Predicting Childhood Mortality Based on Health and Socio-Economic Indicators

Aicha Sidiya, Hanin Alzaher, Razan Almahdi

2023-06-01

Contents

Introduction

This project aims to understand the impact of health and socio-economic factors on child mortality rates. This project compiles data from World Bank, the World Health Organisation, and Unicef including data related to health expenditure, education expenditure, literacy rate, water investment, vaccination etc. We aim to use data visualization and machine learning methods to investigate the variables that highly affect childhood mortality rates. Preliminary results indicate that health expenditure and GDP per capita have a strong influence on mortality rates, although there are significant limitations to the model in its current form.
The rest of the report is organized as follows: section 2 provides a background on childhood mortality predictions, section 3 presents the research question and problem statement that the report aims to answer, section 4 discusses the data used in this project, its sources and provide a brief on the contents of each dataset, section 5 analyzes those datasets and offers a statistical view on the data, section 6 presents the result of the project, section 7 discusses the results of the models used in prediction and their efficacy, and section 8 provides the highlights of the project.

Background

The most used indicator for assessing children's health is under-five mortality. It serves as a gauge for the overall growth of any nation. The likelihood that a youngster would pass away before turning five is referred to as under-five mortality. Worldwide, South Asia and Sub-Saharan Africa have higher rates of under-five mortality [1]. In India, the mortality rate for children under five has decreased from 83 per 1000 live births in 2000 to 42 in 2017. According to reports broken down by state, Uttar Pradesh, Madhya Pradesh, and Chhattisgarh have the highest rates of under-five mortality. Under-five mortality has decreased significantly in these states, but it is still a significant problem for child health in emerging nations like India. To reduce the death rate, it is essential to comprehend the significant contributing factors to childhood mortality, yet this understanding alone is insufficient [2].

Nowadays, public health research makes extensive use of machine learning (ML) approaches. Different machine learning algorithms have been employed to predict and categorize various biomedical and health data. These machine learning (ML) models are capable of discovering non-linear relationships between the target variable and independent variables as well as interactions. To identify the exposures connected to desired health outcomes as well as any possible interactions between those exposures, machine learning techniques can be used. To determine the precise estimation of health data, a variety of machine learning prediction and classification models, including maximum likelihood techniques, decision trees, principle component analysis (PCA), and logistic regression, have been applied. These methods might aid in obtaining an early forecast and insight into the crucial aspects of under-five mortality [2].

The under-five mortality outcome variable was measured as a binary outcome in this study. Therefore, for all models, under-five mortality was calculated as either being alive (coded as 0) or being dead (coded as 1). In this study, individual, household, community, and health service characteristics were employed as predictors (features). Mother and child characteristics made up the individual-level components. Mother's age at birth (20, $>$ 20), education (no education, elementary, secondary, or higher), use of contraception (yes/no), and mother's body mass index (BMI) (underweight/overweight, and normal) are all considered maternal variables. Child factors include the child's sex, birth order (1-2, 3/later), births in the past 5 years, and previous birth interval (2, 2-4, $>$ 4 years), as well as whether the child was wanted (wanted then, wanted later, or not at all) [3].

Important metrics for evaluating the effectiveness of healthcare systems include infant and maternal mortality. In order to reduce preventable death by early intervention, the World Health Organization emphasizes the significance of an effective healthcare system. Health care systems for children and mothers must be accessible, affordable, and readily available as part of early intervention. While numerous studies have evaluated the overt and covert causes of child mortality, there is scant and inconsistent research on the function of policy interventions. Therefore, in the era of reaching the Sustainable Development Goals (SDG), robust

empirical examination of the factors that influence maternal and newborn mortality remains equivocal. In this study, we looked at the impact of health spending on baby and maternal fatalities worldwide from 2000 to 2015. This study adjusted an empirical link between health outcome and health spending for the 2007–2008 financial crisis using panel quantile regression with bootstrapping. We discovered that health spending had a detrimental impact on mortality at all percentiles. Maternal mortality rates fall between 0.09% and 1.91%, and infant mortality rates fall between 0.19% and 1.45%. to fulfill SDG 3's objective of ensuring the wellbeing and healthy lifestyles of all people [3].

Research Question and Problem Statement

Can machine learning models accurately predict mortality rates based on a set of socio-economic and health indicators?

Accurately predicting mortality rates is crucial for understanding population health and making informed policy decisions. Traditional statistical methods may have limitations in capturing complex relationships between various indicators. This research aims to explore the effectiveness of machine learning models in predicting mortality rates based on a dataset consisting of socio-economic and health indicators. By comparing the performance of multiple models, we seek to identify the most accurate and reliable model for mortality rate prediction, which can assist policymakers and healthcare professionals in developing targeted interventions to improve public health outcomes.

Data

This project pulls data from the following sources:

1. The World Bank database of Literacy rate, adult total (% of people ages 15 and above 2022): This dataset describes the contrast of the literate and illiterate individuals of ages 15 and above [4].
2. The World Bank database of Mortality rate, under-5 (per 1,000 live births): This dataset describes the rates of mortality in newborns and toddlers below the age of 5 years old [5].
3. The World Bank database of Water productivity, total (constant 2015 US$ GDP per cubic meter of total freshwater withdrawal): This dataset highlights the freshwater withdrawals in total in the United States [6].
4. The World Bank database of Investment in water and sanitation with private participation (current US$): This dataset shows the amount of financial investment in sanitation in the United States [7].
5. The World Bank database of Current health expenditure (% of GDP): This dataset displays the financial investment in health on a global scale [8].
6. The World Bank database of Current health expenditure per capita (current US$): This dataset refers to the amount on health per capita in the United states [9].
7. The World Bank database of Current education expenditure, total (% of total expenditure in public institutions): This dataset displays the current global financial situation of education within public institutions [10].
8. Our World in Data's database of Income inequality: Gini coefficient, 2019: This dateset sheds light on the global contrast of income according to the Gini coefficient [12].

Analysis

The final data after preprocessing by pivoting all datasets to long with a column for year and the second containing the actual metric the data is representing like mortality rate, literacy rate, health expenditure etc. The data was then merged resulting thus in 4,157 observations, with 12 independent variables. The presence of NAs in some variable columns such as: education expenditure, literacy rate, water invest, GDP-per-capita. Although their presence signified different meanings among the variables. For instance, the literacy rate variable is almost always NA for the country Afghanistan, which entails that this data has not been recorded within said country. On the other hand, the variable water invest is NA for all countries, which means this information was not present in the dataset.

As Figure 1 shows, NAs are dispersed throughout the columns in different concentrations. For some columns such as Health expenditure per, the total amount of NAs is drastically lesser than that of the water invest column, which is majorly composed of NAs. Moreover, columns such as The domestic health expenditure, and mortality rate, show a close range of values for the percentages.

```
## Rows: 4,157
## Columns: 12
## $ country                     <chr> "Afghanistan", "Afghanistan", "Afghanistan~
## $ year                        <chr> "2000", "2001", "2002", "2003", "2004", "2~
## $ mortality_rate              <dbl> 129.3, 125.3, 121.2, 117.0, 112.8, 108.6, ~
## $ health_expenditure          <dbl> NA, NA, 17.00759, 17.81492, 21.42946, 25.1~
## $ health_expenditure_per      <dbl> NA, NA, 9.443391, 8.941258, 9.808474, 9.94~
## $ education_expenditure       <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 81~
## $ domestic_health_expenditure <dbl> NA, NA, 85.37560, 86.06919, 84.52759, 78.9~
## $ literacy_rate               <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ water_invest                <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ water_productivity          <dbl> NA, NA, 0.3725069, 0.4054078, 0.4111407, 0~
## $ immunazation                <int> 24, 33, 36, 41, 50, 58, 58, 63, 64, 63, 66~
## $ GDP_per_capita              <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
```

Figure 1. Glimpse

Figure 2 provides the skimmed dataset with only 12 observations of 17 variables, as seen below:

The variables above have been skimmed, and from the result of the skimming we see that the columns with least to non missing values are Country and Year. Columns such as Mortality rate, Health expenditure, and Domestic health expenditure have a moderate NAs rate. However, Water productivity, Literacy rate, and Education expenditure have a much higher NA rate, with Water invest being the highest. Those values in turn affect the completion rate of the columns, hence we see the Country and Year considered as complete, yet the other columns completion levels vary, based on the amount of missing data. The numerical and statistical means and hists are then also found via computed calculations.

```
## Warning: 'skim_to_wide' is deprecated.
## Use 'skim()' instead.
## See help("Deprecated")
```

Table 1: Data summary

Name	Piped data
Number of rows	4157
Number of columns	12

Table 1: Data summary

Column type frequency:	
character	2
numeric	10
Group variables	None

Variable type: character

skim_variable	n_missing
country	0
year	0

Variable type: numeric

skim_variable	n_missing	mean	sd	hist
mortality_rate	608	39.07	42.14	
health_expenditure	706	959.63	1653.65	
health_expenditure_per	706	6.32	2.83	
education_expenditure	2879	91.00	7.24	
domestic_health_expenditure	706	40.73	19.34	
literacy_rate	3483	83.12	19.35	
water_invest	4004	410438778.30	714234333.74	
water_productivity	1021	70.51	168.45	
immunazation	106	86.86	14.84	
GDP_per_capita	837	18131.01	20409.28	

Figure 2. Skim Result

```
#saving the final data
write.csv(merge_data, "data/my_data.csv")
```

```
#saving the imputed data
write.csv(my_data_imputed, "data/my_data_imputed.csv")
```

Results

What is the trend and pattern of the average mortality rate over the years and are there any significant changes or fluctuations observed?

Figure 3 depicts the average mortality rate over a range of years. The x-axis represents the years, while the y-axis represents the average mortality rate. The plot reveals a trend of the average mortality rate over time. In the earlier years, the average mortality rate was relatively high, gradually declining in subsequent years. However, there was a slight increase in the average mortality rate in 2018. Also, the plot shows a significant and noticeable increase in the average mortality rate in 2020. The sudden surge in the average mortality rate suggests a critical event or influential factor that impacted the population's health during this period.

```
write.csv(imputed_data_mean, "data/imputed_data_mean.csv")
```

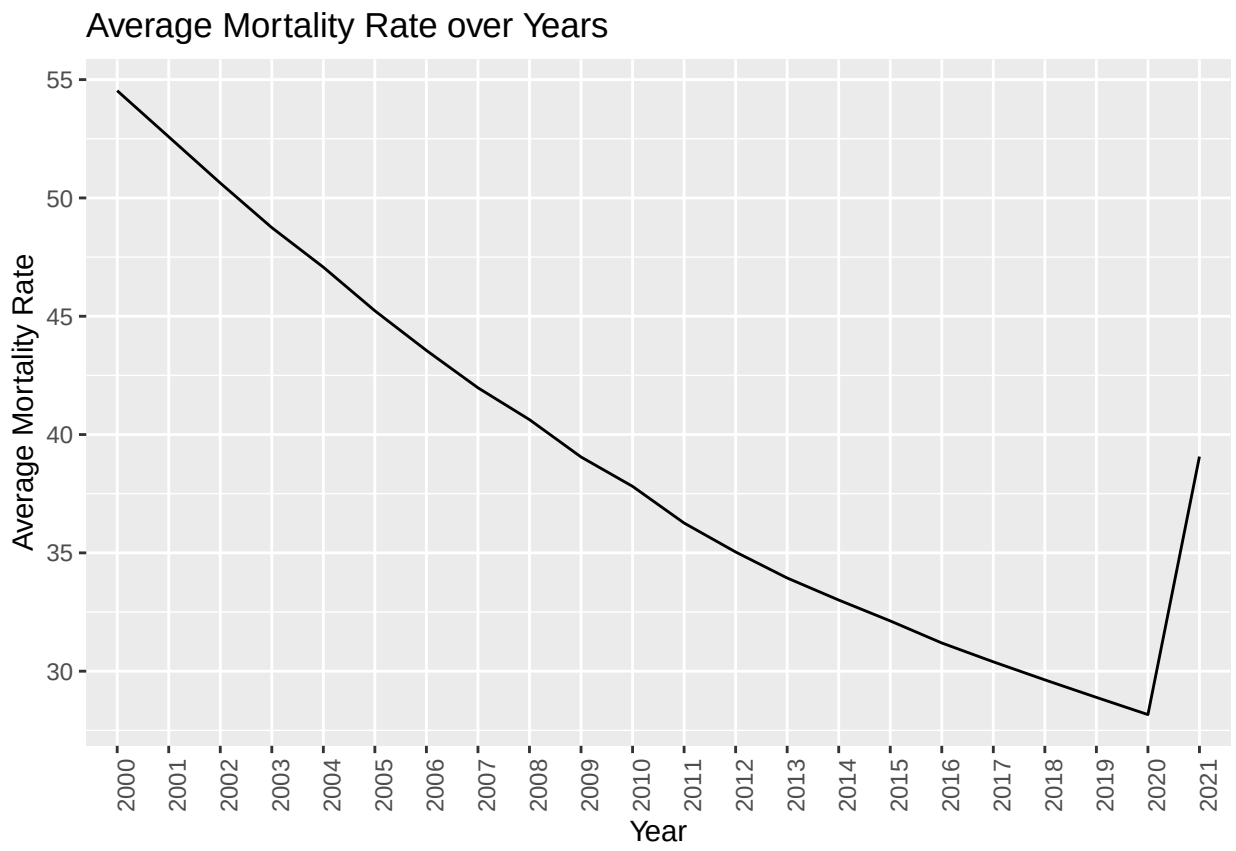

Figure 3. Average Mortality Rate over Years}

What is the relationship between Mortality Rate and Health Expenditure?

Fiure 4 showcases the relationship between Mortality Rate and Health Expenditure. The x-axis represents Health Expenditure, while the y-axis represents Mortality Rate. The plot demonstrates the varying levels of Mortality Rate observed at different levels of Health Expenditure.

Upon closer examination, it becomes apparent that there is an inverse relationship between Mortality Rate and Health Expenditure. As Health Expenditure increases, Mortality Rate tends to decrease, suggesting a potential link between higher healthcare investment and improved health outcomes. However, it is worth noting that this relationship may not be entirely linear, as there are instances where the Mortality Rate remains relatively high despite higher Health Expenditure.

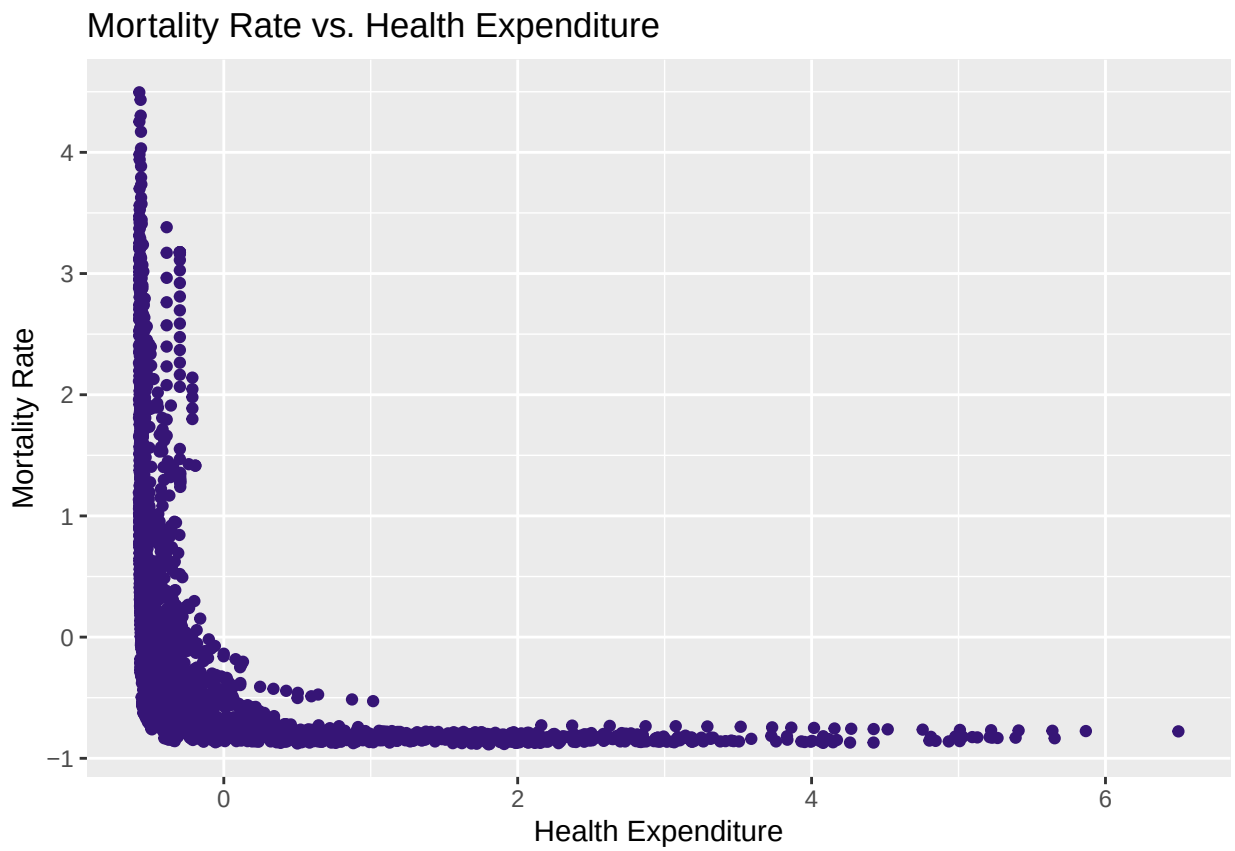

Figure 4.Mortality Rate vs. Health Expenditure

How does Literacy Rate affects Mortality Rate?

Figure 5 illustrates the distribution of Mortality Rate across different levels of Literacy Rates. The x-axis represents the Literacy Rates, while the y-axis represents the Mortality Rate.

Examining the plot, it is evident that there is a discernible pattern in the distribution of Mortality Rate with respect to Literacy Rates. As Literacy Rates increase, there is a tendency for the Mortality Rate to decrease. This suggests a potential correlation between higher literacy levels and lower Mortality Rates, indicating that education and literacy may play a role in improving overall health outcomes.

However, it is important to note that the relationship between Mortality Rate and Literacy Rate may not be solely determined by literacy itself. Other confounding factors such as healthcare access, socioeconomic status, and healthcare utilization may also contribute to the observed distribution.

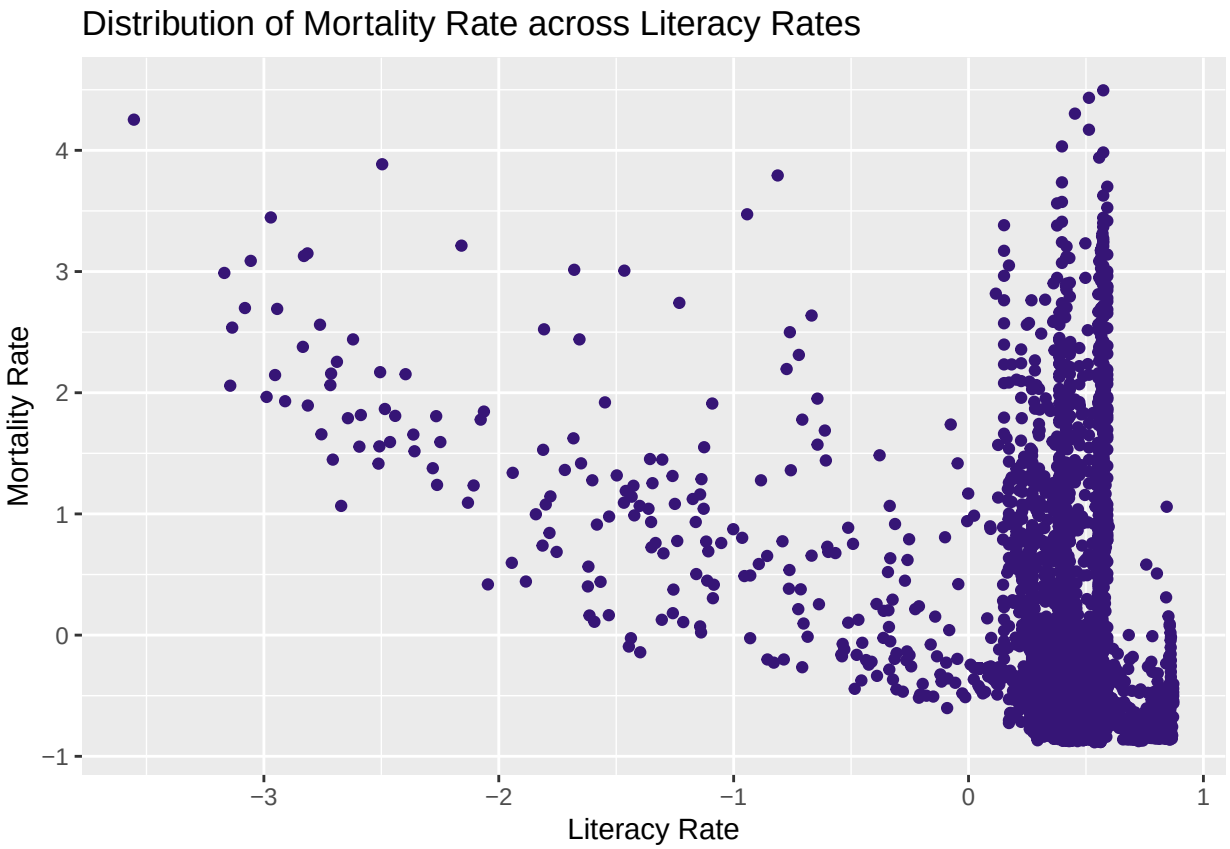

Figure 5. Distribution of Mortality Rate across Literacy Rates

What is the effect of Education Expenditure on Mortality Rate?

Figure 6 represents the relationship between Education Expenditure and Mortality Rate. The x-axis represents Education Expenditure, while the y-axis represents Mortality Rate.

Upon examination of the plot, it is evident that there is an observable pattern indicating the effect of Education Expenditure on Mortality Rate. As Education Expenditure increases, there is a corresponding decrease in Mortality Rate. This suggests a potentially significant and meaningful relationship between higher investments in education and improved health outcomes.

The plot demonstrates that allocating resources towards education expenditure may play a crucial role in reducing Mortality Rate. However, it is important to note that this relationship may not be solely determined by education expenditure itself. As we observe the significant increase of the mortality rate, other factors such as healthcare infrastructure, socioeconomic conditions, and healthcare accessibility might also influence the observed relationship.

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

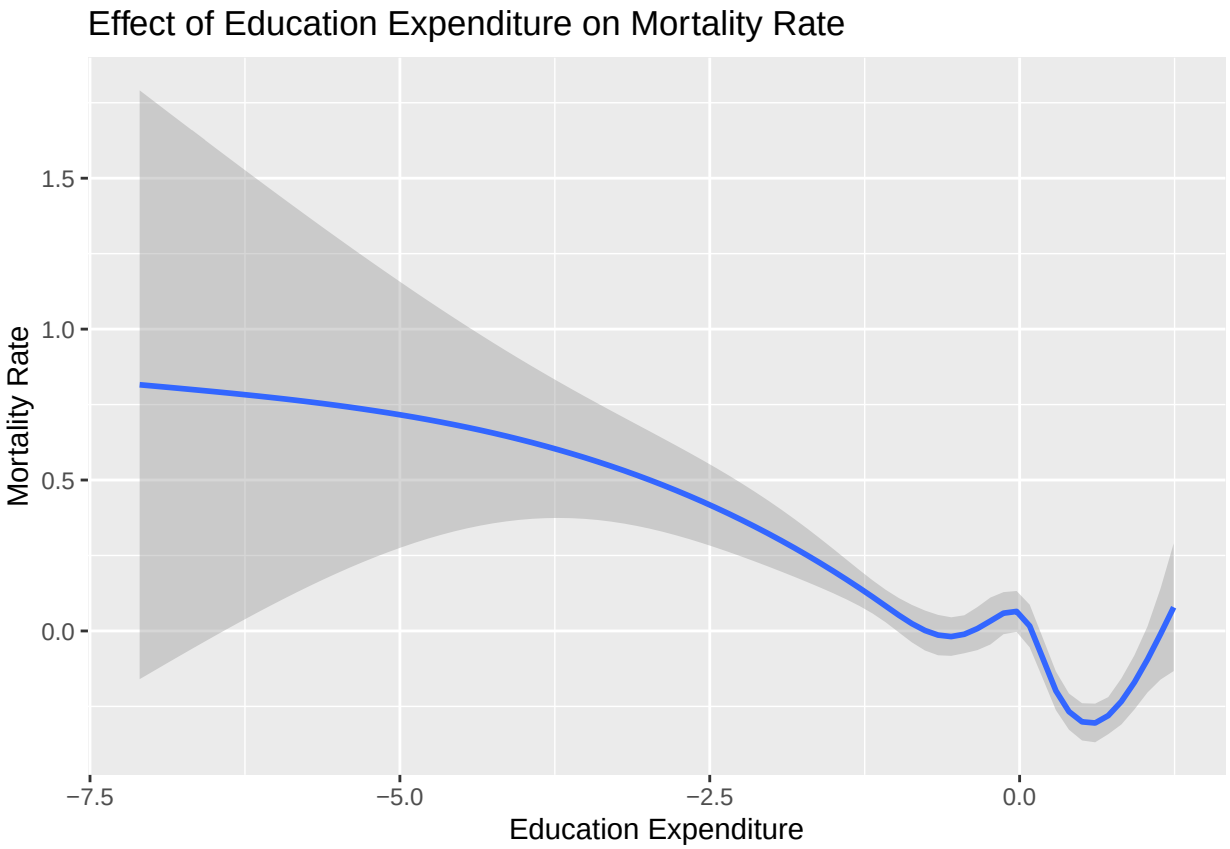

Figure 6. Effect of Education Expenditure on Mortality Rate

What is the effect of GDP per capita on Mortality Rate?

Figure 7 represents the relationship between GDP per capita and Mortality Rate. The x-axis represents GDP per capita, while the y-axis represents Mortality Rate.

It is observalble that GDP per capita highly effects mortality rate as GDP per capita increases, mortalty rate significantly decreases. This suggests a potentially significant and meaningful relationship between higher GDP and improved health.

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

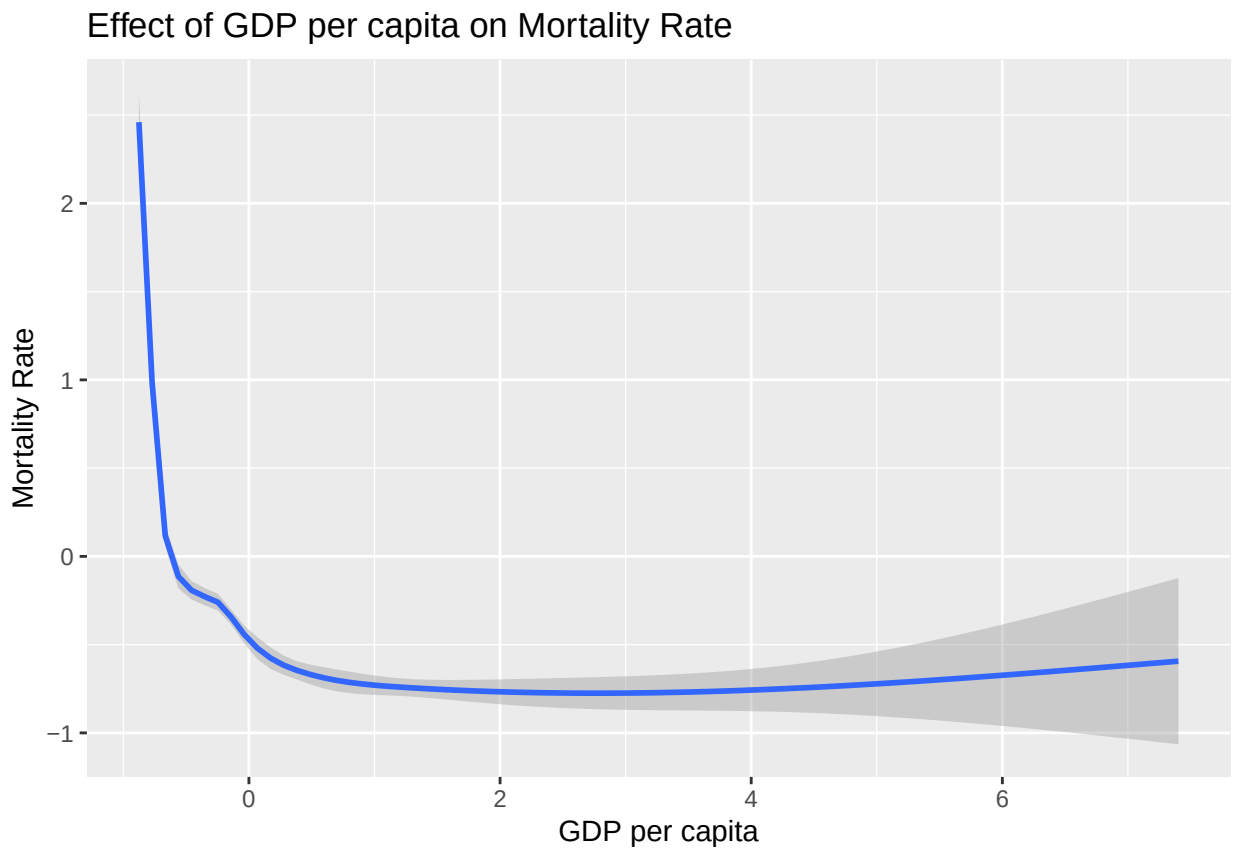

Figure 7. Effect of GDP per capita on Mortality Rate

What is the relationship between Water Productivity and Mortality Rate?

Figure 8 represents the relationship between Water Productivity and Mortality Rate. The x-axis represents Water Productivity, while the y-axis represents Mortality Rate.

Upon examining the plot, it becomes apparent that there is a discernible pattern suggesting a relationship between Water Productivity and Mortality Rate. As Water Productivity increases, there is a tendency for the Mortality Rate to decrease. This implies that higher efficiency and productivity in water usage may be associated with improved health outcomes and reduced Mortality Rates.

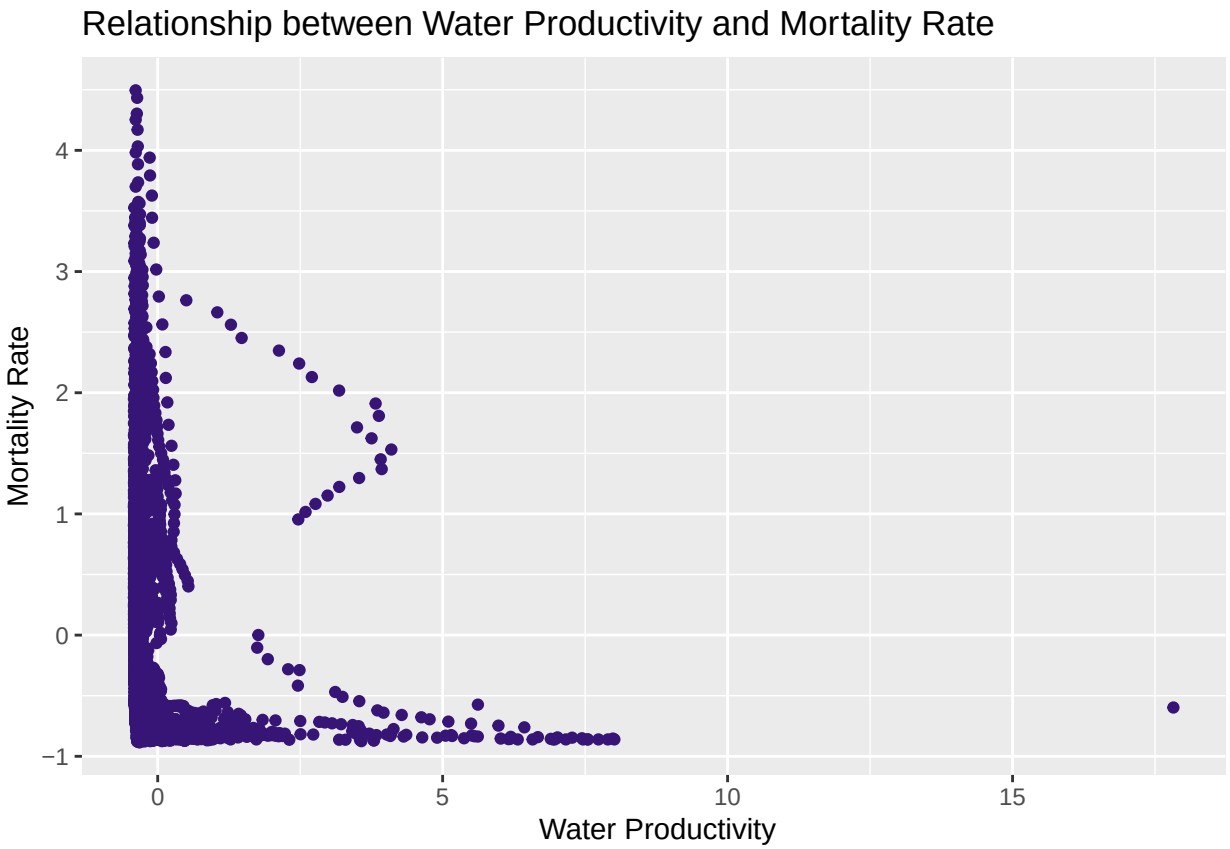

Figure 8. Relationship between Water Productivity and Mortality Rate

What is the relationship between Immunazation and Mortality Rate?

Figure 9 represents the relationship between Immunization and Mortality Rate. The x-axis represents the level of Immunization, while the y-axis represents the Mortality Rate.

Upon examining the plot, a clear pattern emerges, indicating a relationship between Immunization and Mortality Rate. As the level of Immunization increases, there is a corresponding decrease in the Mortality Rate. This suggests that higher rates of Immunization may be associated with lower Mortality Rates, highlighting the potential protective effect of immunization against preventable diseases.

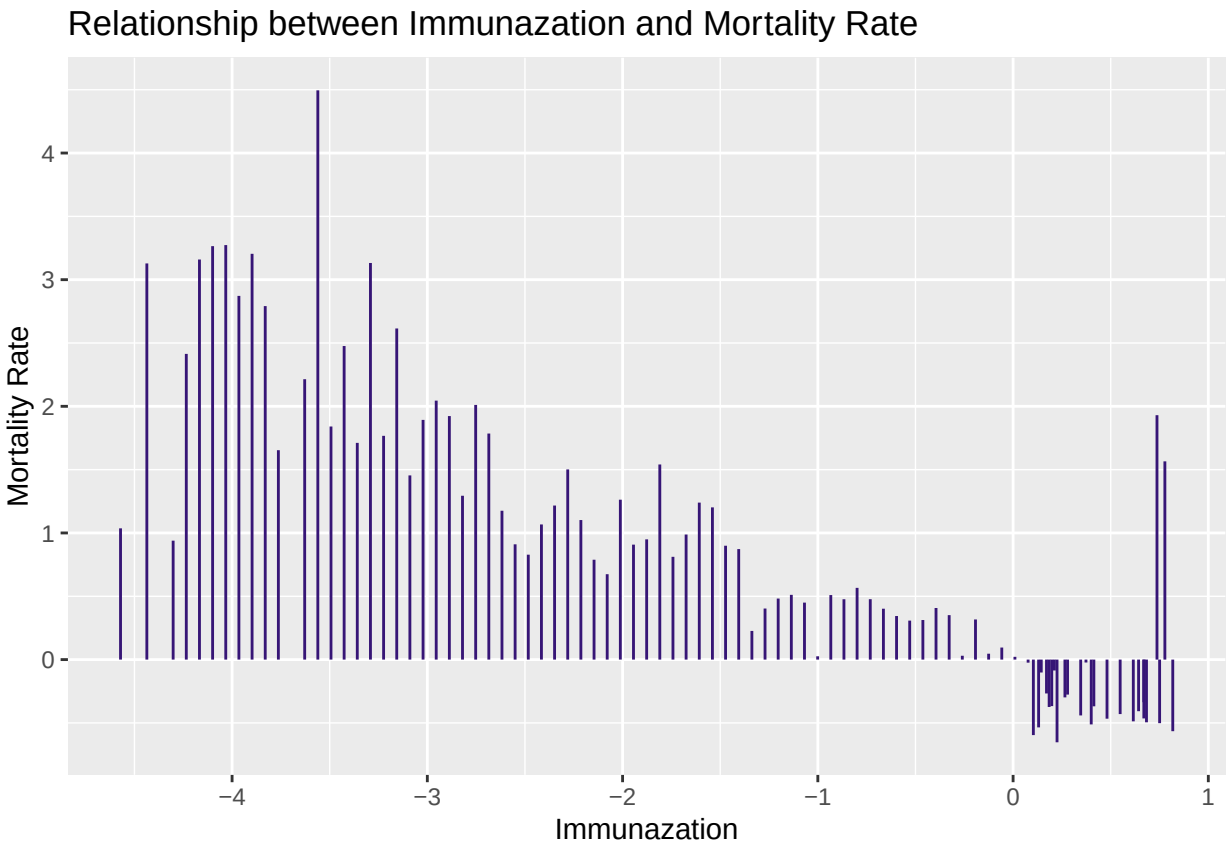

Figure 9. Relationship between Immunazation and Mortality Rate

A seen in the following Figure 10 highest correlation exist between GDP per capita and health expenditure. In terms of mortality rate it has a high correlation with dmestic health expenditure.

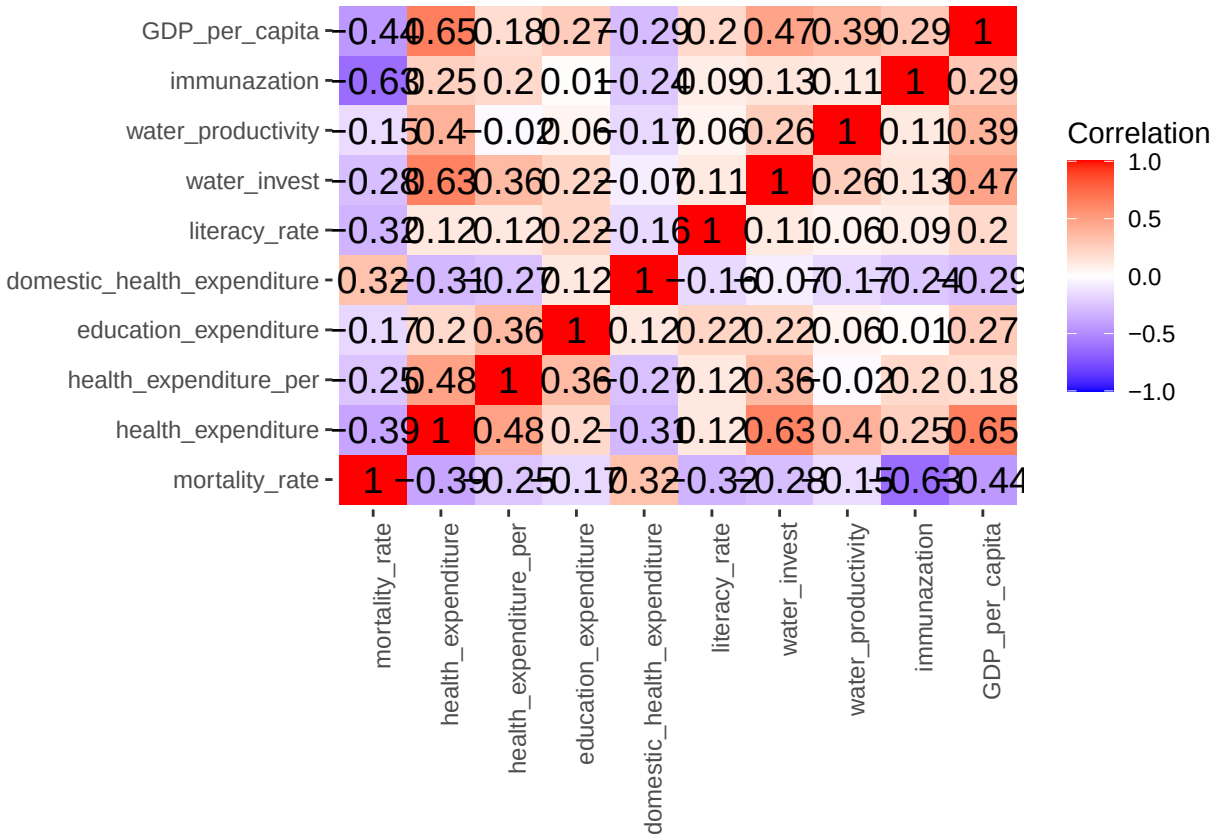

Figure 10. Heatmap of correlation between variables

Discussion

The objective of this project was to analyze the average mortality rate over the years using various statistical and machine-learning models. Before conducting the analysis, the dataset was divided into training and testing data sets, with an 80% and 20% split respectively. This division allowed us to assess the accuracy of the models on unseen data and prevent overfitting. In addition, to defining the features (X) and target variable (Y) which in our context is the mortality rate.

To ensure the repeatability of the results, a seed was set, and the data was further divided into five nearly equal folds. This approach enabled us to evaluate the performance of the models across different subsets of the data and obtain more robust insights.

We employed several models, including Linear Regression, Decision Tree, GBM, XGBoost, and Support Vector Machines (SVM), to predict the average mortality rate. The performance of each model was evaluated using two metrics, namely Root Mean Squared Error (RMSE) and R-squared.

The results of our analysis provide valuable insights into the performance of the models. The model with the lowest RMSE and the highest R-squared is generally considered the best-performing model. These metrics indicate the accuracy and the amount of variance in the dependent variable that can be explained by the independent variables.

After comparing the performance of each model as seen in Figure 11, we found that XGBoost produces the best results with 0.1159076 for RMSE and 0.8703444 for R-squared. Followed by GBM with 0.1665639 for RMSE and 0.8139392 for R-squared. These results indicate their ability to provide accurate predictions and explain a significant portion of the variance.

```
##               Model      RMSE R_squared
## 1 Linear Regression 0.4262095 0.5214246
## 2     Decision Tree 0.4370168 0.5092996
## 3               GBM 0.1665639 0.8139392
## 4           XGBoost 0.1159076 0.8703444
## 5               SVM 0.2509491 0.8703444
```

Figure 11. Model Comparison

In the case of Linear Regression, the model calculates coefficient values for each variable, indicating the magnitude and direction of their impact on the target variable (mortality rate). Variables with larger coefficients are considered more influential. For example, the variable "immunization" was found to have the most significant effect on the model's predictions, suggesting that it strongly contributes to changes in the average mortality rate.

For Decision Tree, it splits the data based on various features to make predictions. By examining the structure of the tree, you can identify the variables that were used most frequently and had the highest information gain or Gini importance in the decision-making process. For instance, variables such as health expenditure and GDP per capita were repeatedly used to split the data, indicating their significant influence on the model's predictions. Also, the same variables have a significant influence on the GBM, SVM, and XGBoost models predictions, these results show that health expenditure and GDP per capita highly affect the mortality rate. In addition, as seen in Figure 4 and Figure 7 higher health expenditure and GDP per capita result in a significant decrease in mortality rate.

Conclusion

In conclusion, our analysis demonstrates the feasibility of using statistical and machine learning models to predict the average mortality rate over the years. The results highlight the potential of certain models, such as GBM, in accurately predicting the mortality rate and providing valuable insights. We effectively demonstrated that variables such that health expenditure and GDP per capita significantly affect mortality rates. In addition, dmestic health expenditure which was demonstrated through the correlation heatmap. However, further exploration and refinement of the models, including feature engineering such as excluding features with alot of null values or features with low correlation factor and low importance, in addition to hyperparameter tuning, may lead to improved performance.

References

1. Bitew, F. H., Nyarko, S. H., Potter, L.,& amp; Sparks, C. S. (2020, November 4). Machine Learning Approach for predicting under-five mortality determinants in Ethiopia: Evidence from the 2016 Ethiopian demographic and Health Survey - genus. SpringerOpen. https://genus.springeropen.com/articles/10.1186/s41118-020-00106-2

2. Owusu, P. A., Sarkodie, S. A., & Pedersen, P. A. (2021, February 24). Relationship between mortality and Health Care Expenditure: Sustainable Assessment of Health Care System. PloS one. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7904168/ #: :text=The%20empirical%20results%20show%20that,wit %20rate

3. Saroj, R. K., Yadav, P. K., Singh, R., & Chilyabanyama, O. N. (2022, September 24). Machine learning algorithms for understanding the determinants of under-five mortality. BioData mining. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9509654/

4. Literacy rate, adult total (% of people ages 15 and above). World Bank Open Data. (n.d.-c). https://data.worldbank.org/indicator/SE.ADT.LITR.ZS

5. Mortality rate, under-5 (per 1,000 live births). World Bank Open Data. (n.d.-d). https://data.worldbank.org/indicator/
6. Water productivity, total (constant 2015 US$ GDP per cubic meter ... - data. (n.d.-b). https://data.worldbank.org/indicator/ER.GDP.FWTL.M3.KD
7. World Bank Open Data. (n.d.-d). https://data.worldbank.org/indicator/IE.PPI.WATR.CD
8. Current health expenditure (% of GDP). World Bank Open Data. (n.d.-a). https://data.worldbank.org/indicator/SH.X
9. Current health expenditure per capita (current US$). World Bank Open Data. (n.d.-b). https://data.worldbank.org/indicator/SH.XPD.CHEX.PC.CD?name_desc=true
10. World Bank Open Data. (n.d.-c). https://data.worldbank.org/indicator/se.xpd.ctot.zs
11. Domestic Private Health Expenditure (% of current ... - world bank data. (n.d.-a). https://data.worldbank.org/indicato
12. Income inequality: Gini coefficient. Our World in Data. (n.d.). https://ourworldindata.org/grapher/economic-inequality-gini-index?country= BRA