Predictive Modelling and Model Evaluation of Child Mortality Risk Among Teen Mothers in Western Kenya

[1st] Wong Hui San
Bachelor of Information Systems (Data Analytics)
Sunway University
Selangor, Malaysia
22034540@imail.sunway.edu.my

[2nd] Ayu Wen Li
Bachelor of Information Systems (Data Analytics)
Sunway University
Selangor, Malaysia
22017867@imail.sunway.edu.my

[3rd] Keertana a/p Subramaniam
Bachelor of Information Systems (Data Analytics)
Sunway University
Selangor, Malaysia
23109614@imail.sunway.edu.my

[4th] Siow Qi Yung
Bachelor of Information Systems (Data Analytics)
Sunway University
Selangor, Malaysia
22053037@imail.sunway.edu.my

Abstract*— Child mortality among teenage mothers remains a pressing public health issue, particularly in developing regions like Western Kenya. Predictive modelling offers a powerful approach to identify at-risk cases early and guide targeted interventions to improve child survival outcomes. Despite existing data on child mortality, there is limited use of predictive methods to identify children at risk of death among teenage mothers in Western Kenya. This study aims to develop predictive models to estimate child mortality risk based on socio-economic and health-related factors, supporting early intervention for high-risk cases. A dataset from Western Kenya containing demographic and psychological variables related to teenage mothers is used. Machine learning algorithms including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Naive Bayes and XGBoost, were applied to build predictive models, and their performance were evaluated using metrics such as accuracy, precision, recall, F1-score, and confusion matrix. The results will be visualized using bar charts and tables. Logistic Regression showed the most balanced performance in predicting child mortality among children born to teenage mothers, with a recall of 66.80% and F1-Score of 67.30%. In contrast, models like Naive Bayes and Random Forest achieved high accuracy but performed poorly in identifying actual deaths. SVM and XGBoost offered moderate recall but low precision. These findings highlight the importance of recall and F1-Score over accuracy in imbalance, high-risk health data. They support the use of balanced models for early intervention and suggest future research into advanced imbalance-handling techniques.

Keywords—Child Mortality Prediction, Under-Five-Mortality, Teen Mothers, Machine Learning Models

I. Introduction

A. Background

Child mortality is commonly measured by the under-5 mortality rate, which represents the likelihood that a child will die before reaching the age of five. [1] explains this rate reflects the number of deaths occurring between a child's first and fifth birthday, compared to the average number of children who are alive during that same timeframe. While [2] defines teenage mothers as females aged 15 to 19 who have either given birth to at least one child or are currently pregnant. This topic is important because child mortality is a serious global health issue, especially in developing countries. Teenage mothers often have limited knowledge of childcare and face greater health, social, and economic challenges compared to older mothers, which can negatively impact the survival and well-being of their children. Therefore, understanding the relationship between teenage motherhood and child mortality is important to implement effective interventions and health policies.

B. Problem Statement

Child death among children of teenage mothers is a major issue that needs to be addressed in Western Kenya. The mortality of children could be due to several causes such as the lack of resources, education, maternal experience, poverty and health issues. Teenagers are usually too young and do not have enough experience in childbirth and childcare, which contributes to the higher risk of death of children under the age of five. Governments often collect data on child mortality, but it is mostly used descriptively. There is a need to shift from descriptive analysis to predictive modelling as it can help to predict further risks, identify high-risk communities and provide targeted support to reduce child mortality.

C. Objectives

The objective of this project is to build predictive modelling that identifies the likelihood of child mortality among children born to teenage mothers in Western Kenya.

The purpose is to analyze factors that may contribute to child mortality, predict the risk of child death, and provide support to early identification of high-risk cases. This can help in providing information towards more successful interventions by health authorities and organizations on reducing child mortality among vulnerable populations.

D. Scope of Study

This study focuses on developing a machine learning-based predictive model to estimate the likelihood of child mortality among teen mothers in Western Kenya. The dataset is leveraged from a publicly available dataset from the Harvard Dataverse titled "Replication Data for: Teen Mothers Report Poor Health and Economic Functioning in Western Kenya." The dataset includes variables such as marital status, education level, wealth index, food and sleep hunger, mental health indicators, and loan burden. The target variable is anychilddead, which is a binary indicator of child death. The study is limited to quantitative data within the dataset and does not include non-teen mothers or external contextual factors. The predictive modelling will use several machine learning algorithms to identify risk factors and assess the probability of child mortality. The analysis is data-driven and does not aim to establish causality but rather to enhance early identification and targeted interventions.

E. Significance of the Study

This study addresses the urgent issue of child mortality, especially among teen mothers, remains a critical public health challenge in Western Kenya. While governments and organizations routinely collect child mortality data, much of it is used for descriptive reporting rather than predictive risk assessment. This study addresses that gap by applying predictive modelling to proactively identify high-risk cases based on measurable socio-demographic and health-related factors. The results can help health authorities, NGOs, and policymakers allocate resources more effectively, implement targeted interventions, and ultimately reduce child mortality. By using data-driven methods, the project contributes to improving maternal and child health outcomes in underserved communities.

II. Literature Review

This section examines existing research on predictive modelling of child mortality among teenage mothers, providing a comprehensive view by focusing on the determinants of child mortality, the statistical and machine learning methods used, and outcomes achieved from these studies. Through critically analyzing past works, this better improves the understanding of factors influencing child mortality in the context of adolescent mothers and addressing possible gaps in the research. The literature reviewed were chosen based on their application to machine learning prediction for health outcomes, published in creditable, peer-reviewed journals and to the topic itself.

A. Review of Articles

Noori et al. [3] found that most previous studies on adolescent pregnancy have grouped all adolescent mothers into a single category and focused mainly on child survival around the time of birth, without considering age-specific differences or long-term impacts. This approach may overlook the significantly higher risks faced by very young adolescents (under 16 years old). Additionally, the role of factors such as place of residence and access to healthcare on these risks remains uncertain. This study used data from Demographic and Health Surveys (DHS) conducted from 2004 to 2018 across 46 countries in sub-Saharan Africa (SSA) and South Asia. It focused on firstborns to mothers aged 25 years or younger. To analyse the data, they used mixed-effects logistic regression to estimate five outcomes which are stillbirth, neonatal mortality, infant mortality, and under-5 mortality. The models were adjusted for key demographic variables, including maternal education and urban versus rural residence, and were further refined by incorporating maternal health-seeking behaviours to examine potential mediating effects. The main findings show that the younger the mother, the higher the risk of child mortality. The study also found that lower use of prenatal care and facility births among younger mothers partly explains their higher rates of neonatal and infant mortality. These results reveal a clear and consistent gradient of increasing child mortality associated with decreasing maternal age, observable across different regions and time periods. However, the limitation of this study is it focuses on maternal characteristics but does not include biological, clinical, and psychological factors like mental health, family support, or nutrition, which could influence child survival.

Although infants born to teenage mothers aged 15 to 19 have the highest mortality rates, Woodall et al. [4] found that previous research has not explored how these rates have changed over time based on key factors such as maternal age, racial or ethnic background, or whether the mothers lived in urban or rural areas. Therefore, they used data from the U.S. National Linked Birth and Infant Death Files (National Vital Statistics System) covering the years 1996 through 2019. The study used Joinpoint regression to analyse trends in infant mortality rates over time and determine whether changes were statistically significant across different racial, ethnic, and geographic subgroups. Additionally, Kitagawa decomposition was used to measure how much of the change in infant deaths was due to shifts in maternal age distribution and changes in age-specific mortality rates (ASMRs), which reflect the risk of infant death by age group. The findings show that overall infant mortality among infants of teenage mothers decreased by 16.7% from 1996 to 1997 through 2018 to 2019. Although most racial/ethnic and urbanization subgroups experienced declines, Black and Hispanic teens continued to have higher birth rates than white teens, with infant mortality rates highest among Black teen mothers and lowest among Hispanic teen mothers. The limitation of this study is it cannot examine how specific personal factors like a mother's health habits, access to prenatal care, or socioeconomic status might directly affect infant mortality. This limits the ability to understand the insight of personal factors.

In low-resource settings, past research focused on prevalence and risk factors, while [5] uses machine learning to accurately identify high-risk children and key predictive patterns. This study used data from the 2019 Ethiopian Demographic and Health Survey to predict under-five child mortality. Five supervised machine learning classifiers were compared, including Random Forest and the J48 decision tree. Additionally, the researchers used Weka v3.8.6 to apply if/then logical association rules to reveal meaningful patterns and relationships between risk factors and under-five child

mortality. The findings show that the Random Forest model achieved the highest accuracy followed by the J48 decision tree. It also found that attributes such as late initiation of breastfeeding, mother's lack of formal education, short birth interval, low maternal wealth, and lack of media exposure significantly increased the risk of child mortality. The limitation of this study is the dataset lacked critical variables like environmental conditions, maternal health behaviours, healthcare access, and child-specific health indicators, which could significantly improve prediction accuracy and explainability.

Bhusal and Khanal [6] tackles the persistent high under-five mortality rates, especially in developing regions, and seeks to uncover key risk factors that may guide targeted interventions and policy efforts. It is a systematic review that aims to identify and synthesize significant determinants of under-five child mortality across low- and middle-income countries. The study was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, and a systematic literature search was carried out across multiple electronic databases, including EMBASE, PubMed, Scopus and Google Scholar. Determinants were grouped into six themes which are socio-economic, maternal, child-related, healthcare utilization, community-level, and paternal factors. Among the 47 factors identified across six themes, the key factors influencing under-five mortality are mother's education, child's size at birth, maternal age at delivery, place of residence, birth interval, child's sex, type of birth (single or multiple), and birth order. These factors were consistently reported in studies from Africa, South-East Asia, the Eastern Mediterranean, and the Western Pacific regions. The study highlights the need to prioritize women's education and access to quality healthcare as central strategies to reduce child mortality. However, the study only described the results from the selected papers and did not combine the data using statistical methods like meta-analysis, which limited the depth of quantitative insights. Additionally, there was a regional imbalance, as most of the included studies (18 out of 22) were from Africa, with limited representation from other low-and-middle-income regions.

Pandey et al. [7] investigated under-five mortality remains high in Uttar Pradesh by utilizing National Family and Health Survey (NFHS-V) data, Uttar Pradesh subset, covering households with children under five during 2019-2021 and compared the performance of Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), Naive Bayes, and traditional Logistic Regression model. The study aims to identify spatial variation and assess how effectively machine learning algorithms can identify the key driving these mortality rates. Logistic Regression achieved the highest accuracy of 79.4% identifying the determinants including breastfeeding status, recent births, child's gender, birth intervals, antenatal care, water source, birth order, and maternal BMI. While the study focused on spatial and maternal-environmental determinants,it did not examine psychological, financial stressors, or teen motherhood limiting its applicability to teen-specific contexts.

Bizzego et al. [8] studied the DHS data from over 20 LMICs collected between 2015 and 2019 to identify key predictors using machine learning algorithms to enhance early detection of at-risk children in order to solve the under-5 child mortality. The researchers utilized Random Forest and XGBoost models to classify child mortality risk. The models achieved high predictive performance (RF AUC is 0.87 and XGBoost AUC is 0.89). SHAP analysis supported the interpretability of results, highlighting key protective and risk factors. However, the authors noted the limited inclusion of psychological variables such as mental health and maternal stress and no explicit focus on teenage mothers or loan burden factors.

Verma and Prasad [9] investigated high infant and child mortality in India due to healthcare access, socioeconomic status, and maternal conditions. Hence, they utilized various AI/ML techniques such as trees, random forest, linear regression, and fuzzy logic for early prediction of infant and child mortality using survey and clinical data. Random forest and ensemble models showed strong predictive potential. However, the study lacked empirical validation and does not address contextual factors like adolescent motherhood, loan burden, or maternal mental health, limiting relevance to specific populations.

There is an inequality between the prevalence of child mortality between developing and developed countries, especially in sub-Saharan Africa such as Zimbabwe [10]. Despite the increasing availability of child mortality data, the application of machine learning for predictive analysis are still in their formative stage. To predict child mortality, this article applies several ML models such as decision tree, random forest, logistic regression and Extreme Gradient Boosting (XGBoost). Through these analyses, the key predictor for child mortality was revealed to be the combination of size of birth, prenatal order, marital condition, the current age of the child, water supply, religion, and residence and wealth index. Secondary data from the 2015 to 2016 Zimbabwe Demography and health survey which includes a sample of over 11,000 households were used to conduct this study. One of the limitations identified in this study is the imbalance of the target variable of B5 which indicates that "child is alive" as there are 5507 positive values (yes) and 299 negative value.

The article [11] acknowledges the increasing interest in utilising AI and machine learning analysis to determine the predictors of child mortality and for better precision. The techniques used in the study included Supervised and Unsupervised machine learning methods, incorporating random forests, decision trees, neural networks, and support vector machines to group and evaluate the data. The findings showed that machine learning models executed more accurate predictions with sensitivity analysis helping to refine model functionality and generalization ability. The article does not specify the source of data collected as it only stated that that dataset included demographic, maternal, healthcare access, and environmental variables to determine the predictors. The study does not include performance comparisons between models, making it difficult to determine which algorithm performed better. Moreover, the dataset was only vaguely described with no clear indication of the source, decreasing credibility and transparency of the research.

Research on the utilisation of machine learning models to predict under 5 mortality risks in Ethiopia [12] remains to be scarce and so this research article focuses on using spatial variations to predict the key sociodemographic factors. The study used three machine learning approaches which encompasses K-nearest neighbours, logistic regressions, random forests and the standard logistic regression model and evaluated the model performance and accuracy with receiver operating characteristics curves. The findings revealed that

random forest has the highest predictive accuracy of 67.2%. It identified key factors such as household size, water access, breastfeeding status, and maternal BMI as key contributors to under-five mortality. The data used originated from the 2016 Ethiopian Demographic and Health Survey. The drawbacks of the study were that the survey did not include deceased mothers, thus may lead to the misinterpretation of the rates of under-five mortality.

The article [13] delves into investigating the accuracy of predictivity of under-five mortality using machine learning models and identifying the significant contributors. Several machine learning methods were applied such as neural network, multivariate logistic regression, random forest, decision tree, Naive Bayes, K- nearest neighbor (KNN), support vector machine (SVM), and ridge classifier. The precision of the models were then evaluated using recall, precision, accuracy, a confusion matrix, F1 score, Cohen's Kappa, and receiver operating characteristics curve (AUROC). The results concluded that the accuracy of the neural network exceeded the other models by 95.96% with the dominant predictors being number of living children, birth size, wealth index, and maternal education. The dataset was derived from the NFHS-IV survey of Uttar Pradesh, which provided state-wise and household-level data on population, health, and nutrition.One of the limitations of this study are coefficient and odds ratio are not included in machine learning models, making it difficult to interpret and identify the significance of factors. Additionally, the survey used for the dataset was non-specific and did not contain objectives regarding under-five mortality.

Fatine and Agustí [14] were interested in predicting the trend and pattern of child mortality as the official systems used to record birth and deaths may not always function properly, therefore making the accuracy of the child mortality rate low. They have used statistical modelling and simulation techniques for prediction. The study showed that LOGIT transformation was not ideal for child mortality estimation, while ILR(Isometric Log-Ratio Transformation) gave better results but more data is required to ensure the accuracy which was difficult for some countries due to data limitations. To fill this gap, they suggested using imputation techniques to handle missing data and improving neonatal mortality data through indirect methods.

K. Harron et al. [15] carried out a study to explore the link between teenage motherhood and infant outcomes to find out how they differ in various social settings. They used statistical analysis techniques like Generalised Linear Model (GLMs) and multilevel model to analyse the trend, together with sensitivity analysis to assess the impact of missing data. The results showed that as the mother's age increased, the chances of the baby having health problems decreased. Teenage mothers were 2 to 4 times more likely to experience child mortality compared to older mothers. Besides, there was also a 25% to 40% of higher risk for teenage mothers to have preterm babies. One of the limitations of this study was that the Socio-Economic Status (SES) measurement could not capture all possible social risk factors related to teenage motherhood. Another limitation was the researchers did not know whether a mother had given birth before, therefore, it was hard to analyse whether the child mortality was due to the inexperience of the mother or due to her age. They have mentioned that they lacked sufficient statistical power and faced issues with data suppression.

A study [16] was conducted with the intention of helping to predict child death rates, which was a major problem in Pakistan. A proper prediction framework using machine learning algorithms was needed to assist the healthcare systems in managing child mortality. The group of researchers built a model by applying supervised machine learning classifiers and evaluated the performance using suitable metrics. They have used predictive mean matching (PMM) to handle missing data. The models used for prediction were Decision Tree, Random Forest, Naive Bayes and Extreme Gradient Boosting. Based on the evaluation, the Random Forest model had the highest prediction accuracy rate of 93.8%. With information gain using smart analysis, they also successfully identified the key risk factors of child mortality. One of the limitations of this study was the possible bias from mother's response when recalling past events in the survey and cause-specific mortality could not be identified from the survey data.

B. Comparison Table

The comparison table is provided in Appendix A.

C. Writing the Literature Review

Several studies have highlighted that there is a strong association between maternal attributes such as young maternal age and under-five mortality. The research also consistently shows that children with teenage mothers, especially mothers that are under 16 years old, face significantly higher mortality risks. Moreover, other significant predictors of under-five mortality include maternal education, short intervals between pregnancies, and limited access to quality healthcare. While traditional statistical methods such as logistic regression were used to identify potential determinants, numerous recent research has explored machine learning models for better accuracy in predictability, including Random Forest and XGBoost to identify determinants of child mortality which includes breastfeeding practices, maternal wealth, healthcare access, household environment, and maternal BMI. Through the use of machine learning models, large and complex datasets could be handled, achieving high AUC scores ranging from 0.87 to 0.94 for Random Forest (RF) as demonstrated by Bizzego et al. (2021). Despite this, many of the studies focus on the models' accuracy but do not explain which specific factors influence the predictions or the weight of the factors influence and few studies incorporate explainability techniques such as SHAP to derive implementable insights from the findings. Moreover, these machine learning approaches also tend to focus more on the performance of the models and variable significance but they do not consider complex real world factors such as environmental, social and cultural nuances.

Several gaps and limitations in the studies could be found. Firstly, psychological and socioeconomic factors which include stress and mental health of the mother, family support, or financial burdens, were overlooked. This diminished the studies' accuracy due to the variables' potential importance in predicting child mortality outcomes. Additionally, numerous articles focusing on machine learning tend to lack transparency in their data source, preparation and model interpretation. This limits the use of these studies in actual, realistic healthcare systems. Thirdly,

the several machine learning models were not evaluated to be deployed and used in clinical or public healthcare environments and few studies explored how child mortality patterns change over time nor do they analyse the causal inference. Thus, future studies can explore the integration of psychosocial and contextual factors in their machine learning models as well as apply interpretable methods such as SHAP to explain the weight of contribution of each input in the models predictions.

D. Summary of Literature Review

The examined literature notes that maternal characteristics, including young maternal age, level of education, birth interval, and access to health care have a significant association with under-five child mortality. Research continues to demonstrate that children having teenage mothers especially under the ages of 16 have greater risks of dying. A number of papers used machine learning methods, such as Random Forest or XGBoost to enhance the accuracy of their predictions and to determine important predicting factors such as breastfeeding practices, wealth of the mother and characteristics of the household. Nonetheless, the studies had some limitations like quality of the data, regional consistency, as most studies focused on Africa and South Asia, and excluding psychological or socioeconomic factors such as family support. Moreover, though machine learning models demonstrated overall good performance, not many scholars spoke specifically about the interpretability of these models, including potential uses of the model in real-life applications to the healthcare system. The findings of these gaps support the consideration that additional research is required, which combines statistics and context knowledge to improve the prediction and decrease child deaths particularly among teen mothers in low-income areas.

III. Research Methodology

This section details the methodology used to build a machine learning model child mortality prediction among young teenage mothers. It describes the measures taken to preprocess the health data, the construction of conceptual and predictive models, model performance evaluation techniques, and the use of visualizations to identify the most accurate and clinically relevant solution.

A. Flow Diagram or Research Framework

The flow diagram in Appendix B presents a structured pipeline for building a machine learning model to predict child mortality among teenage mothers in Western Kenya. The problem statement of this study highlights the disproportionate burden of child mortality faced by the teen mothers because of socio-economic and health vulnerabilities. Hence, this problem was faced with two research questions that focused on identifying the significant predictors and assessing the predictive performance of machine learning models. The objective of this research is to identify which machine learning model most effectively predicts child mortality among teen mothers, enabling early detection of high-risk cases and supporting targeted health interventions.

The literature review grounds the study in prior research. It reveals that teen mothers often experience higher child mortality rates due to limited healthcare access and systemic inequalities [3], [15]. It also shows that ML models like Random Forest and XGBoost have successfully enhanced mortality prediction [5], [7], with key predictors such as maternal education and birth intervals. However, gap remains, particularly in studies that overlook teen-specific stressors like mental health and use poorly balanced datasets, limiting their practical use.

Moving into research methodology, data was collected from a 2021 replication dataset on teenage mothers in Western Kenya (Harvard Dataverse). Data preprocessing included filtering the dataset for teenage mothers (xteenmom values 1 or 2), handling missing values, encoding categorical variables, and normalizing numerical ones when necessary. Feature selection involved isolating independent variables with theoretical or empirical relevance to child mortality. These included socio-demographic factors (marital status, education, wealth index), indicators of food insecurity, mental health metrics (individual items mh1 - mh8, and composite score p75mhscore), past child or infant deaths, and loan burden which captures both health and socio-economic dimensions of vulnerability.

The model development phase tested six machine learning algorithms which are Logistic Regression, Decision Tree, Random Forest, Support Vector Machines (SVM), Naive Bayes, and XGBoost which are trained on selected features. Models were evaluated using metrics such as accuracy, precision, recall, F1 Score, and confusion matrix analysis. Visualizations including performance bar charts and heatmaps were used to interpret results. In the finding's discussion, model comparisons, confusion matrix insights, and feature importance were analyzed. The process concludes with research recommendations, suggesting ML-based screening tools, prioritizing early detection, and calling for future validation with larger or real-time datasets.

B. Research Design

This research applies a quantitative approach using a comparative study design, highlighting the evaluation and comparison of the different machine learning algorithms performance in predicting child mortality among young teenage mothers. By analyzing health data, the study aims to identify the most effective model based on key performance metrics.

C. Steps Involved in the Methodology

1) Identification of research objectives: The primary objective of this study is to compare multiple machine learning algorithms based on predictive performance metrics. This is to determine the most effective and accurate model for estimating child mortality risk among teenage mothers in Western Kenya. By identifying which model performs best, this research hopes to contribute to early identification of high-risk child mortality cases and support targeted health interventions. The broader goal is to support data-based policy decisions on maternal and child health within resource-constrained settings.

2) Literature review and analysis: Relevant studies were retrieved from Google Scholar, ScienceDirect, Wiley Online Library and more using keywords such as "child mortality

prediction," "machine learning," and "teenage mothers." Studies published between 2019 and 2025 were considered. Studies were included if they applied statistical or machine learning techniques to predict child mortality, examined factors influencing child mortality, or focused on child mortality among teenage mothers. This study proposes a comparative analysis of machine learning models using a dataset from Western Kenya. The goal is to identify the most effective model for predicting child mortality among teenage mothers. Most existing studies do not focus on psychological factors and tend to target general populations or conduct only risk factor analyses. This research fills that gap by focusing on the underexplored subgroup of teenage mothers, using a dataset that includes psychological variables and applying predictive classification models. The comparative approach adds further novelty by evaluating multiple algorithms for accuracy and real-world use in public health planning.

3) Criteria for comparison: Multiple machine learning models are used to predict child deaths in teen mothers. Therefore, having suitable evaluation metrics are critical to ensure fairness when analysing the performance of those models. The models will be compared based on five performance metric criteria. Accuracy is how often the models are able to make correct predictions about whether a child will survive or not. Recall, also known as sensitivity, indicates the number of actual deaths of children that have been recognized properly. Precision indicates the number of actual child deaths that are accurately captured in predictions. In other words, out of all the times the model predicted child death, how many predictions were correct? The F1 Score is a balanced measure of recall and precision into a single measure [17]. This can be done effectively in situations where a class imbalance exists for example, when deaths are much lesser than survival. Finally, Confusion Matrix gives an overall picture by including true positives, false positives, false negatives, and true negatives, showing how well the model performs overall.

4) Data collection: The title of this dataset is 'Replication Data for: Teen Mothers Report Poor Health and Economic Functioning in Western Kenya: A Call to Action' [18]. It is a recent dataset published in June 2025 on Harvard Dataverse. The dataset is a real-world dataset collected through surveys with teenage mothers in Western Kenya. The original dataset contains 6,208 cases with 27 variables. In order to predict child mortality among teenage mothers, the data was first filtered to include only rows where 'xteenmom' == 1, indicating teen mother with one child, or 'xteenmom'== 2, indicating teen mother with two or more children. The nine variables selected as the predictor variables are marital status, education, wealth index, household hardship (food/sleep hungry), past child mortality (infant and under-five deaths), mental health indicators (mh1–mh8, p75mhscore), and financial stress (loans). The target variable for this study is 'anychildead' which is a binary indicator representing whether a child has died. These features were chosen based on their potential influence on child health outcomes.

5) Experimental Setup: The experimental setup was conducted using Google Colab, a cloud-based Jupyter Notebook environment that offers a pre-configured Python environment suitable for machine learning tasks. Python 3.10 was used along with essential libraries such as pandas for data manipulation, scikit-learn for machine learning implementation, and matplotlib and seaborn for visualization. The dataset was loaded using the openpyxl engine for Excel compatibility.

6) Implementation of solutions: The implementation began by filtering the dataset using the variable xteenmom to include only teenage mothers. After removing missing values, selected predictors such as education level, wealth index, food insecurity, mental health indicators, and past child deaths were used for modelling. Since the variables were already in numeric or binary format, minimal preprocessing was required aside from normalization for algorithms like SVM. The five models were trained using an 80/20 train-test split, with anychildead as the binary target variable. Model performance was evaluated based on the five key metrics and all models and evaluations were implemented using scikit-learn, with visual results presented through bar charts and heatmaps generated via seaborn and matplotlib.

7) Evaluation and comparison: Each of the machine learning models developed will be evaluated through the five key metrics. These metrics were selected due to the imbalanced dataset in which data regarding the child mortality cases were significantly lower than the non-child mortality data. For model comparison using these metrics, bar charts were utilised for easier assessment through visualisation. Each of the models developed were trained and tested using the same 80/20 train-test split, and their performances were compared directly based on these four evaluation criteria. The results were also validated using the 5-fold cross validation, threshold tuning, SMOTE and GridSearchCV. To ensure consistency across all models, the same dataset and pre-processing methods were used.

8) Validation of results: The model results were validated using stratified 5-fold cross-validation which uses stratification to maintain the initial distribution of every class across each fold, ensuring that each fold has a fair distribution of positive and negative cases. This ensures that the evaluation metrics are stable and not overly dependent on any particular data split. Additionally, the model performance was evaluated using performance metric to account for the class imbalance in the dataset. Moreover, the findings were compared with baseline models such as Logistic Regression and Naive Bayes, serving as performance benchmarks.

D. Tools and Technologies Used

The Python language is used throughout this study to build machine learning models as it is the primary language

taught in our degree program. We have a certain level of knowledge in Python, allowing us to confidently use it and troubleshoot code when necessary. Our team runs the Python code using Google Colab as it is a cloud-based platform that allows multiple collaborators to code together without needing to install any software. This prevents issues like code syncing errors that may happen when compiling different sections of codes. Besides, several libraries were used when coding the predictive models. For example, pandas was used mainly for data manipulation and cleaning, scikit-learn for training the models and evaluating performance, matplotlib and seaborn for generating charts and graphs for visualisation. Furthermore, the dataset was imported using data handling tools from the pandas library, specifically the pd.read_excel() function to read the dataset in excel format.

E. Challenge and Limitations

At the beginning of the assignment, we started by looking for a dataset that we are interested in from the three data sources given by the lecturer. The first challenge arose during the dataset selection process. We are unable to find a dataset with enough variables and a large sample size to support more accurate predictions. Some of the datasets did not provide description or codebooks which are crucial for understanding the meaning of the values in the dataset.

Another challenge we faced was about variable selection. There are many variables included in the dataset, however it is crucial to select only those relevant as predictor variables. We are unsure whether we had selected the appropriate variables or if we had missed out any important variables that could greatly affect the output. We solved this issue by conducting multiple article and literature reviews to identify the potential factors that would affect child mortality among teen mothers.

Furthermore, we faced difficulties in choosing the appropriate evaluation metrics to assess model performance. Not all metrics are suitable to be used because this is a binary classification problem with class imbalance issues. To overcome this, we looked up for the definitions of each metric to better understand when they are applicable and what are their limitations.

F. Summary

This research study follows a quantitative, comparative research methodology to assess the prediction of child mortality among young teenage mothers in Western Kenya using machine learning models. The methodology begins by first defining the research objectives, conducting literature review and selecting relevant variables from a real-world dataset that was published on the Harvard Dataverse. Several machine learning models were developed using the programming language Python through Google Colab. The models that were implemented were Logistic Regression, Decision Tree, Random Forest, SVM, Naive Bayes, and XGBoost, which all included preprocessing steps such as filtering teen mothers and handling any missing data. Additionally, the models were trained using the 80/20 train-test split as well as evaluated based on key metrics which are accuracy, recall, precision, confusion matrix. To visualize comparison between the performance, bar charts were used while the models were validated using the 5-fold cross validation as well as compared with the benchmark models which are Logistic Regression and Naive Bayes.

IV. RESULTS AND DISCUSSION

This research presents the results of multiple machine learning models developed to predict the likelihood of child mortality among children born to teenage mothers in Western Kenya. The performance of each model including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Naive Bayes and XGBoost is assessed using evaluation metrics including accuracy, precision, recall, F1-score, and the confusion matrix. These metrics help to measure not only the overall performance but also how well each model identifies true positive and negative outcomes. A comparative analysis is then conducted to determine the most effective model, with further discussion on its potential application in maternal and child health decision-making.

A. Presenting Results

The quantitative comparison table summarizing the performance metrics of different models is provided in Appendix C.

1) Logistic regression: The confusion matrix shows that the model correctly identified 344 children who survived (true negatives) and 330 actual death cases (true positives). However, it also made 156 false positive predictions, wrongly flagging children as at risk when they were not. Additionally, it missed 164 actual deaths (false negatives), where the model incorrectly predicted survival. The model's accuracy was 0.68. This means 68% of the total predictions made by the model were correct, whether it predicted a child would survive or die. Precision of 0.68 shows that 68% of the cases where the model predicted a child would die were actually correct. In other words, the model does not raise too many false alarms. False alarms can lead to unnecessary concern or misallocation of resources in a public health context. Recall of 0.67 shows that the model correctly identified 67% of actual child death cases. However, in high-impact fields like child health, missing 33% of true death cases (false negatives) could be very harmful, as these at-risk children might not receive timely intervention. The F1 score of 0.67 suggests the model is somewhat balanced between precision and recall, but not strong. A dependable model often has F1 scores at least 0.75 to 0.8 for real-world deployment. In short, while the logistic regression model shows moderate reliability, its current performance may not be sufficient for real-world use in predicting child mortality, where both false positives and missed cases can have serious consequences. Refer to Appendix D, Figure D1 for the confusion matrix of Logistic Regression, and Figure D2 for the bar chart illustrating its performance metrics.

2) Decision Tree: The decision tree model was evaluated using a fixed threshold of 0.35 in order to address the class imbalance issue. After splitting the dataset into 80% for training and 20% for testing, the model produced a confusion matrix which shows that among the 497 true survival cases,

342 were correctly predicted as shown by the true negatives. Meanwhile, the model incorrectly flagged 155 deaths shown by the false positives. The confusion matrix also shows that out of the 47 actual deaths, the model correctly identified 29 cases (true positives) but also misclassified 18 as survivors (false negatives). Moreover, the model achieved an accuracy of 68.19% for predictions and precision of 15.76% which means that out of all the predicted deaths, only 15.76% were correct. This reflects the model's tendency for false alarm as only a small proportion of predicted deaths were correctly identified. The model also had a recall rate of 0.62 which indicates the model was able to correctly identify 62% of under-5 mortality. The F1 score of 0.2511 reflects this trade-off between precision and recall. The use of class weighting and a lowered decision threshold in this model helped increase sensitivity, making the model more cautious in reducing false negatives. Refer to Appendix D, Figure D3 for the confusion matrix of Decision Tree, and Figure D4 for the bar chart illustrating its performance metrics.

3) Random Forest: The evaluation metrics for Random Forest model show that while the overall accuracy is high at 89.71%, this is likely due to class imbalance, as most cases belong to the "no child death" class. However, the precision of 34.48% indicates that when the model predicts a child's death, it is correct only about a third of the time, suggesting many false positives. Even more concerning is the recall of 21.28% meaning the model correctly identifies only about one in five actual child death cases, missing most of them. The F1 score of 26.32%, which balances precision and recall, further confirms that the model struggles to effectively detect the minority class. In summary, despite good accuracy, the model performs poorly in identifying true positive cases and needs improvement, especially if detecting child death is a priority.The confusion matrix shows the model's predictions compared to the actual outcomes. Out of all predictions, 478 cases were correctly predicted as "no child death" (true negatives), and 10 were correctly predicted as "child death" (true positives). However, the model incorrectly predicted 19 cases as "child death" when there were none (false positives), and more critically, it missed 37 actual child death cases (false negatives). This imbalance highlights that the model is good at identifying the majority class ("no child death"), but it struggles to detect the minority class ("child death"), which is crucial in this context. The high number of false negatives is particularly concerning because it means the model is failing to identify most of the cases that actually involve child death which is consistent with the low recall observed earlier. Refer to Appendix D, Figure D5 for the confusion matrix of Random Forest, and Figure D6 for the bar chart illustrating its performance metrics.

4) Support Vector Machine (SVM): Regarding the Support Vector Machine (SVM) model, the accuracy of 65.81% may appears satisfactory, but given the skew of the classes with much less number of child deaths as compared to child survivals, this statistic is misleading. The confusion matrix [[335, 162], [11, 36]] shows that 335 true negatives (the actual correct no child death result) had been identified but there were also 162 false positives (the predicted child death but it did not happen) and 11 false negatives (the actual child death but it was not predicted). The correct prediction of the true child deaths only occurred in 36 cases which implies a low value in precision of 14.01%, indicating that most of the predicted child deaths were not correct and can cause unwarranted worry or resource of healthcare facilities. In this case, recall is 57.68 %, which means that over half of all actual child deaths were captured, a good feature of life-critical predictions. However, the poor overall balance between precision and recall, which is indicated by the low F1-score of 22.53 % limits the practical reliability of the model. Refer to Appendix D, Figure D7 for the confusion matrix of Support Vector Machine, and Figure D8 for the bar chart illustrating its performance metrics.

5) Naive Bayes: The Naive Bayes model achieved an overall accuracy of 91.76%, which initially appears strong. However, given the likely imbalance in the dataset such as more instances of "no child death" than "child death", this high accuracy may not reflect true performance for the minority class. The precision of 53.85% indicates that when the model predicts a child's death, it is correct just over half the time. The recall of 30.43% reveals that the model still misses nearly 70% of actual child death cases. The F1 score of 38.89%, which balances precision and recall, suggests that while the model has improved in identifying the minority class compared to the previous model, it still struggles to capture most true cases of child death. The confusion matrix indicates that 476 true negatives meaning it correctly predicted "no child death" and 14 true positives meaning it correctly predicted "child death" were recorded. Meanwhile, 12 false positives (incorrectly predicted "child death") and 32 false negatives (missed "child death" cases) occurred. This again highlights the model's tendency to correctly classify the majority class while underperforming in identifying the minority class. However, Naive Bayes might be more cautious in predicting child deaths. Refer to Appendix D, Figure D9 for the confusion matrix of Naive Bayes, and Figure D10 for the bar chart illustrating its performance metrics.

6) eXtreme Gradient Boosting (XGBoost): Regarding the XGBoost model, the accuracy score was 78% which at first glance can be considered a rather good result, but when it comes to predicting child mortality, this percentage displays serious shortcomings. A confusion matrix of [[311, 186], [18, 29]] demonstrates that the system assigned most of the observations to the correct class, with no child mortality and misclassified many cases that have occurred child deaths, namely 18 out of 47. As a result, the precision score is 14.86 % which is equivalent to one in every 7 deaths of children being identified accurately. Furthermore, recall is equal to 34.25 %, which indicates that the model fails to identify the majority of real child deaths. Its difficulty in striking the balance between precision and recall is emphasized in the F1-score of 20.61 %. Together, these results indicate that the

model has an acceptable rate of overall accuracy, but it is not effective in predicting important outcomes, which makes it inappropriate to use in high-stakes settings unless it is significantly modified. Refer to Appendix D, Figure D11 for the confusion matrix of XGBoost, and Figure D12 for the bar chart illustrating its performance metrics.

B. Comparative Analysis and Discussion

In all models, interesting trends can be observed. Models with high accuracy like Naive Bayes (91.76%) and Random Forest (89.71%) often suffer the worst recall and F1-score because they appear to be biased toward the majority class of no child deaths. Although there are several techniques used attempting to improve the score by applying standardization and SMOTE and tuned it via GridSearchCV, Naive Bayes was not able to find actual cases of child death at more than 30.43% recall and 38.89% F1-score. Meanwhile, Random Forest could only achieve a recall of 21.28% and F1-score of 26.32%

On the other hand, Logistic Regression did a relatively good job in balancing recall (67.80%), precision (67.90%), and F1-score (67.30%). In fact, it ranks one of the lowest for accuracy (67.80%). Thus, we can see that accuracy is not a reliable metric for performance in imbalanced datasets. Logistic Regression clearly benefits from the use of SMOTE, it was also able to select features (RFE), it adjusts the hyperparameters via GridSearchCV and it managed to set class_weight='balanced' to create samples that better handled class imbalance compared to other models.

The remaining models including Decision Tree, Support Vector Machine, and XGBoost produced similar accuracy scores (67-68%) but had low precision (~14-16%) and F1-scores (~25%) indicating they produced many false positives and missed a high amount of actual death cases. The models did use similar means of addressing imbalance like SMOTE, class_weight='balanced' and tuning, but were still unable to achieve the performance of Logistic Regression, likely due to higher noise sensitivity, risks of overfitting or poorer generalization on a small sample in the minority class.

In general, it appears that linear models that support class imbalance like Logistic Regression performed better when a ranking approach prioritized recall or F1-score over accuracy. Models such as Naive Bayes and Random Forest showed high overall accuracy, but they were only recognizing a few of these actual cases and tended to largely miss them because of bias toward the majority class. As such, these types of models may be inappropriate in a context such as public health where rare outcomes may be serious and need identifying. Logistic Regression, in comparison, was challenged less when capturing these important cases. As a result, Logistic Regression is a better model for predicting child mortality outcomes in imbalanced data.

C. Overall Discussion

The six ML models evaluated in this study show significant variation in their ability to predict child mortality, particularly when dealing with imbalanced data where child death is the minority class. While some models exhibit high overall accuracy such as Naive Bayes (91.76%) and Random Forest (89.71%), this metric can be misleading in imbalance scenarios. High accuracy may simply reflect the model's strength in predicting the dominant class which is "no child death" rather than its effectiveness at identifying the more critical minority class ("child death class"). More informative metrics like precision, recall, and F1 score are therefore essential for evaluating the model's true predictive power in this context.

Models such as Logistic Regression and Naive Bayes provided a relatively balanced trade-off between precision and recall, with F1 scores of 67.30% and 38.89% respectively, indicating moderate ability to identify child death cases. SVM showed the highest recall (57.68%) among all models, meaning it captured more actual child death cases, but its very low precision (14.01%) suggests a high rate of false alarms, which could lead to unnecessary stress or wasted resources. Random Forest and Naive Bayes, despite high accuracy, had the lowest recall scores, 21.28% and 30.43% respectively, signaling poor performance in detecting the minority class. Their low F1 scores, which are 26.32% and 38.89% respectively, further reinforce their limitations for this life-critical task. In summary, none of the models performed satisfactorily in terms of both precision and recall, highlighting the difficulty of predicting rare but crucial events like child mortality. More advanced techniques such as better data balancing, ensemble methods optimized for recall, or domain-specific features may be required to improve model reliability for real-world deployment.

D. Critical Evaluation

The initial objective of this project was to build predictive models that identify the likelihood of child mortality among children born to teenage mothers in Western Kenya. The goal was only to assess key contributing factors but also to predict high-risk cases effectively, enabling early interventions. While the models developed demonstrate varying degrees of success in overall accuracy, a critical issue emerged across all approaches which is the inability to reliably identify the minority class which are actual cases of child death. Models such as Random Forest and Naive Bayes achieved high overall accuracy, but this metric proved insufficient in meeting the core objective, as these models performed poorly on recall, missing a significant portion of actual child death cases.

This outcome highlights a major challenge which is class imbalance significantly undermines the models' ability to detect rare but crucial outcomes, Even the SVM model, which had the highest recall at 57.14% suffered from extremely low precision, creating the risk of too many false alarms. In a health context, both false negatives and false positives carry serious implications that are either by failing to intervene in time or by misleading limited resources. Therefore, while the project has succeeded in demonstrating the feasibility of predictive modelling for child mortality risk, it also exposes the limitations of conventional models in such imbalanced, high-stakes scenarios. Future improvement could involve advanced sampling techniques, tailored cost-sensitive algorithms, or integrating more contextual health and social variables to ensure the model better supports the

early identification of at-risk children in vulnerable communities.

E. Linking to Literature Review

The results achieved by the 6 machine learning models partially align with the literature review in which ensemble models such as Random Forest and XGBoost have strong predictive capabilities as shown by the AUC scores of between 0.87 and 0.94 as demonstrated by Bizzego et al. [8]. However, consistent with the findings of [10], [12], and [16], our recall scores were significantly influenced by the dataset's class imbalance. This highlights the issues that come with relying heavily on accuracy as well as the concerns that were noted by Noori et al. [3] and Woodall et al. [4]. They had stated that specific risks faced by certain groups such as the increased vulnerability of very young adolescent mothers were often overlooked. As a result, these studies often don't fully reflect the deeper, long-term challenges or individual circumstances such as psychological or social issues, that can affect child mortality outcomes. This was reflected in our findings which had shown that while ensemble models performed well overall, false negatives are often not addressed especially in cases of mortality prediction. Our results also support the need to improve sensitivity by threshold tuning and class reweighting.

Furthermore, like several concerns raised in the literature review, our findings reflect the issue of transparency and practical use of machine learning models in public health scenarios. Although the reviewed studies successfully identified the key predictors of under-5 mortality such as age, education, and healthcare access, few studies also included psychosocial factors or explainability tools like SHAP. Our 6 machine learning models were effective in handling large datasets and improving recall. However, our findings still reflect the gap in interpretability and contextual relevance. This supports the argument made by Bizzego et al. [8] that some improvements that could be made to improve the predictive accuracy as well as real world application for child mortality models among teenage mothers can include integrating social and environmental determinants together with explainable AI techniques.

F. Conclude the Results and Discussion

In summary, this study shows the critical importance of selecting appropriate evaluation metrics and modelling when dealing with imbalanced, high-stakes datasets to predict child mortality among children born to teenage mothers. While various models were explored, it shows that high overall accuracy does not translate to meaningful predictive power in identifying the actual cases of child death. Metrics such as recall and F1-score are more appropriate for evaluating model performance in this public health context.

Among the models tested, the Logistic Regression model is the most effective in predicting child mortality compared to other models. It has achieved the best balance across accuracy, recall, precision, and F1-score. Its ability to detect high-risk cases more reliably highlights its suitability for real-world applications where identifying minority outcomes which are cases of child death is crucial. This suggests that linear models, when properly optimized, are better suited for tasks involving rare but life-critical outcomes. It means that Logistic Regression is better at correctly identifying the rare but important cases of child death, which is essential in real-life healthcare decisions because failing to detect such outcomes could have serious consequences. The results of this research provide a valuable starting point for developing early warning systems that support targeted interventions and improve maternal and child health outcomes in vulnerable communities such as Western Kenya.

V. Conclusion and Future Work

A. Summary of Key Findings

The study evaluated six machine learning models which are Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Naive Bayes and XGBoost to predict child mortality among children born to teenage mothers in Western Kenya. The models showed substantial variation in performance, particularly due to dataset imbalance where child deaths were underrepresented. While models like Naive Bayes and Random Forest achieved high accuracy which is above 89%, their low recall scores which are 30.43% and 21.28% respectively revealed poor effectiveness in identifying actual child mortality cases. Logistic Regression demonstrated the best overall balance between precision, which is 67.90%, recall which is 66.80%, and F1 Score which is 67.30%, highlighting its suitability for imbalanced life-critical datasets. Models like SVM and XGBoost showed moderate recall but suffered from very low precision, resulting in high false positive rates.

B. Link Back to Objectives

The primary objective of this research was to develop predictive models that could effectively identify the likelihood of child mortality among children born to teenage mothers in Western Kenya and assist in early intervention. The findings show that while all models were able to process and analyze data, only Logistic Regression came close to meeting the project's goal of balancing sensitivity (recall) with reliability (precision), making it the most promising for identifying high-risk cases. The study also confirmed that class imbalance significantly hampers model performance in identifying rare but critical outcomes. Overall, the project successfully demonstrated the potential of machine learning for early risk detection, while offering practical relevance for public health decision-making.

C. Contribution of the Research

This research adds to the literature by filling a gap in predictive analytics for child mortality in teenage mothers in Western Kenya which is a vulnerable population. It includes less researched psychological and socioeconomic factors and compares several models, noting the superiority of logistic regression. The study identifies key predictors and provides actionable information for the early identification of risk and targeted interventions in low-resource environments.

D. Implications for Practice

The findings are of practical application to public health policymakers and NGOs in that they enable the early detection of high-risk cases of child mortality among adolescent mothers. The predictive model, logistic regression

in particular, can guide targeted interventions, resource allocation, and the adoption of early warning systems in resource-scarce settings. This enhances efforts towards preventing unnecessary child deaths and improving maternal health outcomes in vulnerable groups.

E. Limitations

This study encountered several limitations which may have implicated the effectiveness and strength of the results. One such limitation is the class imbalance of the dataset which has a relatively small number of under-five mortality cases. Although several techniques such as SMOTE and class weighting were used to solve this issue, this limitation decreases the models' ability in generalizing well. Moreover, the dataset used in this study also lacked psychosocial, cultural, and environmental variables such as maternal family support and access to healthcare facilities. Our study also did not include explainability tools such as SHAP, limiting insights on the influence of each variable on child mortality prediction. Finally, another limitation is the lack of computing power such as limited memory and computer resources, as well as data preparation tool limitations to fully clean and transform the dataset. Due to this issue, our study was not able to test every hyperparameter or test out more advanced combinations of models such as ensemble methods like stacking or boosting thoroughly.

F. Recommendation for future research

Some recommendations for future research include incorporating psychosocial and contextual variables to enhance the predictive strength and relevance of machine learning models. Moreover, future analysis should also aim to integrate explainability methods such as SHAP or LIME to help provide details on the reason behind the machine learning model's predictions. This is especially important in real world healthcare situations whereby healthcare workers would need to understand and trust the model's prediction performance in order to make decisions. Finally, future studies should also work with healthcare stakeholders to evaluate models' practicality and potential deployment in real-world health settings.

G. Concluding remarks

This study demonstrates the potential of machine learning in predicting child mortality among teenage mothers by comparing the performance of various algorithms on an imbalanced dataset. While some models had high accuracy but missed many child death cases, and others caught more deaths but made more false predictions, the comparison highlights the trade-offs involved in selecting the most appropriate approach for life-critical tasks. In predicting child mortality, recall and F1-score are more important than just accuracy. From the results, Logistic Regression is the most effective model for identifying high-risk cases. These findings not only guide the selection of more suitable predictive tools but also contribute to the broader field by showing how data-driven models can support targeted public health interventions.

REFERENCES

[1] "Child Mortality - an overview | ScienceDirect Topics," www.sciencedirect.com. https://www.sciencedirect.com/topics/social-sciences/child-mortality

[2] "Metadata Glossary," World Bank DataBank. https://databank.worldbank.org/metadataglossary/world-development-indicators/series/SP.MTR.1519.ZS

[3] N. Noori, J. L. Proctor, Y. Efevbera, and A. P. Oron, "Effect of adolescent pregnancy on child mortality in 46 countries," BMJ Global Health, vol. 7, no. 5, p. e007681, May 2022, doi: https://doi.org/10.1136/bmjgh-2021-007681.

[4] A. M. Woodall, A. K. Driscoll, A. Mirzazadeh, and A. M. Branum, "Disparities in Mortality Trends for Infants of Teenagers: 1996 to 2019," Pediatrics, vol. 151, no. 5, Apr. 2023, doi: https://doi.org/10.1542/peds.2022-060512.

[5] Addisalem Workie Demsash, "Using best performance machine learning algorithm to predict child death before celebrating their fifth birthday," Informatics in Medicine Unlocked, vol. 40, pp. 101298–101298, 2023, doi: https://doi.org/10.1016/j.imu.2023.101298.

[6] M. K. Bhusal and S. P. Khanal, "A Systematic Review of Factors Associated with Under-Five Child Mortality," BioMed Research International, pp. 1–19, Dec. 2022, doi: https://doi.org/10.1155/2022/1181409.

[7] P. Pandey, S. Shukla, N. K. Singh, and M. Kumar, "Predicting child mortality determinants in Uttar Pradesh using Machine Learning: Insights from the National Family and Health Survey (2019–21)," Clinical Epidemiology and Global Health, vol. 32, p. 101949, Feb. 2025, doi: https://doi.org/10.1016/j.cegh.2025.101949.

[8] A. Bizzego et al., "Predictors of Contemporary under-5 Child Mortality in Low- and Middle-Income Countries: A Machine Learning Approach," International Journal of Environmental Research and Public Health, vol. 18, no. 3, p. 1315, Feb. 2021, doi: https://doi.org/10.3390/ijerph18031315.

[9] I. Verma and S. K. Prasad, "A Study on the Factors Affecting Infants' Health-Related Issues and Child Mortality using Machine Learning," vol. 21, pp. 615–624, Aug. 2023, doi: https://doi.org/10.1109/smarttechcon57526.2023.10391399.

[10] E. Mbunge et al., "Application of machine learning techniques for predicting child mortality and identifying associated risk factors," Mar. 2023, doi: https://doi.org/10.1109/ictas56421.2023.10082734.

[11] S. R. Y. and S. Kolla , "Child Mortality Prediction Using Machine Learning Techniques," International Journal of Marketing Management, vol. 12, no. 2, May 2024, doi: https://doi.org/ISSN%202454-5007.

[12] F. H. Bitew, S. H. Nyarko, L. Potter, and C. S. Sparks, "Machine learning approach for predicting under-five mortality determinants in Ethiopia: evidence from the 2016 Ethiopian Demographic and Health Survey," Genus, vol. 76, no. 1, Nov. 2020, doi: https://doi.org/10.1186/s41118-020-00106-2.

[13] R. K. Saroj, P. K. Yadav, R. Singh, and Obvious. N. Chilyabanyama, "Machine Learning Algorithms for understanding the determinants of under-five Mortality," BioData Mining, vol. 15, no. 1, Sep. 2022, doi: https://doi.org/10.1186/s13040-022-00308-8.

[14] F. Ezbakhe and A. PérezFoguet, "Child mortality levels and trends: A new compositional approach," Demographic Research, vol. 43, pp. 1263–1296, Dec. 2020, doi: https://doi.org/10.2307/26967840.

[15] K. Harron et al., "Preterm birth, unplanned hospital contact, and mortality in infants born to teenage mothers in five countries: An administrative data cohort study," Paediatric and Perinatal Epidemiology, vol. 34, no. 6, pp. 645–654, Apr. 2020, doi: https://doi.org/10.1111/ppe.12685.

[16] F. Iqbal, M. Islam Satti, A. Irshad, and M. Asif Shah, "Predictive analytics in smart healthcare for child mortality prediction using a machine learning approach," Central European Journal of Biology, vol. 18, no. 1, Jul. 2023, doi: https://doi.org/10.1515/biol-2022-0609.

[17] "Evaluation Metrics in Machine Learning," GeeksforGeeks, Jul. 15, 2025. https://www.geeksforgeeks.org/machine-learning/metrics-for-machine-learning-model/

[18] Jakubowski and Aleksandra, "Replication Data for: Teen Mothers Report Poor Health and Economic Functioning in Western Kenya: A Call to Action," Harvard Dataverse, Jun. 27, 2025. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/QOPLVI

APPENDIX

Appendix A

Article Review Comparison Table

Reference	Title	Problem Statement	Methodology/ Technique	Findings	Limitation
Noori et al., (2022)	The Effect of Adolescent Pregnancy on Child Mortality in 46 Low- and Middle-Income Countries [3]	Most studies overlook the significantly higher child mortality risk faced by very young adolescents	Used mixed-effects logistic regression	Lower use of prenatal care and facility births among younger mothers has the higher risk of child mortality.	Lack of biological, clinical, and psychological variables to predict child mortality.
Woodall et al., (2023)	Disparities in Mortality Trends for Infants of Teenagers: 1996 to 2019 [4]	Previous research has not explored how infant mortality rates have changed over time based on key factors like maternal age, racial or ethnic background and other more.	Used Joinpoint regression and Kitagawa decomposition analysis	Overall infant mortality among infants of teen mothers decreased by 16.7% from 1996–1997 through 2018–2019	The study did not examine how specific personal factors affect infant mortality which limits potential insights.
Addisalem Workie Demsash, (2023)	Using best performance machine learning algorithm to predict child death before celebrating their fifth birthday [5]	There is a lack of research applying machine learning to accurately predict under-five child mortality and identify key risk factors.	Random Forest and the J48 decision tree were conducted. Weka v3.8.6 is used to reveal meaningful patterns and relationships between risk factors and child mortality.	Random Forest model achieved the highest accuracy Late initiation of breastfeeding, mother's lack of formal education, short birth interval, low maternal wealth, and lack of media exposure significantly increased the child mortality risk.	Lack of important variables like healthcare access, maternal health behaviours, and environmental factors.

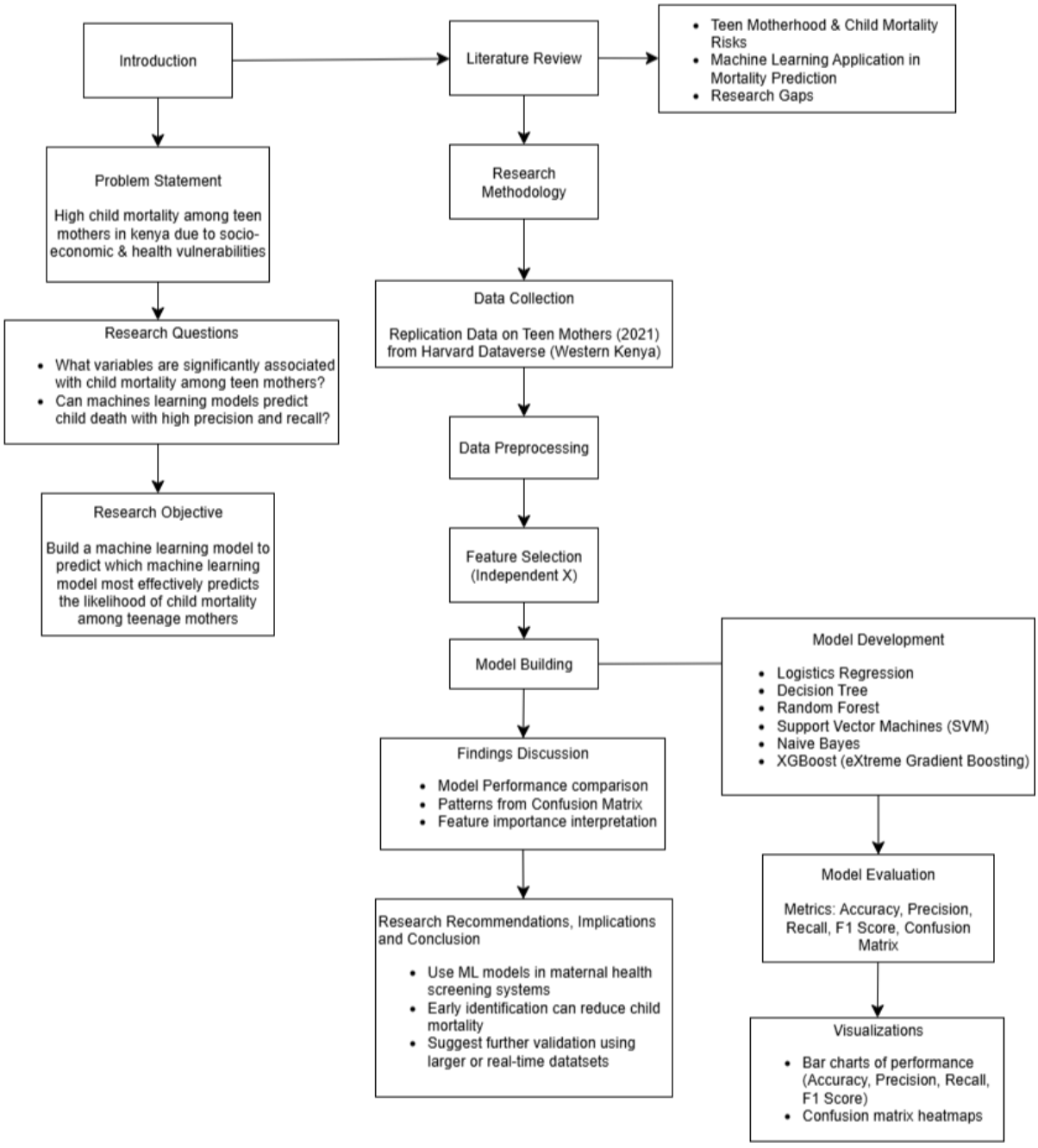
Pandey et al. (2025)	Predicting child mortality determinants in Uttar Pradesh using Machine Learning: Insights from the National Family and Health Survey (2019–21) [7]	Under-five mortality remains unacceptably high in Uttar Pradesh.	Used NFHS-V data to compare performance of Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), Naive Bayes, and traditional Logistic Regression model.	Logistic Regression achieved the highest accuracy (79.4%); Key determinants include breastfeeding status, recent births, child's gender, birth intervals, antenatal care, water source, birth order, and maternal BMI.	Did not examine psychological, financial stressors, or teen motherhood limiting its applicability to teen-specific contexts.
Bizzego et al. (2021)	Predictors of Contemporary under-5 Child Mortality in Low- and Middle-Income Countries: A Machine Learning Approach [8]	Under-5 child mortality continues to pose challenges in low and middle-income countries (LMICs), with traditional methods insufficient to capture complex intervariable-relationships.	Utilized Random Forest and XGBoost models to classify child mortality risk. Used SHAP for interpretability.	The models achieved high predictive performance (RF AUC is 0.87 and XGBoost AUC is 0.89). SHAP analysis supported the interpretability of results, highlighting key protective and risk factors.	Lacked empirical validation and does not address contextual factors like adolescent motherhood, loan burden, or maternal mental health, limiting relevance to specific populations.
Mbunge et al. (2023)	Application of machine learning techniques for predicting child mortality and identifying associated risk factors [10]	The applications of machine learning for predictive analysis are still in their formative stage.	Decision tree, random forest, logistic regression and Extreme Gradient Boosting (XGBoost).	The key predictor for child mortality was the combination of size of birth, prenatal order, marital condition, the current age of the child, water supply, religion, and residence and wealth index.	The imbalance of the target variable of B5 which indicates that "child is alive" as there are 5507 positive values (yes) and 299 negative values (no).
Y. Srinivasa Raju and K. Sivannarayana (2024)	Child Mortality Prediction Using Machine Learning Techniques [11]	There is an increasing interest in utilising AI and machine learning analysis to determine the	Supervised and Unsupervised machine learning methods, incorporating random forests, decision trees, neural	Machine learning models executed more accurate predictions with sensitivity analysis	Does not include performance comparisons between models. Dataset was only vaguely described with no

		predictors of child mortality.	networks, and support vector machines		clear indication of the source.
Bitew et al. (2020)	Machine learning approach for predicting under-five mortality determinants in Ethiopia: evidence from the 2016 Ethiopian Demographic and Health Survey [12]	Research on the utilisation of machine learning models to predict under 5 mortality risks in Ethiopia remains to be scarce.	K-nearest neighbours, logistic regressions, random forests and the standard logistic regression model.	The findings revealed that random forest has the highest predictive accuracy of 67.2%.	The survey did not include deceased mothers, thus may lead to the misinterpretation of the rates of under-five mortality.
Saroj et al. (2022)	Machine Learning Algorithms for understanding the determinants of under-five Mortality [13]	Investigate the accuracy of predictivity of under-five mortality using machine learning models and identifying the significant contributors.	Neural network, multivariate logistic regression, random forest, decision tree, Naive Bayes, K-nearest neighbour (KNN), support vector machine (SVM), and ridge classifier.	The accuracy of the neural network exceeded the other models by 95.96% with the dominant predictors being number of living children, birth size, wealth index, and maternal education.	Coefficient and odds ratio are not included in machine learning models. The survey used for the dataset was non-specific and did not contain objectives regarding under-five mortality.
K. Harron *et al.* (2020)	Preterm birth, unplanned hospital contact, and mortality in infants born to teenage mothers in five countries: An administrative data cohort study [15]	Teenage mothers are more likely to give birth to children with health issues due to social disadvantage. However, how this issue varies across different countries remains unexplored.	Used Generalised Linear Models (GLMs) and Multilevel Model of statistical analysis. Sensitivity analysis is used to evaluate the impact of missing data.	Inverse relationship between maternal age and adverse infant outcomes. The risk of child mortality is 2 to 4 times higher for teenage mothers. Teenage mothers are 25%-40% more risky to have preterm birth.	There is limited socio-economic status measurement, lack of data about first-time motherhood, insufficient statistical power and data suppression due to privacy and protection purpose.
F. Iqbal, M. Islam Satti, A. Irshad, and M.	Predictive analytics in smart	Child death rate is a major issue in	Used Information Gain method for feature ranking.	Random Forest has the highest prediction	There is bias when mothers were recalling

Asif Shah (2023)	healthcare for child mortality prediction using a machine learning approach [16]	Pakistan and there is a lack of generic prediction framework to assess child mortality that could be helpful for healthcare systems in developing countries.	Used Decision Tree, Random Forest, Naive Bayes and Extreme Gradient Boosting	accuracy of 93.8%. Key risk factors related to child mortality were identified using the Information Gain method.	and reporting past events and some cause-specific mortality could not be determined from the survey data.

Appendix B

Flow Diagram or Research Framework



Appendix C

Table to summarize the quantitative results of different machine learning techniques.

Machine Learning Techniques	Accuracy	Precision	Recall	F1-Score	Confusion Matrix
Logistic Regression	0.6780	0.6790	0.6680	0.6730	[344 156] [164 330]
Decision Tree	0.6820	0.1576	0.6170	0.2511	[342 155] [18 29]
Random Forest	0.8971	0.3448	0.2128	0.2632	[478 19] [37 10]
Support Vector Machine (SVM)	0.6581	0.1401	0.5768	0.2253	[335 162] [11 36]
Naive Bayes	0.9176	0.5385	0.3043	0.3889	[476 12] [32 14]
eXtreme Gradient Boosting (XGBoost)	0.7800	0.1486	0.3425	0.2061	[311 186] [18 29]

Appendix D

Model Visualisation Results

Figure D1: Confusion Matrix – Logistic Regression

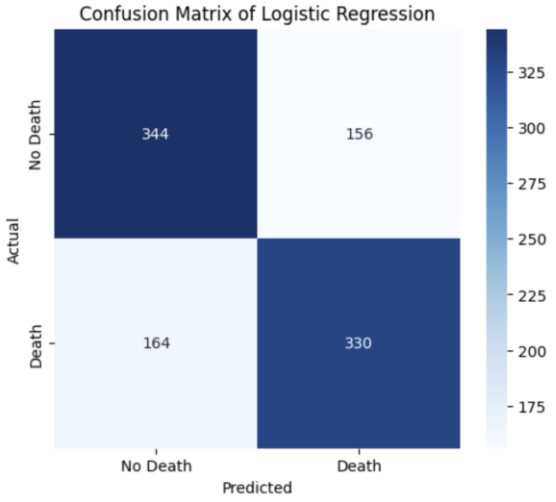


Figure D2: Bar Chart of Performance Metrics – Logistic Regression

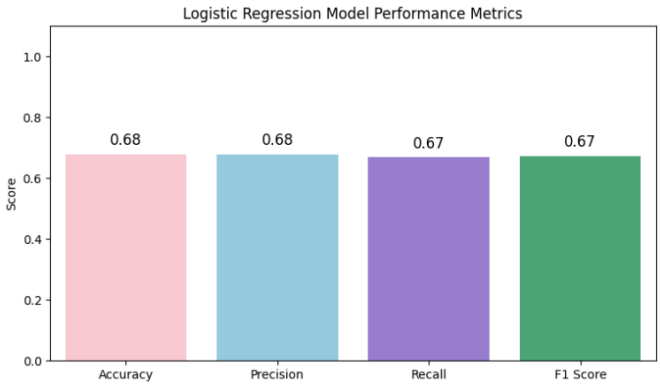


Figure D3: Confusion Matrix – Decision Tree

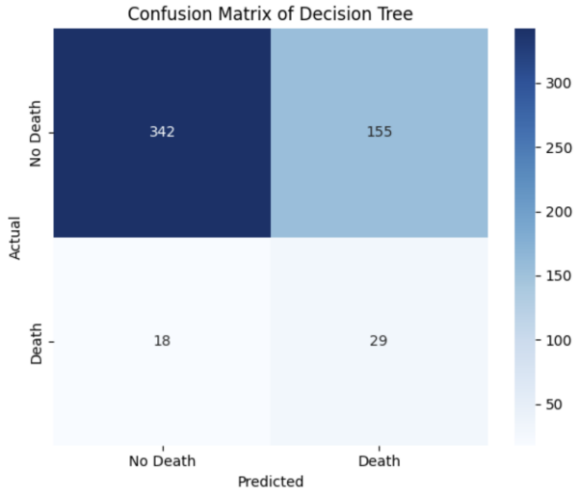


Figure D4: Bar Chart of Performance Metrics – Decision Tree

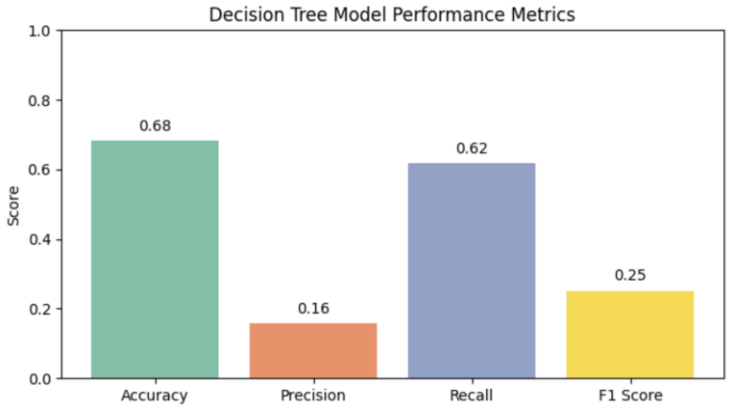


Figure D5: Confusion Matrix – Random Forest

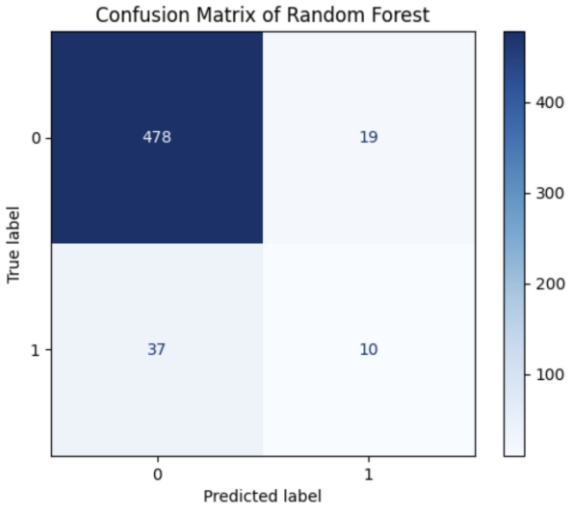


Figure D6: Bar Chart of Performance Metrics – Random Forest

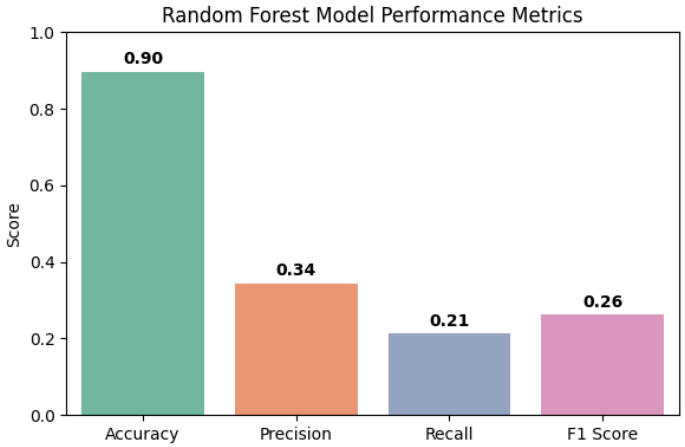


Figure D7: Confusion Matrix – Support Vector Machine (SVM)

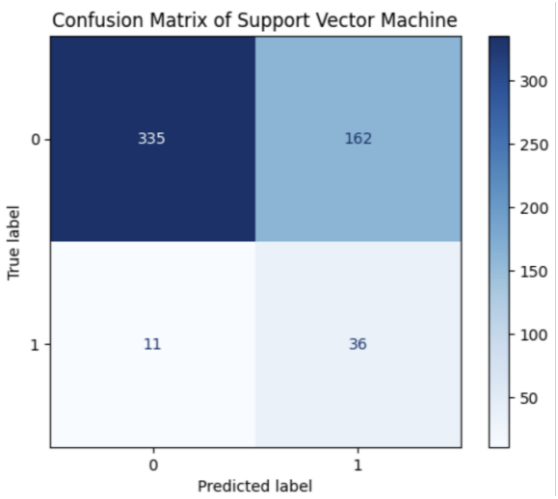


Figure D8: Bar Chart of Performance Metrics – Support Vector Machine (SVM)

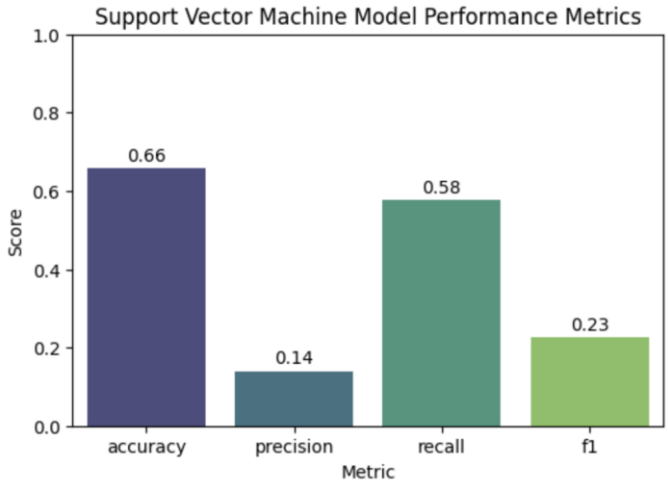


Figure D9: Confusion Matrix – Naive Bayes

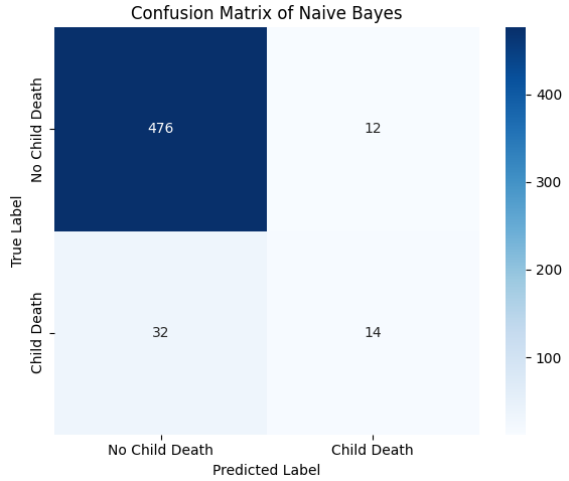


Figure D10: Bar Chart of Performance Metrics – Naive Bayes

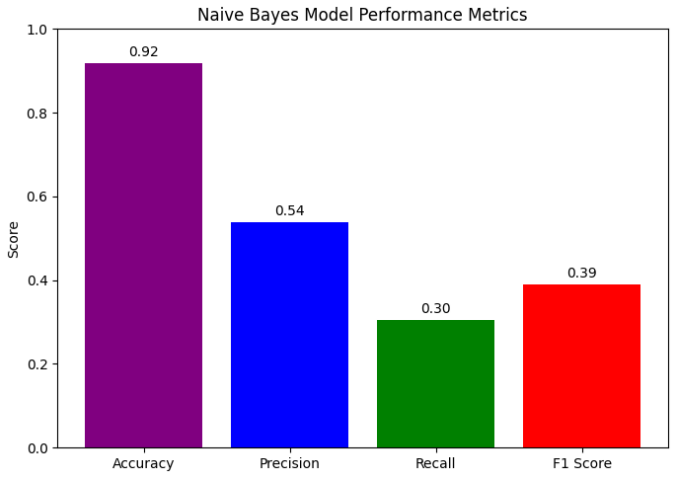


Figure D11: Confusion Matrix – XGBoost

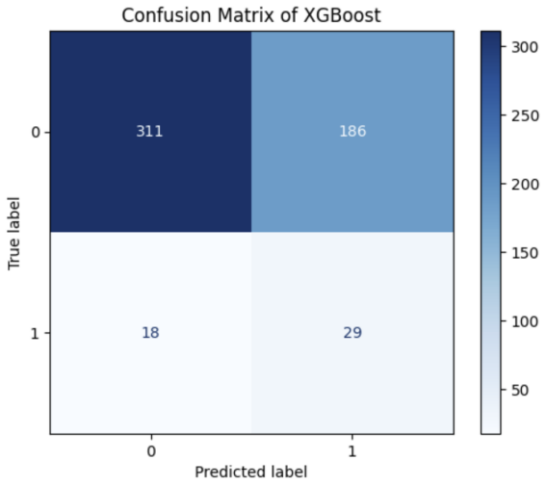


Figure D12: Bar Chart of Performance Metrics – XGBoost