Little Lives, Big Data: Predicting Under-5 Mortality Danger Zones

1. Business Understanding

2. Project Overview

3. Business Objectives

4. Success Criteria

5. Data Understanding

6. Overview

7. Data Description

8. Verifying Data Quality

9. Data Preparation

10. Preprocessing

11. Modeling

12. Evaluation

13. Deployment

Resources
Github repository:https://github.com/sammaash/mortality/tree/main/SARIMAX

Members
Sammy Macharia Sammy.macharia@student.moringaschool.com
Mercy Cherotich mercy.cherotich@student.moringaschool.com
Amadi Growman amadi.growman@student.moringaschool.com
Caroline Ngabu caroline.ngabu@student.moringaschool.com
Gideon Biwott gideon.biwott@student.moringaschool.com
Beryl Saoke beryl.saoke@student.moringaschool.com

PROJECT SUMMARY

In this project, we focus on understanding and addressing the critical issue of infant mortality rates using data-driven approaches. Our goal is to uncover insights that can inform targeted interventions and policies aimed at reducing under-5 mortality rates globally.

We begin by collecting and analyzing comprehensive datasets related to child mortality, including causes of death, demographic factors, and socioeconomic indicators. This data, sourced from various sources such as Kaggle datasets and web-scraped data, forms the foundation of our analysis.

Using Python libraries like pandas, numpy, and scikit-learn, we preprocess the data, ensuring its quality and reliability for analysis. We address challenges such as missing values, inconsistencies, and data type validation to optimize the dataset for further exploration.

Exploratory data analysis (EDA) plays a crucial role in understanding the patterns and trends in under-5 mortality rates. We examine cause-specific mortality data, economic indicators, and demographic factors to identify correlations and insights that can guide our analysis.

For modeling, we employ a variety of techniques, including clustering, logistic regression, and time series analysis. Clustering helps identify high-risk populations, while logistic regression allows us to classify under-5 mortality risk levels based on cause-specific data. Time series analysis, particularly SARIMAX modeling, helps us understand temporal trends and forecast future mortality rates.

Evaluation of our models is conducted using appropriate metrics such as accuracy, precision, recall, and F1-score for classification models, and root mean square error (RMSE) for time series forecasting. Visualizations are created to facilitate interpretation and comparison of results.

By combining meticulous data preparation, advanced modeling techniques, and thorough evaluation, our project aims to provide actionable insights for stakeholders, policymakers, and healthcare professionals. Ultimately, we aspire to contribute to global efforts aimed at reducing under-5 mortality rates and ensuring the well-being of children worldwide.

BUSINESS UNDERSTANDING

Stakeholder: Public Health Authorities, NGOs, Healthcare Organizations.

Under five mortality rate, the number of children who die before reaching the age of five, is a critical indicator of a population's health and well-being. High under five mortality rates are often associated with inadequate healthcare access, poor nutrition, and socioeconomic challenges. Understanding the factors influencing under five mortality rates is essential for policymakers and healthcare providers to implement targeted interventions and improve child survival rates.

Our stakeholders are entities committed to improving child health outcomes and reducing under-5 mortality rates globally. They include public health authorities, non-governmental organizations (NGOs), and healthcare organizations dedicated to addressing child mortality as a public health priority.

BUSINESS PROBLEM.

Under-5 mortality is a significant global health concern. High under-five mortality reflects the loss of young lives representing a significant humanitarian tragedy. It often indicates weaknesses in healthcare infrastructure, inadequate access to essential health services, and disparities in healthcare delivery. Moreover, the loss of young lives deprives communities of their future workforce and potential contributors to economic development. Notably, many of the causes of under-five mortality, such as infectious diseases, malnutrition, and inadequate maternal healthcare, are preventable and treatable with cost-effective interventions. Despite under 5 mortality rate halving since 2000, it is still quite high at 4.9 million globally or 38 per 1000 live births. Moreover, there is a huge disparity in low- and middle-income countries where the under five mortality rates are almost twice as high as the global rates at 67 – 74 deaths per 1000 live births (World Health Organization, 2023).

MAIN OBJECTIVE;

Is to reduce Under-5 Mortality Rates and Improve Child Health Outcomes

PROJECT SPECIFIC OBJECTIVES

1.Estimate regional under-5 mortality rates to identify high-risk areas.

2.Analyze demographic and health indicators to understand factors contributing to under-5 mortality.

3.Develop predictive models to identify areas prone to high under-5 mortality.

4.Provide actionable insights to guide targeted interventions and resource allocation for reducing under-5 mortality.

5.Raise awareness about the importance of addressing under-5 mortality and advocating for effective healthcare policies and interventions.

Data Understanding.

The datasets utilized in this project include demographic, healthcare, and socio-economic indicators from regions with varying under five mortality rates. These datasets provide valuable insights into the determinants of child mortality and aid in developing targeted interventions. While the data is comprehensive, data quality verification procedures are necessary to address inconsistencies and missing values. Data cleaning and validation techniques will ensure the integrity and reliability of the analysis.

Dataset Summary:

1.causes_death_children_under5.csv:
 This dataset provides information on causes of death in children under 5 years old.

2.gdp.csv:
Contains data on Gross Domestic Product (GDP), offering insights into economic factors.

3.child_deaths_by_age.csv:
Includes data on child deaths categorized by age groups.

4.child_mortality_by_sex.csv:
 Provides information on child mortality rates categorized by sex.

5.poverty_and_child_mortality.csv:
Contains data on poverty levels and child mortality rates.

6.per_capita_total_expenditure_on_health_vs_child_mortality.csv:
Includes data on per capita total expenditure on health versus child mortality rates.

Feature Relevance:
The dataset provides essential information for analyzing mortality rates.Cause-specific mortality data,Economic indicators (GDP),Child mortality rates by age,sex,Poverty levels and health expenditure. These features are crucial for understanding patterns and trends in under-5 mortality rates and identifying factors contributing to higher mortality rates in certain areas.

Problem Relevance:
The dataset aligns with the project's objective of analyzing under-5 mortality rates and identifying high-risk areas. By analyzing the mortality rate data and demographic information, we can assess the prevalence of under-5 mortality in different regions and identify areas with disproportionately high mortality rates. This helps address the stakeholder's objective of prioritizing interventions and resource allocation to reduce under-5 mortality rates.

Duration Coverage:
The dataset covers a time period from the year 1990 to 2019, providing historical data on under-5 mortality rates. This historical perspective is valuable for understanding trends over time and assessing the effectiveness of past interventions. It also allows for the identification of long-term patterns and changes in mortality rates, aiding in future decision-making and policy planning.

Analytical Opportunities:

With the dataset's comprehensive coverage and detailed mortality rate data, there are opportunities for advanced analytical techniques such as time series analysis and spatial analysis. These techniques can help uncover temporal and spatial patterns in under-5 mortality rates, identify hotspots of mortality, and forecast future mortality trends. Additionally, the dataset allows for comparisons between regions and demographic groups, enabling targeted interventions and policy recommendations.

Data Preparation

Preprocessing

Preprocessing steps involve cleaning and transforming the data to prepare it for analysis. Techniques such as imputation of missing values, normalization of variables, and feature engineering may be applied to enhance the quality of the data.

Cluster Analysis

We used KMeans clustering to create a binary target variable for under-5 mortality risk, distinguishing between high and low risk

Classification

We evaluated several classification models for predicting under-5 mortality risk, including logistic regression, support vector machines (SVM), naive Bayes, and random forest classifiers Correlation Analysis You calculated the correlation matrix to explore the strength and direction of relationships between variables in the dataset.

Time Series Analysis

converted the 'Year' column to datetime format and set it as the index then plotted the trend of under-5 mortality rates for each country from 1990 to 2020.

Stationarity

Identifying Stationarity We checked the stationarity of the time series data using the Augmented Dickey-Fuller (ADF) test and visual inspection of ACF and PACF plots.

Seasonal Decomposition

We applied seasonal decomposition to decompose the time series data into trend, seasonal, and residual components. Differencing We applied differencing to make the time series data stationary, where necessary.

Modeling & Forecasting

Analysis Approach
Statistical and machine learning techniques will be employed to analyze the data and identify significant factors influencing under five mortality rates. Regression analysis, classification models, and clustering techniques may be utilized to uncover patterns and relationships within the data.We used the fitted SARIMA models to forecast future values of the under-5 mortality rate for each country.

Evaluation

Model performance was evaluated based on metrics such as accuracy, sensitivity, specificity, and predictive power. The effectiveness of the models in predicting under five mortality rates and identifying influential factors will be assessed to guide decision-making processes.

Findings

Regional Distribution Highest under-5 mortality rates are concentrated in Sub-Saharan Africa and Southeast Asia, while Australia, North America, and Europe have the lowest rates. Country Specifics Top 5 countries with the highest rates are from Sub-Saharan Africa (West and Central Africa), while the lowest rates are in Europe (Nordic and Baltic countries). Leading Causes Lower respiratory infections, diarrheal diseases, preterm births, neonatal encephalopathy, trauma, and congenital birth defects are the main causes. Gender Disparity Male children under 5 are at a higher risk than females. Trend Analysis Mortality rates have generally reduced since 1990, except for early neonatal mortality which remains high as of 2019. Critical Year 1994

saw the highest under-5 mortality, with Rwanda recording the highest rate

Deployment

Implementation
The findings and recommendations derived from the analysis will be communicated to stakeholders through reports, presentations, and interactive dashboards. Policymakers and healthcare providers can utilize this information to implement targeted interventions and improve child health outcomes in affected regions.

Conclusion

In conclusion, this project aims to provide valuable insights into under five mortality rates and identify strategies to reduce child mortality. By leveraging data-driven approaches and advanced analytics techniques, policymakers and healthcare providers can develop evidence-based interventions and allocate resources effectively to improve child health outcomes and promote overall well-being in affected regions. The successful completion of this project will contribute to achieving Sustainable Development Goal 3, which aims to ensure healthy lives and promote well-being for all at all ages.

RECOMMENDATIONS
1.Governmental Commitment and Leadership:

Healthcare providers and policymakers should prioritize children's health and well-being at the forefront of their agendas. Strong governmental commitment and leadership are crucial in driving initiatives to improve child health outcomes.

2.Community-Based Health Initiatives:

Implement community-based health initiatives with a focus on mother and child health components. Engage local communities in promoting healthy practices and providing access to essential healthcare services.

3.Health Insurance Coverage:

Establish or improve health insurance programs to ensure comprehensive health coverage for all citizens, especially vulnerable populations such as children and mothers. Programs like Rwanda's Mutuelle de Santé can serve as models for expanding health insurance coverage.

4.Investments in Healthcare Infrastructure:

Allocate resources towards building and remodeling healthcare infrastructure, including hospitals, clinics, and primary healthcare centers. Investing in modern medical facilities can improve access to quality healthcare services for children and families.

5.Emphasis on Preventative Healthcare:

Shift focus towards preventative healthcare initiatives, including immunization drives, nutrition programs, and family planning services. Prioritize early intervention and preventive measures to reduce the burden of preventable diseases and improve child health outcomes.

6.Partnerships and Collaborations:

Foster partnerships with international organizations, non-governmental organizations (NGOs), and other stakeholders to leverage expertise, resources, and best practices. Collaborative efforts can strengthen healthcare systems and enhance the delivery of child health services.