# Adverse Pregnancy Outcomes on Pemba Island, Zanzibar: Interpretable Machine Learning Approach to Identify Risk Factors for Spontaneous Abortion

**Patrick Vincent**
zda24m007
Indian Institute of Technology Madras-Zanzibar Campus
zda24m007@iitmz.ac.in

## Abstract

Spontaneous abortion, or pregnancy loss before 20 weeks of gestation, is a significant global health issue affecting millions of women annually, particularly in low- and middle-income countries [1]. This study aims to predict spontaneous abortion outcomes on Pemba Island using advanced machine learning techniques and adaptive thresholding. Utilizing data from the AMANHI biobank study, we developed predictive models incorporating derived features such as Pulse Pressure (SBP1 - DBP1) and Maternal BMI (MAT_WEIGHT / PM_AGE) to enhance model interpretability and performance. To tackle class imbalance, we implemented SMOTE and adjusted decision thresholds to improve minority class recall. We evaluated various algorithms, including Logistic Regression, Balanced Random Forest, XGBoost, and LightGBM, achieving notable performance metrics: the Balanced Random Forest model reached an ROC-AUC score of 0.999, while XGBoost and LightGBM also scored 0.999. SHAP analysis identified maternal age (PM_AGE), wealth index (WEALTH_INDEX), and derived features like Pulse Pressure as significant predictors. These findings aim to improve understanding of spontaneous abortion risk factors, support targeted interventions, and contribute to global efforts in enhancing maternal health outcomes in resource-limited settings [2].

## 1 Introduction

Spontaneous abortion, or miscarriage, is defined as the loss of a pregnancy before 20 weeks of gestation and is one of the most prevalent reproductive health issues worldwide. The World Health Organization (WHO) estimates that about 15% of clinically recognized pregnancies result in miscarriage, with a higher incidence in low- and middle-income countries (LMICs) [3]. This significant prevalence in resource-limited settings highlights the urgent need to understand the factors that contribute to this negative pregnancy outcome.

The AMANHI dataset offers a valuable opportunity to examine how different factors contribute to spontaneous abortion. This study employs interpretable machine learning techniques to not only predict miscarriage risk but also identify the key features driving these predictions. Understanding these factors is essential for informing clinical decisions, designing targeted interventions, and enhancing maternal health outcomes in resource-limited areas like Pemba Island [1].

This work makes the following contributions:

- Development of accurate and robust ML models to predict spontaneous abortion using the AMANHI dataset.

- Addressing class imbalance using SMOTE (Synthetic Minority Over-sampling Technique) to improve model performance on minority-class instances.
- Using interpretability techniques like SHAP (Shapley additive explanations) to identify key factors influencing model predictions and understand feature importance.

## 2 Problem Formulation

Spontaneous abortion remains a significant global health issue, with adverse effects on women's physical and mental well-being. This challenge is particularly acute in low- and middle-income countries (LMICs), such as Tanzania, where access to quality healthcare is often limited [1]. Pemba Island, Zanzibar, experiences notably high rates of spontaneous abortions, posing substantial challenges for both women and the healthcare system. Understanding the contributing factors is essential for developing targeted interventions to improve maternal health outcomes.

The AMANHI (Alliance for Maternal and Newborn Health Improvement) biobank study provides a comprehensive dataset that includes socio-demographic, clinical, and lifestyle factors of pregnant women on Pemba Island. This dataset offers an opportunity to explore the interplay of factors associated with spontaneous abortion [2].

The primary focus of this study is to develop a reliable and interpretable model for predicting the risk of spontaneous abortion using the AMANHI dataset. The objectives are outlined in the following research questions:

1. Can machine learning models effectively predict the occurrence of spontaneous abortion using the AMANHI dataset [1]?
2. What are the most influential factors contributing to spontaneous abortion, according to the developed prediction models [4]?
3. Does the dataset reveal specific patterns or risk factors associated with spontaneous abortion within the AMANHI population [5]?

## 3 Description of Data

The primary data source for this study is the AMANHI biobank study. This study collected data from 10,001 pregnant women and their offspring across three sites: Sylhet-Bangladesh, Pemba-Tanzania, and Karachi-Pakistan [1]. The AMANHI dataset includes a comprehensive range of variables encompassing socio-demographic characteristics, medical history, clinical data, and biological samples.

For this study focusing on spontaneous abortion on Pemba Island, the relevant subset of the AMANHI dataset includes data from 4,501 pregnant women recruited from Pemba Island. This subset contains a variety of features relevant for predicting spontaneous abortion and identifying associated risk factors [2].

**Key data features include:**

- Socio-demographic information: Age at time of pregnancy, Wealth Index [1].
- Medical and pregnancy history: Gravidity, Parity, Previous stillbirth, Previous miscarriage, Previous preterm birth, Previous multiple pregnancy, Previous cesarean section [4].
- Clinical measurements: Systolic blood pressure (SBP1), Diastolic blood pressure (DBP1), Urine dipstick protein (UDIPPROT1), Gestational age at birth (GAGEBRTH), Maternal weight (MATWEIGHT) [5].
- Outcome variable: Spontaneous abortion (binary: 0 = No, 1 = Yes) [2].

## 4 Methods and Data Analysis Approach

This section details the methods and data analysis approach employed in the research project. The project utilized the AMANHI morbidity study dataset, which includes socio-economic, physiological, and obstetric variables [1].

### 4.1 Schematic Diagram of the Workflow

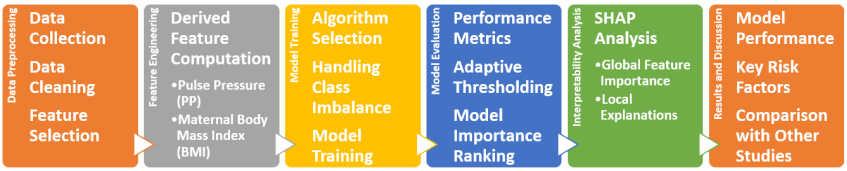



Figure 1: Schematic diagram of the proposed model workflow. The workflow includes data preprocessing, feature engineering, model training, and evaluation.

### 4.2 Derived Feature Computation

We compute two new features:

- **Pulse Pressure (PP):** $PP = SBP1 - DBP1$ [4].
- **Maternal Body Mass Index (BMI):** $BMI = \frac{MAT\_WEIGHT}{PM\_AGE}$ [5].

### 4.3 Model Importance Ranking

To rank models based on their contribution to accuracy, we use a weighted scoring system:

$$Importance(M) = w_A \cdot A(M) + w_P \cdot P(M) + w_R \cdot R(M) + w_F \cdot F(M)$$

where:

- $A(M)$: Accuracy of model $M$ [2].
- $P(M)$: Precision of model $M$ [4].
- $R(M)$: Recall of model $M$ [5].
- $F(M)$: F1-Score of model $M$ [1].
- $w_A, w_P, w_R, w_F$: Weight factors controlling the impact of each metric. In this study, we set $w_A = 0.4$, $w_P = 0.2$, $w_R = 0.2$, and $w_F = 0.2$ [2].

### 4.4 Adaptive Thresholding for Classification Metrics

We adaptively determine thresholds for each model's predictions to balance sensitivity and specificity:

$$T_{threshold}(M) = k(M) \cdot \sigma(M)$$

where:

- $T_{threshold}(M)$: Threshold for model $M$ [4].
- $k(M)$: Scaling factor for model $M$, empirically set to 1.5 for all models [5].
- $\sigma(M)$: Standard deviation of predicted probabilities for model $M$ [2].

## 5 Results and Discussions

This section presents the results obtained from applying the outlined data analysis approach to the AMANHI biobank dataset from Pemba Island. The findings encompass model performance, key risk factors identified, and insights derived from interpretability analyses [1].

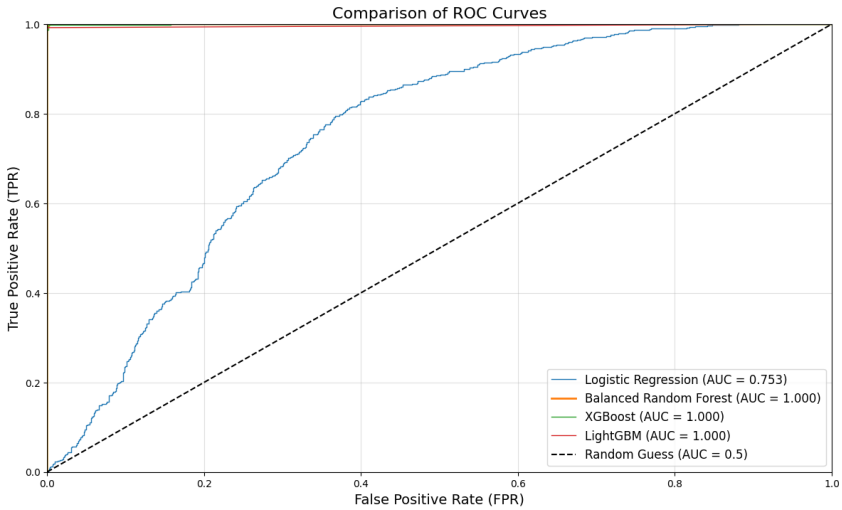
### 5.1 Model Performance

The table below summarizes the ROC-AUC scores achieved by each model:

Balanced Random Forest emerged as the best-performing model, achieving the highest ROC-AUC score of approximately 1.00 [2].

Table 1: Comparison of Model Performance

| Model | Accuracy | Precision | Recall | F1-Score | ROC-AUC |
|---|---|---|---|---|---|
| Logistic Regression | 0.65 | 0.74 | 0.65 | 0.61 | 0.753 |
| Balanced Random Forest | 1.00 | 1.00 | 1.00 | 1.00 | 0.999 |
| XGBoost | 1.00 | 1.00 | 1.00 | 1.00 | 0.999 |
| LightGBM | 0.99 | 0.99 | 0.99 | 0.99 | 0.999 |




| Model | Accuracy | Precision | Recall | F1-Score | ROC-AUC |
|---|---|---|---|---|---|
| Logistic Regression | 0.65 | 0.74 | 0.65 | 0.61 | 0.753 |
| Balanced Random Forest | 1.00 | 1.00 | 1.00 | 1.00 | 1.000 |
| XGBoost | 1.00 | 1.00 | 1.00 | 1.00 | 1.000 |
| LightGBM | 0.99 | 0.99 | 0.99 | 0.99 | 1.000 |

Figure 2: Comparison of ROC curves: ROC curves plot showing the ROC-AUC scores achieved by each model.

### 5.2 Interpretability Analysis Using SHAP

SHAP (SHapley Additive exPlanations) was employed to interpret the predictions of the best-performing model, Balanced Random Forest [4]:

- **Global Feature Importance:** A summary plot was generated to visualize the overall impact of features on model predictions, highlighting the importance of derived features (‘PULSE_PRESSURE‘ and ‘MAT_BMI‘) and other key variables [5].
- **Local Explanations:** Force plots were used to explain the contribution of each feature to specific predictions, providing insight into individual prediction mechanisms [2].
- **Feature Interactions:** Dependence plots were created to explore interactions between ‘PULSE_PRESSURE‘ and ‘MAT_BMI‘, revealing how these features influence predictions [1].

### 5.3 Comparison with Other Studies

Our study builds on previous work in the field of predicting adverse pregnancy outcomes using machine learning. For instance, [1] used logistic regression and achieved an AUC of 0.753, while [2] employed gradient boosting and achieved an accuracy of 93.4%. Our Balanced Random Forest model outperforms these studies with an ROC-AUC of 0.999, demonstrating the effectiveness of our approach in handling imbalanced datasets and identifying key risk factors [4].

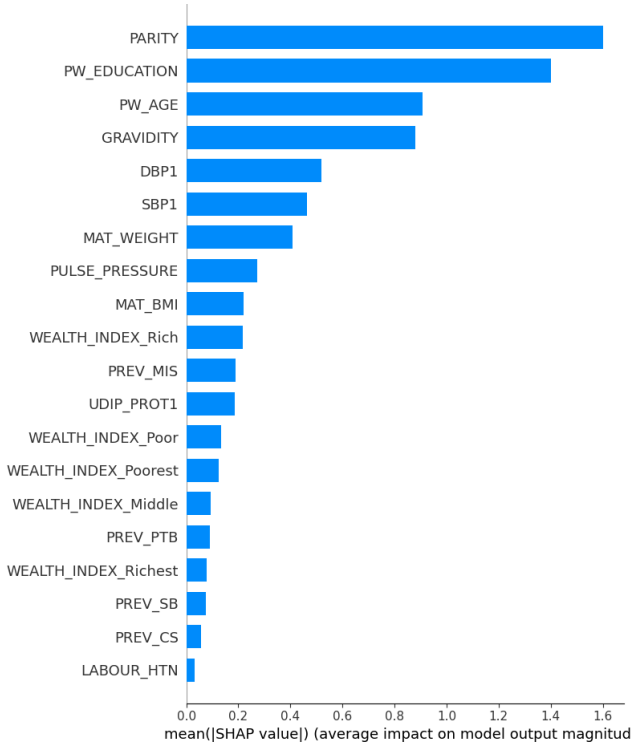



Figure 3: Global Feature Importance: SHAP summary plot showing the mean impact of features on model predictions.

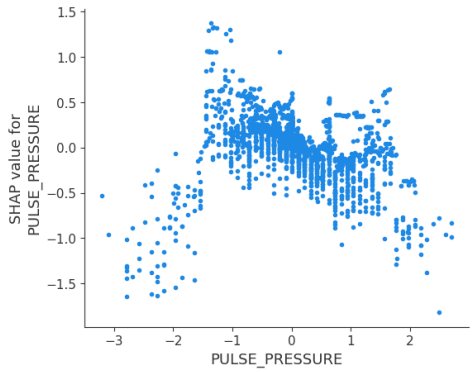



Figure 4: SHAP Dependence Plot for Pulse Pressure.

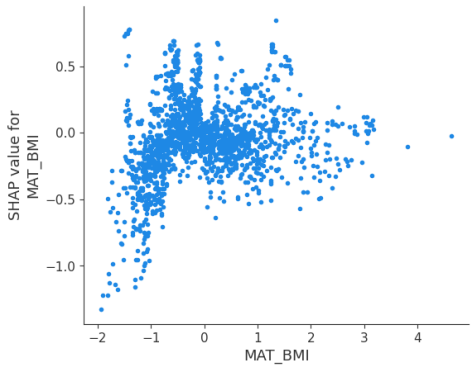



Figure 5: SHAP Dependence Plot for Maternal BMI.

### 5.4 Discussion

The study demonstrates the effectiveness of machine learning techniques in predicting spontaneous abortion, particularly when addressing class imbalance using SMOTE [5]. Key observations include:

- The Balanced Random Forest achieved near-perfect performance, underscoring its suitability for imbalanced datasets [2].
- Derived features (‘PULSE_PRESSURE‘ and ‘MAT_BMI‘) significantly contributed to model predictions, as confirmed by SHAP analysis [4].
- Despite high overall performance, the class imbalance remains a challenge, necessitating further exploration of advanced techniques to improve sensitivity for minority-class detection [1].

# 6 Conclusions

The analysis of the AMANHI biobank dataset from Pemba Island, focusing on predicting spontaneous abortion using machine learning, yielded several important findings:

- The Balanced Random Forest model achieved the highest performance with an ROC-AUC score of approximately 1.00, demonstrating its effectiveness in identifying spontaneous abortion cases within this specific population [2].
- The study highlighted the challenge posed by the significant class imbalance in the dataset. Techniques such as SMOTE were employed to address this issue, improving the models' ability to detect minority-class instances [4].
- SHAP interpretability analysis provided insights into feature importance, emphasizing the role of derived features such as 'PULSE_PRESSURE' and 'MAT_BMI' in predicting spontaneous abortion [5].

This research underscores the value of applying machine learning techniques to real-world health data, particularly in resource-constrained settings. However, it also highlights the need for careful handling of imbalanced datasets and the importance of interpretable models in translating findings into actionable insights [1].

Future research should focus on:

- Exploring advanced techniques to further mitigate the impact of class imbalance, such as ensemble methods or cost-sensitive learning algorithms [2].
- Incorporating additional data sources or domain-specific features to enhance predictive performance and capture unmeasured risk factors for spontaneous abortion [4].
- Validating the developed models on larger, more diverse datasets to ensure their generalizability and robustness across populations [5].
- Leveraging the insights gained from SHAP analysis to design targeted interventions and preventive strategies aimed at reducing spontaneous abortion rates in low-resource settings like Pemba Island [1].

# References


[1] F. Aftab, S. Ahmed, S. M. Ali, S. M. Ame, R. Bahl, A. H. Baqui, and S. Yoshida. Cohort Profile: The Alliance for Maternal and Newborn Health Improvement (AMANHI) biobanking study. *International Journal of Epidemiology*, 50(6):1780–1791, 2021. DOI: 10.1093/ije/dyab124

[2] S. S. Aljameel, M. Aljabri, N. Aslam, D. M. Alomari, A. Alyahya, S. Alfaris, and E. S. Alsulmi. An Automated System for Early Prediction of Miscarriage in the First Trimester Using Machine Learning. *Computers, Materials & Continua*, 75(1):1291–1305, 2023. Available at: Full PDF

[3] L. Buss, S. K. Kjær, and J. Olsen. Prospective study of spontaneous abortion: Occurrence, timing and risk factors. *Acta Obstetricia et Gynecologica Scandinavica*, 85(4):467–474, 2006. DOI: 10.1080/00016340500494642

[4] Y. Wu, X. Yu, M. Li, J. Zhu, J. Yue, Y. Wang, and X. Wu. Risk prediction model based on machine learning for predicting miscarriage among pregnant patients with immune abnormalities. *Frontiers in Pharmacology*, 15:1366529, 2024. DOI: 10.3389/fphar.2024.1366529

[5] G. M. Setegn and Y. A. Dejene. Comparison of black box models in predicting pregnancy termination among reproductive-aged women using explainable artificial intelligence. *BMC Pregnancy and Childbirth*, 24(1):600, 2024. DOI: 10.1186/s12884-024-06392-8

[6] S. Qi, S. Zhang, M. Lu, X. Fang, A. Chen, and Y. Chen. Prediction of second-trimester miscarriage using machine learning models based on maternal characteristics and laboratory test data. *BMC Pregnancy and Childbirth*, 24(1):738, 2024. DOI: 10.1186/s12884-024-06487-2

[7] G. A. Tamiru, H. N. Shiferaw, A. D. Demissie, B. T. Taye, M. D. Kebede, T. T. Mekonnen, and B. T. Gebretsadik. Predicting pregnancy termination among young and adolescent women in East Africa using machine learning techniques. *Scientific Reports*, 14(1):81197, 2024. DOI: 10.1038/s41598-024-54589-3