CONTENTS

INTRODUCTION

Child mortality is a key indicator of a country's socio-economic development and healthy population development. While the overall burden of child mortality has decreased substantially in recent decades, it remains high, especially in low-income and developing countries. In 2022, approximately 4.9 million children under the age of five died globally from a variety of causes that can be prevented, such as infectious diseases, complications of preterm birth, and malnutrition. Achieving Sustainable Development Goal 3 (SDG-3) – which seeks to reduce under-five mortality to at least 25 deaths per 1,000 live births by 2030 – is a global challenge.

Statistically speaking, there are several factors influencing child survival outcomes. Previous studies have indicated the importance of socio-economic status, maternal education, access to health-care facilities, birth intervals, and environmental variables. Statistical modeling will facilitate the quantification of these associations and the identification of high-risk groups for interventions, which can help in the design of interventions.

India has made rapid progress in the burden of under-five mortality, with the Under-5 Mortality Rate (U5MR) falling from 45% per 1,000 live births in 2014 to 32% per 1,000 in 2020, although disparities persist across regions, socio-economic groups, and gender.

While logistic regression and survival analysis have traditionally been the dominant statistical approaches for analysing child mortality determinants, their usefulness has grown diminished due to the complexity and size of more recent datasets (e.g., the National Family Health Survey, NFHS), which require more flexible modeling. Aspects of machine learning (ML) methods are useful in terms of nonlinear regression models, large feature spaces, and improved predictive performance (i. e. fewer assumptions about distribution).

Using NFHS data, we propose two machine learning algorithms for the classification of children’s survival outcome: Logistic Regression, K-Nearest Neighbours (KNN), Random Forest, Gradient Boosting, XGBoost, and Support Vector Machine (SVM). To address the class imbalance problem (there were more surviving children than dying), we use the Synthetic Minority Oversampling Technique (SMOTE) during preprocessing.

To evaluate model performance, we analysed various statistical measures: accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (ROC-AUC). Not only was predictive performance assessed in the study, but also the features responsible for child survival have been identified which provide methodological insights as well as practical implications for public health policy.

By combining machine learning techniques with traditional statistical research methods this project is aiming to contribute to a new area of applied statistics in public health in which we focus on how advanced modelling approaches can help to understand and reduce child mortality.

LITERATURE REVIEW

Child mortality remains a critical public health challenge globally and particularly in low- and middle-income countries like India. According to the World Health Organization (WHO) and UNICEF, factors such as poor maternal health, limited access to healthcare, low socioeconomic status, and inadequate birth spacing significantly contribute to under-five mortality rates in India (WHO, 2022; UNICEF, 2015).

Traditional statistical methods such as logistic regression and survival analysis have been widely used to study child mortality determinants (Saroj et al., 2018). **These approaches consistently highlight that younger maternal age, poor household wealth, low education level, short preceding birth intervals, and rural residence are associated with higher risk of child death** (Wahl et al., 2023; Dusabe, 2016).

In recent years, machine learning (ML) techniques have gained attention in public health research for their predictive accuracy and ability to capture complex non-linear interactions. Studies like Shukla et al. (2020) and **Brahma & Mukherjee (2022)** demonstrated the effectiveness of models such as Random Forest, Naïve Bayes, and Support Vector Machines in predicting perinatal and child mortality with high accuracy. For instance, **Shukla et al. achieved over 90% accuracy using ML in resource-limited settings.** However, **these studies also highlight trade-offs between model complexity and interpretability, especially in black-box models like neural networks** (Ramakrishnan et al., 2021).

A study focused on Uttar Pradesh by Saroj et al. (2018) emphasized that **birth intervals shorter than 24 months, lower maternal age, and limited antenatal care significantly increased mortality risk.** Their results align with **findings from UNICEF (2022)**, which states that **increasing birth spacing to over 36 months could reduce infant mortality by up to 51%**

Research Gap

While previous studies have shown the potential of ML models in predicting child mortality, most were limited to regional datasets or hospital-based data. Few studies have applied modern ML models to large-scale, nationally representative surveys such as the NFHS. Moreover, many of the existing works did not incorporate model interpretation techniques, which are essential for translating predictions into policy-relevant insights. This study aims to address these gaps by applying interpretable machine learning—specifically XGBoost with SHAP analysis—on the NFHS dataset to both predict child mortality and identify its most significant contributing factors.

Background and Theory

1. Dataset Balance or Imbalance

In machine learning, the balance of a dataset plays a crucial role in how well a model can learn and generalize.

- **Balanced Dataset:**
  A dataset is considered balanced when all classes are represented equally or nearly equally. This ensures that the model can learn patterns from all categories without any bias.
  Example (binary classification):
  - Class A: 500 samples
  - Class B: 500 samples
- **Imbalanced Dataset:**
  When one or more classes dominate the dataset while others are underrepresented, it is called imbalanced. This can cause the model to favor the majority class and ignore the minority class.
  Example:
  - Class A: 950 samples
  - Class B: 50 samples
- **Why It Matters:**
  - Biased Predictions: Models perform poorly on the minority class.
  - Misleading Metrics: High accuracy may hide poor performance on minority classes.
  - Real-World Impact: In critical areas (e.g., healthcare), ignoring the minority class can have serious consequences.
- **Techniques to Handle Imbalanced Data:**
  - Resampling (oversampling minority or undersampling majority)
  - Synthetic data generation **(**e.g., SMOTE)
  - Class weighting in algorithms

2. Missing Values

Missing values are the absence of data in certain fields. They are often represented as NaN (Not a Number).

- **Reasons for Missing Data:**
  - Data corruption or loss.
  - Failure to record values due to human error.
  - Intentional skipping of information.
  - Participant refusal (item nonresponse).
- **Handling Missing Values:**
  - Deletion: Removing rows or columns with missing values.
  - Imputation: Replacing missing values with estimated ones:
    - Mean/Median/Mode imputation
    - K-Nearest Neighbors (KNN) imputation
    - Model-based imputation

3. Outliers

An outlier is a data point that significantly differs from the rest of the observations.

- **Causes of Outliers:**
  - Measurement errors
  - New, unusual observations
  - Experimental errors
- **Outlier Detection and Handling:**
  - Trimming: Removing outliers from the analysis.
  - Capping: Setting boundaries to limit outlier effects.
  - Discretization (Binning): Grouping data into categories.

- **Outlier Detection Techniques:**
  - **For Normal Distributions**:

        - Data points beyond mean ± 3×(standard deviation) are considered outliers.
    - **For Skewed Distributions:**
        - Using the Inter-Quartile Range (IQR) rule: Outliers are points below Q1 − 1.5×IQR or above Q3 + 1.5×IQR.
    - **Percentile-Based Approach:**
        - Values beyond 1st or 99th percentile are flagged as outliers.
    - **Box Plot Method**:
        - Visualization tool to detect outliers based on the spread of the data.

4. Correlation

Correlation measures the strength and direction of a linear relationship between two variables.

- **Correlation Coefficient (r):**
    - Range: −1 to +1
    - Positive Correlation ($0 < r \leq 1$): Both variables move in the same direction.
    - Negative Correlation ($-1 \leq r < 0$): Variables move in opposite directions.
    - No Correlation ($r \approx 0$): No linear relationship between the variables.
- **Visualization of Correlation:**
    - Pairplot: Scatterplots for all pairs of features to detect relationships.
    - Heatmap: Displays the correlation matrix with colors indicating the strength and direction of correlations.

5. Feature Scaling

Feature Scaling is crucial to ensure that all variables contribute equally to the model's performance, especially for distance-based models like KNN and SVM.

- **Types of Scaling Techniques:**
    - **Absolute Maximum Scaling:**
      Scaling data between −1 and 1 by dividing each entry by the maximum absolute value.

$$X_{\text{scaled}} = \frac{X_i - \max(|X|)}{\max(|X|)}$$

- **Min-Max Scaling:**
  Rescales features to a fixed range, usually [0, 1].

$$X_{\text{scaled}} = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}}$$

- **Normalization:**
  Shifts and rescales data based on the mean and range.

$$X_{\text{scaled}} = \frac{X_i - X_{\text{mean}}}{X_{\max} - X_{\min}}$$

- **Standardization:**
  Centers the data to have a mean of 0 and a standard deviation of 1.

$$X_{\text{scaled}} = \frac{X_i - X_{\text{mean}}}{\sigma}$$

Where σ is the standard deviation.

- Method Used in Project:
  In this project, Standardization was applied to scale numeric variables.

6. Encoding of Categorical Variables:

Most machine learning algorithms require numerical input. Encoding transforms categorical variables into a numerical format.

1) Types of encoders:

- **Label Encoding:**
  Label Encoding is a technique used to convert categorical variables into numerical values by assigning each unique category a specific integer. It is mainly applied to ordinal variables where the order or ranking between the categories has meaning. For example, consider the feature "Education Level" with categories: "No education," "Primary," "Secondary," and "Higher." Using Label Encoding, these categories can be mapped to numerical values such as 0, 1, 2, and 3, respectively**.**

- **One-Hot Encoding:**
  One-Hot Encoding is a method of converting categorical variables into a binary matrix format, where each category is represented by a separate column. It is commonly used for nominal variables where no intrinsic order exists between categories. For example, if the feature "Place of Residence" has categories "Urban" and "Rural," One-Hot Encoding would create two new binary columns: "Urban" and "Rural." If a sample belongs to the "Urban" category, it would be encoded as (1,0); if it belongs to "Rural,"

it would be encoded as (0,1). One-Hot Encoding generates a new feature matrix. This encoding prevents models from assuming any ordinal relationship between categories and ensures that each category is treated independently, making it very effective for algorithms sensitive to numerical magnitude or order.

6. **Model Used:**

For predicting child mortality, several machine learning classification algorithms were implemented, each offering unique strengths in handling different types of data structures and complexities. The models used are as follows:

• **Logistic Regression:**

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, logistic regression is a predictive analysis. It is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables. Logistic regression uses the logistic function (also called the sigmoid function) to map the output of the linear combination of features to a probability value between 0 and 1.

$$P(y=1|x)=\frac{1}{1+e^{-(b_0+b_1x_1+\cdots+b_nx_n)}}$$

Where;

1. P(y=1|x) is the probability of the target variable being 1 given input x.
2. e is the base of the natural logarithm.
3. b0,b1,...,bn are the coefficients of the logistic regression model.
4. x1,...,xn are the independent variables (features).

The logistic regression model is trained using optimization techniques such as maximum likelihood estimation (MLE) or gradient descent. The goal is to find the optimal coefficients that maximize the likelihood of the observed data. In binary classification, logistic regression uses a decision boundary to classify the input into one of the two classes based on the predicted probability. By default, the decision boundary is set at 0.5, but it can be adjusted depending on the specific requirements of the problem.

• **K Nearest Neighbour:**

KNN is a straightforward supervised learning algorithm used for:

- Classification (e.g., "Is this tumor malignant?")
- Regression (e.g., "Predicting house prices")

How It Works:

- Memorize the Data: Unlike other algorithms, KNN doesn't "learn" during training—it just stores the dataset (*lazy learner*).
- Predict New Points: For a new data point:
    - Calculate distances (e.g., Euclidean distance) to all stored points.
    - Pick the 'k' closest neighbors (e.g., k=5). 18

Key Characteristics:

- No Training Phase: Works directly with data (fast training, slower predictions).
- Hyperparameter 'k': Small k = sensitive to noise; large k = smoother but less precise.
- Feature Scaling Matters: Distance-based, so normalize features (e.g., use StandardScaler).
- Curse of Dimensionality: Struggles with high-dimensional data (many features).

When to Use?

- Small to medium datasets.
- Simple baseline model.
- Interpretability matters (you can *see* why a prediction was made).

Example: If k=3 and 2 neighbors are "Malignant" while 1 is "Benign," the prediction is "Malignant."

- ## SVM:

SVM is a supervised learning model for classification (and regression) that finds the optimal hyperplane separating classes with the maximum margin.

Key Concepts:

- **Hyperplane**
    - A decision boundary (e.g., a line in 2D, plane in 3D) that separates classes.
    - Defined by: $w^T x + b = 0$

      where:
        - w = weight vector (normal to the hyperplane),
        - b = bias term.
- **Support Vectors**
    - The closest data points to the hyperplane (they "support" the margin).
- **Margin**
    - The distance between the hyperplane and the nearest data points of either class.
    - SVM maximizes this margin for better generalization.

- **Random Forest Classifier:**
  Random Forest is an ensemble learning method that builds multiple decision trees and merges their results to obtain a more accurate and stable prediction. It reduces overfitting and improves generalization by using techniques like bootstrapping and feature randomness. It is highly effective for handling datasets with complex relationships and missing values.

- **Gradient Boosting Classifier:**
  Gradient Boosting is an ensemble technique that builds models sequentially, where each new model tries to correct the errors made by the previous ones. It optimizes a loss function using gradient descent and produces strong predictive models by combining many weak learners (typically decision trees). It is powerful but can be prone to overfitting if not properly tuned.

- **XGBoost Classifier:**
  XGBoost (Extreme Gradient Boosting) is an advanced implementation of gradient boosting algorithms designed for speed and performance. It includes regularization techniques (L1 and L2) to avoid overfitting and supports parallel processing, making it highly efficient for large datasets. XGBoost is widely used in machine learning competitions due to its high predictive accuracy.

7. Evaluation Metrics:

To evaluate classification models, multiple metrics are used:

- **Accuracy**

Accuracy is the most common metric to be used in everyday talk. Accuracy answers the question **“Out of all the predictions we made, how many were true?”**

$$accuracy = \frac{true\ positives + true\ negatives}{true\ positives + true\ negatives + false\ negatives + false\ positives}$$

- **Recall**

Recall focuses on how good the model is at finding all the positives. Recall is also called the true positive rate and answers the question, "**Out of all the data points that should be predicted as true, how many did we correctly predict as true?**"

$$recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

- **F1 Score**

F1 Score is a measure that combines recall and precision. As we have seen, there is a trade-off between precision and recall; F1 can therefore be used to measure how effectively our models make that trade-off.

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

- **AUC-ROC Curve:**

  The AUC-ROC curve is a popular metric for evaluating the performance of a classification model, especially in imbalanced datasets.

  - **ROC (Receiver Operating Characteristic) Curve** is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

    It is created by plotting:

    - **True Positive Rate (Recall)** on the Y-axis against
    - **False Positive Rate (FPR)** on the X-axis at various threshold settings.

Mathematical Formulas:

- **True Positive Rate (TPR) / Recall:**

$$\text{TPR} = \frac{TP}{TP + FN}$$

- **False Positive Rate (FPR):**

$\text{FPR} = \frac{FP}{FP+TN}$

Where:

- TP = True Positive
- FP = False Positive
- TN = True Negative
- FN = False Negative

- **AUC (Area Under the Curve):**
- The **AUC** is the area under the ROC curve.
- AUC provides an aggregate measure of the model's ability to distinguish between the positive and negative classes across all possible classification thresholds.
- The value of AUC lies between 0 and 1:
  - **AUC = 1** → Perfect model.
  - **AUC = 0.5** → Model performs no better than random guessing.
  - **AUC < 0.5** → Model performs worse than random guessing.

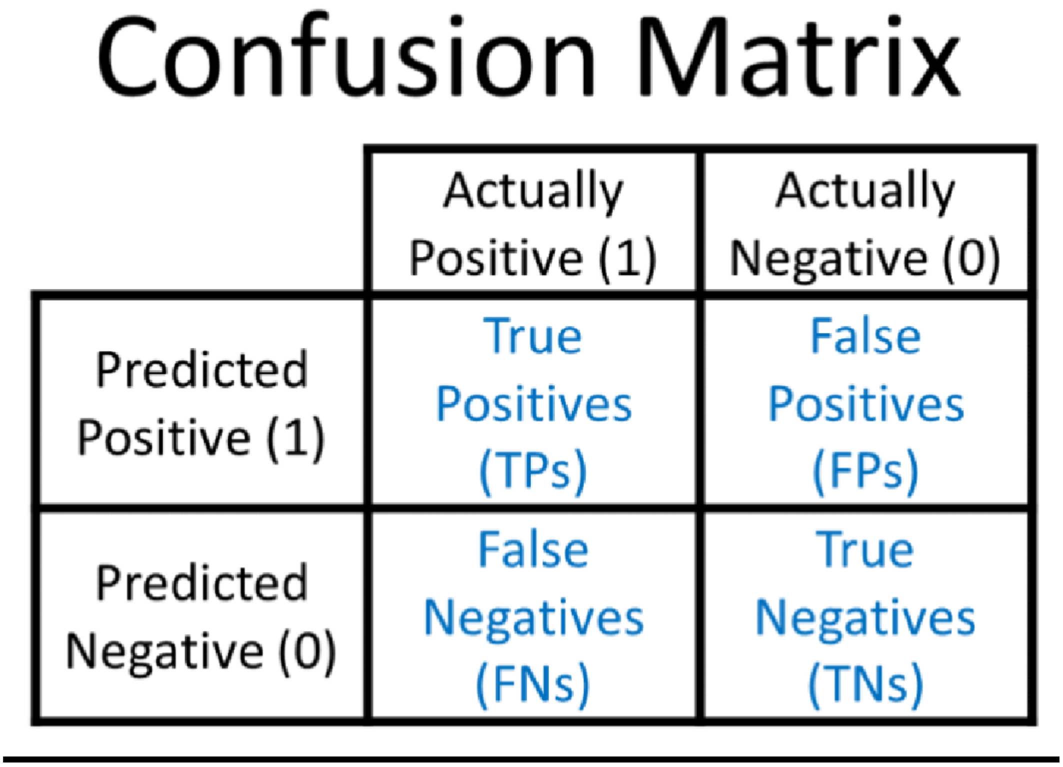
Interpretation:

- A higher AUC value indicates a better-performing model.
- A model with an AUC closer to 1 has a high ability to distinguish between the positive and negative classes.
- ROC curves and AUC scores are particularly useful when the classes are imbalanced because they focus on the model's performance over a range of thresholds instead of relying on a single fixed threshold.

- **Confusion Matrix:**
  A confusion matrix is a table used to evaluate the performance of a classification model by comparing the actual and predicted classes. It provides a comprehensive breakdown of how many predictions were correct and where errors occurred. It helps not only to calculate the accuracy but also other important evaluation metrics like precision, recall,

and F1-score.
For binary classification problems, the confusion matrix consists of four values:

- **True Positive (TP):** Cases where the model correctly predicted the positive class.
- **True Negative (TN):** Cases where the model correctly predicted the negative class.
- **False Positive (FP):** Cases where the model incorrectly predicted the positive class (Type I error).
- **False Negative (FN):** Cases where the model incorrectly predicted the negative class (Type II error).



Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

DATA DESCRIPTION

This section provides an overview of the variables used for the classification task aimed at predicting child mortality.

The dataset for this project is sourced from the National Family Health Survey (NFHS), which is part of the Demographic and Health Surveys (DHS) Program. The NFHS collects detailed information on health, population, and nutrition indicators across India. The data can be accessed through the official DHS Program website: https://dhsprogram.com.

Features:

- **Mother's Age:** Age of the respondent mother at the time of the interview (in years).
- **Total Children Ever Born:** Total number of children the mother has given birth to.
- **Preceding Birth Interval:** Time (in months) between the birth of the deceased child and the previous birth.
- **Birth Order**: Order of birth of the child (1st, 2nd, etc.).
- **Child's Sex:** Sex of the child (Male or Female).
- **Child Age at Interview**: Age of the child (in months) at the time of interview.
- **Education Level:** Highest level of education attained by the respondent (No education, Primary, Secondary, Higher).
- **Wealth Index**: Economic status of the household categorized as Poorest, Poorer, Middle, Richer, and Richest.
- **Place of Residence:** Urban or Rural residence classification.
- **Religion:** Religious affiliation of the respondent (Hindu, Muslim, Christian, Sikh, Buddhist, Jain, Jewish, Parsi, No Religion, Other).
- **Caste:** Social classification (Caste, Tribe, No caste, Don't Know).
- **Contraceptive Method Used:** Type of contraceptive method used at the time of the survey (No method, Folkloric, Traditional, Modern).
- **Marital Status**: Marital status of the respondent (Never in union, Currently in union/living with a man, Formerly in union/living with a man).
- **Mother's Date of Birth (Century Month Code):** Used to calculate mother's age.
- **Interview Date (Century Month Code):** Used to calculate child's or mother's age at interview.
- **Child's Date of Birth (Century Month Code):** Used to derive age-related variables.
- **Age at Death:** Age (in days/months/years) at which the child died (used for filtering or dropped).

Target Variable

- **Child Mortality**: Binary outcome variable indicating whether the respondent has experienced the death of a child (Yes = 1, No = 0).

METHODOLOGY

The objective of the project is to develop classification models for the prediction of the risk of child mortality through socio-demographic and health-related variables and model development. The methodology used for the project is as follows:

Data Understanding:

The dataset was carefully analysed to obtain the distribution of variables, identify types of features (categorical or numerical), and generate basic summary statistics.

Missing Value Imputation:

Missing values were handled appropriately. Specifically, missing values in key variables (such as **age_at_death, age_at_death_months**, and **preceding_birth_interval**) were imputed with 0, particularly where they corresponded to the first child. Irrelevant variables were removed as necessary.

Feature Selection:

Relevant socio-demographic variables (education, wealth index, place of residence, religion, caste, method of contraception, and marital status) were selected as predictors using domain knowledge and their relevance to child mortality.

Data Splitting:

To objectively assess model performance, the dataset was divided into training and testing sets (80% training, 20% testing). The training set was used for model fitting, and the testing set was reserved for final evaluation.

Handling Class Imbalance:

Recognizing that child mortality events are relatively rare compared to survival events, the training set was balanced using the Synthetic Minority Over-sampling Technique (SMOTE) to prevent model bias towards the majority class.

Model Building:

Several classification algorithms were implemented to predict child mortality, including:

- Logistic Regression
- K-Nearest Neighbors (KNN)
- Random Forest Classifier
- Gradient Boosting Classifier
- XGBoost Classifier
- Support Vector Machine (SVM)

Model Evaluation:

The performance of the models was evaluated using various metrics:

- Accuracy
- Precision
- Recall
- F1-Score
- ROC-AUC Score

Feature Importance Analysis:

For tree-based models (Random Forest, Gradient Boosting, and XGBoost), feature importance scores were calculated to determine which variables were most strongly associated with child mortality.

Model Comparison and Selection:

All models were compared based on the evaluation metrics, and the best-performing model for predicting child mortality was selected.

DATA PREPROCESSING

Data preprocessing is a critical step in preparing the dataset for analysis, as it ensures that the data is clean, structured, and suitable for the classification models. Below are the detailed steps followed for data preprocessing in this study:

1. Handling Missing Data:

Three columns in the dataset, **age_at_death**, **age_at_death_months**, and **preceding_birth_interval**, had missing values. Upon investigating the dataset, it was observed that these missing values primarily corresponded to children with a **birth order of 1** (firstborn children). Since these children do not have prior birth or death-related data, it was assumed that these missing values represented a lack of prior information. Therefore:

- The missing values in the **age_at_death** and **age_at_death_months** columns were imputed with **0**, representing the absence of data for the firstborn.
- The **preceding_birth_interval** column was not required for the analysis of child mortality and was thus **removed** from the dataset, as it would not contribute to the prediction of the target variable.

2. Encoding Ordinal and Nominal Variables:

To prepare the categorical variables for machine learning algorithms, different encoding techniques were used based on the nature of the variable:

- **Ordinal variables** (variables with an inherent order) like **education_level** and **wealth_index** were encoded using **Label Encoding**. This method assigns a unique integer to each category based on the order, making it suitable for models that take into account the ordinal relationship between categories.
- **Nominal variables** (variables without any intrinsic order) such as **place_of_residence**, **religion**, **caste**, **contraceptive_method**, **marital_status**, and **child_sex** were encoded using **One-Hot Encoding**. This method creates binary columns for each category, ensuring that all categorical data is transformed into a format compatible with machine learning algorithms.

3. Target Variable Transformation:

In the original dataset, the target variable for child mortality was encoded such that **1 represented "alive"** and **0 represented "death"**. Since the focus of this analysis is to predict child mortality, the target variable was **reversed** to align with the research goal:

- **1 now represents death** (child mortality event),
- **0 represents alive** (no child mortality).

This transformation ensures that the target variable is meaningful for the classification task, where the model predicts whether a child is alive or has died.

4. Feature Scaling:

Many machine learning algorithms, particularly those that rely on distance metrics (e.g., KNN, SVM), require optimal feature scaling. To address this, **feature scaling** was applied to the numerical features in the dataset. This was done by **standardizing** the data (i.e., transforming it to have a **mean of 0** and a **standard deviation of 1**). Standardization helps ensure that no single feature dominates the learning process due to its larger scale, allowing the models to learn more effectively.

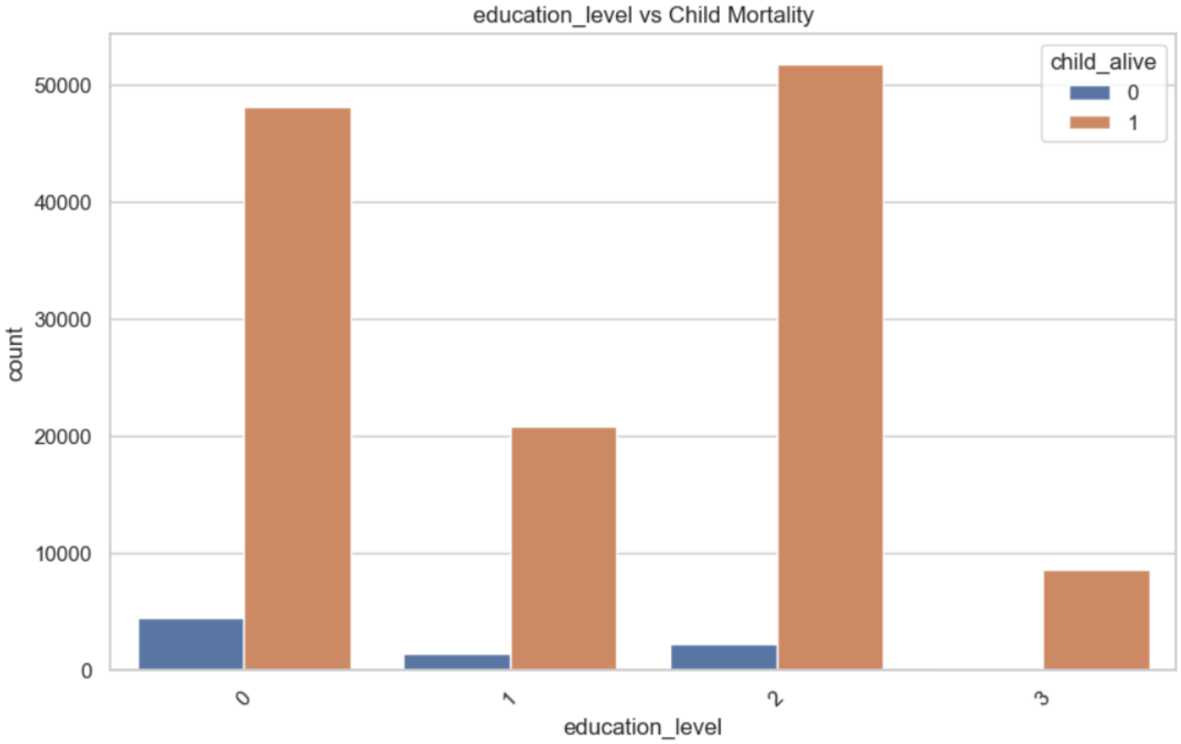
5. Data Splitting:

The dataset was divided into **80% training** and **20% testing**. The training set was used to train the machine learning models, while the testing set served as an independent dataset for model evaluation. This splitting ensures that the model is tested on data it has not seen during training, providing a realistic estimate of its performance on new, unseen data.

EXPLANATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) was conducted to understand the underlying structure of the data, identify patterns, detect anomalies, and explore relationships between variables, particularly focusing on factors associated with child mortality.

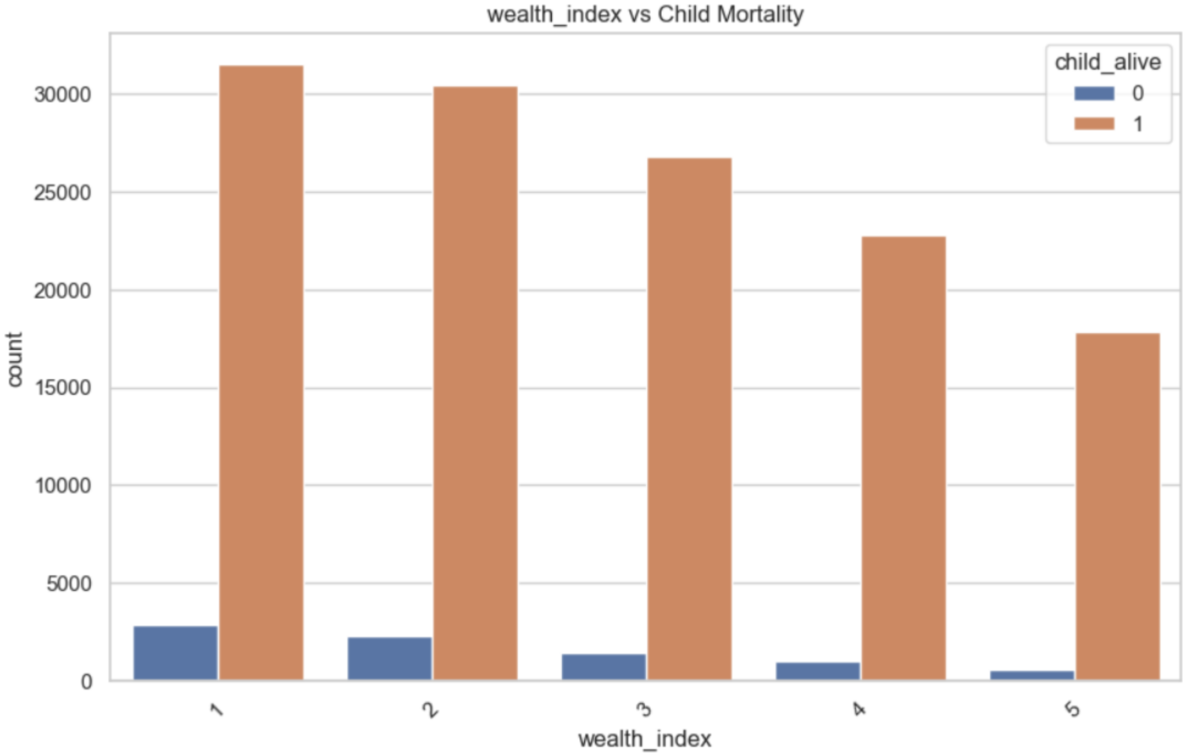
Graph 1: **Education Level vs Child Mortality**



Interpretation:

The graph for education level indicates that child mortality is significantly higher among mothers with no education or only primary education compared to those with higher education levels. As education increases, particularly towards secondary and higher education, the proportion of child deaths decreases. This trend suggests that maternal education plays a critical role in child health outcomes, likely by improving knowledge about healthcare practices, nutrition, and access to medical services.

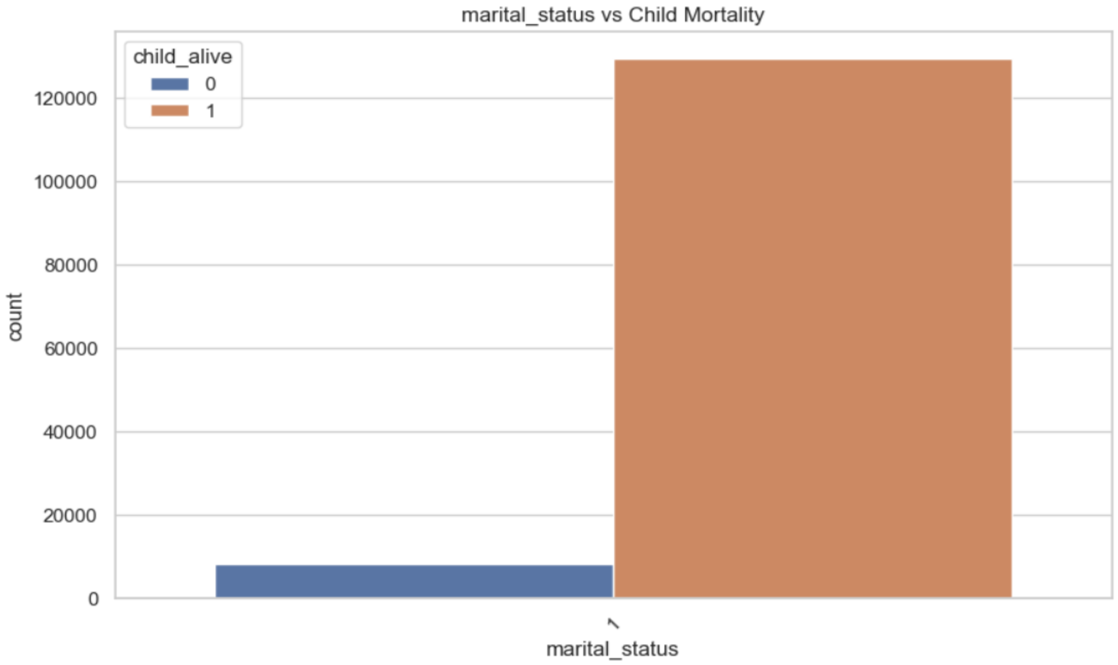
Graph 2: **Wealth Index vs Child Mortality**



Interpretation:

The wealth index plot shows a clear inverse relationship between household wealth and child mortality. Children from the poorest households experience the highest mortality rates, whereas those from the richest households show a much lower risk. This pattern highlights the importance of socioeconomic factors in child survival, as wealthier families are likely to have better access to healthcare, sanitation, and overall living conditions that support child health.

Graph 3: **Marital Status vs Child Mortality**

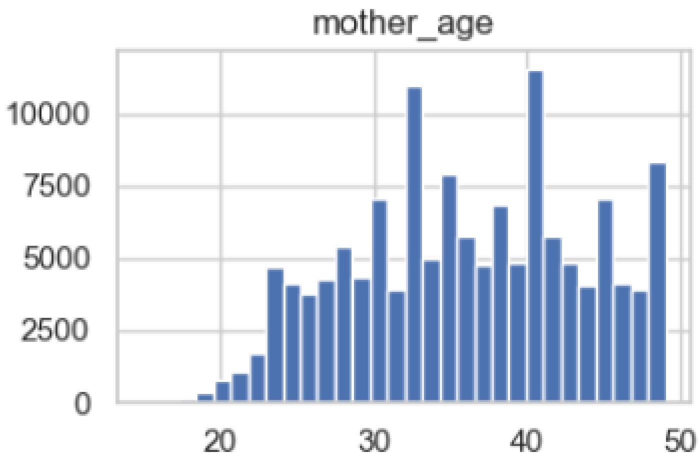


Interpretation:

The graph illustrating marital status reveals that the majority of mothers are currently in a union (married or living with a partner), and within this group, most children are alive. Although some mortality exists across all categories, stable marital relationships seem to be associated with lower child mortality, possibly because of better emotional, financial, and caregiving support that benefits both mothers and their children.

Histogram for some numerical features:

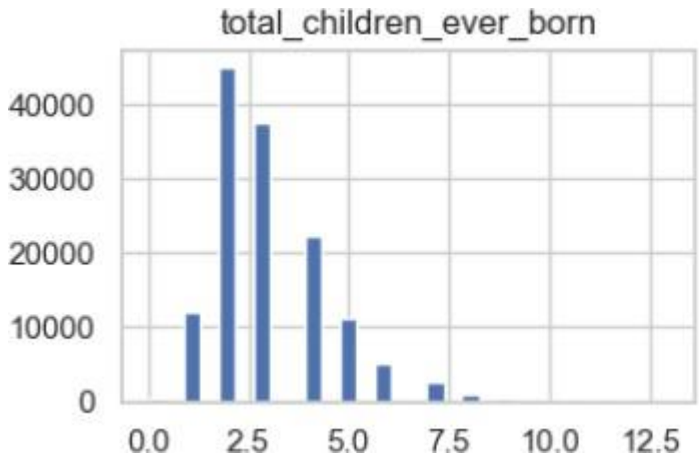
Graph 4: **Mother's Age**



Interpretation:

The histogram of mothers' age shows that most mothers in the dataset gave birth between the ages of 25 and 45, with notable peaks around 30 and 40 years. Very young mothers (below 20) and older mothers (above 45) are comparatively fewer. The distribution is slightly right-skewed, indicating that higher maternal ages are less frequent. This age pattern provides useful insights into the reproductive behavior within the population studied.

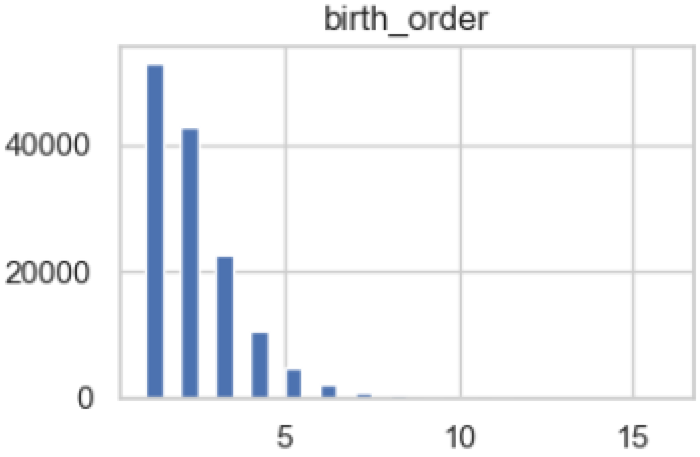
Graph 5: **Total Children Ever Born**



Interpretation:

The histogram for total children ever born indicates that most mothers have between 2 to 4 children, with a sharp peak at 2 children. The frequency steadily declines as the number of children increases, and very few mothers have more than 6 children. The distribution is strongly right-skewed, suggesting that having a large number of children is relatively rare in this dataset. This pattern reflects a trend toward smaller family sizes.

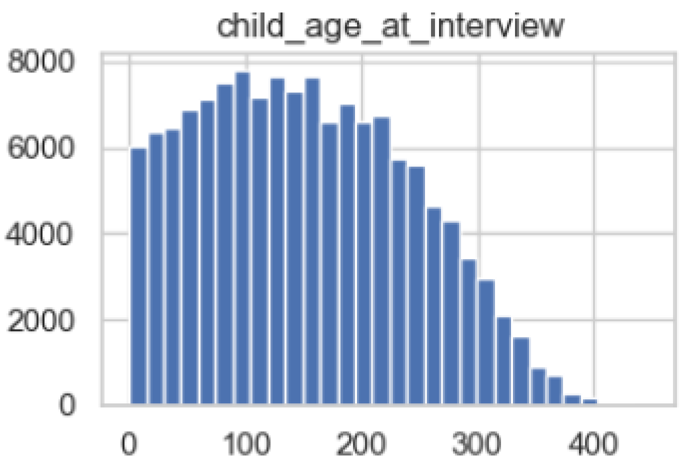
Graph 5: **Birth Order**



Interpretation:

The histogram for birth order shows that most children are first or second born, with the frequency sharply declining as the birth order increases. Very few observations are recorded for birth orders higher than five. The distribution is heavily right-skewed, indicating that higher birth orders are relatively rare in the sample. This trend reflects a preference for smaller family sizes in the population studied.

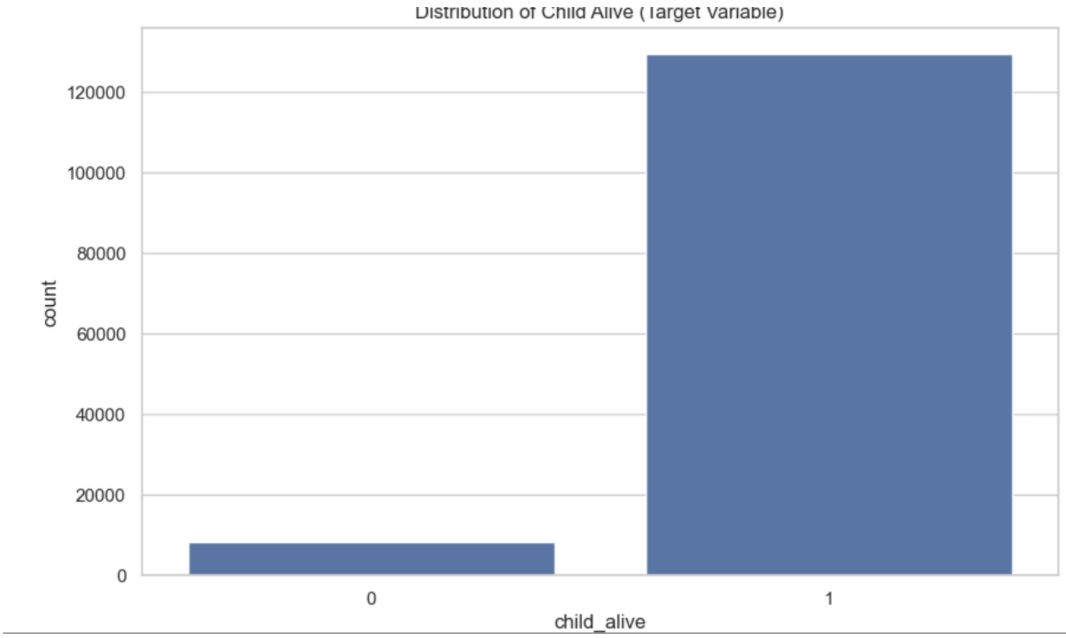
Graph 6: **Child Age at Interview**



Interpretation:

The histogram of **child_age_at_interview** shows a fairly uniform distribution up to around 150 months (about 12 years), after which the frequency gradually declines. The largest number of children are between 50 to 150 months old, suggesting a concentration of children aged roughly 4 to 12 years at the time of the survey. After 150 months, the number of observations steadily drops, with very few children above 300 months (about 25 years), likely representing older adolescents or young adults still captured in the household. This distribution indicates that the sample mainly consists of younger children and early adolescents, which aligns well with typical child health and mortality analyses focused on early life stages.

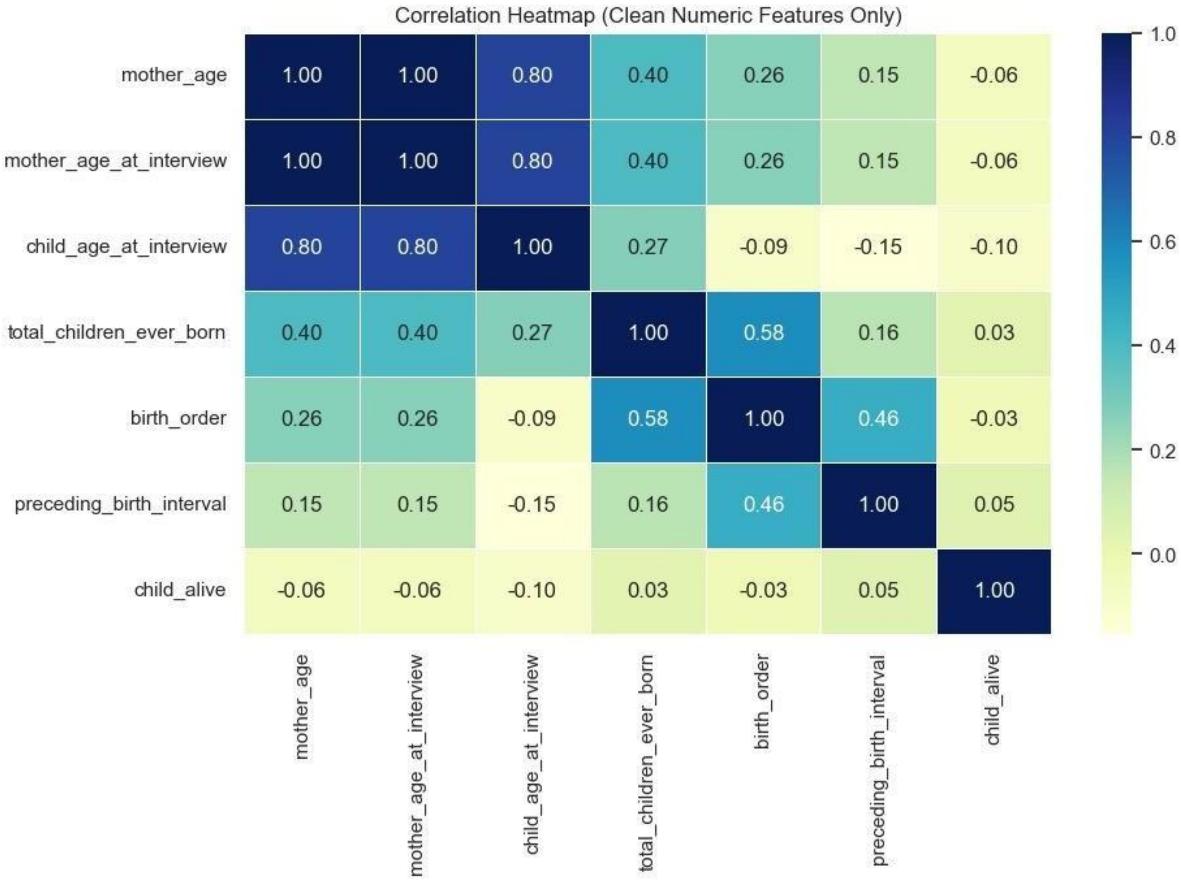
Graph 7: **Distribution of Child Alive**



Interpretation:

The bar plot of the **child_alive** target variable shows a significant class imbalance in the dataset. A vast majority of the children (coded as 1) are alive at the time of the survey, while a much smaller proportion (coded as 0) are deceased. This distribution is expected in population data, where child mortality is relatively rare compared to survival. However, this imbalance should be carefully addressed during modeling, as it can affect the performance of classification algorithms by biasing predictions towards the majority class.

Graph 7: **Correlation Matrix**



Interpretation:

The correlation heatmap displays the relationships between the clean numeric features. Strong positive correlations are observed between mother_age, mother_age_at_interview, and child_age_at_interview, indicating that these age-related variables are closely linked, as expected. Total_children_ever_born also shows a moderate positive correlation with birth_order. However, the target variable child_alive exhibits very weak correlations with all features, suggesting that child mortality is influenced by more complex interactions beyond simple linear relationships with individual variables. Overall, multicollinearity among features appears manageable except among age variables, which may require attention in modeling.

RESULT OF MODEL PERFORMANCE

1. Logistic Regression:

Table 1: Confusion matrix of logistic Regression

Actual	Predicted	
	0	1
0	17536	8352
1	629	1014

- **True Positive (TP)** = 1014
- **True Negative (TN)** = 17536
- **False Positive (FP)** = 8352
- **False Negative (FN)** = 629
- **Model Accuracy** = 67.38%
- **ROC-AUC Score** = 69.03%

2. KNN:

Table 2: Confusion matrix of KNN

Actual	Predicted	
	0	1
0	22230	3658
1	1129	514

- **True Positive (TP)** = 514
- **True Negative (TN)** = 22230
- **False Positive (FP)** = 3658
- **False Negative (FN)** = 1129

- **Model Accuracy** = 82.61%
- **ROC-AUC Score** = 61.68%

3. SVM:

Table 3: Confusion matrix of SVM

Actual	Predicted	
	0	1
0	17580	8308
1	629	1014

- **True Positive (TP)** = 1014
- **True Negative (TN)** = 17580
- **False Positive (FP)** = 8308
- **False Negative (FN)** = 629
- **Model Accuracy** = 67.54%
- **ROC-AUC Score** = 69.36%

4. Random Forest:

Table 4: Confusion matrix of Random Forest

Actual	Predicted	
	0	1
0	24756	1132
1	1285	358

- **True Positive (TP)** = 358
- **True Negative (TN)** = 24756
- **False Positive (FP)** = 1132

- **False Negative (FN)** = 1285
- **Model Accuracy** = 91.22%
- **ROC-AUC Score** = 72.82%

5. Gradient Boost:

Table 4: Confusion matrix of Random Forest

Actual	Predicted	
	0	1
0	22130	3758
1	873	770

- **True Positive (TP)** = 770
- **True Negative (TN)** = 22130
- **False Positive (FP)** = 3758
- **False Negative (FN)** = 873
- **Model Accuracy** = 83.18%
- **ROC-AUC Score** = 76.15%

6. XGBoost:

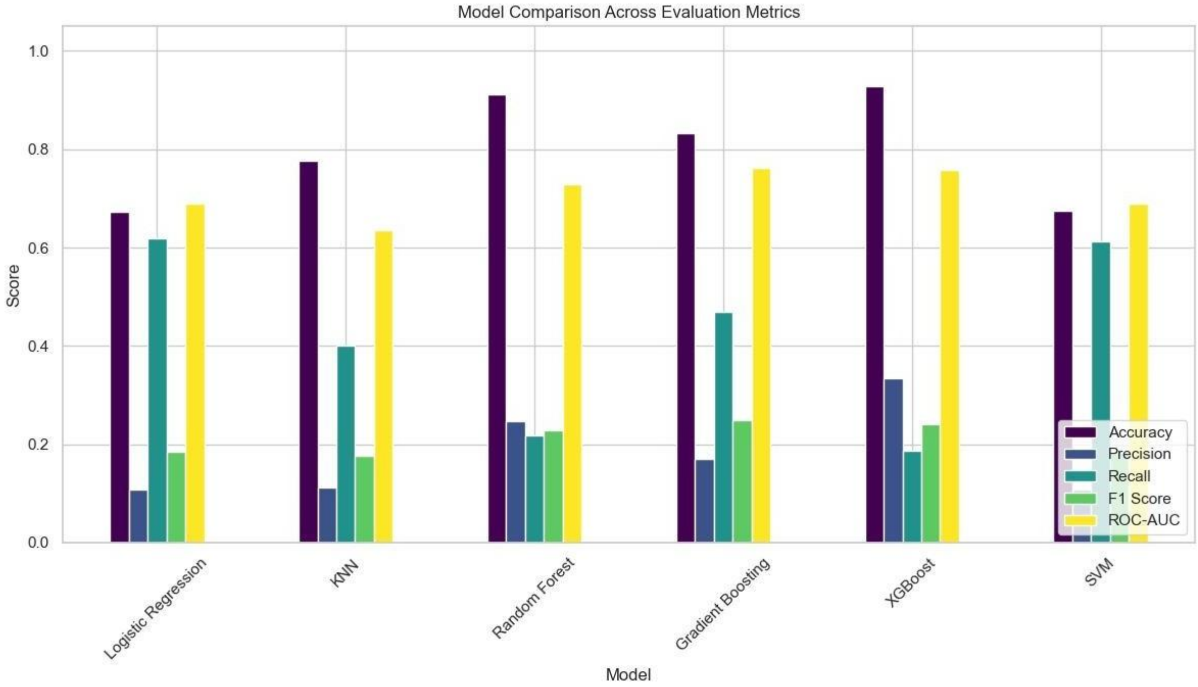
Table 4: Confusion matrix of Random Forest

Actual	Predicted	
	0	1
0	25277	611
1	1335	308

- **True Positive (TP)** = 308

- **True Negative (TN)** = 25277
- **False Positive (FP)** = 611
- **False Negative (FN)** = 1335
- **Model Accuracy** = 92.93%
- **ROC-AUC Score** = 75.75%

Graph 1: A comparison between Evaluation Metrics of various models



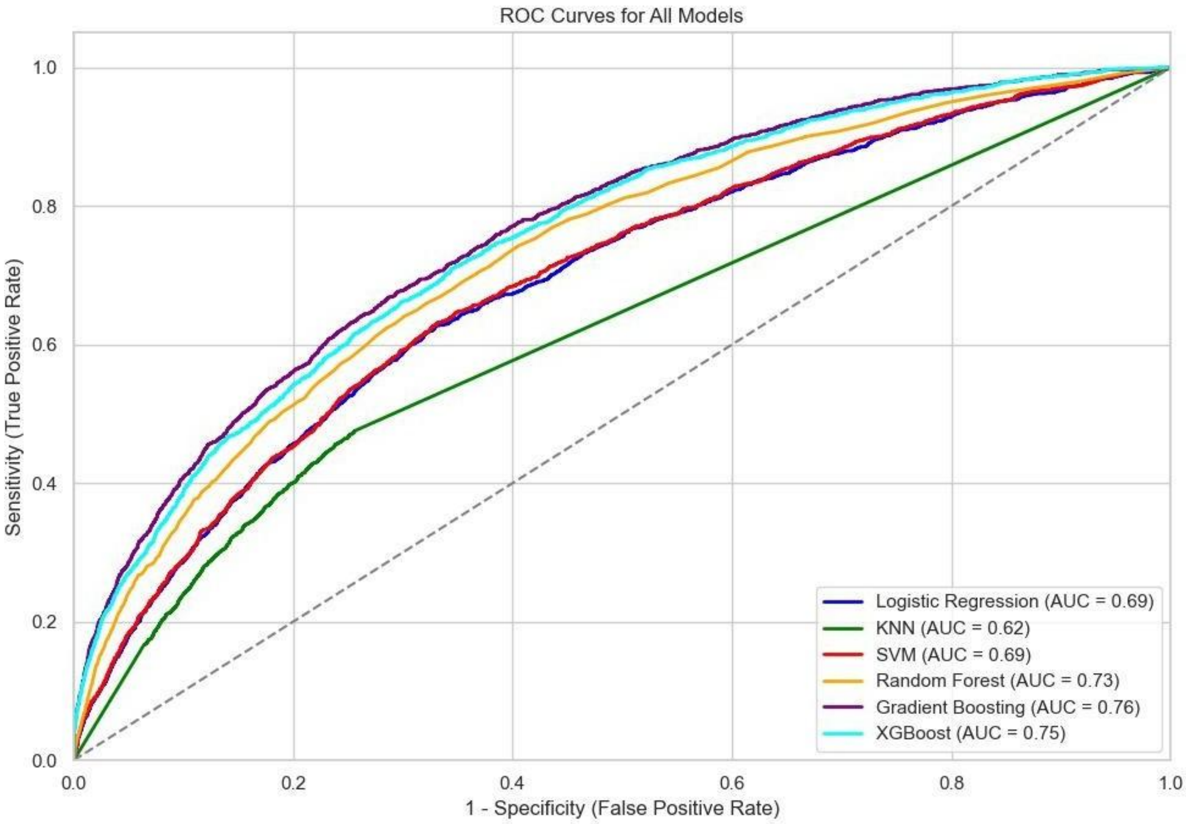
Interpretation:

This bar chart visually compares six machine learning models—Logistic Regression, KNN, Random Forest, Gradient Boosting, XGBoost, and SVM—across five key evaluation metrics: Accuracy, Precision, Recall, F1 Score, and ROC-AUC.

- XGBoost and Random Forest show the highest overall accuracy and ROC-AUC, suggesting strong general performance.
- Logistic Regression and SVM offer balanced performance across all metrics.
- KNN lags in recall and F1 score, indicating weaker classification on the positive class.
- Gradient Boosting performs consistently well, especially in ROC-AUC.

This comparative plot helps select the best model based on whether precision, recall, or overall discrimination (ROC-AUC) is the primary concern**.**

Graph 2: ROC Curves for All Models



Interpretation:

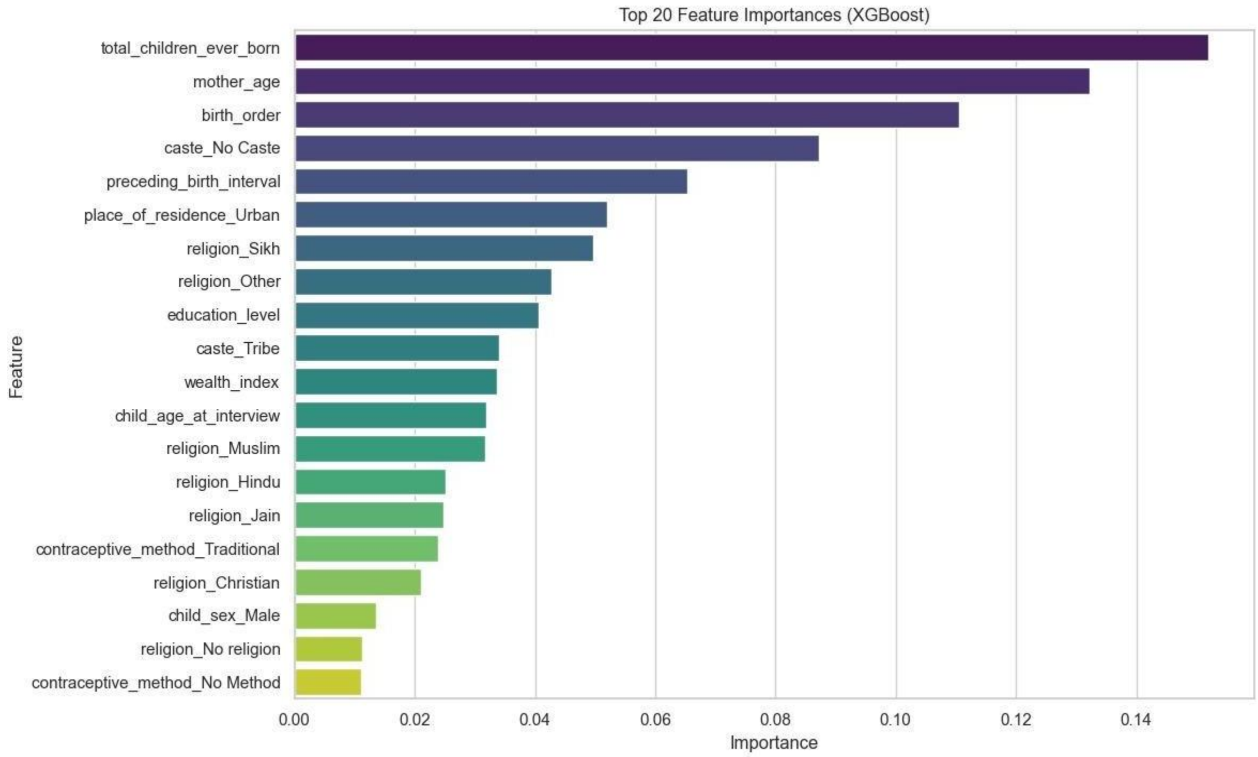
This ROC curve comparison illustrates the classification performance of six different machine learning models:

- XGBoost (AUC = 0.75) and Gradient Boosting (AUC = 0.76) perform the best in terms of discriminating between classes.
- Random Forest also performs well with an AUC of 0.73.
- Logistic Regression and SVM yield moderate performance (AUC = 0.69 each).
- KNN performs the worst with an AUC of 0.62, indicating weaker ability to distinguish between the classes.

The ROC curve plots True Positive Rate vs. False Positive Rate, where curves closer to the top-left indicate better model performance. The diagonal dashed line represents random chance (AUC = 0.5). This chart helps visually compare how well each model separates positive and negative outcomes.

FEATURE IMPORTANCE

The bar plot below illustrates the top 20 features ranked by their importance in the XGBoost model based on gain, reflecting each feature's contribution to improving model performance.



The most important features were:

- **total_children_ever_born**: This feature had the highest importance, indicating that higher parity is strongly associated with increased risk of child mortality.
- **mother_age**: Younger maternal age was a significant predictor of mortality, consistent with existing demographic literature.
- **birth_order and preceding_birth_interval:** These fertility-related indicators further highlighted the importance of birth spacing and order in child survival outcomes.
- **caste_No Caste and place_of_residence_Urban:** These social and geographic features played substantial roles, suggesting that both caste affiliation and urban/rural context affect mortality risk.
- **education_level and wealth_index**: These socioeconomic variables reinforced the link between higher education, economic advantage, and improved child health outcomes.

Religious identity, sex of the child, and contraceptive method also showed modest but non-negligible contributions, suggesting that cultural and behavioural factors.

CONCLUSION

In this study, we investigated the application of statistical and machine learning methods for binary classification of child mortality using the National Family Health Survey (NFHS) dataset. The analysis incorporated a range of categorical and continuous predictors related to maternal characteristics, household demographics, and child-level attributes.

Extensive data preprocessing was conducted, including missing value treatment, feature engineering, label encoding, and standardization. To address significant class imbalance in the response variable, Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training set.

A comparative modeling approach was undertaken using several classification algorithms — namely, Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, Gradient Boosting, and Extreme Gradient Boosting (XGBoost). Performance was evaluated using multiple metrics: accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). Among all models, XGBoost exhibited superior predictive ability, yielding the highest AUC and F1-score, and was therefore selected as the final model.

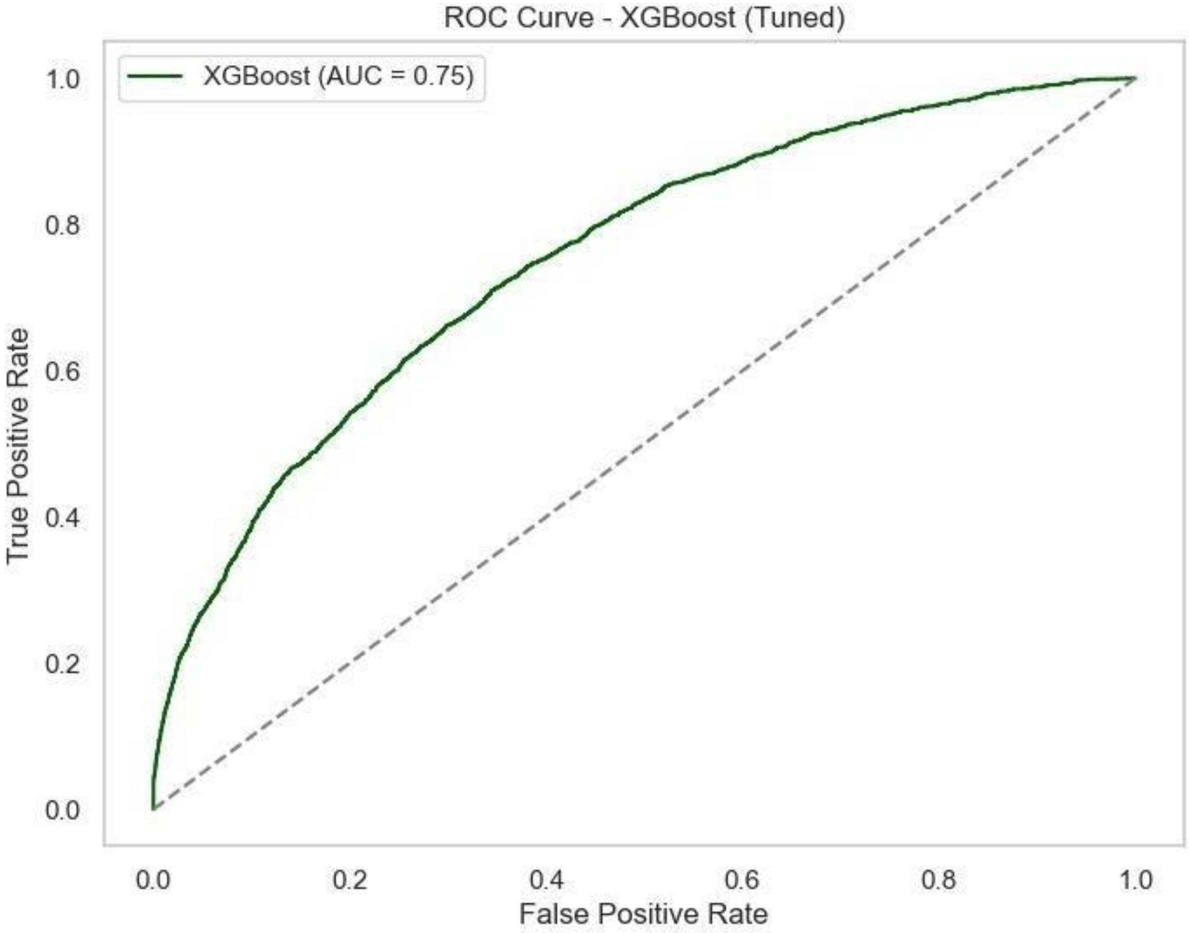
To interpret the influence of individual predictors on model outcomes, SHAP (SHapley Additive exPlanations) values were computed. These revealed that variables such as maternal age, total children ever born, preceding birth interval, education level, and household wealth index had the most substantial marginal impact on the probability of child mortality.

The results underscore the potential of supervised learning models in extracting statistically significant and policy-relevant patterns from large-scale demographic data. While the model demonstrates strong discriminative power, future extensions may include the incorporation of spatial features, temporal components, or ensemble stacking for improved generalization.

DISCUSSION

This study examined the determinants of child mortality in India using the NFHS dataset and applied several machine learning algorithms, including **Logistic Regression, KNN, SVM, Random Forest, Gradient Boosting, and XGBoost**. To address the **severe class imbalance (~6% mortality cases),** we applied SMOTE, which improved the learning process and enhanced performance metrics for the minority class.

XGBoost emerged as the best-performing model with **an accuracy of 93.12%, and a ROC-AUC of 0.7496,** outperforming traditional models in both recall and overall discriminative power. While Logistic Regression achieved the highest precision, it failed to capture enough true positives (i.e., mortality cases), underscoring the need for more flexible and robust methods in imbalanced classification settings.



Our results are consistent with prior literature. Previous studies have emphasized the importance of fertility-related variables, such as mother's age, birth spacing, and number of children, as strong predictors of under-five mortality. We observed the same, with SHAP and model-based importance rankings highlighting mother's age, total children ever born, preceding birth interval, and wealth index as dominant features.

Studies have also noted that **birth intervals shorter than 24 months are associated with a significantly higher mortality risk due to maternal depletion and sibling competition**—findings which were reflected in our own model outputs.

Additionally, similar to literature reports, we found that proper preprocessing (e.g., encoding, scaling, outlier treatment) and class imbalance correction (e.g., SMOTE) were essential in obtaining robust model performance. As emphasized in public health research, evaluation metrics beyond accuracy, such as recall and F1-score, are critical. Our model selection prioritized these metrics, reflecting the serious consequences of false negatives in child mortality prediction.

However, challenges remain. Like prior studies, we observed that high predictive accuracy on a survey dataset may not directly translate to field-level application without further validation. Moreover, complex models such as XGBoost, though powerful, require interpretability tools like SHAP to translate results into policy-relevant insights.

In conclusion, this project not only confirms findings from earlier child mortality studies but also demonstrates the added value of ensemble machine learning and interpretable AI in extracting actionable insights from large-scale survey data. With further refinement, such as incorporating antenatal care, geographic indicators, and longitudinal data, this framework can aid in identifying vulnerable populations and informing targeted interventions.

REFERENCES

1. World Health Organization. (n.d.). SDG Target 3.2: Newborn and child mortality. Retrieved from https://www.who.int/data/gho/data/themes/topics/sdg-target-3_2-newborn-and-child-mortality
2. UNICEF. (2015). UNICEF Annual Report 2015. New York. Retrieved from https://www.unicef.org/reports/unicef-annual-report-2015
3. Shukla, V. V., Eggleston, B., Ambalavanan, N., et al. (2020). Predictive modeling for perinatal mortality in resource-limited settings. JAMA Network Open, 3(11), e2026750. https://doi.org/10.1001/jamanetworkopen.2020.26750
4. UNICEF. (n.d.). Child mortality. Retrieved from https://data.unicef.org/topic/child-survival/under-five-mortality
5. World Bank. (n.d.). Mortality rate, under-5 (per 1,000 live births). Retrieved from https://data.worldbank.org/indicator/SH.DYN.MORT
6. Press Information Bureau, Government of India. (n.d.). Retrieved from https://pib.gov.in/PressReleasePage.aspx?PRID=1861710
7. Wahl, B., Nama, N., Pandey, R. R., et al. (2023). Neonatal, infant, and child mortality in India: progress and future directions. Indian Journal of Pediatrics, 90(S1), 1–9. https://doi.org/10.1007/s12098-023-04834-z
8. Saroj, K. R., Murty, K. N., & Kumar, M. (2018). Survival analysis for under-five child mortality in Uttar Pradesh. International Journal of Research and Analytical Reviews, 5(3).
9. Ramakrishnan, R., Rao, S., & He, J-R. (2021). Perinatal health predictors using artificial intelligence: a review. Women's Health, 17. https://doi.org/10.1177/17455065211046132
10. Sheikhtaheri, A., Zarkesh, M. R., Moradi, R., & Kermani, F. (2021). Prediction of neonatal deaths in NICUs: development and validation of machine learning models. BMC Medical Informatics and Decision Making, 21(1), 131. https://doi.org/10.1186/s12911-021-01497-8
11. Ali, G., Mijwil, M. M., Adamopoulos, I., Buruga, B. A., Gök, M., & Sallam, M. (2024). Harnessing the potential of artificial intelligence in managing viral hepatitis. Journal of Big Data, 2024, 128–163. https://doi.org/10.58496/MJBD/2024/010
12. Brahma, D., & Mukherjee, D. (2022). Early warning signs: targeting neonatal and infant mortality using machine learning. Applied Economics, 54(1), 57–74. https://doi.org/10.1080/00036846.2021.1958141
13. Dusabe, J. (2016). Determinants of Under-5 Mortality in Rwanda (Master's thesis, University of Nairobi, School of Economics).