Abstract

This paper analyzes and discusses the Life Expectancy Dataset by following the data science cycle, focusing on cleaning, preprocessing, and exploring the data to uncover relationships between the variables and the target metric, life expectancy. The dataset, derived from real-world scenarios, includes features such as socioeconomic indicators, healthcare access, and disease prevalence.

The analysis involves categorizing the dataset into thematic groups, addressing missing data, and investigating variable relationships through statistical and visual techniques. Key observations highlight the strong correlation between life expectancy and factors like schooling, income composition of resources, and HIV/AIDS prevalence. These findings emphasize the disparities between developed and developing nations.

This study reflects on the data's complexity, exploring the challenges encountered and the steps taken to produce meaningful insights. The discussion is grounded in a human-centered perspective, ensuring relevance to real-world contexts while adhering to the objectives set for this project.

Table of Contents

Introduction

Life expectancy is a vital indicator of a population's overall health, development, and well-being. A complex interplay of socio-economic, health-related, and environmental factors influences it. Understanding these factors can provide actionable insights for policymakers and health organizations aiming to improve the quality of life across nations. This project focuses on analyzing the *Life Expectancy Dataset* (KumarRajarshi, n.d.), which offers a wealth of information on demographic, economic, and health-related variables collected for 193 countries from 2000 to 2015.

The dataset, compiled from the World Health Organization (WHO) and United Nations repositories, contains 22 features and 2938 rows, including life expectancy as the target variable. These features are categorized into socio-economic factors, healthcare access indicators, mortality-related metrics, and immunization rates, offering a comprehensive view of the factors influencing life expectancy. The dataset provides a unique opportunity to investigate real-world trends and patterns that can inform decisions in public health and development sectors.

The primary goal of this project is to explore, analyze, and discuss the *Life Expectancy Dataset* using the "Data Science Cycle" framework. The analysis encompasses data cleaning, exploration, visualization, feature correlation, and modeling to understand the impact of various factors on life expectancy. Special attention is given to socio-economic and health-care-related variables due to their strong correlations with the target variable.

To achieve this, the data was thoroughly cleaned, and exploratory data analysis (EDA) was performed to uncover patterns and relationships among the variables. Models such as Random Forest and Artificial Neural Networks (ANNs) were trained to predict life expectancy, and advanced interpretation techniques, including SHAP values and Random Forest feature importances, were employed to understand the contributions of individual features.

Through this project, we aim to address key questions, such as the relationship between life expectancy and factors like income composition, schooling, and disease prevalence. Additionally, we investigate how these factors differ across developing and developed countries. By discussing these patterns and insights, this project provides a deeper understanding of life expectancy's determinants and how they vary globally.

2. Data Overview

2.1 Dataset Description

The dataset used for this analysis is sourced from the World Health Organization (WHO) and is owned by [Name of the Individual or Entity]. It contains data on life expectancy and various related factors across multiple countries over the years 2000 to 2013. The dataset includes 22 features, which can be grouped into four broad categories:

- Socio-Economic Factors:
    - Country: Name of the country.
    - Year: The year of data collection (2000–2013).
    - Status: Indicates whether the country is developed or developing.
    - GDP: Gross Domestic Product per capita in USD.
    - Income composition of resources: A measure of income-based human development.
    - Schooling: Average number of years of schooling.
- **Healthcare Access**:
    - percentage expenditure: Expenditure on health as a percentage of GDP.
    - Hepatitis B: Immunization coverage among one-year-olds (%).
    - Polio: Immunization coverage for polio among one-year-olds (%).
    - Diphtheria: Immunization coverage for diphtheria, tetanus toxoid, and pertussis among one-year-olds (%).
- **Lifestyle Factors**:
    - Alcohol: Per capita alcohol consumption (in liters of pure alcohol).
    - BMI: Average body mass index of the population.
    - thinness 1-19 years: Prevalence of thinness among children and adolescents aged 10–19 years.
    - thinness 5-9 years: Prevalence of thinness among children aged 5–9 years.
- **Mortality and Disease Factors**:
    - Adult Mortality: Probability of dying between ages 15 and 60 per 1,000 population.
    - infant deaths: Number of infant deaths per 1,000 live births.
    - under-five deaths: Number of under-five deaths per 1,000 live births.
    - HIV/AIDS: Deaths caused by HIV/AIDS per 1,000 live births.
    - Measles: Reported cases of measles per 1,000 population.

The target variable for this analysis is Life expectancy, representing the average age a person is expected to live in a specific country and year.

2.2 Data Cleaning

To ensure the dataset's quality and reliability, the following steps were undertaken:
Handling Missing Data
Missing values were identified using the Python function df.isnull().sum. The features with missing values and their respective counts are:

- Life expectancy: 10 missing values
- Adult Mortality: 10 missing values
- Alcohol: 194 missing values
- Hepatitis B: 553 missing values
- BMI: 34 missing values
- Polio: 19 missing values
- Total expenditure: 226 missing values
- Diphtheria: 19 missing values
- GDP: 448 missing values
- Population: 652 missing values
- thinness 1-19 years: 34 missing values
- thinness 5-9 years: 34 missing values
- Income composition of resources: 167 missing values
- Schooling: 163 missing values

These missing values were imputed using sklearn's SimpleImputer, replacing them with the mean value of each respective column. This approach was chosen to maintain the data's statistical integrity while minimizing information loss.

2.3 Identifying and Handling Outliers

Outliers were identified using box plots, focusing on features with continuous numerical values. The Interquartile Range (IQR) method was employed to detect and address these outliers. Observations lying beyond 1.5 times the IQR above the third quartile or below the first quartile were considered outliers and replaced with the mean value of the respective column to ensure the dataset remained representative and avoided potential biases.
By thoroughly addressing missing data and outliers, the dataset was prepared for subsequent exploratory data analysis and modeling.

3. Exploratory Data Analysis (EDA)

3.1 Overview

The exploratory data analysis (EDA) aimed to uncover patterns, relationships, and insights within the dataset to provide a deeper understanding of the factors influencing life expectancy. This section presents key visualizations and findings, highlighting trends and statistical characteristics of the data. Each visualization is referenced by its figure number (e.g., Fig. 1).

3.2 Key Visualizations and Insights

3.2.1 Dataset Description Table

A descriptive summary of the dataset (Fig. 1) provided the following insights:

- **Distribution of Life Expectancy**: The mean life expectancy is 69.38 years, with a standard deviation of 9.29 years, indicating significant variation.
- **Range of Life Expectancy**: Values range from a minimum of 44.80 years to a maximum of 89 years, illustrating substantial diversity.
- **Quartiles as Distribution Markers**: Quartiles (25th percentile: 63.42 years, median: 72.00 years, 75th percentile: 75.60 years) reflect the data's spread and central tendencies.

	count	mean	std	min	25%	50%	75%	max
Year	2938.0	2.007519e+03	4.613841e+00	2000.00000	2004.000000	2.008000e+03	2.012000e+03	2.015000e+03
Life expectancy	2938.0	6.937729e+01	9.291395e+00	44.80000	63.425000	7.200000e+01	7.560000e+01	8.900000e+01
Adult Mortality	2938.0	1.534855e+02	1.035598e+02	1.00000	74.000000	1.440000e+02	2.180000e+02	4.540000e+02
infant deaths	2938.0	1.098732e+01	1.429768e+01	0.00000	0.000000	3.000000e+00	2.200000e+01	5.500000e+01
Alcohol	2938.0	4.589804e+00	3.894867e+00	0.01000	1.092500	4.160000e+00	7.380000e+00	1.658000e+01
percentage expenditure	2938.0	2.365719e+02	2.987520e+02	0.00000	4.685343	6.491291e+01	4.415341e+02	1.092155e+03
Hepatitis B	2938.0	8.706782e+01	9.227711e+00	59.00000	80.940461	8.700000e+01	9.600000e+01	9.900000e+01
Measles	2938.0	5.139636e+02	9.198642e+02	0.00000	0.000000	1.700000e+01	3.602500e+02	2.419592e+03
BMI	2938.0	3.832125e+01	1.992768e+01	1.00000	19.400000	4.300000e+01	5.610000e+01	8.730000e+01
under-five deaths	2938.0	1.421582e+01	1.830841e+01	0.00000	0.000000	4.000000e+00	2.800000e+01	7.000000e+01
Polio	2938.0	8.842749e+01	1.089945e+01	51.00000	82.550188	9.300000e+01	9.700000e+01	9.900000e+01
Total expenditure	2938.0	5.803315e+00	2.152645e+00	0.37000	4.370000	5.938190e+00	7.150000e+00	1.171000e+01
Diphtheria	2938.0	8.856220e+01	1.045148e+01	51.00000	82.324084	9.300000e+01	9.700000e+01	9.900000e+01
HIV/AIDS	2938.0	5.163104e-01	6.559171e-01	0.10000	0.100000	1.000000e-01	8.000000e-01	1.800000e+00
GDP	2938.0	4.086574e+03	3.782663e+03	1.68135	580.486996	3.116562e+03	7.483158e+03	1.778942e+04
Population	2938.0	6.459943e+06	6.531159e+06	34.00000	418917.250000	3.675929e+06	1.275338e+07	3.122588e+07
thinness 1-19 years	2938.0	4.347556e+00	3.385446e+00	0.10000	1.600000	3.400000e+00	6.600000e+00	1.530000e+01
thinness 5-9 years	2938.0	4.362577e+00	3.416209e+00	0.10000	1.600000	3.400000e+00	6.600000e+00	1.550000e+01
Income composition of resources	2938.0	6.553188e-01	1.541045e-01	0.25300	0.554000	6.620000e-01	7.720000e-01	9.480000e-01
Schooling	2938.0	1.217028e+01	2.850434e+00	4.70000	10.500000	1.210000e+01	1.410000e+01	1.970000e+01

Figure 1

3.2.2 Distribution of Numerical Columns

The distribution of all numerical columns was examined (Fig. 2). Key findings include:

- Most features are not normally distributed, likely due to regional and geographical diversity in the dataset.
- The majority of distributions are right-tailed, with a few exceptions exhibiting left-tailed behavior.
- The target variable, life expectancy, is almost normally distributed but with a higher standard deviation (9 years) than typical normal distributions.

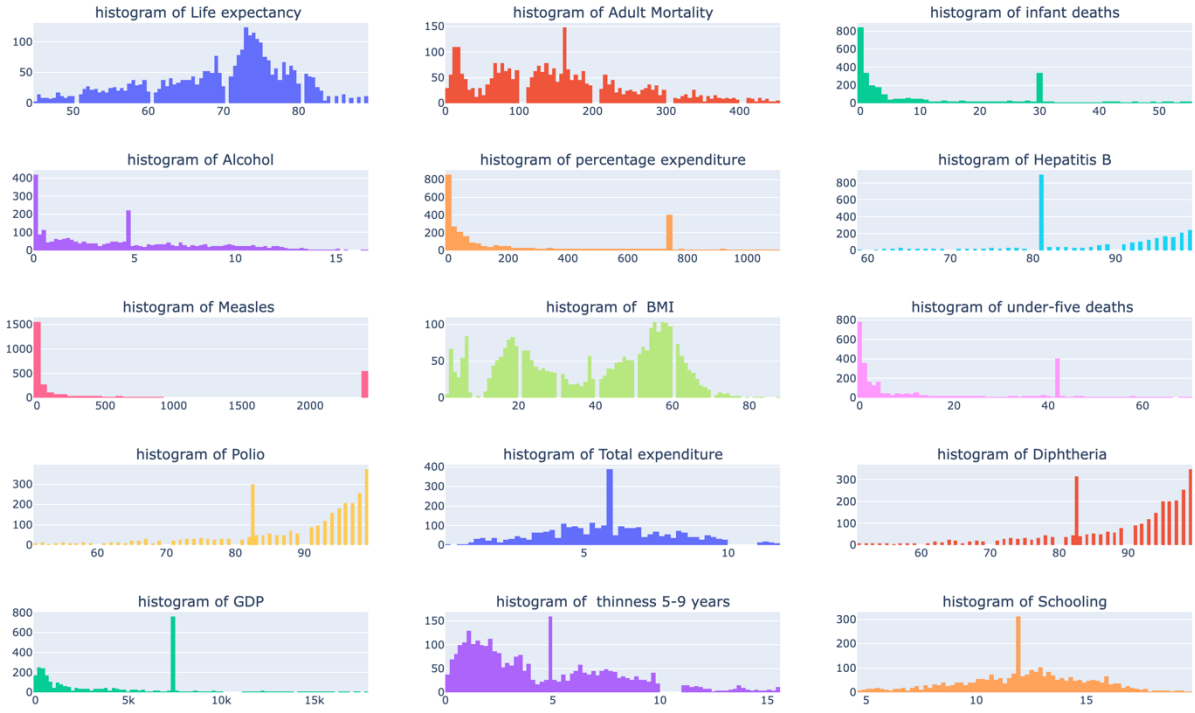

Figure 2

3.2.3 Distribution of Country Status

A pie chart (Fig. 3) illustrated the distribution of the Status feature:

- **Developing Countries**: Comprise 82.6% of the dataset, indicating a dominant representation.
- **Developed Countries**: Account for 17.4% of the dataset, representing a minority.

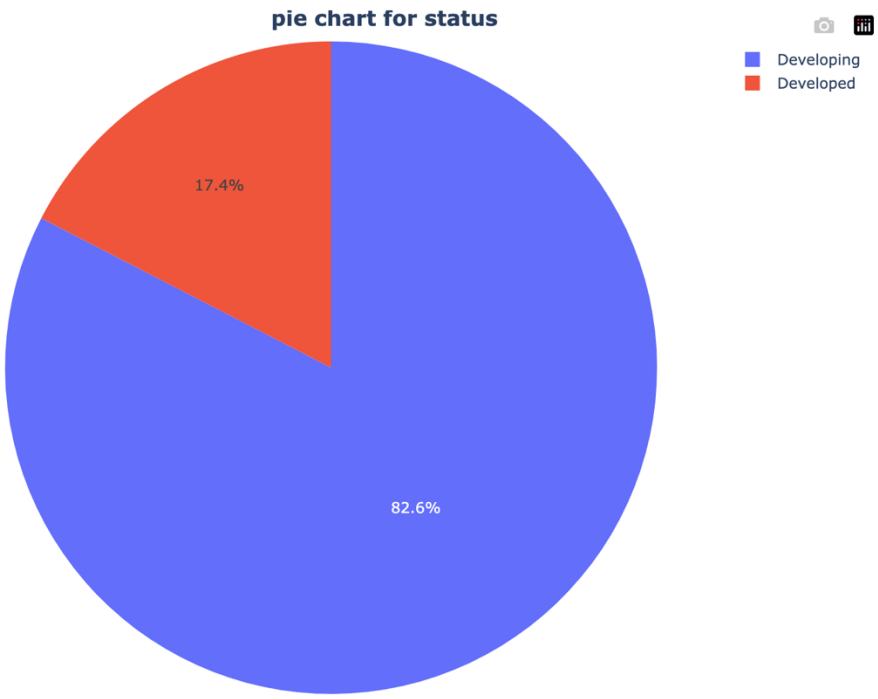

Figure 3

3.2.4 Correlation Analysis

Correlation heatmaps for the four feature categories (socio-economic, healthcare access, lifestyle, and mortality/disease) revealed key relationships (Fig. 4):

- **Highest Correlation with Life Expectancy**: Schooling (0.85) and Income composition of resources (0.72).
- **Negative Correlation**: Adult Mortality (-0.61) and HIV/AIDS (-0.78).
- **High Predictor Correlations**: Features such as Schooling and Income composition of resources or Thinness 5-9 years and Thinness 1-19 years showed strong internal correlations.

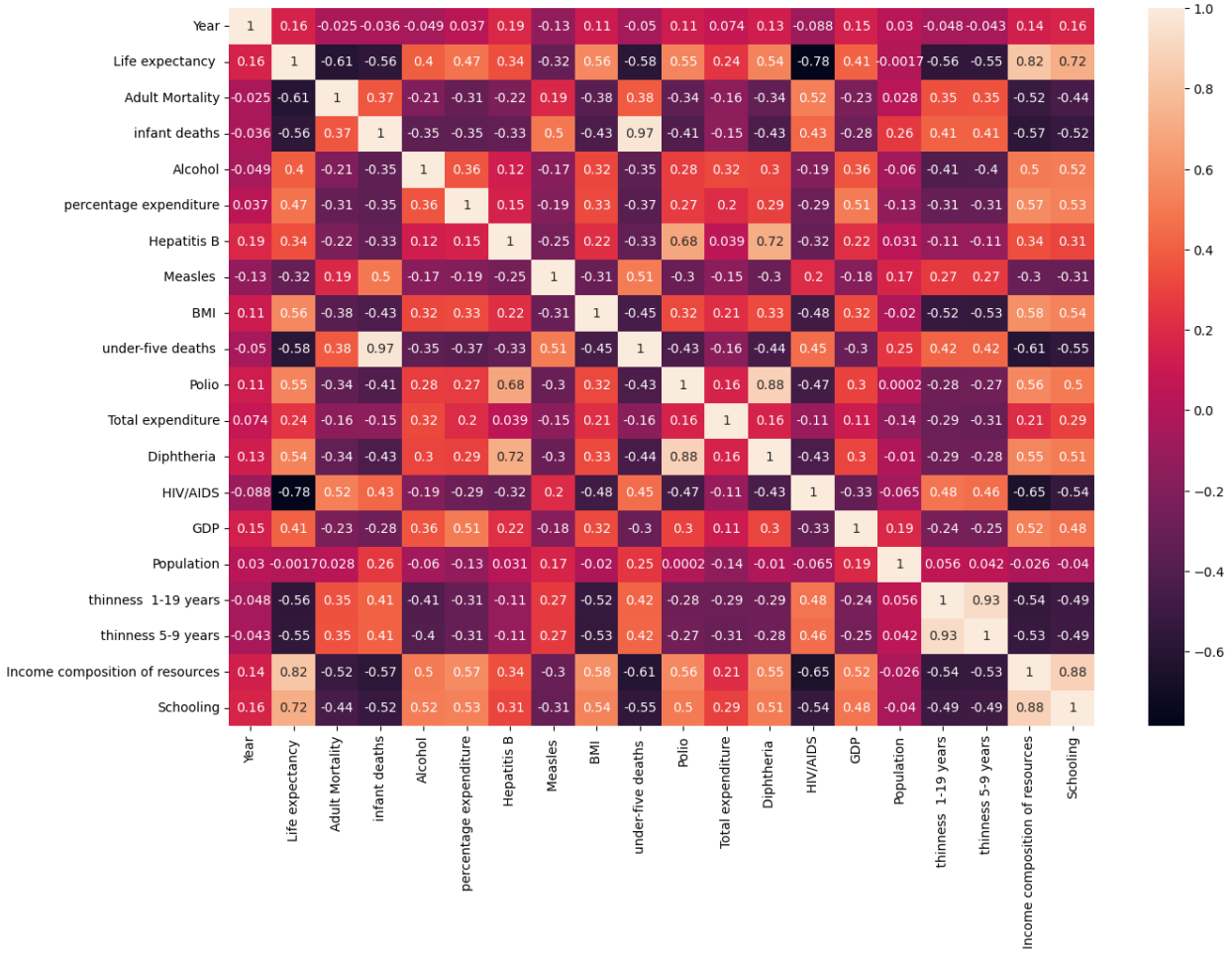

Figure 4

3.2.5 Violin Plot of Life Expectancy by Country Status

A violin plot (Fig. 5) compared life expectancy distributions across country statuses:

- **Developing Countries**: Show a broader range (44.8 to 89 years) but cluster densely around the third quartile (74 years).
- **Developed Countries**: Exhibit higher consistency, with a median of 79.2 years and narrower interquartile range (Q1: 76.8, Q3: 81.7).

Figure 5

3.2.6 Scatter Plot of Predictors vs. Life Expectancy

Scatter plots (Fig. 6) revealed trends between predictors and life expectancy:

- Positively correlated features (e.g., Schooling) showed upward trends with increasing life expectancy.
- Negatively correlated features (e.g., HIV/AIDS, Adult Mortality) displayed downward trends as life expectancy improved.

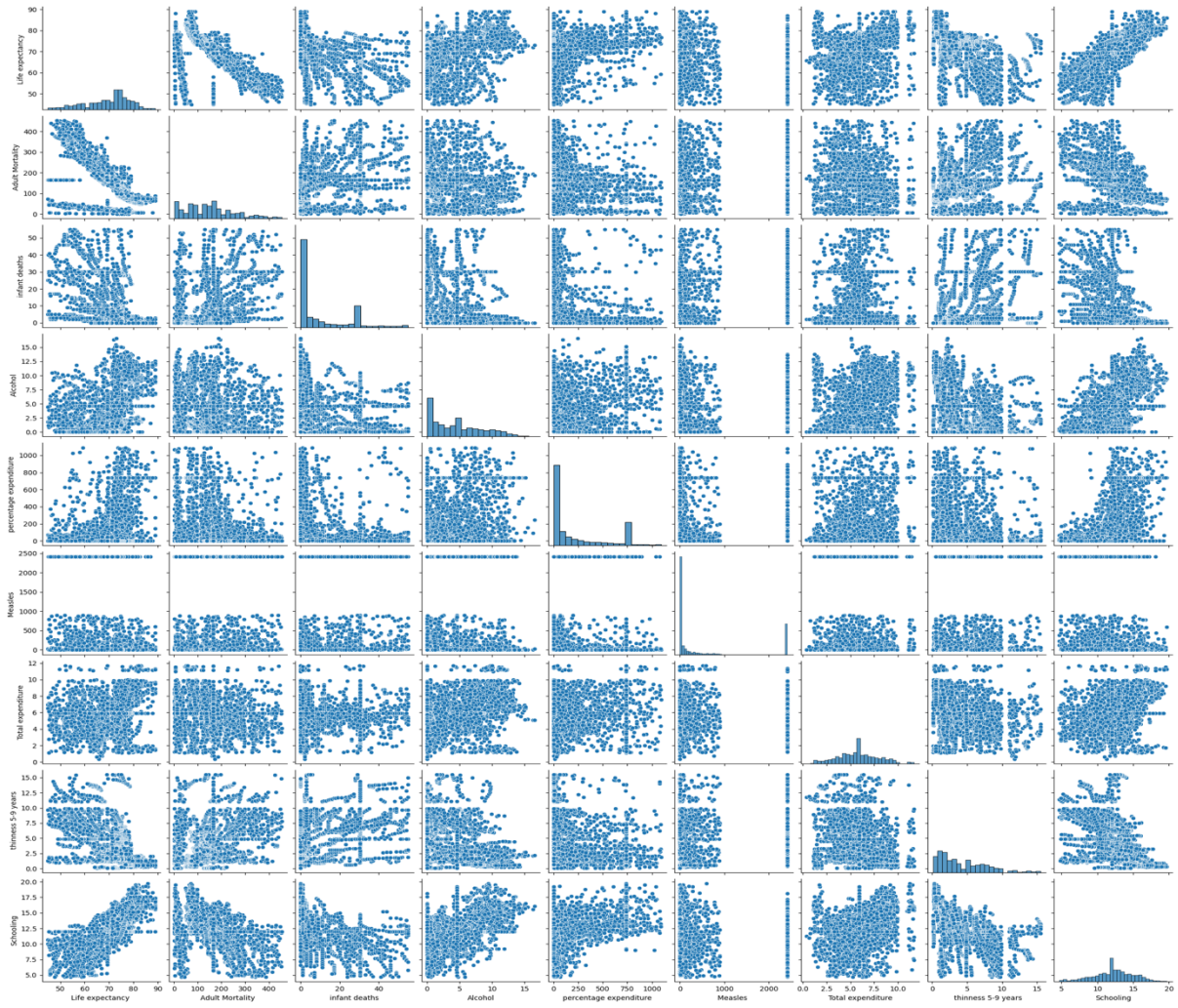

Figure 6

3.2.7 Geographic Visualizations

1. **Global Life Expectancy by Country**: A world map (Fig. 7) colored by life expectancy and circle sizes based on Percentage expenditure highlighted:
    - High total expenditure correlates with high life expectancy, except in cases like South Africa and Equatorial Guinea.

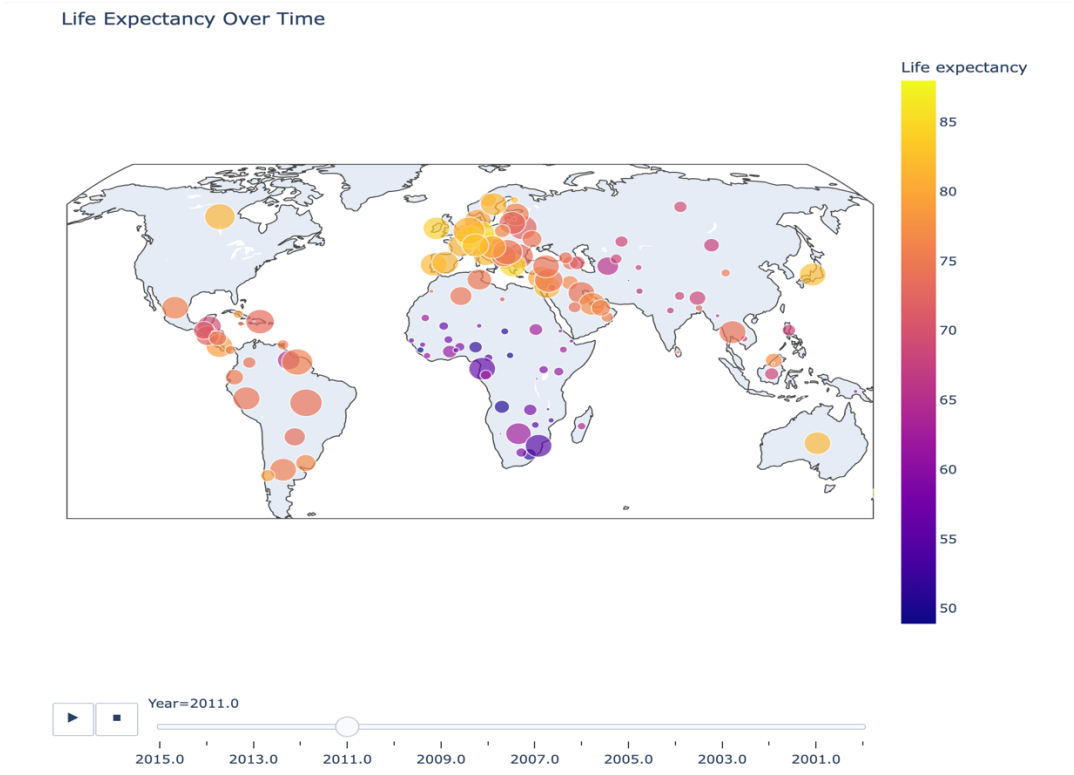

Figure 7

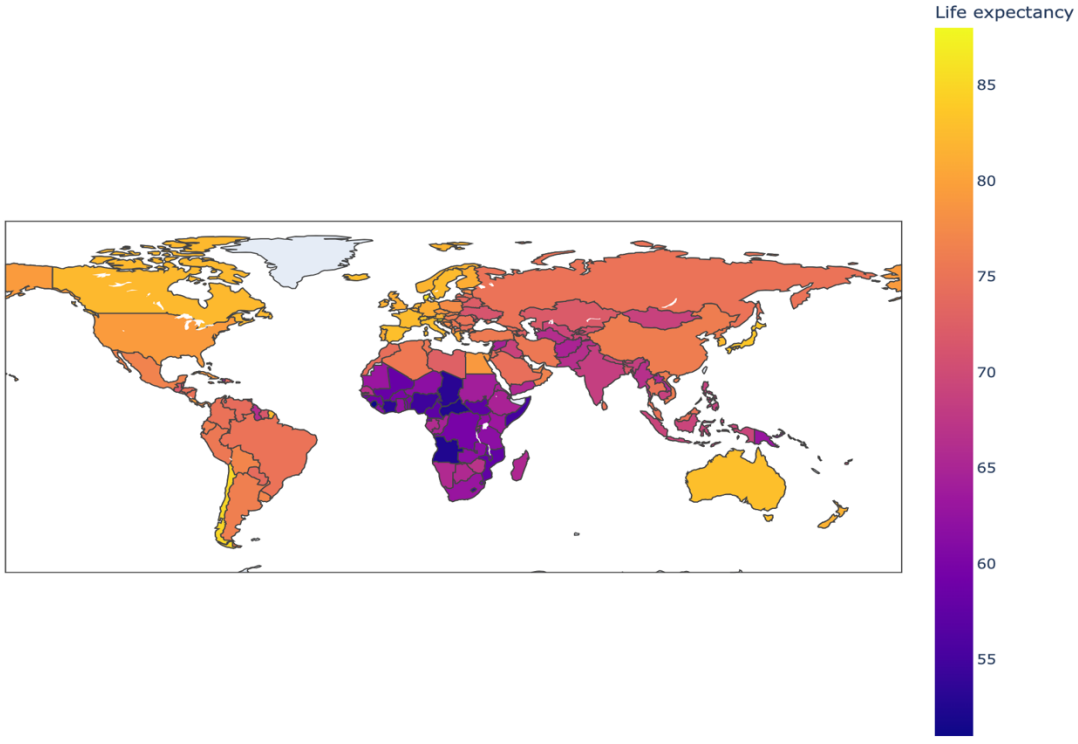

Figure 8

2. **Animated Life Expectancy by Year**: A dynamic visualization (Fig. 8) showed life expectancy trends (2000–2015), emphasizing consistently lower values in Afghanistan and sub-Saharan Africa.

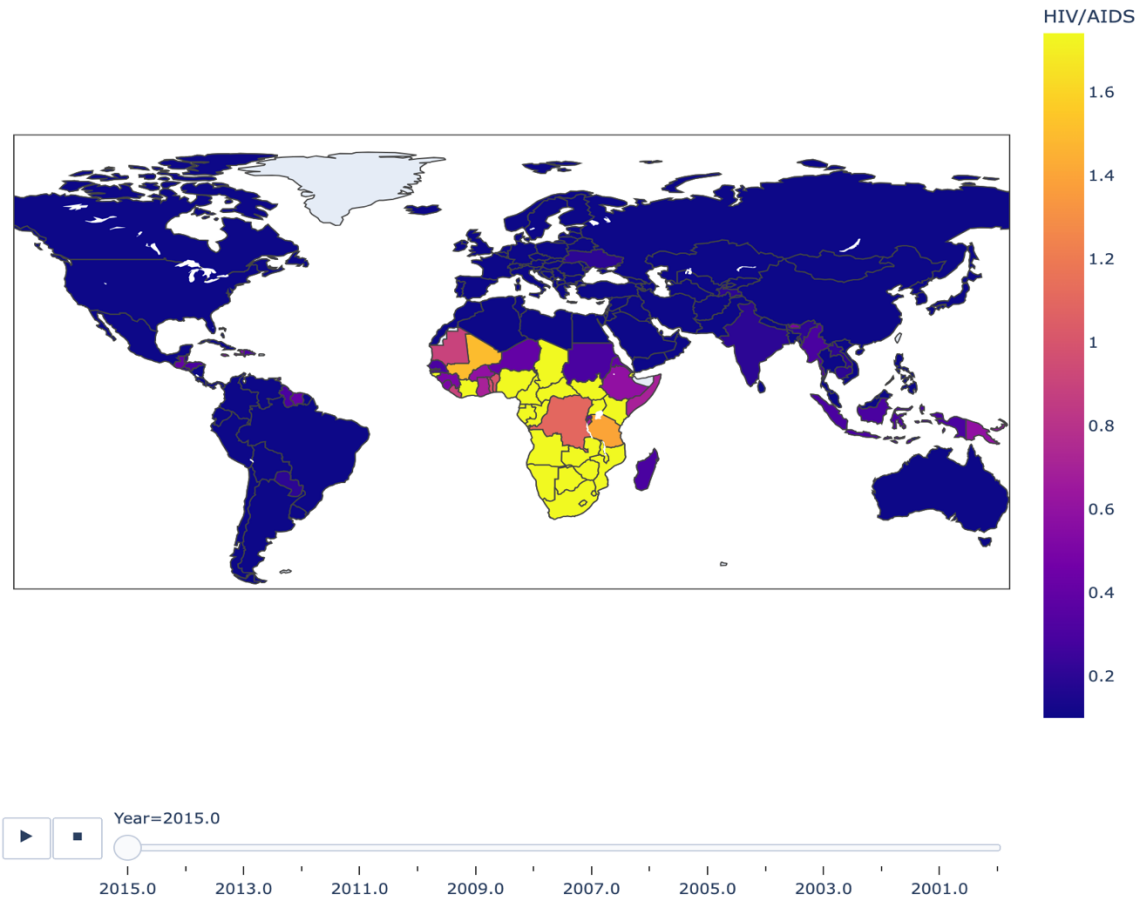

Figure 9

3. **HIV/AIDS Impact**: A similar map colored by HIV/AIDS (Fig. 9) depicted an inverse relationship, with high HIV/AIDS prevalence in countries with low life expectancy.

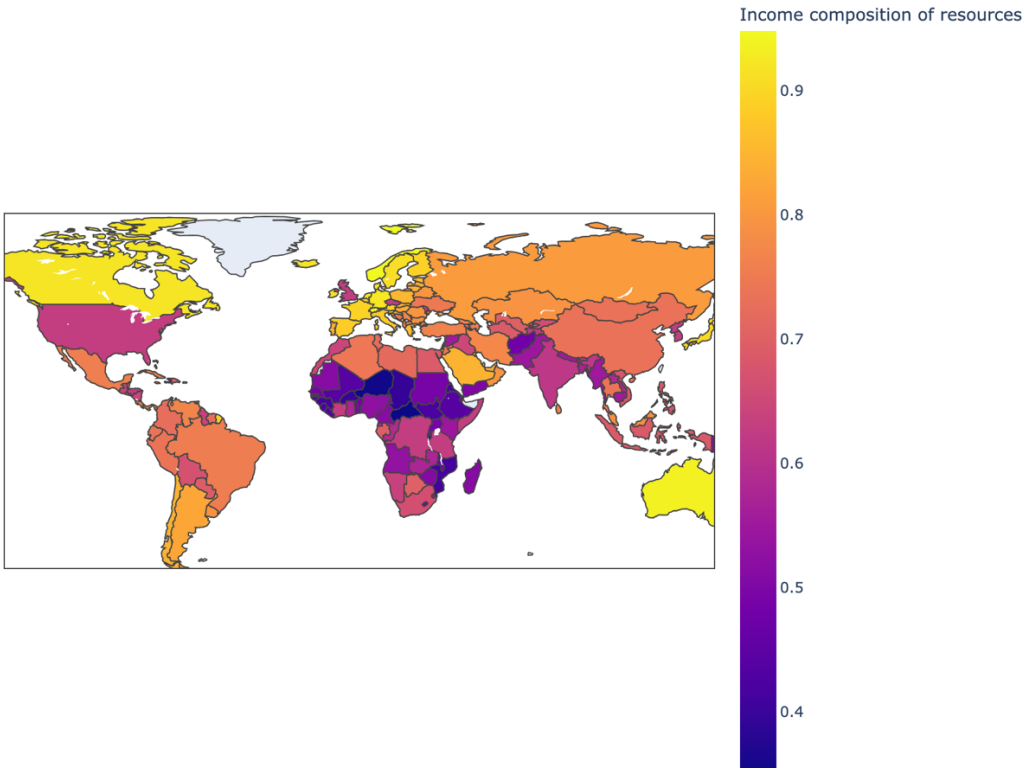

Figure 10

4. **Income Composition of Resources**: Another map (Fig. 10) mirrored life expectancy patterns, affirming a positive correlation with the target variable.

3.2.8 Temporal Trends in Life Expectancy

A line plot animated by country (Fig. 11) illustrated life expectancy trends over time:

- **General Increase**: Most countries experienced improvements in life expectancy from 2000 to 2015, reinforcing positive global health developments.

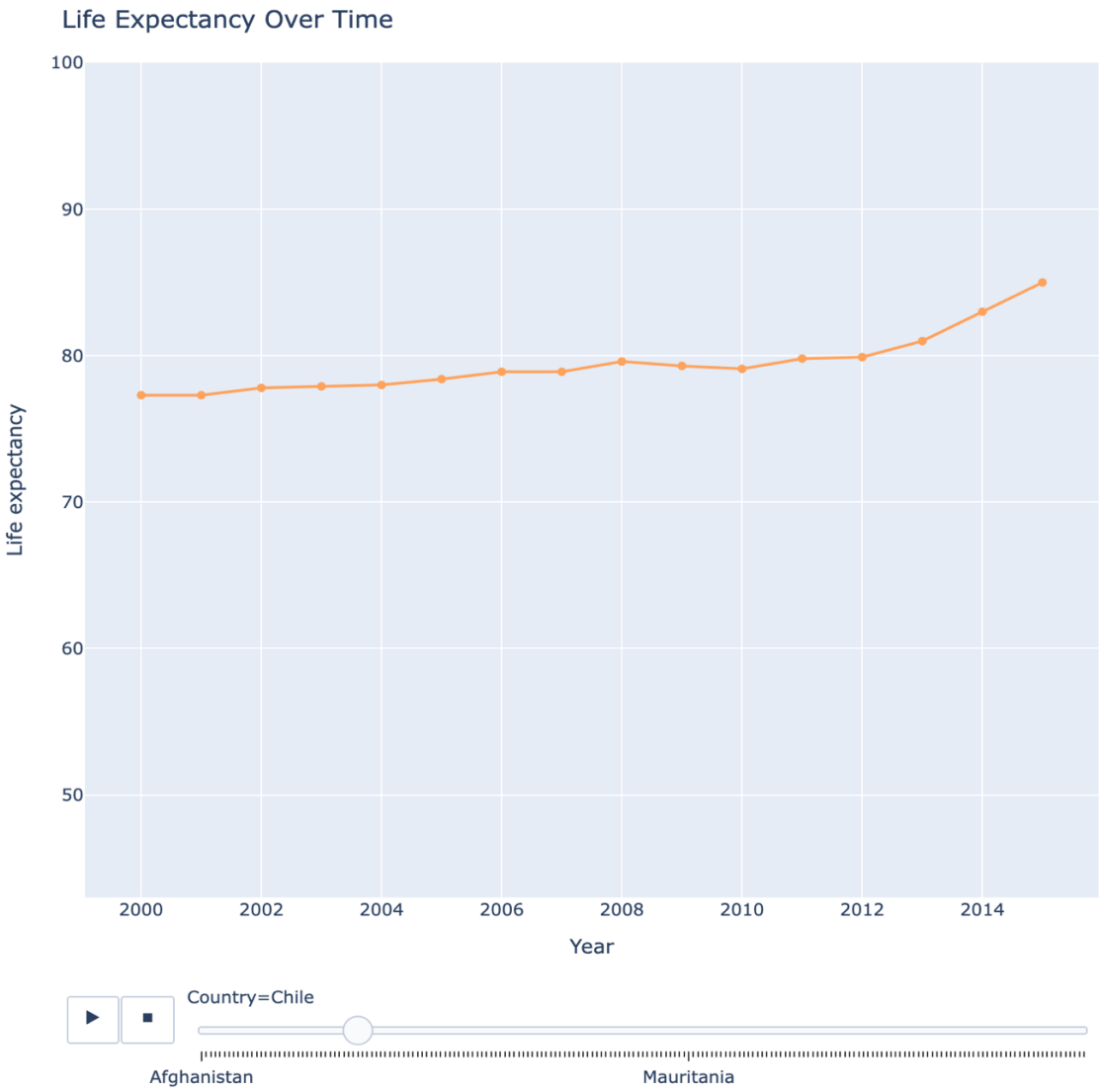

Figure 11

3.3 Conclusion

The EDA uncovered critical relationships and patterns within the dataset, providing valuable context for the subsequent modeling process. Each visualization contributed unique insights, from highlighting socio-economic disparities to illustrating temporal and geographical trends in life expectancy.

4. Data Preprocessing

4.1 Overview

Data preprocessing is a crucial step in preparing the dataset for machine learning models. The goal was to handle categorical data, scale numerical features, and split the data for training and testing to ensure effective and fair model evaluation.

4.2 Techniques and Justifications

Label Encoding:

- Categorical features (Country and Status) were transformed into numerical values using label encoding. This method was chosen because the dataset included categorical variables that needed to be represented numerically for compatibility with machine learning algorithms.

Feature Scaling:

- Numerical features were scaled using Min-Max Scaling, normalizing their values to a range between 0 and 1. This was necessary to reduce the impact of varying feature scales, ensuring that all features contributed equally to the model's learning process. Scaling also helps certain models (e.g., gradient-based algorithms) converge more efficiently.

Train-Test Split:

- The dataset was divided into training and testing subsets, with 70% of the data allocated for training and 30% for testing. This approach was chosen to evaluate the model's generalization performance on unseen data. The split ratio balances sufficient data for training while reserving enough data for testing.

4.3 Conclusion

These preprocessing steps prepared the dataset for modeling by addressing compatibility, scaling, and evaluation needs. This ensures a robust foundation for building and evaluating predictive models.

Modeling

5.1 Overview

This section discusses the training and evaluation of two predictive models: Random Forest and Artificial Neural Network (ANN). These models were chosen to balance interpretability and the ability to capture complex, non-linear relationships in the data. The evaluation metrics include MAE, MSE, RMSE, R^2, and Adjusted R^2.

5.2 Model Selection and Architecture

- ### Random Forest
  Random Forest was selected due to its robustness against overfitting and its ability to model non-linear relationships efficiently. The model was trained with default hyperparameters, except for n_estimators, which was set to 200 to enhance stability and performance.
- ### Artificial Neural Network (ANN)
  An ANN model was chosen to leverage its capability of capturing intricate feature interactions. After experimenting with multiple architectures and hyperparameters, the following structure was finalized:

    - Three hidden layers, each with 64 neurons and ReLU activation.
    - One output layer with a single neuron and a linear activation function for regression.
    - The Adam optimizer and a Mean Absolute Error (MAE) loss function were used. Early stopping was applied to monitor validation loss, with a patience of 10 epochs to prevent overfitting.

5.3 Model Training and Evaluation

- ### Random Forest Performance
  The Random Forest model achieved excellent performance, with low error metrics and high R^2 values:
    - **MAE**: 1.26
    - **MSE**: 5.28
    - **RMSE**: 2.30
    - **R^2**: 0.94
    - **Adjusted R^2**: 0.94

- ### ANN Performance
  The ANN model demonstrated slightly lower performance compared to Random Forest:
    - **MAE**: 1.98
    - **MSE**: 10.14

- **RMSE**: 3.19
- **R^2**: 0.88
- **Adjusted R^2**: 0.88

5.4 Overfitting Analysis

To assess overfitting, the training and validation losses of the ANN were plotted (figure 12). The plot indicated that both losses approached zero, suggesting minimal overfitting and strong generalization on the validation data.

Figure 12

5.5 Results and Comparison

The Random Forest model outperformed the ANN across all metrics. While the ANN captured the non-linear relationships to some extent, Random Forest provided superior predictive power with lower error rates and higher R^2 scores.

Tableau 1

Model	MAE	MSE	RMSE	R2	Adjusted R2
Random Forest	1.26	5.28	2.30	0.94	0.94
ANN	1.98	10.14	3.19	0.88	0.88

5.6 Conclusion

The Random Forest model emerged as the better performer for predicting life expectancy, demonstrating its suitability for this dataset. The ANN, while slightly less accurate, remains a viable option for future experimentation and refinement.

6. Interpretation and Insights

6.1 Feature Importance Analysis

Random Forest Feature Importance

Using the Random Forest model's feature importance scores, several key factors influencing life expectancy emerged:

Top Contributors:

- **HIV/AIDS**: The most influential feature, contributing 57.4% to the model's predictions.
- **Income composition of resources**: A significant socio-economic factor (15.5%).
- **Adult Mortality**: Strongly correlated with life expectancy (12.7%).

Moderate to Minor Contributors:

- Features like Schooling, BMI, and Under-five deaths showed moderate importance but were less influential compared to the top contributors.
- Variables such as Status (developing or developed) had minimal influence, with an importance of only 0.04%.

ANN Feature Importance Using SHAP

To interpret the ANN, Shapley values were used. The SHAP summary plot (Fig. 13) revealed a different order of feature importance:

1. **Income composition of resources**
2. **Adult Mortality**
3. **HIV/AIDS**
4. **GDP**
5. **Status**

Other features such as Diphtheria, Thinness (5-9 years), and Percentage expenditure also showed meaningful contributions.

The differences in feature rankings between Random Forest and ANN highlight the models' distinct mechanisms in identifying patterns.

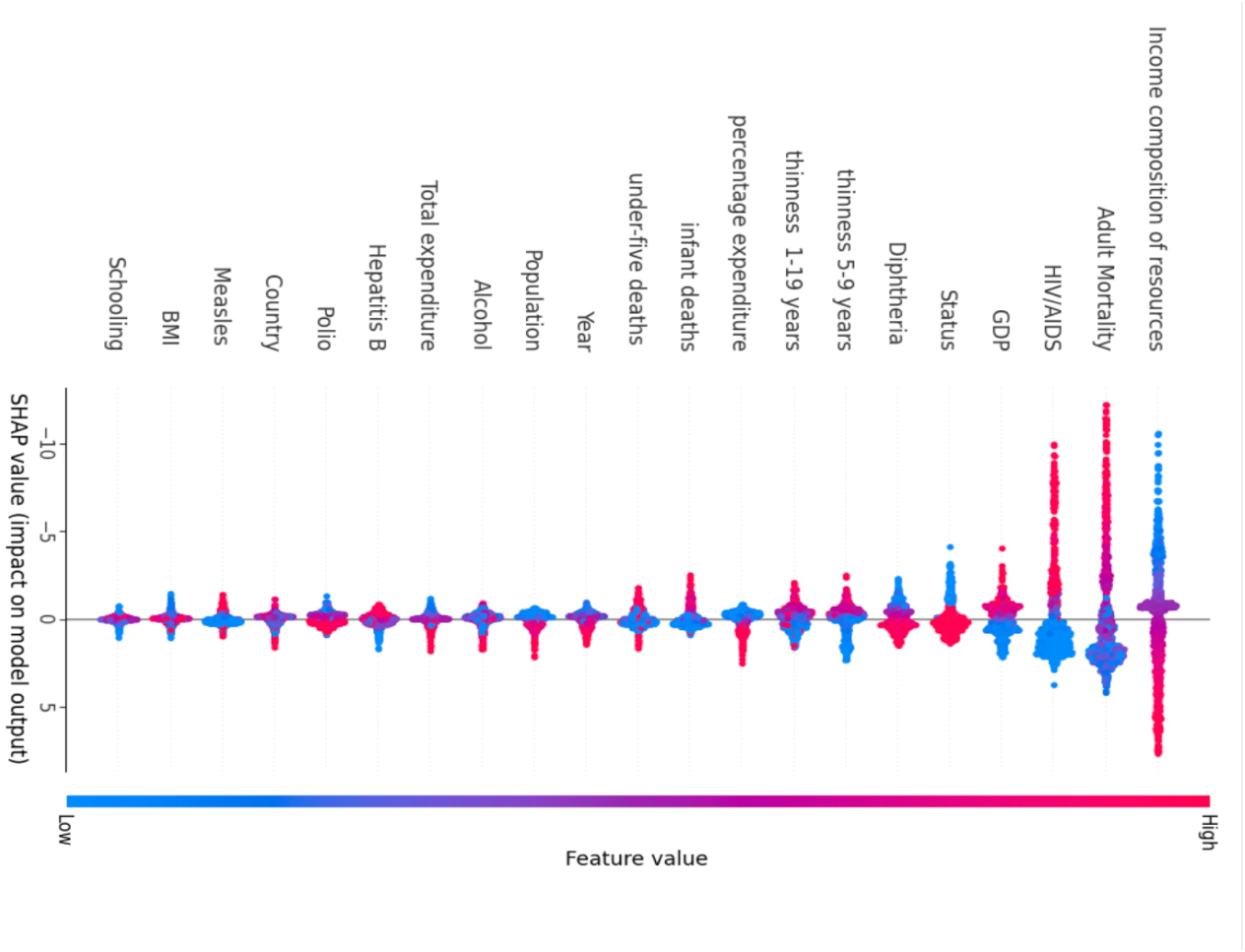

Figure 13

6.2 Individual Prediction Analysis

To enhance transparency, a SHAP force plot (Fig. 14) was used to explain one specific prediction. Key observations include:

- **Features Increasing the Prediction**:
    - **Adult Mortality** (0.3): A significant driver that pushed the prediction higher.
    - **Hepatitis B** (0.225): Contributed positively but to a lesser extent.
- **Features Decreasing the Prediction**:
    - **HIV/AIDS** (0.9659): The most impactful feature reducing the predicted life expectancy.
    - **Diphtheria** (0.5625) and **Polio** (0.4167): Also played roles in decreasing the prediction.

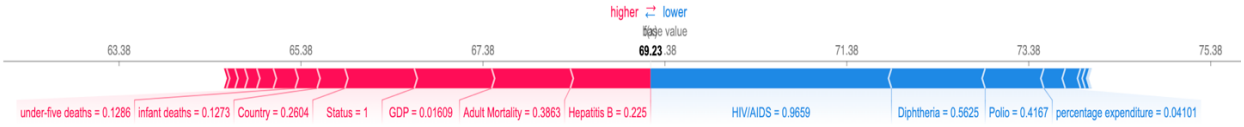

Figure 14

The force plot provided a clear visualization of how each feature contributed to the specific prediction, offering valuable insights into the ANN's behavior.

6.3 Comparative Insights

- Both models highlighted the critical role of socio-economic and healthcare-related features, such as Income composition of resources and Adult Mortality.
- While Random Forest emphasized HIV/AIDS as the most influential factor, the ANN ranked Income composition of resources higher.
- The alignment on certain features but divergence on their relative importance underscores the complementary nature of these models.

6.4 Real-World Implications

These insights can guide policymakers in prioritizing interventions:

- Targeting HIV/AIDS prevalence and improving healthcare infrastructure could yield substantial improvements in life expectancy.
- Socio-economic enhancements, such as better income distribution and education (as captured by Schooling), remain critical to sustainable health outcomes.

Conclusion

This project aimed to explore the factors influencing life expectancy using machine learning models while emphasizing interpretability and human-centered insights. By leveraging Random Forest and Artificial Neural Networks (ANNs), we were able to achieve strong predictive performance, with the Random Forest model outperforming the ANN in terms of evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 score.

The feature importance analysis revealed that factors like HIV/AIDS, income composition of resources, and adult mortality have the most significant impact on life expectancy. The interpretability techniques, including feature importance from Random Forest and SHAP analysis for the ANN, provided transparent insights into how each feature contributed to the predictions, allowing us to link data-driven findings with real-world contexts. For instance, high HIV/AIDS rates were identified as a critical negative driver of life expectancy, emphasizing the importance of public health interventions in this area.

While the models performed well, the project faced challenges such as handling imbalanced features and interpreting the contribution of less impactful features. Future studies could explore the inclusion of additional data sources, advanced model optimization techniques, and a deeper investigation into regional disparities.

Overall, this project demonstrates the potential of machine learning in understanding complex health-related phenomena and underscores the importance of integrating transparent and ethical practices when applying data science to sensitive human issues.

References

- KumarRajarshi. (n.d.). Life Expectancy Dataset. Retrieved from https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who/data

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

- Abadi, M., Agarwal, A., Barham, P., et al. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Retrieved from https://www.tensorflow.org/

- Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*. Retrieved from https://github.com/slundberg/shap

- Waskom, M. (2021). Seaborn: Statistical Data Visualization. *Journal of Open Source Software, 6*(60), 3021. https://doi.org/10.21105/joss.03021

- Plotly Technologies Inc. (2015). Collaborative Data Science. Montréal, QC. Retrieved from https://plotly.com/python/