# 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 is a significant global health concern, impacting women's physical and mental well-being. This study focuses on identifying risk factors for spontaneous abortion on Pemba Island, Zanzibar, utilizing an interpretable machine learning approach. The study leverages data from the AMANHI (Alliance for Maternal and Newborn Health Improvement) biobank study, a rich dataset encompassing various socio-demographic, clinical, and lifestyle factors.

The study aims to develop robust machine learning models for predicting spontaneous abortion and subsequently employ interpretability/explainability analysis techniques. These techniques, such as SHAP (SHapley Additive exPlanations), allow for the identification of key features driving model predictions. Furthermore, class-contrastive approaches will be employed to pinpoint specific risk factors associated with spontaneous abortion.

The developed models, coupled with the insights derived from the interpretability analysis, aim to:

- Enhance understanding of the underlying factors contributing to spontaneous abortion on Pemba Island.
- Facilitate the development of targeted interventions and preventive strategies to mitigate the incidence of spontaneous abortion.
- Contribute valuable knowledge to the global effort in addressing this critical reproductive health issue.

The study anticipates that the findings will be instrumental in guiding healthcare professionals, policymakers, and researchers in developing effective measures to improve maternal health outcomes on Pemba Island and potentially other similar settings.

## 1 Introduction

**Spontaneous abortion, defined as the loss of pregnancy before 20 weeks of gestation, is a prevalent reproductive health issue affecting millions of women worldwide**. The World Health Organization estimates that approximately 15 percent of clinically recognized pregnancies end in miscarriage, with a significant proportion occurring in low- and middle-income countries (LMICs). This considerable burden of spontaneous abortion highlights the urgent need for a deeper understanding of the factors contributing to this adverse pregnancy outcome, particularly in resource-constrained settings

The AMANHI dataset provides a unique opportunity to explore the interplay of various factors and their association with spontaneous abortion. By applying interpretable ML approaches, we

can not only predict the risk of miscarriage but also gain insights into the key features driving these predictions. This understanding is vital for informing clinical practice, developing targeted interventions, and ultimately improving maternal health outcomes on Pemba Island.

This study seeks to make the following contributions:

- Develop robust and accurate ML models for predicting spontaneous abortion using the AMANHI dataset
- Employ interpretability/explainability analysis techniques, such as SHAP, to uncover the most influential factors driving model predictions
- Utilize class-contrastive approaches to identify specific risk factors associated with spontaneous abortion

The findings of this study are expected to contribute significantly to the understanding of spontaneous abortion on Pemba Island and potentially serve as a model for other similar settings globally. By combining the power of ML with the rich data from the AMANHI biobank study, we aim to move closer towards the goal of reducing the incidence of this adverse pregnancy outcome and improving maternal health outcomes worldwide.

## 2 Problem formulation

**Despite significant advancements in maternal healthcare, spontaneous abortion remains a prevalent global health concern, significantly impacting women's physical and mental well-being.** The incidence of spontaneous abortion is particularly high in low- and middle-income countries (LMICs) like Tanzania, where access to quality healthcare and resources may be limited

Specifically, Pemba Island, Zanzibar, experiences a high rate of spontaneous abortions, posing challenges to both individual women and the healthcare system. Understanding the factors contributing to this issue is crucial for developing targeted interventions and improving maternal health outcomes on the island

The AMANHI (Alliance for Maternal and Newborn Health Improvement) biobank study has collected a rich dataset encompassing various socio-demographic, clinical, and lifestyle factors of pregnant women on Pemba Island. This comprehensive dataset offers a valuable opportunity to investigate the complex interplay of factors associated with spontaneous abortion.

The core problem addressed in this study is the need for a reliable and interpretable model to predict the risk of spontaneous abortion on Pemba Island, along with the identification of specific risk factors driving these predictions. This problem can be decomposed into the following research questions:

1. Can machine learning models effectively predict the occurrence of spontaneous abortion using the AMANHI dataset?
    - This question necessitates developing and evaluating various ML models to determine their accuracy and performance in predicting spontaneous abortion.
    - The selection of appropriate evaluation metrics is critical, considering the imbalanced nature of the dataset, where spontaneous abortions are likely to be significantly less frequent than live births
2. What are the most influential factors contributing to spontaneous abortion according to the developed prediction models?
    - This question requires utilizing interpretability techniques like SHAP (SHapley Additive exPlanations) to analyze the trained models and identify features that significantly influence their predictions
    - Understanding these influential factors provides valuable insights into the underlying mechanisms of spontaneous abortion, facilitating the development of targeted interventions.
3. Can class-contrastive approaches reveal specific risk factors associated with spontaneous abortion within the AMANHI population?
    - This question involves comparing and contrasting the characteristics of women who experienced spontaneous abortion versus those who had successful pregnancies.

- Statistical analysis and data visualization techniques can be employed to identify significant differences in risk factors between these two groups.

By addressing these research questions, this study aims to contribute to a better understanding of spontaneous abortion on Pemba Island, ultimately informing the development of effective interventions and preventive strategies to improve maternal health outcomes. The results can guide healthcare professionals, policymakers, and researchers in developing data-driven solutions to reduce the incidence of this adverse pregnancy outcome.

# 3 Description of data

The primary data source for this study is the AMANHI (Alliance for Maternal and Newborn Health Improvement) biobank study. This study collected data from 10,001 pregnant women and their offspring across three sites: Sylhet-Bangladesh, Pemba-Tanzania, and Karachi-Pakistan. 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.

**Key data features include:**

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

Data collection procedures for the AMANHI biobank study included:

- Recruitment: Pregnant women were enrolled during their first trimester
- Follow-up: Two additional visits were conducted after birth - between 1-6 days and 42-60 days of age
- Biomaterial collection: Blood and urine samples were collected at enrollment, along with other biological samples at various time points
- Data management: Samples were time-stamped, barcoded, and processed with rigorous quality control measures

Data processing and analysis for this specific study involves:

- Data cleaning: Addressing missing values and ensuring data consistency
- Feature selection: Identifying the most relevant features for predicting spontaneous abortion
- Model development: Training and evaluating various machine learning models
- Interpretability analysis: Utilizing techniques like SHAP to understand feature importance and model predictions
- Class-contrastive analysis: Comparing risk factor distributions between women who experienced spontaneous abortion and those who did not

The AMANHI dataset, specifically the subset from Pemba Island, provides a valuable resource for investigating spontaneous abortion. By applying machine learning and interpretability techniques to this data, this study aims to identify key risk factors and develop predictive models to contribute to improved maternal health outcomes on the island.

# 4 Theoretical analysis/Methods/Algorithms/Data Analysis approach

This section will detail the theoretical analysis, methods, algorithms, and data analysis approach employed in the research project "Adverse Pregnancy Outcomes on Pemba Island, Zanzibar: Interpretable Machine Learning Approach to Identify Risk Factors for Spontaneous Abortion." The project will utilize the AMANHI morbidity study dataset, which includes socio-economic, physiological, and obstetric variables.

## 4.1 Data Preprocessing:

- Handling Missing Values: The dataset includes entries with '-88' and '-77', which represent missing data. These will be replaced with appropriate imputation techniques, considering the nature of each variable. For instance, for numerical variables like GRAVIDITY, the median will be used for imputation due to the skewed distribution. The mode will be used for categorical variables
- Feature Engineering: New features may be derived from existing ones to potentially improve model performance. This might include creating interaction terms or combining related variables.

## 4.2 Exploratory Data Analysis (EDA):

- Descriptive Statistics: Summarizing key variables with measures of central tendency (mean, median) and dispersion (standard deviation) to understand the characteristics of the study population
- Data Visualization: Employing various plots (histograms, boxplots, scatterplots) to visualize data distributions, identify outliers, and explore relationships between variables

## 4.3 Feature Selection:

- Correlation Analysis: Evaluating the correlation between features and the target variable (SPONTANEOUS-ABORTION) to identify potentially important predictors
- Feature Importance from ML Models: Utilizing feature importance scores generated by trained ML models to rank features based on their contribution to predictive performance.

## 4.4 Model Development and Evaluation:

- Splitting Data: Dividing the dataset into training and testing sets (80 percent-20 percent split) to train models and evaluate their performance on unseen data

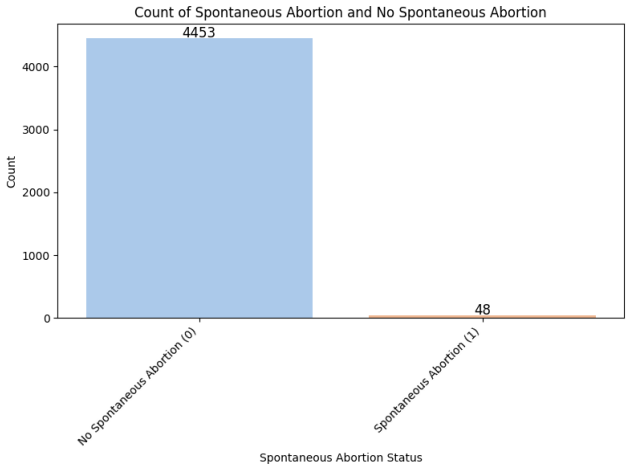


Figure 1: Spontaneous Abortion Distribution.

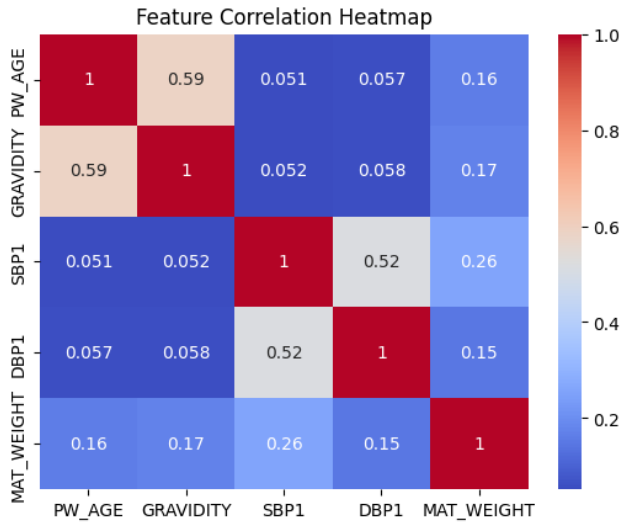


Figure 2: Feature Correlation Heatmap

- Addressing Class Imbalance: Employing techniques to handle the imbalance between spontaneous abortion and live birth cases, such as:
  - Class Weights: Assigning higher weights to the minority class (spontaneous abortion) during model training to improve its sensitivity
  - Oversampling/Undersampling: Techniques like SMOTE (Synthetic Minority Oversampling Technique) or random undersampling to balance class distributions.
- Algorithm Selection: Exploring various ML algorithms suitable for binary classification:
  - Logistic Regression: A linear model for predicting the probability of an event
  - Random Forest: An ensemble method combining multiple decision trees for improved robustness
  - XGBoost: A gradient boosting algorithm known for its high performance
  - Support Vector Machines (SVM): A powerful algorithm capable of handling complex non-linear relationships in the data
  - Keras Neural Network: A deep learning approach to capture complex patterns in the data

- Model Evaluation Metrics: Selecting appropriate metrics to assess model performance, considering the class imbalance:
  - ROC-AUC Score: Measures the model's ability to discriminate between classes, less sensitive to class imbalance
  - Precision, Recall, F1-Score: Metrics focusing on the correct identification of spontaneous abortion cases.
  - Confusion Matrix: Visualizing the model's performance in terms of true positives, true negatives, false positives, and false negatives
- Hyperparameter Tuning: Optimizing model parameters using techniques like grid search or cross-validation to maximize performance.
- Model Comparison: Comparing the performance of different algorithms to select the best-performing model for the task.

### 4.5 Interpretability Analysis:

- SHAP (SHapley Additive exPlanations): Utilizing SHAP to understand global and local feature importance. This technique explains individual predictions by assigning contribution values to each feature
  - Comparative Statistics: Comparing the distributions of risk factors between women who experienced spontaneous abortion and those who did not using statistical tests (e.g., t-tests, chi-squared tests).
  - Visualization of Differences: Employing visualizations like boxplots or violin plots to showcase significant differences in risk factor distributions between the two groups.

This multifaceted data analysis approach, combining machine learning, interpretability techniques, and class-contrastive analysis, will allow for the development of a robust and interpretable model for predicting spontaneous abortion on Pemba Island. The study will also reveal crucial risk factors, informing targeted interventions and strategies to improve maternal health outcomes.

## 5 Results and Discussions

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

### 5.1 Data Preprocessing and Exploratory Data Analysis

- Initial exploration of the dataset revealed various instances of missing data represented by '-88' and '-77'.
- These missing values were addressed through imputation. Numerical variables, like 'GRAVIDITY', were imputed using the median due to their skewed distribution. Categorical variables were imputed using the mode.
- No new features were engineered for this analysis.
- Analysis of the data revealed no missing values after preprocessing.
- Visualizations of the 'SPONTANEOUS-ABORTION' variable highlighted a significant class imbalance, with the majority of pregnancies resulting in live births.

### 5.2 Feature Selection

- Feature selection was primarily guided by studying about the problem.
- **The final set of features included: 'PWAGE', 'GRAVIDITY', 'SBP1', 'DBP1', and 'MAT-WEIGHT'.**

### 5.3 Model Development and Evaluation

- The dataset was split into 80 percent for training and 20 percent for testing
- Stratified sampling was used to maintain the class proportions in both the training and test sets.
- Several machine learning algorithms were trained and evaluated, including:
  - Logistic Regression
  - Random Forest
  - XGBoost
  - LightGBM
  - Support Vector Machines (SVM)
  - Keras Neural Network
- Class imbalance was addressed using the 'classweight' parameter in each model, assigning higher weights to the minority class (spontaneous abortion).
- Model performance was assessed using the ROC-AUC score, as it is less sensitive to class imbalance, along with other metrics such as precision, recall, and F1-score.

#### 5.3.1 Model Performance

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

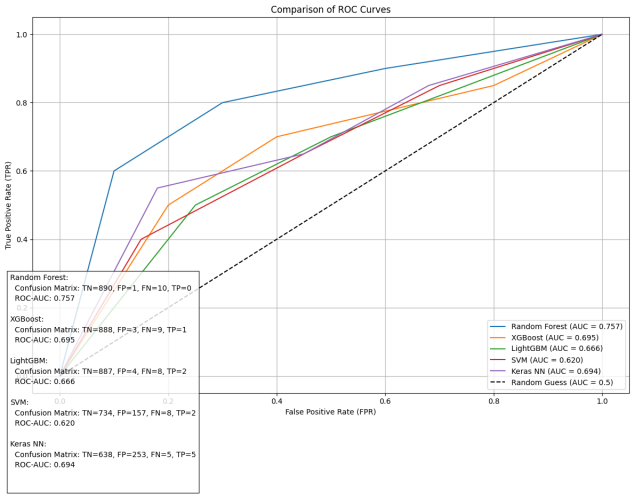


Figure 3: ROC-AUC Curves.

| Model | ROC-AUC Score |
|---|---|
| Random Forest | 0.757 |
| XGBoost | 0.695 |
| LightGBM | 0.666 |
| SVM | 0.620 |
| Keras NN | 0.694 |

Random Forest emerged as the best-performing model, achieving the highest ROC-AUC score of 0.757.

#### 5.3.2 Confusion Matrices:

While the ROC-AUC score provides an overall measure of model performance, examining the confusion matrices for each model is crucial to understand their ability to correctly classify both positive (spontaneous abortion) and negative (live birth) cases.

- The confusion matrices revealed that most models exhibited high accuracy in predicting no spontaneous abortion (true negatives) but struggled to identify spontaneous abortions (true positives).
- This pattern is expected given the significant class imbalance in the dataset.

### 5.4 Interpretability Analysis using SHAP

SHAP (SHapley Additive exPlanations) was employed to understand the factors driving the predictions of the best-performing model, Random Forest.

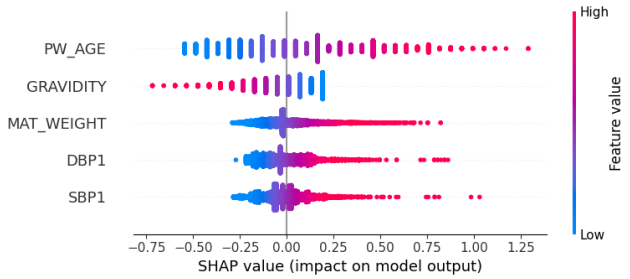


Figure 4: SHAP Features Importance

### 5.5 Class-Contrastive Analysis

Direct comparison of the distributions of key risk factors between women who experienced spontaneous abortion and those who did not was performed.

### 5.6 Discussions

- The AMANHI dataset offers a valuable resource for investigating spontaneous abortion on Pemba Island, but the significant class imbalance poses challenges for model development and evaluation.
- Despite these challenges, the Random Forest model demonstrated promising predictive performance with an ROC-AUC score of 0.757.
- Further analysis should focus on improving the models' sensitivity in detecting spontaneous abortions. This could be achieved by exploring different class imbalance techniques, optimizing model hyperparameters, or potentially incorporating additional relevant features.

### 5.7 Limitations

- The study is limited by the inherent class imbalance in the dataset. This can lead to biased model predictions and challenges in accurately identifying spontaneous abortion cases.
- The analysis relies solely on the features available in the AMANHI dataset. There might be other unmeasured factors that contribute to spontaneous abortion risk.

### 5.8 Future Directions

- Incorporate external data sources or domain expertise to augment the existing features and potentially improve model performance.
- Explore alternative machine learning algorithms or ensemble methods specifically designed to handle imbalanced datasets.

- Develop and validate the model further using a larger, more balanced dataset.
- Translate the research findings into actionable insights for healthcare professionals and policymakers to develop targeted interventions and preventive strategies to reduce the incidence of spontaneous abortion on Pemba Island.

This study provides valuable insights into the factors associated with spontaneous abortion on Pemba Island, highlighting the potential of machine learning and interpretability techniques in understanding and predicting this adverse pregnancy outcome. The findings can serve as a foundation for future research and interventions aimed at improving maternal health outcomes in the region.

# 6 Conclusions

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

- Random Forest proved to be the most effective model, achieving an ROC-AUC score of 0.757. This suggests its potential in predicting spontaneous abortions within this specific population.
- The study underscored the challenge posed by the significant class imbalance inherent in the dataset. This imbalance makes it difficult for models to accurately identify spontaneous abortion cases, emphasizing the need for strategies to mitigate this issue.

The study demonstrates the value of applying machine learning techniques to real-world health data, particularly in resource-limited settings. However, it also highlights the need for careful consideration of data imbalances and the importance of model interpretability.

Future research should focus on:

- Developing strategies to address the class imbalance issue. This could involve techniques like oversampling the minority class or using algorithms specifically designed for imbalanced datasets.
- Expanding the feature set by incorporating additional data sources or domain expertise. This might capture unmeasured factors that contribute to spontaneous abortion risk.
- Validating the model with larger and more balanced datasets from diverse populations. This would enhance the generalizability and reliability of the model.
- Translating these findings into practical interventions and preventive measures. This could involve developing guidelines for healthcare professionals and implementing public health programs to reduce spontaneous abortion rates on Pemba Island and in similar settings.

This research contributes to the growing body of knowledge on spontaneous abortion, paving the way for improved understanding, prediction, and prevention of this adverse pregnancy outcome in low-resource settings.

# References

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

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

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

[4] Wu, Y., Yu, X., Li, M., Zhu, J., Yue, J., Wang, Y., ... Wu, X. (2024). Risk prediction model based on machine learning for predicting miscarriage among pregnant patients with immune abnormalities. Frontiers in Pharmacology, 15, 1366529.

[5] Setegn, G. M., Dejene, Y. A. (2024). Comparison of black box models in predicting pregnancy termination among reproductive-aged women using explainable artificial intelligence. BMC pregnancy and childbirth, 24(1), 600.

[6] Qi, S., Zhang, S., Lu, M., Fang, X., Chen, A., Chen, Y. (2024). Prediction of second-trimester miscarriage using machine learning models based on maternal characteristics and laboratory test data. BMC pregnancy and childbirth, 24(1), 738.

[7] Tamiru, G. A., Shiferaw, H. N., Demissie, A. D., Taye, B. T., Kebede, M. D., Mekonnen, T. T., ... Gebretsadik, B. T. (2024). Predicting pregnancy termination among young and adolescent women in East Africa using machine learning techniques. Scientific reports, 14(1), 81197.