ISSN Online: 3006-4708
ISSN Print: 3006-4694

SOCIAL SCIENCE REVIEW ARCHIVES
https://policyjournalofms.com

Early Childhood Deaths Prediction using Machine Learning and VAR Model

Fadia Shah[1], Yasir Shah[2], Faiza Shah[3], Imran Shahid[4], Aftab Hussain Tabasam[5]

[1] Shaheed Zulfikar Ali Bhutto Institute of Science and Technology, Islamabad, Pakistan.
Email: fadiashah13@yahoo.com
[2] School of Business, Zhengzhou University, Zhengzhou 450001, China.
Email: yasirshah_pk@yahoo.com
[3] School of Political Science and Public Administration, Henan Normal University, China.
Email: faizashah55@gmail.com
[4] School of Political Science and Public Administration, Henan Normal University, China.
Email: imranzzu87@gmail.com
[5] Business Administration, University of Poonch Rawalakot, Pakistan.
Email: aftabtabasam@upr.edu.pk

DOI: https://doi.org/10.70670/sra.v3i4.1266

Abstract

Health-related issues are always very serious. From 1st years there are always policies have been set to improve the health and education of individuals. This issue is more serious in developing countries. One of the problems of infant mortality due to lack of facilities, medicine, and financial concerns are key challenges in various parts of the world. All of these are counted under the area of standard development goals (SDG). International organizations such as the World Health Organization (WHO) are always making efforts to improve life standards every year. They always generate health reports publicly so that the annual improvements can be analyzed and new reforms can be made. The serious issue identified for children is the infant mortality rate, as well as child casualties, particularly under 5 years old. There could be many reasons for child mortality or infant death. They include socioeconomic factors. This issue is serious in both developing and non-developing countries. However, countries facing poor economic conditions are suffering more. The proposed study includes artificial intelligence(AI)-based algorithms under supervised learning, specifically using Naïve Bayes (NB) and XGBoost algorithms, which are considered highly efficient for this domain to identify the child death rate under age 5. Another statistical model was used for the comparative study. The third algorithm is the Vector Auto Regression (VAR) model, which is also famous for identifying regressive patterns in the data. The dataset was split into training and testing subsets, employing data balancing techniques such as Synthetic Minority Over-Sampling technique (SMOTE) for qualitative data generation. Machine learning classifiers, naïve Bayes, and extreme gradient boosting were used for the results deduction. A comparative analysis via a confusion matrix was performed for the performance evaluation. The results reveal the chances of health impairments when a child belongs to a certain statistical frame. The accuracy and precision of the results indicate the performance.

Keywords: Infant Casualty, Vector Auto Regression (VAR), Machine learning, Sustainable Development Goals (SDG), Predictive Analysis

Introduction

Health issues are global highlights and are crucial for developing countries. The key focus of world-recognized organizations is to improve the survival rate of infants by utilizing the available resources and services at their best. One way is to determine the number of infant deaths in a region and set health care reforms [1]. Despite global progress, child mortality remains a significant health challenge in developing nations. Policies and infrastructure maintenance have been upgraded, considering formal and informal ways, such as improving maternal awareness by educating people, especially those with low income and childbearing families, making efforts for the availability and access to clean water, and training and guidelines for appropriate healthcare management. Consequently, there seems to be a gradual reduction in infant casualties [2]. Ideally, it must ensure remarkable improvements, but the problem is not solved by these corrective actions. It is a fact that more than five million infants died in a few years. The exact number is high worldwide [3]. To tackle these problems, the United Nations set up goals called Sustainable Development Goals (SDGs) and announced them globally. The goal is to stop unnecessary child deaths within a reasonable timeframe.

This issue is taken very seriously, and policies have been planned to address these problems. Although many countries have made efforts to reduce child mortality, the highest child deaths are still found in countries such as Africa, Somalia, and Sudan. Some countries in Asia also have higher numbers, especially those with low income [4]. This clearly indicates that they failed to arrange suitable resources to overcome life-threatening situations. Some countries have shown a decline in death rates; thus, the corrective actions seem to be minimally effective. This study highlights the worldwide problem of infants dying early and stresses the importance of developing nations focusing on changing factors that contribute to child deaths. This is to meet the goals set by the SDG by 2030.

Health Demographics

Over the past six decades, developing countries have faced significant difficulties in economic growth. Commonly, there seems to be economic instability. There seem to be the least efforts marked under health and education reforms. These regions are unable to feed the population with food and other necessities of life. Therefore, the education and health sectors are suffering. While most countries have extended and strengthened healthcare and education facilities to the possible level, assuming this may increase the lifespan, there is still a need to cope with serious challenges in maternal and child health care; therefore, more improvement in this area demands more attention [5]. Lack of health facilities, approach towards health facilities, health awareness and education, provision of sufficient training, and systematic mechanisms cause the early death of a child, whose life can be saved. In this regard, figure 1 illustrates the numbers of children, young kids, and children under five dying, which remain very high. This shows that we need to act quickly to improve the situation.

Despite a decrease in the number of infant deaths, the situation is still not satisfactory. The rate decreased annually, but again, there was no stability in the picture. The condition seems beyond control, which was not expected if the corrective actions were useful. This means that there are not many significant policies, but challenges still exist that may lead to uncontrollable conditions [6].Numerous factors are responsible for fatal consequences in infants. One of these issues is the low latency rate. Families with no formal and traditional awareness and academic education often face higher chances of infant care. Simple unawareness or negligence may cause a small issue to become much bigger; this highlights the importance of the educational impact on one household as well as regional behavior. Similarly, the mother's age at birth is another significant factor, with mothers under 20 years facing increased likelihoods of child mortality, especially in children under five years. Another reason is the higher frequency of childbirth. Regions with seven or more children in a family are also associated with the possibility of child deaths. Additionally, wealth status contributes to this issue [7], with individuals in the lowest wealth quintile experiencing higher mortality rates. This may occur due to poor diet, hygiene, health facilitation, and numerous other factors. Ultimately, the result is always the lack of survival of the child. Understanding and addressing the complex factors influencing child mortality are essential for developing effective interventions and policies to improve child health outcomes in the country. Infant/childcare educates mothers and provides supplements to prevent adverse prenatal outcomes, while skilled health practitioners decrease the chances of health complexities caused by unprofessional health risks or even health hazards. Similarly, in the case of childbirth, postnatal care includes infant immunization and maternal education [8] on proper weaning practices, collectively contributing to the reduction of child mortality. These are also carried out professionally when there is health education.

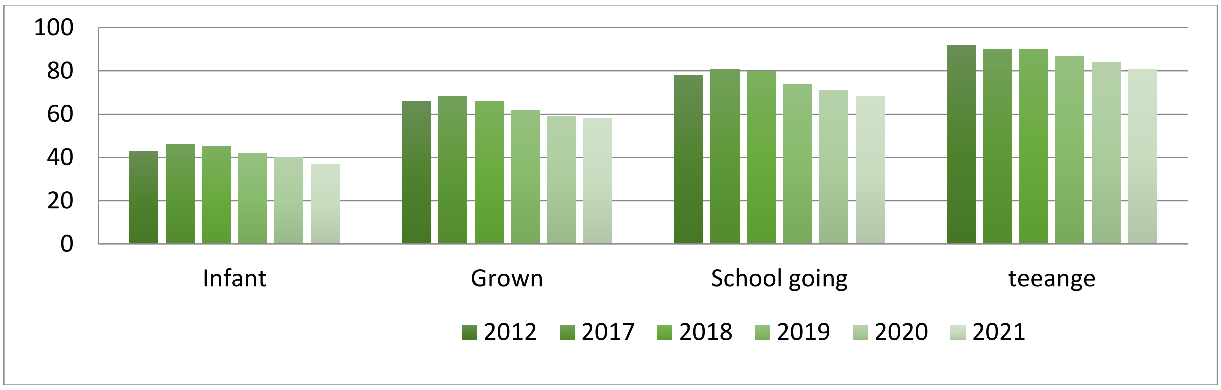


Figure 0: Under developing countries child health

The Statistical analysis of health indicators [9] represented as above in figure 1, indicates that there is a strong association between the death rate and that of a region. The graphs further indicate the infant, grown child, school-going child, and teenager as data samples. The gradual decrease in death rates from 2012 to 2021, with some intervals, reveals that there is a decrease in the death rates of these age groups, but there is still room for improvement.

Infant Health Hazards and Consequences

Since child mortality is a serious issue, various researchers have attempted to identify the real cause and data on child mortality. Research conducted by a scholar [9] has shown his research to deeply identify factors associated with child mortality. This study further used AI techniques to predict early child mortality by considering historical data. The research includes globally available data provided by healthcare reform agencies and selected various countries to support the data analysis. Study [10] is again in favor of the domain to identify child death considering previous data in such a way when there are no changes in the parameter. The scholar also used data analysis and advanced techniques to support his research. The study is considered important as it provides insights and specifically claims that available facilities in the health sector can save the child when maternal health is considered important in the beginning. Proper maternal care, medication, health awareness, and literacy towards improving the quality of life play an important role in saving the life of a child. This means that maternal awareness plays a significant role in saving the lives of neonates. If it continues, it is the responsibility of the parents to protect the child, especially in the early 5 years of his life.

In Asia, researchers have identified socioeconomic factors as the primary contributors to child mortality, highlighting their effects on child health. They utilized logistic regression (LR) and Principal Component Analysis (PCA) to analyze a health demographic dataset, revealing a complex interplay of influences on child health. The study [11] differentiated between national-level determinants, such as healthcare access, medical expertise, poverty, literacy, health policies, and welfare programs, and household-level factors. This thorough investigation offers a detailed perspective on the various elements affecting child mortality, providing essential information for targeted interventions and policy development in the future. Additionally, another innovative study established a link between child mortality and household air pollution (HAP) by applying multivariate logistic regression to country-level data. This study highlighted the association between cooking fuel consumption and child mortality [12]. Clean fuels, such as liquid petroleum gas and electricity, exhibited a protective effect, whereas polluting fuels, such as wood and animal dung, were correlated with increased child mortality risk. Similarly, maternal care, pregnancy health care, and awareness emerged as protective measures against child mortality. This study emphasizes the importance of appropriate antenatal care visits and interventions to improve public health outcomes for mothers and newborns.

Another study [13] calculated trends in child mortality using a deterministic approach and demographic health surveys. It was found that factors such as age at, total children born, and births in the last five years aimed to identify the impact of these factors on child death trends throughout the time span. This study highlighted the dynamic nature of child death factors, emphasizing the need for region-specific and time-sensitive interventions.

While these studies contribute in literature studies, as well as found with highlighting limitations. A study on household air pollution [14] in specific regions noted its country-level scope and the importance of further investigation at the regional level. These diverse studies underscore the multidimensional nature of child mortality factors [30]. The importance of responding based on variations in different areas and specific circumstances. Understanding these factors through predictive analytics not only informs effective public health strategies but also highlights that child mortality trends are always changing; therefore, continuous research is needed to tackle new challenges.

Problem and Research Objectives

The importance of studies related to child mortality analytics is crucial for regional studies. This is because the findings reflect the socioeconomic, health, and educational importance of these factors and their influence on various parts of the country. It is common to observe that wealthy countries have successful models for health, education, and development sectors. The health care parameter standards set by the WHO and UNICEF are continuously making efforts to improve survival rates [15] as well as quality of life. In the context of elevated infant mortality rates in the Asian subcontinent, prior investigations have predominantly relied on statistical methodologies to discern the socioeconomic and demographic factors associated with child mortality. Despite the increased use of AI to determine child mortality prediction in select studies [16], a notable omission persists regarding the essential determinants influencing under-five mortality. These include, but are not limited to, parameters such as the number of children under five in a household, frequency of antenatal care visits, total number of children ever born, and location of delivery.

This study aims to apply a statistical analysis of implementing supervised machine learning algorithms for child mortality rate. The scope further includes exploring feature-selection techniques for child mortality [17]. It aims to reason the influence of newly considered features for child mortality prediction. Similarly, the optimal feature set for predicting child mortality was obtained using a technique called Information Gain. The central idea of the proposed research is to enhance results through the incorporation of low-income factors of infant mortality within the existing dataset. The implementation of supervised machine learning classifiers for the prediction of mortality has useful statistical analysis. Assessment of the influence of identified risk factors on the overall performance of the model.

Literature Review

This section emphasizes the significance of the existing literature in the domain of healthcare statistical analysis, particularly focusing on the health of children under five years of age and those attending school. It includes an overview of predictive analytics techniques relevant to child healthcare, including a brief examination of data mining methods used to uncover hidden patterns [28]. Further studies are needed to analyze the statistical techniques aimed at identifying critical risk factors associated with mortality among children of school-going age. Further studies extensively discuss the datasets employed in previous studies. Finally, the later part of the literature review provides a concrete overview of the research gaps and significant challenges existing within this field, laying the groundwork for the study's objective of enhancing predictive models for under-five child mortality. This involves integrating risk factors and employing advanced statistical and predictive analytic techniques.

Fatal consequences of children [18] looked closely at how predictive analytics can be used to predict when children might die. We learn from different studies about the methods used to do so.

One study [19] compared different types of machine learning programs. They looked at those that followed straight lines and those that did not. Some well-defined algorithms for data analysis and pattern identification use machine learning algorithms. They looked at different things, such as how children were born, the mother's race, if there were twins, how much the baby weighed when born, and other health-related factors. This study highlights the importance of using predictive analytics to prevent child mortality.

Another study conducted by researchers [20] examined the number of children who died in Paraná, Brazil, and used computer models to predict this. They examined healthcare survey data and used a method to measure the accuracy of their predictions in percentages. They compared different computer models, such as linear regression and support vector machine (SVM) algorithms. They found that MLP was the best at making accurate predictions with the least amount of error. This study [29] emphasizes the importance of having precise predictive models to understand and address child mortality.

A study comparing underdeveloped countries in Central Asia showed how things to prevent children from dying were effective on one side [21]. Considering another region, they focused on using modern computer methods to predict child mortality. They wanted to help healthcare workers know when to act quickly to save children's lives, especially in places where there are not many healthcare workers. A study [22] concentrated on identifying high-entropy features for a low-age child death predictive model. In low-income and financially dependent countries, a study [23] compared two different ways of predicting whether a child under five might die. They found it difficult to use these methods with small sets of data and when the data had many different factors to consider. This suggests that different approaches may be necessary to handle healthcare data depending on the location and type of data available. Another study examined different ways to determine how breastfeeding affects whether a baby is breastfed [24,25]. They used a method called data mining to explore a large amount of information in large databases. Their goal was to improve breastfeeding practices and lower the number of children who die, showing how helpful predictive analytics can be for the health of mothers and children [26]. Although these studies offer helpful information, they also highlight some challenges, such as problems with getting data ready, limited use of the findings, and the importance of healthcare experts working together with data analysts. This shows how predictive analytics in child mortality research is changing and requires more progress and teamwork to solve healthcare problems.

Proposed Model Data Preprocessing

The datasets collected for the current study are available in an open-access directory. This dataset contains features such as age, maternal health, education, and family economic conditions. There are a total of 13380 record sets. Some of these belong to the above-mentioned dataset, and other records include the Get Well Clinic and Maternal Health Center. Organizations that are actively working on SDG believe in qualitative statistical health care analysis. They execute the work of collecting population health data from developing nations.

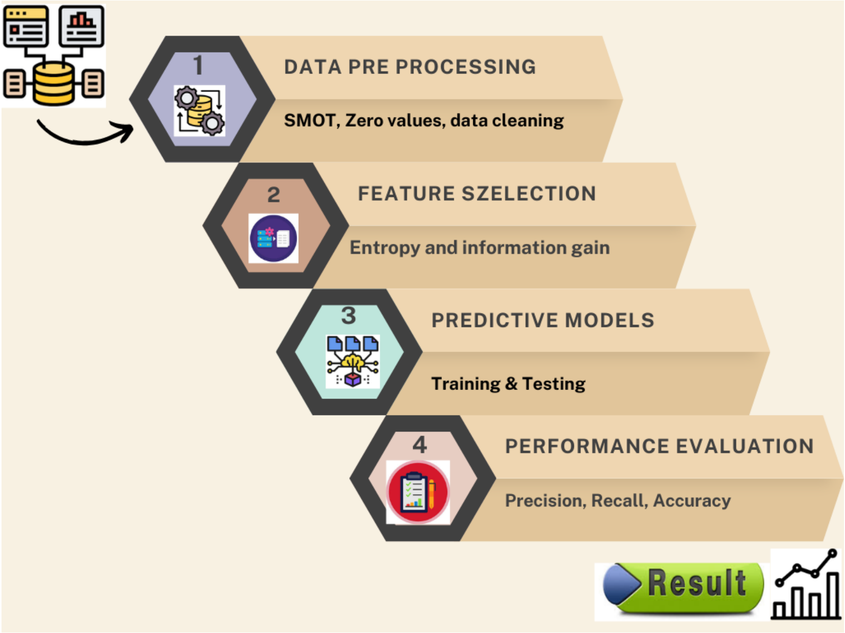


Figure 2: Process model

Figure 2 illustrates the system execution. The process model is initiated by selecting a dataset. The data were then processed to convert the quantitative data into qualitative data. This includes deleting incomplete values, but with a minimum effect. To determine the mean values that generate an accurate impact, the data after this preprocessing are reliable. Data preprocessing is crucial for obtaining accurate results. Qualitative data are preferred for scientific health care analysis; therefore, only relevant data were selected. Irrelevant data are noise that causes anomalies. The process involves cleaning, normalization, feature selection, and handling errors, duplicates, outliers, and missing values in the dataset to ensure reliable outcomes [25].

Table 1: Process flow and outcomes

Input	Processing	Outcomes
Data set	**Methodology initialization, confirming data set, data preprocessing and cleaning. Data purity for qualitative results.**	**Qualitative Data**
Qualitative Data	Feature selection via threshold and entropy. Choosing most valuable features to generate remarkable output.	Data subset
Data subset	Model training and testing with the ratio of 70 and 30. The data execution and processing.	Train and Test data
Classifier Algorithms	The XGB and Naïve Bayes classifiers execution and result generation.	Confusion Matrix
Results	Performance measurement through accuracy, precision and F1 scores.	Performance calculations

There are numerous methods for handling missing data. One approach involves removing attributes and records with a high number of missing values. Another method is to label the missing values as unknown. Precise imputation methods include using the mean, mode, and median. The mode is suitable for categorical variables, the mean is for continuous features, and the median is for handling outliers. Table 1 indicates the process flow with respect to the input and outcome of the module. Machine learning algorithms, such as Predictive Mean Matching (PMM), offer a more sophisticated approach. PMM reduces bias by replacing missing values with real values obtained from similar data. For categorical features, Logistic Regression was implemented to impute binary classes.

Missing Values Replacement

This study used data from the 2018 Pakistan Demographic and Health Survey, with 13798 instances and 17 independent features. The key features determine the information that is gained. This indicates which columns are most essential for the experiment and which have the least impact.

Table 2: Missing values and interpretation

Features	Missing Values	Algorithm
Diet Rich Families	33	Logistic Regression
Educated Families	31	Logistic Regression
Regular Hospital Consultancy	28	Predictive Mean Matching
Families with Income Above Average	21	Predictive Mean Matching
Neonatal Jaundice	4	Logistic Regression
Family with Medical Awareness	1	Logistic Regression

From Table 2, it is observed that three independent features have 4372 (35.0 per cent) missing values. Preceding birth interval had 3007 (24.1 per cent) missing values, the presence of diarrhea had 701 (5.6 per cent) missing values, and father’s education had 131 (1.0 per cent) missing values. Missing values were replaced with multiple imputations using IBM SPSS Statistics 22.

Feature Selection

Another important aspect of the model is feature selection. There are 21 features in the dataset, but it is useless to use all of them. This is time-consuming, and often, all the columns are not very helpful in concluding the results. Therefore, choosing the columns that provide the most useful results is an entropy calculation, and this approach is used for the predictive algorithm. Feature selection is vital for identifying quality features in a dataset, reducing dimensionality, and preventing overfitting. Information Gain, which measures entropy, helps evaluate the impact of each attribute on the output variable. High-entropy features boost system performance, whereas low-entropy features can be ignored. This aids in testing hypotheses and predictive models, as shown in figure 3, the country-wise data analysis, by focusing on significant features and removing less important ones. The data used for this purpose were globally available [31] on Kaggle, including countries and child death information.

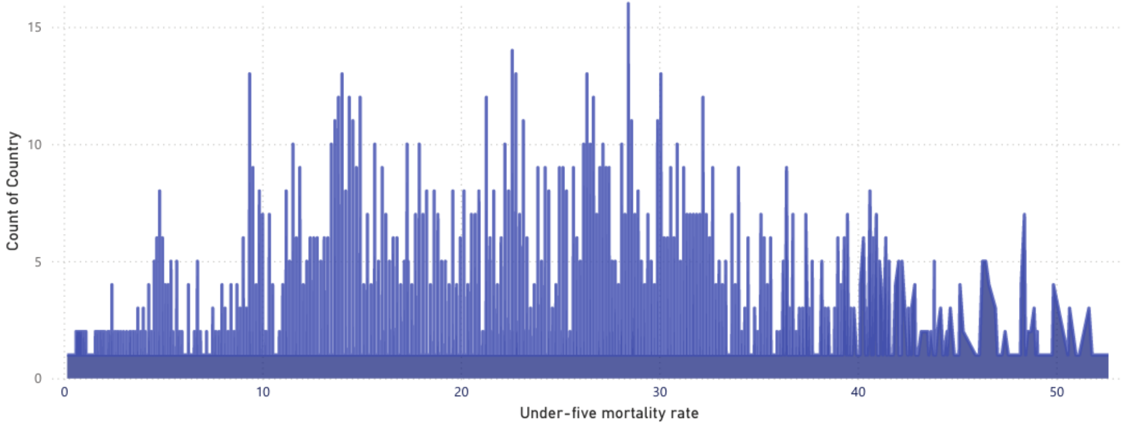


Figure3: Regional child mortality

The concept of information gain is determined through the above table, which indicates crucial features for predicting under-five child mortality, as depicted in figure 3. This means that the features that are more impactful are considered. Although postnatal checkups and mother education are not significant for information gain. Therefore, the columns with the highest entropy were chosen. This histogram illustrates the distribution of under-five mortality rates (deaths per 1,000 live births) across a group of countries, showing that the majority of nations have a relatively low rate, as indicated by the tallest bar on the left, representing approximately 15 countries, while progressively fewer countries suffer from higher mortality rates, as seen by the successively shorter bars extending to the right.

ML Classifiers

ML algorithms can be used to train machines for result generation. Supervised learning is a type of learning in which a labeled dataset is used. Among these columns, one is the target column, which contains the known results [27]. In the proposed research, three classifiers are used: Decision Tree (DT), Extreme Gradient Boosting (XGB), and Naïve Bayes (NB). These are standard algorithms that are efficient in terms of the results. Among the algorithms, feature selection serves as a crucial tool for selecting pertinent features within a dataset, and is applicable across various models, from linear regression to complex algorithms such as XGBoost. This method not only validates the model's correctness but also fortifies it by focusing solely on essential variables, thereby enhancing the comprehension of the model's underlying logic [32]. By utilizing feature importance, less significant parameters can be identified and potentially eliminated, leading to improved training efficiency and comparable or better performance than the baseline.

Extreme Gradient Boosting (XGB)

XGBoost is a highly efficient machine learning algorithm based on the gradient boosting framework, which is optimized for both speed and performance. XGBoost builds an ensemble of decision trees sequentially, where each new tree corrects the errors made by the previous trees to improve the predictive performance. One of the main advantages of this algorithm is that it easily works on large datasets. This research is highly capable because it can manage the problem of overfitting. In addition, complicated datasets can be handled. It can efficiently manage missing values. The results generated were highly reliable. It can generate unseen patterns from early child disease identification and mortality datasets.

Naïve Bayes (NB)

Naïve Bayes (NB) encompasses a set of probabilistic classification algorithms based on Bayes' Theorem. The "naïve" aspect refers to the assumption that the features involved in the classification are independent of each other, which simplifies the computation process. These classifiers calculate the probability of each class given the input features and select the class with the highest probability as the predicted outcome of the model. There are several variations of Naïve Bayes, including Gaussian Naïve Bayes for continuous data, Multinomial Naïve Bayes for discrete data (often used in text classification), and Bernoulli Naïve Bayes for binary data. This algorithm is popular because of its simplicity, speed, and effectiveness, particularly in managing large and high-dimensional datasets, making it suitable for tasks such as spam detection, sentiment analysis, and document classification. Despite the independence assumption, Naïve Bayes frequently produces strong results in practical applications.

Naïve Bayes is particularly effective for predicting under-five child mortality for several reasons. First, it is adept at handling categorical data, which are prevalent in health-related datasets that include demographic information, health indicators, and socioeconomic variables. Its probabilistic approach allows for the incorporation of various risk factors, aiding the assessment of mortality likelihood based on these inputs. Naïve Bayes is computationally efficient, allowing for the rapid analysis of large datasets, which is crucial in public health research, where timely insights are essential. The algorithm's straightforward nature also facilitates easy interpretation of the results, enabling stakeholders to understand the factors influencing child mortality. One aspect is that even with the assumption of feature independence, Naïve Bayes often provides reliable predictions, offering valuable insights that can inform interventions and policies aimed at reducing child mortality rates.

Decision Tree (DT)

Decision tree algorithms are known for their efficiency, primarily because of their simplicity and ease of interpretation, which allows users to understand the decision-making process through a clear visual format. As non-parametric models, they do not depend on specific assumptions regarding data distribution and can effectively handle both numerical and categorical variables without extensive pre-processing. Decision trees automatically identify the most significant features, making them robust against outliers and effective in noisy environments. They are also scalable, making them suitable for large datasets, and can be used for both classification and regression tasks. Additionally, decision trees serve as the foundation for powerful ensemble methods, such as Random Forests and Gradient Boosting, which enhance predictive accuracy and reduce the risk of overfitting. Their ability to manage missing values efficiently and provide rapid predictions further contributes to their popularity in various domains.

Decision tree algorithms are particularly useful for predicting child mortality because of their clear visual representation and capability to manage complex, non-linear datasets. They can identify critical risk factors, do not require data normalization, and are resilient to outliers, making them instrumental in shaping public health strategies and improving child health outcomes.

Statistical Analysis Model Vector Auto Regression (VAR)

The Vector Autoregression (VAR) model is a statistical method used to analyze multivariate time series data, emphasizing the linear relationships among various time series variables. Unlike univariate models that focus on a single time series, VAR examines multiple variables simultaneously, offering a more comprehensive understanding of their interconnections over time. In this model, each variable is expressed as a linear function of its own past values and the past values of all other variables included in the analysis [33].

This approach is particularly advantageous in areas such as econometrics and finance, where it is used for forecasting and exploring the dynamic relationships between variables such as the GDP, inflation, and interest rates. It is also effective for impulse response analysis, which examines how a shock to one variable affects others over time, and for variance decomposition, which assesses the contribution of each variable to the overall forecast error variance. In essence, the VAR model serves as a powerful tool for understanding the complex relationships within multivariate time-series data, making it crucial for both analysis and forecasting.

When applied to predicting early child survival, the VAR model is particularly useful because it captures the intricate interrelationships among various factors that influence child health. By analyzing time-dependent data, it provides insights into how aspects such as maternal health, nutrition, healthcare access, and socioeconomic conditions interact and impact child survival rates. The model's ability to conduct impulse response analysis allows researchers to assess the effects of sudden changes in one variable on child survival, while its forecasting capabilities enable predictions based on historical data, which are vital for developing informed health policies in the region. The requirement for stationarity in the VAR model ensures that the relationships being analyzed remain stable over time, thereby enhancing the reliability of the predictions. Overall, its capacity to evaluate the influence of multiple factors makes the VAR model a valuable resource for improving early child survival outcomes.

Results & Discussion

The proposed research results are obtained using the Naïve Bayes algorithm for its ease and speed in predicting the class of the test dataset. Naïve Bayes performs effectively even with small amounts of data and is recognized as one of the simplest algorithms for predicting results. The Naïve Bayes

Classifier was trained using 70% of the training data, and the remaining 30% was utilized for testing and assessing the model's generalization. For the applied results, the dataset was split into training and testing sets, with 70% training and 30% testing categories. Out of 13380 there are 9366 instances in the training data and 4014 instances in the testing data. These instances are preprocessed, missing values are interpreted, and finally, the evaluation is performed by the XGB and NB classifiers.

Comparison of Information Gain

Table 3 displays the outcomes of the Extreme Gradient Boosting classifier on a balanced dataset, featuring the top 17 features ranked by information gain, along with the top 15 features ranked by information gain.

Table3: Results with Information Gain

	DT	XGB	NB	VAR
Child death	0.783	0.935	0.812	0.857
Maternal Health	0.922	0.983	0.934	0.935
Maternal Education	0.893	0.894	0.801	0.793
Family Income	0.890	0.953	0.834	0.828

The information gain through feature analysis was observed during the experimentation. Among the 17 chosen features, the XGBoost classifier identified 231 children with health hazards out of 322. Similarly, the accurately predicted healthy children were 2734from 3327 record sets. The confusion matrix further identified that the false negatives were 34 children, which was true otherwise.

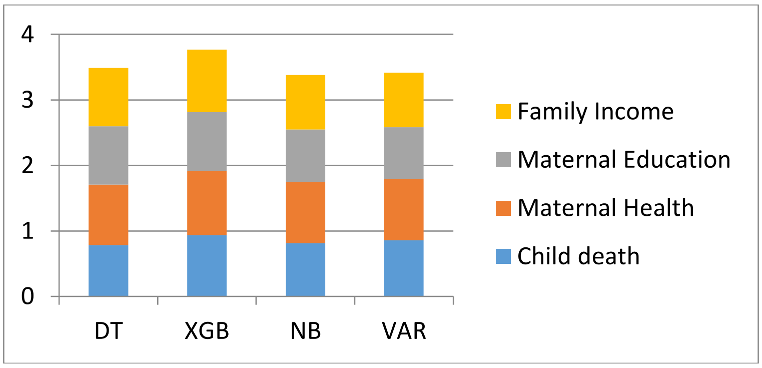


Figure 4: Information gain

Figure 4 indicates that the major entropy is calculated by maternal health at the highest level, followed by maternal education as the second most important factor that leads to better chances of a child's survival, child health as the third contributing factor, and family income with a low proportion in information gain. The false positives were 324 children who were alive but predicted to be dead by the XGBoost model. Later, it was also found that the false negatives were correctly found, but the false positives increased after the information gain analysis. Thus, the results with selected features are better in the case of the Naïve Bayes Classifier to predict mortality among children below the age of five.

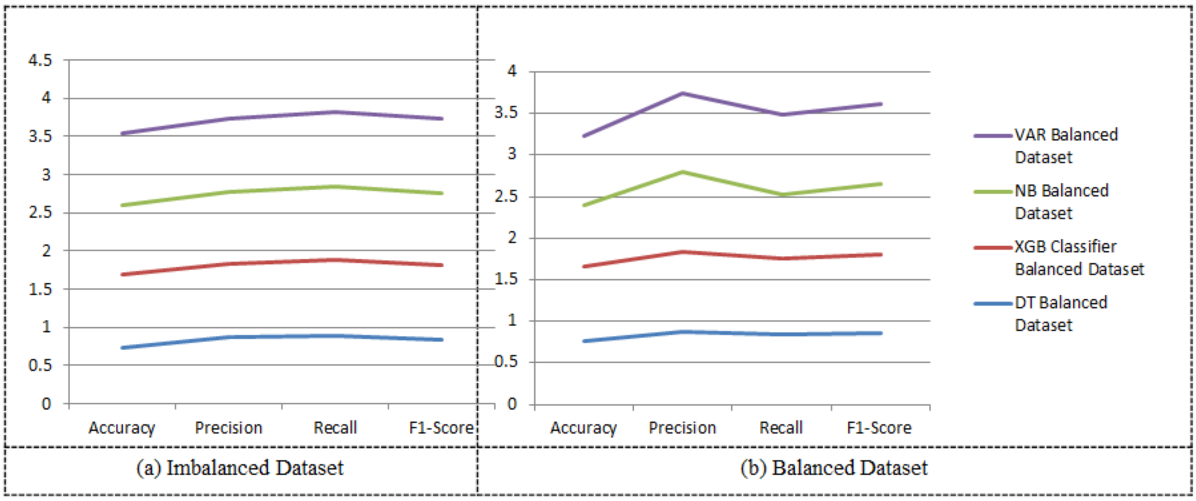
Performance Evaluation

Gradient Boosting Classifier was trained on 70 per cent of training data and 30 per cent of data used for testing and to check generalization of model. Table 3 shows the results of the extreme gradient boosting classifier for the imbalanced and balanced datasets. Various performance metrics are used to evaluate the Extreme Gradient Boosting Classifier such as Accuracy, Precision, Recall and F1-Score.

Table4: Algorithms Performance

	DT		XGB Classifier		NB		VAR	
	Imbalanced Dataset	Balanced Dataset	Imbalanced Dataset	Balanced Dataset	Imbalanced Dataset	Balanced Dataset	Imbalanced Dataset	Balanced Dataset
Accuracy	0.735	0.763	0.954	0.892	0.915	0.745	0.943	0.836
Precision	0.865	0.871	0.959	0.973	0.952	0.953	0.955	0.938
Recall	0.883	0.846	0.994	0.911	0.959	0.769	0.974	0.957
F1-Score	0.832	0.863	0.976	0.941	0.955	0.851	0.966	0.952

From the above table, we can elaborate that on imbalanced data, the extreme gradient boosting classifier performed well, but when we applied Smote, Accuracy, Recall and F1 score decreased. Before applying SMOTE, correctly predicted dead children were ‘50’ out of ‘201’ and correctly predicted alive children were ‘4322’ out of ‘3847.’ False negatives were ‘234’ children who were dead but predicted to be alive by the extreme gradient boosting model. False positives were ‘38’ children who were alive but predicted dead by the extreme gradient boosting model. After applying SMOTE, correctly predicted dead children were ‘275’ out of the ‘330’ and correctly predicted alive children were ‘3557’ out of the ‘3659.’ Figure 5 shows the performance when using balanced and imbalanced datasets. The left side of Figure 5,(a), shows the results using imbalanced data, but (b) the graph is plotted after preprocessing and clean data. False-negative children who were dead but predicted to be alive by the extreme gradient boosting classifier. False positives were ‘312’ children who were alive but predicted dead by the extreme gradient boosting classifier. We observed that false positives increased, whereas false negatives decreased after applying the synthetic minority over-sampling technique (SMOTE).



(a) Imbalanced Dataset

(b) Balanced Dataset

Figure 5: Balanced and Imbalanced dataset performance analysis

It is concluded that on imbalanced data, the naïve Bayes classifier performed well, but when we applied Smote, Accuracy, Recall and F1 score decreased. Figure 5 above indicates that before applying

SMOTE, correctly predicted dead children were ‘34’ out of ‘275’ and correctly predicted alive children were ‘3399’ out of ‘4385.’ False negatives are ‘234’ children who were dead but predicted to be alive by the naïve Bayes model. False positives are ‘221’ children who were alive but predicted dead by the naïve Bayes model. After applying SMOTE, correctly predicted dead children were ‘67’ out of the ‘198’ and correctly predicted alive children were ‘3487’ out of the ‘3635.’ False negatives are ‘230’ children who were dead but predicted to be alive by the naïve Bayes classifier. False positives are ‘956’ children who were alive but predicted dead by the naïve Bayes classifier. It is observed that false positives are increased whereas false negatives are reduced after applying data refining techniques.

Conclusion

This research paper explores the domain of child health reform statistical analysis and highlights some of the core challenges in this domain. This study was conducted to overcome the gap in determining the child death rate as a challenge in data presentation. The Maternal, Newborn, and Child Health Program is a fast-flowing, continuous, and large data source. This study identified that the risk factors associated with infant mortality are significant for the prediction of child mortality. The Research used ML algorithms ie Extreme Gradient Boosting Classifier (XGB) and Naïve Bayes Classifier (NB). The models were trained on the training data and tested on hidden data. The performances of the three classification models, XGB Classifier, Naive Bayes (NB), and the Proposed Model, were compared on both imbalanced and balanced datasets. The results show that the XGB Classifier achieved an accuracy of 0.954 on the imbalanced dataset and 0.892 on the balanced dataset. The model demonstrated a precision of 0.959 and 0.973 on the imbalanced and balanced datasets, respectively, with a recall of 0.994 and 0.911 for the imbalanced and balanced datasets, respectively, leading to F1-scores of 0.976 and 0.941, respectively. In contrast, the Naive Bayes model achieved an accuracy of 0.915 and 0.745 on the imbalanced and balanced datasets, respectively, with a precision of 0.952 and 0.953 on the imbalanced and balanced datasets, respectively. The recall values were 0.959 and 0.769 for the imbalanced and balanced datasets, respectively, resulting in F1-scores of 0.955 and 0.851, respectively. The Proposed Model showed an accuracy of 0.943 and 0.836 on the imbalanced and balanced datasets, respectively, achieving a precision of 0.955 and 0.938 on the imbalanced and balanced datasets, respectively. The recall for this model was 0.974 on the imbalanced dataset and 0.957 on the balanced dataset, yielding F1-scores of 0.966 and 0.952, respectively. Overall, the models performed better on balanced datasets, with the Proposed Model exhibiting the highest performance in terms of precision, recall, and F1-score on the balanced dataset. Additionally, data were collected from mothers with living children, excluding those with deceased children. A limitation of the study is the lack of longitudinal data, and some respondents may have changed their place of residence since their child's birth at the time of the survey.

References:

Hasan F, Tantawi ME, Haque F, Foláyan MO, Virtanen JI. Early childhood caries risk prediction using machine learning approaches in Bangladesh. BMC Oral Health. 2025 Jan 8;25(1):49.

Hasan F, Tantawi ME, Haque F, Foláyan MO, Virtanen JI. Early childhood caries risk prediction using machine learning approaches in Bangladesh. BMC Oral Health. 2025 Jan 8;25(1):49.

Ding Y, Deng A, Qi TF, Yu H, Wu LP, Zhang H. Burden and trend prediction of ischemic heart disease associated with lead exposure: Insights from the Global Burden of Disease study 2021. Environmental Health. 2025 Dec;24(1):1-9.

S. F., et al. AI Based Cardiovascular Disease Prediction using Ensemble Learning. Research Journal for Social Affairs 3 (5), 211-223.2025.

Espinola-Sánchez M, Campaña-Acuña A, Urrunaga-Pastor D, Maguiña JL, Jumpa M, Ugarte-Ubillus O. Impact of Comprehensive Health Insurance affiliation on mortality in children under one year: an analysis of the Demographic and Health Survey 2010–2022 in Peru. Frontiers in Public Health. 2025 Jan 23;12:1405244.

Huang L, Zhang D, Liu M. Global trends in refractive disorders from 1990 to 2021: insights from the global burden of disease study and predictive modeling. Frontiers in Public Health. 2025 Mar 26;13:1449607.

Abdelouahed M, Yateem D, Amzil C, Aribi I, Abdelwahed EH, Fredericks S. Integrating artificial intelligence into public health education and healthcare: insights from the COVID-19 and monkeypox crises for future pandemic readiness. InFrontiers in Education 2025 Apr 17 (Vol. 10, p. 1518909). Frontiers Media SA.

S. F., et al. Hadoop with Wavelet support for medical big data. 2021 18th International Computer Conference on Wavelet Active Media.

Khan I, Gunwant DF. “Revealing the future”: an ARIMA model analysis for predicting remittance inflows. Journal of Business and Socio-economic Development. 2025 Apr 1;5(2):155-70.

Wang Z, Ma K, Zhu Y, Li Z, Li S. Predictive value of myocardial markers for early postoperative mortality in children with congenital heart disease. Pediatric cardiology. 2025 Feb;46(2):324-31.

Rees CA, Haggie S, Florin TA. Narrative review of clinical prediction models for paediatric community acquired pneumonia. Paediatric respiratory reviews. 2025 Jan 16.

S. F., et al. Broad big data domain via medical big data. 2017 4th International Conference on Systems and Informatics (ICSAI), 732-737

Samuel O, Zewotir T, North D. Application of machine learning methods for predicting under-five mortality: analysis of Nigerian demographic health survey 2018 dataset. BMC Medical Informatics and Decision Making. 2024 Mar 25;24(1):86.

Gou H, Song H, Tian Z, Liu Y. Prediction models for children/adolescents with obesity/overweight: A systematic review and meta-analysis. Preventive Medicine. 2024 Feb 1;179:107823.

Reza TB, Salma N. Prediction and feature selection of low birth weight using machine learning algorithms. Journal of Health, Population and Nutrition. 2024 Oct 12;43(1):157.

Ekundayo F, Nyavor H. AI-driven predictive analytics in cardiovascular diseases: Integrating big data and machine learning for early diagnosis and risk prediction. International Journal of Research Publication and Reviews. 2024;5(12):1240-56.

Gray MM, Malay S, Kleinman LC, Stange KC, Borawski EA, Shein SL, Slain KN. Child opportunity index and hospital utilization in children with traumatic brain injury admitted to the PICU. Critical care explorations. 2023 Feb 1;5(2):e0840.

Cotache-Condor C, Rice HE, Schroeder K, Staton C, Majaliwa E, Tang S, Rice HE, Smith ER. Delays in cancer care for children in low-income and middle-income countries: development of a composite vulnerability index. The Lancet Global Health. 2023 Apr 1;11(4):e505-15.

Alzakari SA, Alhussan AA, Qenawy AS, Elshewey AM, Eed M. An enhanced long short-term memory recurrent neural network deep learning model for potato price prediction. Potato Research. 2025 Mar;68(1):621-39.

Zhao M, Huang X, Zhang Y, Wang Z, Zhang S, Peng J. Predictive value of the neutrophil percentage-to-albumin ratio for coronary atherosclerosis severity in patients with CKD. BMC cardiovascular disorders. 2024 May 28;24(1):277.

Yang X, Liu C, Liu Y, He Z, Li J, Li Y, Wu Y, Manyande A, Feng M, Xiang H. The global burden, trends, and inequalities of individuals with developmental and intellectual disabilities attributable to iodine deficiency from 1990 to 2019 and its prediction up to 2030. Frontiers in Nutrition. 2024 Jun 17;11:1366525.

Onsay EA, Rabajante JF. Measuring the unmeasurable multidimensional poverty for economic development: Datasets, algorithms, and models from the poorest region of Luzon, Philippines. Data in Brief. 2024 Apr 1;53:110150.

Li J, Jia H, Liu Z, Xu K. Global, regional and national trends in the burden of low bone mineral density from 1990 to 2030: A Bayesian age-period-cohort modeling study. Bone. 2024 Dec 1;189:117253.

Shen H, Zhao H, Jiang Y. Machine learning algorithms for predicting stunting among under-five children in Papua New Guinea. Children. 2023 Sep 30;10(10):1638.

Akter S, Voumik LC, Rahman MH, Raihan A, Zimon G. GDP, health expenditure, industrialization, education and environmental sustainability impact on child mortality: Evidence from G-7 countries. Sustainable Environment. 2023 Dec 31;9(1):2269746.

Duong SQ, Elfituri MO, Zaniletti I, Ressler RW, Noelke C, Gelb BD, Pass RH, Horowitz CR, Seiden HS, Anderson BR. Neighborhood childhood opportunity, race/ethnicity, and surgical outcomes in children with congenital heart disease. Journal of the American College of Cardiology. 2023 Aug 29;82(9):801-13.

Xiao L, Zhang Y, Xu X, Dou Y, Guan X, Guo Y, Wen X, Meng Y, Liao M, Hu Q, Yu J. Predictive model for early death risk in pediatric hemophagocytic lymphohistiocytosis patients based on machine learning. Heliyon. 2023 Nov 1;9(11).

Hackelöer M, Schmidt L, Verlohren S. New advances in prediction and surveillance of preeclampsia: role of machine learning approaches and remote monitoring. Archives of gynecology and obstetrics. 2023 Dec;308(6):1663-77.

Zhuo J, Harrigan N. Low education predicts large increase in COVID-19 mortality: the role of collective culture and individual literacy. Public Health. 2023 Aug 1;221:201-7.

Al-Tashi Q, Saad MB, Muneer A, Qureshi R, Mirjalili S, Sheshadri A, Le X, Vokes NI, Zhang J, Wu J. Machine learning models for the identification of prognostic and predictive cancer biomarkers: a systematic review. International journal of molecular sciences. 2023 Apr 24;24(9):7781.

Favril L, Yu R, Geddes JR, Fazel S. Individual-level risk factors for suicide mortality in the general population: an umbrella review. The Lancet Public Health. 2023 Nov 1;8(11):e868-77.

Assaduzzaman M, Al Mamun A, Hasan MZ. Early prediction of maternal health risk factors using machine learning techniques. In2023 international conference for advancement in technology (ICONAT) 2023 Jan 24 (pp. 1-6). IEEE.

Eslami M, Pourghazi F, Khazdouz M, Tian J, Pourrostami K, Esmaeili-Abdar Z, Ejtahed HS, Qorbani M. Optimal cut-off value of waist circumference-to-height ratio to predict central obesity in children and adolescents: A systematic review and meta-analysis of diagnostic studies. Frontiers in Nutrition. 2023 Jan 4;9:985319.