." *Hypertension* 72, no. 1 (July): 24–43.

- <https://doi.org/10.1161/HYPERTENSIONAHA.117.10803/ASSET/68AB27F7-67F4-4487-B0E4-735091C6D909/ASSETS/IMAGES/LARGE/24FIG03.JPG>.
- Brownlee Jason. 2020. "Feature Importance and Feature Selection With XGBoost in Python - MachineLearningMastery.Com." August 27, 2020. <https://machinelearningmastery.com/feature-importance-and-feature-selection-with-xgboost-in-python/>.
- Bucher, Valentina, Alexandra Roddy Mitchell, Pia Gudmundsson, Jessica Atkinson, Nicole Wallin, Joline Asp, Maria Sennström, et al. 2024. "Prediction of Adverse Maternal and Perinatal Outcomes Associated with Pre-Eclampsia and Hypertensive Disorders of Pregnancy: A Systematic Review and Meta-Analysis." www.thelancet.com.
- Chen, Junjun, Yuelong Ji, Tao Su, Ma Jin, Zhichao Yuan, Yuanzhou Peng, Shuang Zhou, et al. 2022. "Prediction of Adverse Outcomes in De Novo Hypertensive Disorders of Pregnancy: Development and Validation of Maternal and Neonatal Prognostic Models." *Healthcare (Switzerland)* 10, no. 11 (November): 2307. <https://doi.org/10.3390/HEALTHCARE10112307/S1>.
- Cho, Hunyong, Jane She, Daniel De Marchi, Helal El-Zaatari, Edward L. Barnes, Anna R. Kahkoska, Michael R. Kosorok, and Arti V. Virkud. 2024. "Machine Learning and Health Science Research: Tutorial." *Journal of Medical Internet Research* 26, no. 1 (January). <https://doi.org/10.2196/50890>.
- Dadelszen, Peter von, Jennifer M. Menzies, Beth Payne, and Laura A. Magee. 2009. "Predicting Adverse Outcomes in Women with Severe Pre-Eclampsia." *Seminars in Perinatology* 33, no. 3 (June): 152–57. <https://doi.org/10.1053/J.SEMPERI.2009.02.009>.
- Dalianis, Hercules. 2018. "Evaluation Metrics and Evaluation." *Clinical Text Mining*, 45–53. https://doi.org/10.1007/978-3-319-78503-5_6
- Feleke, Sefineh Fenta, Zelalem Alamrew Anteneh, Gizachew Tadesse Wassie, Anteneh Kassa Yalew, and Anteneh Mengist Dessie. 2022. "Developing and Validating a Risk Prediction Model for Preterm Birth at Felege Hiwot Comprehensive Specialized Hospital, North-West Ethiopia: A Retrospective Follow-up Study." *BMJ Open* 12, no. 9 (September). <https://doi.org/10.1136/bmjopen-2022-061061>.
- Gao, Yongqi, Zhongji Wu, Simin Liu, Yiwen Chen, Guojun Zhao, and Hui Ping Lin. 2023. "Identification of Key Genes in the Pathogenesis of Preeclampsia via Bioinformatic Analysis and Experimental Verification." *Frontiers in Endocrinology* 14, no. July (July): 1190012. <https://doi.org/10.3389/FENDO.2023.1190012/BIBTEX>.
- Gedefaw, Getnet, Birhan Alemnew, and Asmamaw Demis. 2020. "Adverse Fetal Outcomes and Its Associated Factors in Ethiopia: A Systematic Review and Meta-Analysis." *BMC Pediatrics* 20, no. 1 (June). <https://doi.org/10.1186/s12887-020-02176-9>.
- Hang, Yuwei, Yan Zhang, Yan Lv, Wenbin Yu, and Yi Lin. 2021a. "Electronic Medical Record Based Machine Learning Methods for Adverse Pregnancy Outcome Prediction." In , 17. SPIE-Intl Soc Optical Eng. <https://doi.org/10.1117/12.2581720>.
- Hassen, Hamid Y., Seifu H. Gebreyesus, Bilal S. Endris, Meselech A. Roro, and Jean Pierre Van Geertruyden. 2020. "Development and Validation of a Risk Score to Predict Low Birthweight Using Characteristics of the Mother: Analysis from Bunmap Cohort in Ethiopia." *Journal of Clinical Medicine* 9, no. 5 (May). <https://doi.org/10.3390/jcm9051587>.
- Hossain Riyad, Timmer Douglas, and Moya Hiram. 2021. "(PDF) Machine Learning Model Optimization with Hyper-Parameter Tuning Approach." *Nternational Conference on Advanced Engineering, Technology and Applications*, August (August). https://www.researchgate.net/publication/354495368_Machine_learning_model_optimization_with_hyper-parameter_tuning_approach.
- Hutcheon, Jennifer A., Sarka Lisonkova, and K. S. Joseph. 2011. "Epidemiology of Pre-Eclampsia and the Other Hypertensive Disorders of Pregnancy." *Best Practice & Research. Clinical*

- Obstetrics & Gynaecology* 25, no. 4: 391–403. <https://doi.org/10.1016/J.BPOBGYN.2011.01.006>.
- Jaleta, Debela Dereje, and Diriba Kumara Abdisa. 2022. “Predictors of Adverse Perinatal Outcome among Women Who Gave Birth at Medical Center of Southwest Ethiopia: A Retrospective Cohort Study.” *BMJ Open* 12, no. 3 (March): e053881. <https://doi.org/10.1136/BMJOPEN-2021-053881>.
- Jaleta, Debela Dereje, Tadesse Gebremedhin, and Mulusew Gerbaba Jebena. 2021. “Perinatal Outcomes of Women with Hypertensive Disorders of Pregnancy in Jimma Medical Center, Southwest Ethiopia: Retrospective Cohort Study.” *PLoS ONE* 16, no. 8 (August): e0256520. <https://doi.org/10.1371/JOURNAL.PONE.0256520>.
- Jiang, Li, Kun Tang, Laura A. Magee, Peter von Dadelszen, Alec Ekeroma, Xuan Li, Enyao Zhang, and Zulfiqar A. Bhutta. 2022. “A Global View of Hypertensive Disorders and Diabetes Mellitus during Pregnancy.” *Nature Reviews. Endocrinology* 18, no. 12 (December): 760. <https://doi.org/10.1038/S41574-022-00734-Y>.
- Kang, Hyun. 2013. “The Prevention and Handling of the Missing Data.” *Korean Journal of Anesthesiology* 64, no. 5 (May): 402. <https://doi.org/10.4097/KJAE.2013.64.5.402>.
- Khedagi, Apurva M., and Natalie A. Bello. 2020. “Hypertensive Disorders of Pregnancy.” *Cardiology Clinics* 39, no. 1 (February): 77. <https://doi.org/10.1016/J.CCL.2020.09.005>.
- Kumar, Manisha, V. Ravi, Deepika Meena, Kanika Chopra, Shilpi Nain, and Manju Puri. 2021. “Predictive Model for Late Stillbirth Among Antenatal Hypertensive Women.” *Journal of Obstetrics and Gynaecology of India* 72, no. Suppl 1 (August): 96. <https://doi.org/10.1007/S13224-021-01561-3>.
- LEEMAN, LAWRENCE, LEE T. DRESANG, and PATRICIA FONTAINE. 2016. “Hypertensive Disorders of Pregnancy.” *American Family Physician* 93, no. 2 (January): 121–27. <https://www.aafp.org/pubs/afp/issues/2016/0115/p121.html>.
- Li, Jia Hang, Shu Xia Guo, Ru Lin Ma, Jia He, Xiang Hui Zhang, Dong Sheng Rui, Yu Song Ding, et al. 2024. “Comparison of the Effects of Imputation Methods for Missing Data in Predictive Modelling of Cohort Study Datasets.” *BMC Medical Research Methodology* 24, no. 1 (December): 1–9. <https://doi.org/10.1186/S12874-024-02173-X/TABLES/3>.
- Magee, Laura A., Peter von Dadelszen, Joel Singer, Terry Lee, Evelyne Rey, Susan Ross, Elizabeth Asztalos, et al. 2016. “Can Adverse Maternal and Perinatal Outcomes Be Predicted When Blood Pressure Becomes Elevated? Secondary Analyses from the CHIPS (Control of Hypertension In Pregnancy Study) Randomized Controlled Trial.” *Acta Obstetrica et Gynecologica Scandinavica* 95, no. 7 (July): 763–76. <https://doi.org/10.1111/aogs.12877>.
- Mammaro, Alessia, Sabina Carrara, Alessandro Cavaliere, Santina Ermito, Angela Dinatale, Elisa Maria Pappalardo, Mariapia Militello, Rosa Pedata, and Alessandra Cacciatore. 2009. “Hypertensive Disorders of Pregnancy.” *Journal of Prenatal Medicine* 3, no. 1: 1. <https://pmc.ncbi.nlm.nih.gov/articles/PMC3279097/>.
- Mersha, Amanual Getnet, Tadesse Melaku Abegaz, and Mohammed Assen Seid. 2019. “Maternal and Perinatal Outcomes of Hypertensive Disorders of Pregnancy in Ethiopia: Systematic Review and Meta-Analysis.” *BMC Pregnancy and Childbirth* 19, no. 1 (December): 1–12. <https://doi.org/10.1186/S12884-019-2617-8/FIGURES/6>.
- MIZUNO, Satoshi, Satoshi NAGAIE, Junichi SUGAWARA, Gen TAMIYA, Taku OBARA, Mami ISHIKURO, Shinichi KURIYAMA, et al. 2024. “Early Prediction of Hypertensive Disorders of Pregnancy toward Preventive Early Intervention.” *AJOG Global Reports*, November (November), 100383. <https://doi.org/10.1016/j.xagr.2024.100383>.
- Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar. 2012. *Foundations of Machine Learning*.
- Muche, Achenef Asmamaw, Likelesh Lemma Baruda, Clara Pons-Duran, Rohera Olana Fite, Kassahun Alemu Gelaye, Alemayehu Worku Yalew, Lisanu Tadesse, et al. 2024. “Prognostic

- Prediction Models for Adverse Birth Outcomes: A Systematic Review.” *Journal of Global Health* 14, no. October (October): 04214. <https://doi.org/10.7189/JOGH.14.04214>.
- Ngusie, Habtamu Setegn, Shegaw Anagaw Mengiste, Alemu Birara Zemariam, Bogale Molla, Getanew Aschalew Tesfa, Binyam Tariku Seboka, Tilahun Dessie Alene, and Jing Sun. 2024. “Predicting Adverse Birth Outcome among Childbearing Women in Sub-Saharan Africa: Employing Innovative Machine Learning Techniques.” *BMC Public Health* 24, no. 1 (December). <https://doi.org/10.1186/s12889-024-19566-8>.
- Oumer, Abdu, Ahmed Muhye, Imam Dagne, Nesredin Ishak, Ahmed Ale, and Abiyot Bekele. 2021. “Utilization, Determinants, and Prospects of Electronic Medical Records in Ethiopia.” *BioMed Research International* 2021. <https://doi.org/10.1155/2021/2230618>.
- Poye Laney. 2023. “What Are The Risk Factors For Preeclampsia? An Updated Research Perspective.” February 28, 2023. <https://www.preeclampsia.org/the-news/health-information/what-are-the-risk-factors-for-preeclampsia-an-updated-research-perspective>.
- Payne, Beth A., Jennifer A. Hutcheon, J. Mark Ansermino, David R. Hall, Zulfiqar A. Bhutta, Shereen Z. Bhutta, Christine Biryabarema, et al. 2014. “A Risk Prediction Model for the Assessment and Triage of Women with Hypertensive Disorders of Pregnancy in Low-Resourced Settings: The MiniPIERS (Pre-Eclampsia Integrated Estimate of RiSk) Multi-Country Prospective Cohort Study.” *PLOS Medicine* 11, no. 1 (January): e1001589. <https://doi.org/10.1371/JOURNAL.PMED.1001589>.
- Rana, Sarosh, Camille E Powe, Saira Salahuddin, Stefan Verlohren, Frank H Perschel, Richard J Levine, Kee-Hak Lim, Julia B Wenger, Ravi Thadhani, and ; S Ananth Karumanchi. 2012. “Vascular Medicine Angiogenic Factors and the Risk of Adverse Outcomes in Women With Suspected Preeclampsia.” <https://doi.org/10.1161/CIRCULATIONAHA>.
- Radparvar, Ashley A., Kavita Vani, Kevin Fiori, Sonali Gupta, Patricia Chavez, Molly Fisher, Garima Sharma, Diana Wolfe, and Anna E. Bortnick. 2024. “Hypertensive Disorders of Pregnancy: Innovative Management Strategies.” *JACC: Advances* 3, no. 3 (March). <https://doi.org/10.1016/J.JACADV.2024.100864>.
- Rawat Tara, and Khemchandani Vineeta. 2017. “(PDF) Feature Engineering (FE) Tools and Techniques for Better Classification Performance” 8, no. 2 (April). <https://doi.org/10.21172/ijiet.82.024>.
- Roberts, James M., Gail Pearson, Jeff Cutler, and Marshall Lindheimer. 2003. “Summary of the NHLBI Working Group on Research on Hypertension During Pregnancy.” *Hypertension* 41, no. 3 I (March): 437–45. <https://doi.org/10.1161/01.HYP.0000054981.03589.E9>.
- Sarker, Iqbal H. 2021. “Machine Learning: Algorithms, Real-World Applications and Research Directions.” *SN Computer Science* 2, no. 3 (May): 1–21. <https://doi.org/10.1007/S42979-021-00592-X/METRICS>.
- Sarno, Laura, Ricardo Ricci Lopes, Kedong Song, Chong Qiao, Dongying Zheng, Xinyu Hao, Muhammmad Khan, et al. n.d. “OPEN ACCESS EDITED BY Comparison of Machine Learning and Logistic Regression as Predictive Models for Adverse Maternal and Neonatal Outcomes of Preeclampsia: A Retrospective Study.”
- Shah, Taram, Muhammad Khan, Mumtaz Ali, and Bilal Khan. 2018. “Critical Analysis of Six Frequently Used Classification Algorithms.” *University of Swabi Journal* 2, no. 2 (April): 36–40. <https://www.uoswabi.edu.pk/data/downloads/1124.pdf>.
- Sheikhtaheri, Abbas, Mohammad Reza Zarkesh, Raheleh Moradi, and Farzaneh Kermani. 2021. “Prediction of Neonatal Deaths in NICUs: Development and Validation of Machine Learning Models.” *BMC Medical Informatics and Decision Making* 21, no. 1 (December). <https://doi.org/10.1186/s12911-021-01497-8>.
- Syoun, Fisseha Hailemariam, Girmatsion Fisseha Abreha, Dessalegn Massa Teklemichael, and Mebrahtu Kalayu Chekole. 2022. “Fetomaternal Outcomes and Associated Factors among Mothers with Hypertensive Disorders of Pregnancy in Suhul Hospital, Northwest Tigray, Ethiopia.” *Journal of Pregnancy* 2022. <https://doi.org/10.1155/2022/6917009>.

- Tadese, Mesfin, Wogene Asefa Damesa, Gebeyehu Shumet Solomon, Girma Wogie Fitie, Yohannes Moges Mitiku, Saba Desta Tessema, and Agizew Endale. 2024. "Prevalence and Determinants of Adverse Perinatal Outcomes of Preeclampsia with Severe Features at Two Selected Public Hospitals in Addis Ababa, Ethiopia." *Frontiers in Pediatrics* 12, no. March (March): 1345055. <https://doi.org/10.3389/FPED.2024.1345055/BIBTEX>.
- Tesfa, Endalamaw, Endalkachew Nibret, Solomon Tebeje Gizaw, Yohannes Zenebe, Zewdie Mekonnen, Sefealem Assefa, Mulatu Melese, Netsanet Fentahun, and Abaineh Munshea. 2020. "Prevalence and Determinants of Hypertensive Disorders of Pregnancy in Ethiopia: A Systematic Review and Meta-Analysis." *PLOS ONE* 15, no. 9 (September): e0239048. <https://doi.org/10.1371/JOURNAL.PONE.0239048>.
- Teshome, Firanol, Workneh Gidi, Se-Woon Choe, Jude Kong, and Gelan Ayana. n.d. "Machine Learning for Early Prediction of Preterm Birth." <https://ssrn.com/abstract=4952965>.
- Thi Huyen Anh, Nguyen, Nguyen Manh Thang, and Truong Thanh Huong. 2024a. "Maternal and Perinatal Outcomes of Hypertensive Disorders in Pregnancy: Insights from the National Hospital of Obstetrics and Gynecology in Vietnam." *PloS One* 19, no. 1: e0297302. <https://doi.org/10.1371/journal.pone.0297302>
- Torres-Torres, Johnatan, Salvador Espino-y-Sosa, Raigam Martinez-Portilla, Hector Borboa-Olivares, Guadalupe Estrada-Gutierrez, Sandra Acevedo-Gallegos, Erika Ruiz-Ramirez, et al. 2024. "A Narrative Review on the Pathophysiology of Preeclampsia." *International Journal of Molecular Sciences* 25, no. 14 (July): 7569. <https://doi.org/10.3390/IJMS25147569/S1>.
- Ugochinyere, :, and Vivian Ukah. n.d. "THE RISKS OF ADVERSE OUTCOMES (MATERNAL AND FOETAL) IN HYPERTENSIVE DISORDERS OF PREGNANCY."
- Ukah, U. Vivian, Dane A. De Silva, Beth Payne, Laura A. Magee, Jennifer A. Hutcheon, Helen Brown, J. Mark Ansermino, Tang Lee, and Peter von Dadelszen. 2018. "Prediction of Adverse Maternal Outcomes from Pre-Eclampsia and Other Hypertensive Disorders of Pregnancy: A Systematic Review." *Pregnancy Hypertension* 11, no. January (January): 115–23. <https://doi.org/10.1016/J.PREGHY.2017.11.006>.
- Umesawa, Mitsumasa, and Gen Kobashi. 2016. "Epidemiology of Hypertensive Disorders in Pregnancy: Prevalence, Risk Factors, Predictors and Prognosis." *Hypertension Research* 40:3 40, no. 3 (September): 213–20. <https://doi.org/10.1038/hr.2016.126>.
- Varoquaux, Gael, and Olivier Colliot. 2023. "Evaluating Machine Learning Models and Their Diagnostic Value." *Neuromethods* 197: 601–30. https://doi.org/10.1007/978-1-0716-3195-9_20/TABLES/2.
- Yahaya, Haphsheitu, Queen Esther Adeyemo, and Augustine Kumah. 2024. "Adverse Perinatal Outcomes and Their Associated Determinants in Sub-Saharan Africa." *Journal of Medicine, Surgery, and Public Health* 3, no. August (August): 100124. <https://doi.org/10.1016/J.GLMEDI.2024.100124>.
- Zheng, Alice, and Amanda Casari. 2018. "Feature Engineering for Machine Learning PRINCIPLES AND TECHNIQUES FOR DATA SCIENTISTS."
- Zheng, Dongying, Xinyu Hao, Muhanmmad Khan, Lixia Wang, Fan Li, Ning Xiang, Fuli Kang, et al. 2022. "Comparison of Machine Learning and Logistic Regression as Predictive Models for Adverse Maternal and Neonatal Outcomes of Preeclampsia: A Retrospective Study." *Frontiers in Cardiovascular Medicine* 9, no. October (October): 959649. <https://doi.org/10.3389/FCVM.2022.959649/BIBTEX>.

Annex

A: Table One: List of final attributes selected for modelling

No	Attribute Name
1	Age
2	BMI
3	PR
4	RR
5	Blood Group
6	Gestational Age
7	WBC
8	Neu
9	Lym
10	Mono
11	eos
12	baso
13	RBC
14	hemo
15	hema
16	mcv
17	mch
18	mchc
19	mpv
20	plt
21	appearance
22	ph
23	sg
24	prot
25	glu
26	ket
27	blood
28	Urobili
29	Nitrite
30	urea
31	AST
32	ALT
33	ALP
34	Total bilirubin
35	SBP
36	DBP
37	Anemia
38	Presentation
39	Malnutrition
40	Type of HDP

41	Birth type
42	Outcome

B: Table Two: Performance of models using different imputation methods

Model	Imputation Method	Precision(%)	Recall(%)	F1-Score(%)	Accuracy(%)
LR	Median	80	52	63	89
	*MICE	84	54	65	89
	KNN	83	52	64	89
NB	Median	22	78	34	43
	MICE	22	78	34	43
	*KNN	22	78	35	44
SVM	Median	72	26	38	84
	*MICE	72	26	38	84
	KNN	72	26	38	84
DT	Median	52	53	52	82
	MICE	54	54	54	83
	*KNN	56	57	57	84
XGBoost	*Median	85	55	67	90
	MICE	84	51	63	89
	KNN	81	52	64	89
RF	Median	95	51	66	90
	MICE	95	50	65	90
	*KNN	95	52	67	90

*Models and imputation methods selected for hyperparameter tuning

C: Table Three: Performance of models using different tuning methods

Model	Hyperparameter Tuning	Precision(%)	Recall(%)	F1-Score(%)	Accuracy(%)
LR	Grid Search	84	52	64	89
	*Bayesian Optimization	85	52	64	89
NB	*Grid Search	51	60	55	82
	Bayesian Optimization	51	60	55	82
SVM	*Grid Search	81	50	62	88
	Bayesian Optimization	71	48	57	87
DT	Grid Search	73	36	48	85
	*Bayesian Optimization	62	38	47	84
XGBoost	Grid Search	89	54	67	90
	*Bayesian Optimization	91	56	69	91
RF	*Grid Search	93	52	66	90
	Bayesian Optimization	91	50	65	90

*Models selected as best from the two tuning methods