CRISP-DM Documentation - Global Life Expectancy Analysis and Prediction

Leon Baumgärtner
Cem Ashbaugh

1 Business Understanding

The following chapter outlines the objectives and success criteria for both Business and Data Mining aspects.

1.1 Data Source and Scenario

Data Source Description: The "Life Expectancy (WHO)" dataset available on Kaggle, curated from data by the World Health Organization, spans from 2000 to 2015 covering 193 countries. It features health and development indicators such as life expectancy, mortality rates, health expenditures, and also encompasses socioeconomic factors like education levels, economic status, and population demographics. This comprehensive data can assist with evaluating global health patterns and shaping evidence-based policies.

Business Analytics Scenario: A think tank founded by the WHO seeks to leverage this dataset to develop predictive models for estimating life expectancy across regions. The aim is to provide actionable insights for healthcare providers and policymakers, enabling them to enhance health strategies and public health policies. Using the CRISP-DM framework, the project will involve data cleaning, exploratory analysis, feature engineering, and machine learning. The deliverables include a predictive model, a detailed analytical report, and insightful visualizations, offering customers the tools to make informed decisions and improve healthcare outcomes globally.

1.2 Business Objectives

The primary objective is to leverage the "Life Expectancy (WHO)" dataset to develop predictive models and generate actionable insights for key figures in healthcare, insurance, and public policy.

Specific Goals:

(1) **Model Development:** Build accurate predictive models to estimate life expectancy based on health and socioeconomic indicators.
(2) **Insight Provision:** Provide healthcare providers and insurers with actionable insights to better align services with demographic needs and expectations.

Expected Outcomes:

- A predictive model, comprehensive analytical report, and insightful visualizations.
- Improved healthcare strategies, refined insurance risk assessments, and more effective public health policies.

1.3 Business Success Criteria

(1) **Accuracy of Models:** Success will be measured by the predictive accuracy of the models, assessed using suitable metrics. The goal is to achieve results that indicate strong predictive performance, with errors kept below an acceptable threshold and high explanatory power of the model. Robustness across regions and demographic groups will also be evaluated to ensure generalizability.
(2) **Customer Satisfaction:** Success will also be gauged by the practical utility of insights for healthcare providers, insurers, and policymakers, measured through structured feedback mechanisms such as surveys, workshops, or direct adoption rates of analytics-driven recommendations.

1.4 Data Mining Goals

(1) **Model Development:** Upon preprocessing of the dataset, a suitable regression model for predicting life expectancy will be determined through experiments with different modeling alternatives.
(2) **Insight Provision:** Through understanding of the different variables and their interconnections, their influence on life expectancy will be quantified. This includes descriptive as well as visual analyses as well as feature importances for the prediction of life expectancy.

1.5 Data Mining Success Criteria

(1) **Model Development:** The models will be evaluated using R^2 and RMSE. The model with the highest R^2 and lowest RMSE will be chosen. The R^2 shall be at least 90%.
(2) **Insight Provision:** A list of the most influential predictors of life expectancy will be produced using e.g. the Gini importance in tree-based models, as well as visualizations and summery statistics. Feedback from stakeholders on these materials shall be at least 80% positive.

1.6 AI Risk Considerations

The to be developed system does not fall under the forbidden category in Art. 5 of the EU AI Regulation. Since influence on government policies regarding healthcare (e.g. hospital infrastructure and vaccination) is not out of the question, it may however fall under the high risk classification in Art. 6, particularly *Management and operation of critical infrastructure*. This must be discussed with legal experts.

2 Data Understanding

At this stage, the goal is to understand the dataset thoroughly, which includes descibing it statistically and visually, analysing data quality and more.

2.1 Attribute Types and Semantics

It is crucial to be aware of the meaning and representation of data attributes for all subsequent tasks. Therefore, attribute semantics are interpreted in this report to have four relevant descriptive dimensions: Real-world interpretation, value interpretation, data type and unit of measurement. In table 4, each attribute is described using these dimensions. Due to the extensive number of features, the table is to be found in the appendix.

2.2 Statistical Properties of the Dataset

2.2.1 Categorical Features.

- The **Year** column contains the years 2000 to 2015 for all countries except Cook Islands, Dominica, Marshall Islands,

Monaco, Nauru, Niue, Palau, Saint Kitts and Nevis, San Marino, and Tuvalu (each of those only occurs once in 2013).
- The **Country** column contains 193 unique countries that occur exactly 16 times each, once for each unique year, with the above-mentioned exceptions.
- The **Status** column stays constant for a country across all years, meaning a country is either developing or developed in the dataset. There are 161 developing and 32 developed countries in the dataset.

2.2.2 Numerical Features. Some basic statistics (minimum, maximum and median) for all numerical columns (except for year, which is considered categorical for this purpose) are shown in tables 1 and 2. They will however not be commented here, since the main abnormalities are discussed under data quality.

2.2.3 Correlations. Since most features are numerical, the dataset can be well described using a correlation matrix (see JupyterNotebook). The pearson correlation is used, which measures linear relationships. The most relevant correlations include the following:

- Life expectancy is correlated with schooling (0.75), adult mortality (-0.7), BMI (0.57), HIV/AIDS (-0.56), and thinness (-0.48).
- Thinness from 1-19 years and 5-9 years are highly correlated (0.94) and might contain redundant information.
- Schooling and the Income composition of resources (human development index) are highly correlated with 0.8, indicating that the index calculation heavily weighs education.
- Under-five deaths and infant deaths have a 1.0 correlation, indicating that they are redundant. Since no metadata is provided on what is considered an infant by the dataset creators, this is likely and should be verified.
- Vaccine coverage for Diphtheria, Hepatitis B, and Polio are strongly related amongst each other with correlations of at least 0.49.

2.3 Data Quality Aspects

2.3.1 Missing values: effects and reasons. First, it should be noted that the WHO aggregates different data sources with varying availability depending on the country and other factors. All values (cells) in the dataset contain aggregate values like means and percentages over whole country populations. [5]

They also represent concepts that are theoretically valid in any country (the concept of GDP, vaccination coverage etc.). Therefore, if a value is missing, in this dataset, it is generally due to data availability and sharing by source countries, not because a value does not exist or does not logically make sense. With that in mind, these are the missing value counts for the columns: Population: 652, Hepatitis B: 553, GDP: 448, Alcohol: 194, Income composition of resources: 167, Schooling: 163, Total expenditure: 226, BMI: 34, thinness 1-19 years: 34, thinness 5-9 years: 34, Diphtheria: 19, Polio: 19, Life expectancy: 10, Adult Mortality: 10. The columns that are not mentioned have zero missing values.

For reference, 652 missing values in the Population column is a proportion of 22.2 % of missing values. Such a high percentage can have a considerable negative effect if an attribute is used as an independent variable in inference, since much information is lost when dropping the datapoints and reliable imputation is more challenging. It should also be noted that 10 values are missing in the target column life expectancy, which means there are no labels available for those rows as is.

2.3.2 Uneven class distributions of categorical attributes. As mentioned above, generally, each country is represented 16 times in the dataset for each year from 2000 to 2015. The only exception are the 10 identified countries that only have data for the year 2013. Of all countries, 161 are classified as Developing in the dataset, representing the fact that there are fewer developed countries. Instead of using the country column as a feature, a continent column will be introduced and commented in following sections.

2.3.3 Value plausibility. We'll get an overview of value plausibility using tables 1 and 2. Values that might arguably be extreme are the maximum of 89 years for life expectancy (such values are estimated by the WHO [5]) and 723 for adult mortality out of 1000. A maximum BMI of 87 as well as a minimum of 1 over a population of a country should also be further evaluated, since it would mean the whole population is heavily obese and extremely underweight, respectively. A population of 34 inhabitants and 0 schooling years should also be more colosely evaluated.

Values that do not logically make sense are 1800 infant deaths out of 1000 inhabitants (same for 2500 for under-five deaths) and 212183 measles cases per 1000 inhabitants (as measles can only be contracted once).

2.3.4 Outliers. The outlier count using the IQR method with a scaling factor of 1.5 per feature are the following (unmentioned columns have 0 outliers): Measles: 542, HIV/AIDS: 542, under-five deaths: 394, percentage expenditure: 389, GDP: 365, Diphtheria: 298, Population: 294, Polio: 279, Hepatitis B: 254, thinness 5-9 years: 96, thinness 1-19 years: 89, Adult Mortality: 82, Income composition of resources: 130, Schooling: 44, Total expenditure: 32, Life expectancy: 10.

There is a significant number of outliers for some columns. These insights build a foundation for dealing with the outliers in section 3, with special care to the ones that do not logically make sense.

2.3.5 Previous data provenance and cleaning. It should be noted that the WHO life expectancy dataset has undergone some previous processing before it was accessed in this project. The health data is from the WHO and the economic data from the UN. The WHO uses transformation methods to aggregate data from member countries and make it comparable [5]. The health- and economic data were merged by the dataset creators and shared on Kaggle. As the core data is published by major international institutions, a certain trustworthiness can be assumed. According to the metadata on the dataset on Kaggle, no preprocessing was done by the dataset creators except the merging.

2.4 Visual Exploration of Data

Often times, only using descriptive statistics omits some of the information of a dataset. Therefore, some relevant visualizations are included here. A more comprehensive visual analysis can be found in the JupyterNotebook.

Figure 1 shows the distribution of the dependent variable Life expectancy. It centers around a mean of about 70 years and is skewed to the right, indicating a higher frequency of larger values. Figure 2 shows boxplots for polio vaccination coverage by country status. The distributions of both lend towards higher values, with developed countries having a higher median and fewer outliers in the low percentages, meaning that in developed countries, larger

Statistic	LE	AM	ID	Alc	PE	HepB	Meas	BMI	UD
Min	36.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0
Median	72.0	144.0	3.0	4.0	65.0	92.0	17.0	44.0	4.0
Max	89.0	723.0	1800.0	18.0	19480.0	99.0	212183.0	87.0	2500.0

Table 1: Descriptive statistics for dataset columns (Part 1). Abbreviations: LE - Life Expectancy, AM - Adult Mortality, ID - Infant Deaths, Alc - Alcohol, PE - Percentage Expenditure, HepB - Hepatitis B, Meas - Measles, BMI - Body Mass Index, UD - Under-five Deaths.

Statistic	Polio	TE	Dip	HIV	GDP	Pop	Thin 1-19	Thin 5-9	ICR	Sch
Min	3.0	0.0	2.0	0.0	2.0	3.4×10^1	0.0	0.0	0.0	0.0
Median	93.0	6.0	93.0	0.0	1767.0	1.386542×10^6	3.0	3.0	1.0	12.0
Max	99.0	18.0	99.0	51.0	119173.0	1.293859×10^9	28.0	29.0	1.0	21.0

Table 2: Descriptive statistics for dataset columns (Part 2). Abbreviations: Polio - Polio Immunization Coverage, TE - Total Expenditure, Dip - Diphtheria Immunization Coverage, HIV - HIV/AIDS Death Rate, GDP - Gross Domestic Product, Pop - Population, Thin 1-19 - Thinness for 1-19 Years, Thin 5-9 - Thinness for 5-9 Years, ICR - Income Composition of Resources, Sch - Schooling Years.

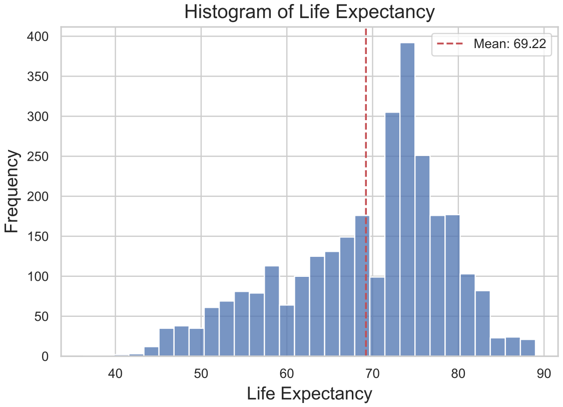

Figure 1: Histogram of the dependent variable Life expectancy.

proportions of the population are vaccinated against polio. The scatter plot in figure 3 shows the relationship of life expectancy and schooling years. We can now visually verify the linear correlation of 0.75 stated above, as higher values of schooling years are clearly associated with a higher life expectancy. The aformentioned extreme values of zero schooling years are also visible, which will be addressed in Chapter 3.

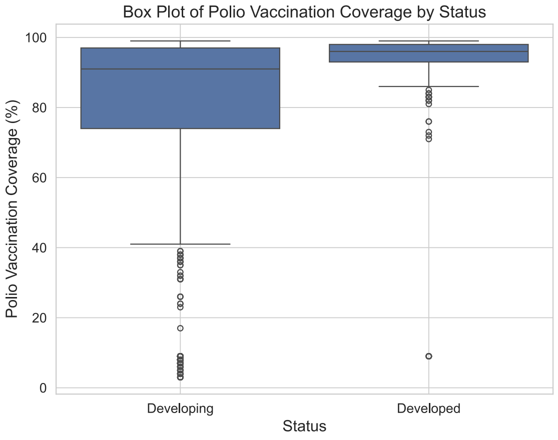

Figure 2: Boxplots of Polio vaccination coverage by country status.

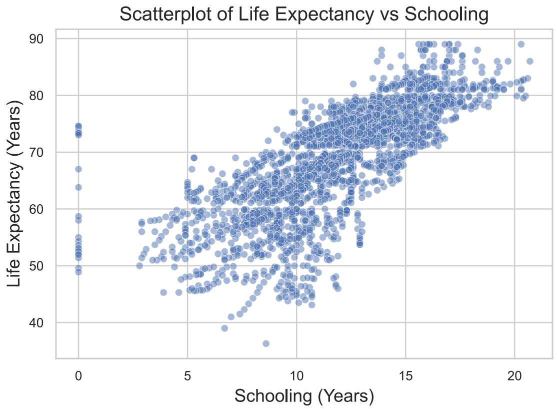

Figure 3: Scatterplot of Life expectancy and Schooling.

2.5 Ethically Sensitive or Underrepresented Data

Some small countries are not evenly represented in the dataset (only once in 2013), neglecting the information those countries might have contributed. Predictions would also be less effective for them. However, apart from that, all member countries of the WHO, which represent almost all countries in the world, are represented in the dataset.

Furthermore, by performing analysis on this dataset and carelessly communicating results, unintended stereotypes could be enforced, e.g. if analyses result in higher illness numbers for the african continent. In general, the overrepsesentation of developing countries could influence the findings disproportionately.

In addition to the imbalanced categorical class distributions mentioned above, there are some skewed numerical distributions. For example, GDP has mostly small values < 10000, with some very large values up to 120000. This will be addressed by dealing with outliers. Transformations like box-cox could be applied to handle such cases. As we deal with a regression task, over- or undersampling of the dependent variable is not applicable here.

2.6 Potential Risks and Bias

Risks and biases associated with the data and its analysis can significantly impact the outcomes of the study.

Risks include the potential for wrong or incomplete deductions based on the data regarding predictors for life expectancy. This can drive bad decisions by government agencies and other institutions. To mitigate such risks, transparency throughout the entire data mining process is crucial.

Biases in the dataset may influence the analysis as well. Inferences might not reflect developed countries as effectively due to underrepresentation. However, it is arguable whether this is a sampling bias since almost all countries are included, and the dataset realistically reflects the fact that there are fewer developed countries. This might also be interpreted as historical bias, as the classification of developing may inherently reflect historical conditions, as the data mirrors the real-world state, even if it is sampled correctly.

Lastly, confirmation bias is a concern, as features might be selected based on prior beliefs about what affects life expectancy. This can lead to the risk of overlooking other potentially relevant factors.

Some questions arise that should be answered by external experts:

- What is considered an infant for the infant deaths column? Is there a slightly different meaning from under-five deaths or are they indeed redundant?
- Is there a sensible meaning/explanation for values much larger than 1000 for features with a scale out of 1000 inhabitants like under-five deaths?
- How were the features chosen? Are there other theoretically interesting features that were not included because of a lack of data or similar?

2.7 Actions Required for Data Preparation

- As there is a considerable amount of missing values, they will preferably be imputed instead of deleted with suitable values based on the identified distributions of and correlations between variables. If it turns out to be sensible, a small amount of datapoints might also be dropped. The same is planned for outliers. This will also address the illogical values at the ends of the distributions identified above.
- If scale-sensitive modeling techniques are used, numerical features will be normalised. Categorical attributes will be encoded as numbers.
- High-correlation features, like under-five deaths and infant deaths, may be removed.
- New features might be introduced to represent the data more effectively for prediction, such as a continent feature.

3 Data Preparation

In this chapter, the steps taken to prepare the dataset for analysis are documented in detail. The necessary pre-processing actions, considerations for additional steps, options for derived features, and potential external data sources are systematically analyzed to ensure data quality and alignment with the business objectives.

3.1 Pre-Processing Actions

3.1.1 ***Handling Missing Values.*** A major focus was addressing missing values, which were thoroughly analyzed at both the column and country levels. Specific columns exhibited varying levels of missing data, and these were handled using tailored imputation techniques based on their importance and the extent of missingness. The treated columns were grouped into the following categories:

- Critical Variables: These variables are directly linked to health outcomes and were prioritized during the imputation process. `Life expectancy` and `Adult Mortality` were treated using median imputation, which is robust to outliers and ensures the data's distribution is preserved. This approach avoids introducing bias, making it suitable for variables that significantly influence the analysis.
- Low Missing Rates: Columns with relatively few missing values were also treated with median imputation, given its computational efficiency and ability to maintain the integrity of the data distribution. This category included `Polio`, `Diphtheria`, `BMI`, `thinness 1-19 years`, `thinness 5-9 years`, `Income composition of resources`, and `Schooling`.
- Moderate Missing Rates: Columns with a higher percentage of missing data required a more advanced imputation strategy. For `Alcohol`, `Total expenditure`, `GDP`, `Hepatitis B`, and `Population`, K-Nearest Neighbors (KNN) imputation was applied. This method estimates missing values by considering relationships between similar rows in the dataset, preserving the multivariate structure of the data. A value of $k = 5$ (five nearest neighbors) was used for this imputation.

Some countries were identified as having no values in specific columns, such as `Population`, `Hepatitis B`, and `GDP`, for the entire period covered by the dataset. Addressing these gaps comprehensively was beyond the scope of this project; however, it is recommended for future work to merge external datasets to enhance data quality and completeness. After applying the imputation methods, a validation step confirmed that no missing values remained in the treated columns, ensuring a consistent dataset for subsequent analysis.

3.1.2 ***Outlier Handling.*** Outlier detection and treatment were essential steps to ensure the robustness of the dataset for subsequent analysis and modeling. Outliers were identified using the interquartile range (IQR) method, a well-established approach that does not rely on assumptions about the data distribution. The thresholds for detecting outliers were calculated as follows:

$$\text{Lower bound} = Q1 - 1.5 \times \text{IQR}, \quad \text{Upper bound} = Q3 + 1.5 \times \text{IQR}$$

where $Q1$ is the 25th percentile, $Q3$ is the 75th percentile, and IQR is the interquartile range ($Q3 - Q1$).

For each numerical column, values below the lower bound or above the upper bound were considered outliers. These extreme values were capped at the respective thresholds to reduce their influence while preserving the data's overall structure. This method ensures that outliers are treated consistently across the dataset without discarding potentially valuable information.

To evaluate the impact of outlier handling, boxplots were created for each numerical column before and after capping. These visualizations demonstrated that capping effectively mitigated the effect of extreme values while maintaining the overall variability of the data. For example:

- Effective Treatment: Columns like `Schooling` exhibited significant improvement in their distribution, with reduced extreme values that were unlikely to represent real-world scenarios.

- Challenging Cases: Some variables, such as `HIV/AIDS`, showed high inherent variability, where outlier treatment had limited impact. In these cases, alternative approaches may be more appropriate in future work.

In addition to column-level outlier detection, population variability was analyzed in-depth. High Coefficient of Variation (CoV) values and extreme Year-over-Year (YoY) percentage changes in the `Population` column highlighted potential data quality issues or real-world phenomena, such as migration or conflicts. For instance:

- Countries with CoV values exceeding 200% or implausibly low population values (e.g., below 100) were flagged for further investigation.
- YoY percentage changes greater than ±500% were capped to mitigate the effect of reporting inconsistencies or extreme demographic shifts.

While the applied methods effectively reduced the impact of outliers in many cases, some limitations were noted. For example, outlier capping may not adequately address systemic data issues or extreme variability in certain features. These cases were documented, and recommendations for alternative strategies, such as domain-specific thresholds or machine learning-based anomaly detection, were provided for future work.

By systematically addressing outliers, the dataset was prepared for robust analysis, with reduced bias from extreme values and enhanced reliability for modeling purposes.

3.1.3 ***Dropped Columns.*** As part of the pre-processing, columns with a correlation coefficient of at least 0.85 to other features were identified and dropped to reduce redundancy and avoid multicollinearity. The following columns were removed based on these criteria:

- `percentage expenditure`: This column was dropped due to a linear correlation of 0.9 with `GDP`. The decision to retain `GDP` over `percentage expenditure` was made because the dataset already includes `total expenditure`, which provides a more comprehensive measure of spending.
- `thinness 5-9 years`: This column was removed due to a correlation of 0.94 with `thinness 1-19 years`. The latter was retained because it represents a broader age range and is likely more informative for health-related analysis.
- `infant deaths`: This column was dropped as it exhibited a perfect correlation (1.0) with `under-five deaths`. `under-five deaths` was retained, as the metadata does not clearly define what is counted as an infant death, making `under-five deaths` a more reliable and interpretable feature.

By removing these columns, the dataset was streamlined, and potential issues with redundancy and multicollinearity were mitigated, ensuring a cleaner and more interpretable feature set for analysis.

3.1.4 ***Encoding of Categorical Variables.*** To prepare the categorical variables for analysis, appropriate encoding methods were applied:

1. One-Hot-Encoding for `Status`: The `Status` column, which indicates whether a country is "Developed" or "Developing," was one-hot encoded. This resulted in two binary columns: one for `Developed` and one for `Developing`. This approach ensures that the categorical variable is represented numerically without introducing ordinal assumptions.

2. Grouping and Encoding of `Country`: The `Country` column contains 193 unique countries, making direct one-hot encoding impractical due to the high dimensionality it would introduce. Instead, countries were grouped by their respective continents (e.g., Europe, Asia, Africa, North America, South America, Oceania), reducing the number of categories to six. These grouped regions were then one-hot encoded, resulting in six binary columns. This approach balances the preservation of geographic context while reducing complexity in the feature space.

3.1.5 ***Scaling of Numerical Variables.*** To standardize numerical variables and ensure comparability, the following scaling methods were applied:

- **StandardScaler:** Centers features by removing the mean and scaling to unit variance, effective for normally distributed data. The transformation is defined as:
$$z = \frac{x - \text{mean}}{\text{std}} \tag{1}$$
where x is the original value, mean is the mean of the feature, and std is the standard deviation. This ensures the resulting values have a mean of 0 and a variance of 1.
- **MinMaxScaler:** Scales features to a fixed range, typically [0, 1], preserving the original distribution shape. The transformation is defined as:
$$x' = \frac{x - \min}{\max - \min} \tag{2}$$
where x is the original value, and min and max are the minimum and maximum values of the feature. This method ensures that all values are scaled between 0 and 1.
- RobustScaler: Uses the median and interquartile range (IQR), making it robust to outliers.
- MaxAbsScaler: Scales values by the maximum absolute value, ensuring a range of [-1, 1], ideal for sparse data.

Exclusion of Categorical Variables: One-hot encoded variables, such as `Status_Developed`, `Status_Developing`, and continent-based binary columns (e.g., `Continent_Asia`, `Continent_Europe`), were excluded from the scaling process. These variables are binary (0 or 1) and do not require normalization or standardization, as scaling them could distort their categorical nature.

This multi-scaling approach ensures that numerical features are appropriately standardized for the analysis, while categorical variables remain unaltered to retain their interpretability and integrity.

3.2 Unapplied Pre-Processing Steps

Several pre-processing steps were considered but not implemented during the data preparation phase for the following reasons:

- **Data Cleansing:** Extensive cleansing addressed many issues, but gaps such as incomplete data for `Population`, `Hepatitis B`, and `GDP` were left unaddressed, as filling these would require external data sources beyond this project's scope.
- **Transformations:** Logarithmic or power transformations for variables like `GDP` and `Total expenditure` were considered to correct skewness but were ultimately not used because exploratory analysis indicated minimal impact on the analysis outcomes.
- **Binning:** Binning continuous variables like `Life expectancy` and `BMI` into categories was avoided to preserve data granularity, which is crucial for model interpretability.

- **Outlier Removal:** Instead of removing outliers entirely, they were capped using the IQR method in columns like `Population` and `HIV/AIDS` to preserve extreme values that might signify significant events.
- **Feature Removal:** Columns missing significant data were retained rather than removed to preserve their potential value in future analyses with additional data sources.
- **Transcoding:** Encoding of categorical variables was deferred, as most were already in formats suitable for analysis, with further encoding planned only if required by specific algorithms.

Conclusion: The steps above were omitted to maintain data integrity, reduce complexity, and ensure result interpretability. The iterative nature of the CRISP-DM process allows for revisiting preprocessing steps based on insights from later phases, ensuring continuous improvements in data quality and model performance.

3.3 Options for Derived Features

Several options for derived features were considered during the data preparation phase. Their potential impact on the analysis is discussed below:

- **Life Expectancy Grouping:** Categorizing life expectancy into "Low," "Medium," and "High" was considered for easier comparison but rejected due to loss of detail crucial for regression analysis.
- **Mortality Ratios:** A ratio of adult mortality to infant deaths was explored to assess health system effectiveness. Not implemented due to time constraints and indirect relevance to main goals.
- **Normalized Health Expenditure:** Normalizing total expenditure by population for per-capita health spending analysis was considered. High potential but postponed to focus on more direct variables.
- **Population Growth Rate:** Year-over-year population change was considered to highlight demographic trends. Not used due to data variability and inconsistencies.
- **Interaction Terms:** Potential interaction terms between income, schooling, and GDP were analyzed to gauge their collective impact on life expectancy. Omitted to simplify the feature set and due to uncertain added value.
- **Combining Redundant Features:** Merging similar columns like Hepatitis B, Diphtheria, and Polio was considered to streamline the dataset but decided against to preserve information clarity and feature interpretability.

Conclusion: Derived features were evaluated for relevance, interpretability, and potential insights. Options such as `Life Expectancy Grouping` and `Normalized Health Expenditure` were not implemented to maintain simplicity and focus on core objectives but are documented for possible future exploration.

3.4 Options for External Data Sources

To enhance the analysis and better address business objectives, we explored integrating external data sources to enrich the dataset with valuable features:

- **Economic Indicators:** Include metrics like `Unemployment Rate`, `Inflation Rate`, and `Gini Index` to capture economic conditions influencing life expectancy.
- **Environmental Data:** Add factors such as `Air Quality Index`, `CO`$_2$ `Emissions`, and `Access to Clean Water` to evaluate environmental health impacts.
- **Healthcare Metrics:** Incorporate `Physicians per Capita`, `Hospital Beds`, and `Vaccination Rates` to assess healthcare quality and accessibility.
- **Demographic and Epidemiological Data:** Integrate `Age Distribution`, `Urbanization Rate`, and disease prevalence rates from sources like WHO to refine health analyses.
- **Education and Political Context:** Include `Literacy Rates`, `Higher Education Enrollment`, and `Political Stability Index` to explore the role of education and governance on health outcomes.

Conclusion: Integrating external data sources was beyond this project's scope but could add value in future analyses. Metrics like environmental and healthcare data may help explore regional disparities in life expectancy and are documented as potential avenues for future enrichment.

4 Modeling

In this section, the necessary steps for a initial model training are outlined, including algorithm selection, hyperparameter optimization, and consideration for performance measurement.

4.1 Initial Data Mining Algorithm Selection

There is a large collection of traditional and modern machine learning algorithms that can be used for the task at hand, which is life expectancy regression with tabular data. One model family in particular has received much attention in previous years in machine learning competitions and industry, which are gradient-boosted decision trees such as XGBoost. They are widely accepted as the recommended option for real-life tabular data problems. [7]

More recently, researchers are working to close the gap between deep neural network approaches and gradient-boosted decision trees for tabular data with some success. The idea is to introduce neural network architectures that are especially suitable for tabular data. [2, 8]

However, Ravid Shwartz-Ziv and Armon [7] point out that deep neural network approaches for tabular data are to be taken with a grain of salt, since they are more challenging to optimize on a new dataset compared to XGBoost. We therefore decide to use a gradient-boosted decision tree model in this project, in particular the Python library Scikit-learn's implementation of the Gradient Boosting Regressor [6].

4.2 Hyperparameter Considerations

The Gradient Boosting Regressor combines the strengths of decision tree ensambling and gradient descent, leading to a wide variety of hyperparameters controlling this process.

The most widely used hyperparameters are the following:

- Learning rate: Controls the contribution of a new tree's predictions to the existing model's predictions.
- Number of estimators: The number of trees used in the prediction.
- Max depth: The number of splits from the root to the leaf node for each tree, which controls overfitting.
- Loss function: The loss function used in gradient descent.

- Other parameters to control overfitting, like the minimum number of samples to perform a split or in a leaf node.

We argue that the most direct impact on the learning process and therefore performance is exercised by:

- Learning rate, since it is the scaling factor by which the previous predictions are "moved" to minimize the loss.
- Number of estimators, since it controls the total number of iterations in which the model learns by adjusting previous predictions based on the gradient of the loss function.
- Max depth, since it controls the size of each tree and therefore model complexity (mainly to control overfitting).

This is verified by obtaining importances on the target prediction for the three hyperparameters mentioned above. To this end, 100 random combinations of values are generated, with the intervals [0.01, 0.25), [100, 300) and [3, 6) for learning rate, number of estimators and max depth, respectively. The exact generated values are provided in the JupyterNotebook that accompanies this report. For each combination, 5-fold cross-validation is performed using the combined training and validation set and the mean RMSE across the five folds is calculated. A seed of 42 is used for drawing the folds and hyperparameter values using scikit-learn's RandomizedSearchCV methods to ensure producing the same values every time. To get a measure of hyperparameter importance, the variance in performance explained by each one is calculated using the formula $\text{Importance}_h = \text{Var}\left(\mathbb{E}[\text{RMSE} \mid x]\right)$, where h is the hyperparameter and $\mathbb{E}[\text{score} \mid x]$ is the mean RMSE conditioned on a specific value of the hyperparameter.

The results are 0.045, 0.031 and 0.015 for learning rate, number of estimators and max depth, respectively. Accordingly, the learing rate is chosen for optimization.

4.3 Data Split

The data is split randomly into three subsets for training, validation and testing while preserving the temporal component of the Year column. This ensures that we can simulate testing the performance of a model trained on past data on future data. This is meaningful since the prediction of life expectancy using new data collected for a country is of interest. Years 2000 to 2010 are used for the training set, 2011 to 2013 for the validation set and 2014 to 2015 for the test set, which corresponds roughly to a 70/20/10 split. This ensures a balance of sufficient amounts of data to train the model with still enough remaining for reliable hyperparameter tuning and testing.

A decision against a true time series forecast for individual countries was made mainly due to the limited historical depth of the dataset, with only 16 years for each country.

4.4 Hyperparameter Optimization and Model Training

The same methodology for attaining hyperparameter-performance pairs is used as in hyperparameter optimization, with the only difference being that only the learning rate is adjusted. The learning rate that corresponds with the lowest R^2 value is determined as 0.21. This value is used to fit the Gradient Boosting Regressor on the training set. The resulting RMSE and R^2 are 2.487 and 0.915 for the validtation set, respectively.

4.5 Performance Metrics

The Coefficient of variance R^2 and the Root Mean Squared Error RMSE offer a great insight into the model performance. The R^2 explains how much of the variance in the target life expectancy is explained by the model, in this case about 91.2%. The RMSE is the square root of the average differences between the true values and predictions. Since it is conveniently in the same unit as the target variable, in this case years of life expectancy, it complements the R^2 effectively by providing additional interpretability of the predictions. This means the average error in the predictions is roughly 2.5 years.

The relationship of the learning rate with the performance metrics RMSE and R^2 is depicted in Figure 4. A better performance, as represented by a lower RMSE and higher R^2, is clearly associated with larger learning rate, with the performance growing especially quickly approaching a value of about 0.05. The generally higher performances in hyperparameter optimization compared to the actual model training might be explained by the fact that in the first case, the effective training fold size is about 72% of the original data while the actual training split constitutes almost 70%.

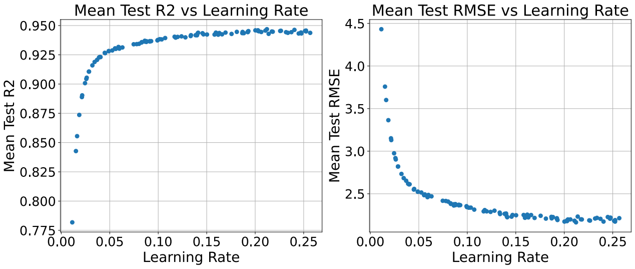

Figure 4: Relationship of the learning rate with RMSE and R^2.

5 Evaluation

This section evaluates the Gradient Boosting Regressor model's performance through detailed metrics such as R^2 Score, RMSE, MAE, and MSE, and assesses its predictive accuracy on historical and recent data. It includes visual representations like scatter plots for a direct comparison of predicted versus actual values and discusses the impact of re-training with an augmented dataset. The aim is to verify the model's effectiveness and identify areas for improvement.

5.1 Apply the final model

5.1.1 Objective. The objective of this analysis was to assess the performance of the Gradient Boosting Regressor model, trained on historical data from 2000 to 2010, when applied to the test dataset spanning the years 2014 to 2015.

5.1.2 Method. The model's predictive accuracy was evaluated using two key metrics: **$\mathbf{R}^2$ Score (Coefficient of Determination)**, which indicates the proportion of variance in the dependent variable predictable from the independent variables, and **Root Mean Squared Error (RMSE)**, which reflects the average magnitude of the prediction errors.

5.1.3 Results and Analysis. The evaluation yielded an **$\mathbf{R}^2$ Score** of 0.8941, a **Root Mean Squared Error (RMSE)** of 2.7080, a **Mean Absolute Error (MAE)** of 2.0483, and a **Mean Squared Error (MSE)** of 7.3330. These metrics confirm that the model explains approximately 89.41% of the variance within the test dataset, with an average prediction error of 2.7080 units.

Figure 5 illustrates the relationship between the model's predictions and the true values. Each point represents a single prediction, where the x-axis corresponds to the actual (true) values and the y-axis shows the predicted values. The diagonal line represents perfect predictions ($y = x$). The clustering of points around the diagonal indicates that the model predictions are generally close to the true values, demonstrating its accuracy.

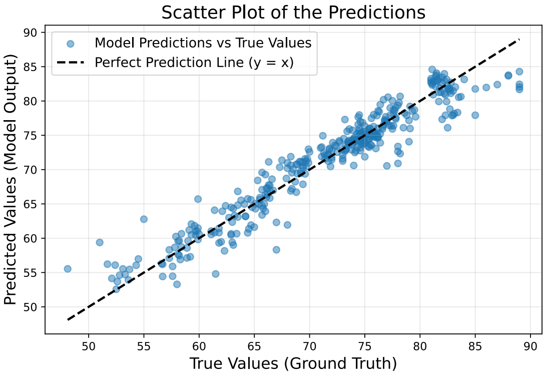

Figure 5: Scatter plot comparing True Values and Model Predictions. The diagonal line represents ideal predictions where predicted values perfectly match actual values. The spread of points indicates the degree of prediction error, with most points clustering near the diagonal, demonstrating the model's strong performance.

5.1.4 Conclusion. The evaluation reveals robust predictive performance, with the model effectively capturing the trends and variations in the dataset. The scatter plot (Figure 5) visually reinforces this by showing a strong alignment between predictions and actual values. Considering these results, future enhancements might involve exploring additional feature engineering techniques or alternative model configurations to further reduce prediction errors and improve accuracy.

5.2 Re-train the model

5.2.1 Objective. The objective of this evaluation was to assess the performance of the Gradient Boosting Regressor model when retrained using both training and validation datasets from 2000 to 2013 and subsequently tested on data from 2014 to 2015.

5.2.2 Method. The model was re-trained with the identical hyperparameters previously used, ensuring consistency in the model's configuration. The expanded dataset aimed to enhance the model's learning and generalization capabilities.

5.2.3 Results and Analysis. The re-trained model achieved an overall R^2 Score of 0.9188, a RMSE of 2.3712, a MAE of 1.7340, and a MSE of 5.6226.

These results indicate an improvement over the model trained solely on the training set, reflecting enhanced predictive accuracy and reliability. The clustering of predicted values near their true counterparts highlights the model's improved capability to generalize effectively.

5.2.4 Reflection. This improved performance, evidenced by a higher R^2 Score and a lower RMSE, MAE and MSE, demonstrates the benefits of training the model with a more comprehensive dataset. The consistent alignment of predictions to actual values supports the model's enhanced ability to generalize on unseen data.

5.2.5 Conclusion. The use of both training and validation datasets for re-training the Gradient Boosting Regressor has significantly improved its performance on the test data, underlining the importance of utilizing as much relevant data as possible for training to enhance the model's robustness and accuracy. Future efforts should continue to explore the use of expanded training datasets, alongside further hyperparameter optimization and feature engineering, to optimize the model's performance in real-world scenarios.

5.3 Perfomance Identification

5.3.1 Objective. The objective of this task is to identify and document the state-of-the-art performance achieved by others using the same WHO dataset available on Kaggle, or similar datasets. The goal is to contextualize the model's performance by comparing it to benchmarks in peer-reviewed literature or other analyses.

Baseline and Comparison Models. Baseline models were implemented to assess performance relative to the trained models.

The **mean baseline** predicted the average life expectancy across all data points, achieving an MAE of 6.8508, an MSE of 69.2623, and an R^2 Score of 0.0000. This demonstrates no ability to capture variance in the data, as the model simply predicts the same value for all samples regardless of their features.

A **group-based baseline** predicted the mean life expectancy per continent, significantly improving upon the mean baseline. This approach achieved an MAE of 4.2243, an MSE of 28.9876, and an R^2 Score of 0.5815, indicating a better capacity to account for regional differences in life expectancy.

While these baselines establish minimal performance standards, they are substantially outperformed by advanced machine learning models such as Random Forest and XGBoost. These models achieved RMSE values as low as 1.98 and R^2 Scores up to 0.9609, demonstrating their ability to leverage complex interactions within the data for improved predictive accuracy.

5.3.2 i) State-of-the-Art Performance. Several studies have utilized the Kaggle WHO Life Expectancy dataset or similar datasets to predict life expectancy and analyze its determinants. Below is a detailed summary of notable studies and their reported performance.

Study 1: "An Application of a Supervised Machine Learning Model for Predicting Life Expectancy [4]". This study utilized the Kaggle WHO Life Expectancy dataset, covering multiple countries and health, socioeconomic, and behavioral indicators. The authors employed the eXtreme Gradient Boosting (XGBoost) algorithm to predict life expectancy. Features were preprocessed to handle missing values, and hyperparameters were fine-tuned using cross-validation. The study reported a MAE of 1.554 and a RMSE of 2.402, setting a high benchmark for models using this dataset. XGBoost significantly outperformed other models like Random Forest (RF) and Artificial Neural Networks (ANN) in terms of prediction accuracy, with GDP, immunization rates, and mortality rates highlighted as key predictors.

Study 2: "Analyzing Implications of Various Social Factors on Life Expectancy [3]". Using the WHO dataset, this study analyzed the effects of social factors on life expectancy across 193 countries. Multiple machine learning models, including Multiple Linear Regression, Decision Tree, Random Forest, XGBoost, AdaBoost, and Support Vector Machines (SVM), were tested to predict life expectancy. The best performance was achieved by Random Forest, with an R^2 Score of 0.9609 and RMSE of 1.98, slightly outperforming XGBoost (R^2

Score: 0.955, RMSE: 2.125). Other models included Multiple Linear Regression (R^2: 0.9479, RMSE: 2.290), SVM (R^2: 0.946, RMSE: 2.24), AdaBoost (R^2: 0.896, RMSE: 3.233), and Decision Tree (R^2: 0.88, RMSE: 4.289). Random Forest emerged as the top-performing model, with key predictors identified as healthcare expenditure, schooling, and immunization rates.

Study 3: GitHub Project on "Life Expectancy Prediction Using Advanced Machine Learning Models [1]". This project implemented a variety of machine learning models on the Kaggle WHO dataset, including Linear Regression, Random Forest Regressor, Gradient Boosting Regressor, and XGBoost Regressor. The reported performance metrics were as follows: Linear Regression achieved an RMSE of 0.459074 and an R^2 Score of 0.802415, demonstrating reasonable performance but lower accuracy compared to ensemble models. Random Forest Regressor significantly outperformed Linear Regression, with an RMSE of 0.217848 and an R^2 Score of 0.955507. Gradient Boosting Regressor delivered competitive results, achieving an RMSE of 0.250281 and an R^2 Score of 0.941272. XGBoost Regressor was the top-performing model in this project, with an RMSE of 0.211332, an R^2 Score of 0.958128, and a cross-validation R^2 Score of 0.9611. These results underline the efficacy of ensemble methods, particularly XGBoost and Random Forest, in handling the complexity of the dataset and achieving state-of-the-art performance.

Subgroup Performance Analysis. In addition to overall model performance, subgroup correlations and error metrics were analyzed to evaluate the model's performance across different continents. The overall performance metrics were as follows:

Region	MAE	MSE	RMSE	R^2 Score
Africa	2.4407	9.6224	3.1020	0.7779
Asia	1.8650	6.5018	2.5499	0.7716
Europe	2.1584	8.3339	2.8868	0.5832
North America	1.8003	5.1891	2.2780	0.6498
Oceania	1.5914	3.5545	1.8853	0.8988
South America	1.4576	3.9342	1.9835	0.7445
Overall	**2.0483**	**7.3330**	**2.7080**	**0.8941**

Table 3: Segmented Error Metrics and Overall Model Performance

The overall correlation between predicted and true values was 0.9586, indicating high predictive accuracy. Subgroup correlations were 0.9134 for Africa, 0.8978 for Asia, 0.8079 for Europe, 0.9180 for North America, 0.9842 for Oceania, and 0.8910 for South America. These results suggest that the model performs consistently across most regions, with slightly lower correlations and R^2 scores observed in Europe and South America. In contrast, Oceania demonstrated the highest correlation and R^2 score, reflecting the model's strong predictive capability for this subgroup. This analysis underscores the importance of evaluating subgroup performance to identify regional variations and potential areas for improvement.

5.3.3 *Conclusion.* Tree-based algorithms, particularly XGBoost and Random Forest, have proven effective in managing the complexity of the Kaggle WHO dataset and achieving state-of-the-art performance, with R^2 Scores up to 0.8941 and RMSE as low as 2.0483. These models outperform simpler approaches like Multiple Linear Regression and SVM, which, while emphasizing interpretability, fall short in predictive accuracy. Baseline models, such as the mean and group-based baselines, significantly underperform advanced methods, further emphasizing the value of machine learning techniques in leveraging this dataset effectively.

Although this project and similar GitHub repositories and Kaggle notebooks provide practical benchmarks, they are not peer-reviewed and may lack the rigorous validation and standardized methodologies typical of academic publications. Nonetheless, such implementations offer valuable insights into the potential of ensemble-based approaches.

The performance of the Gradient Boosting Regressor, while competitive, could be further improved by exploring additional hyperparameter optimizations, such as tuning the number of trees, learning rates, and maximum depths. Implementing grid search cross-validation could allow for a more exhaustive exploration of the hyperparameter space, potentially uncovering configurations that yield even better predictive accuracy. These steps, combined with the demonstrated efficacy of ensemble methods, underscore the opportunity for further enhancing predictive models on this dataset.

5.4 Comparison of Performance with Business Objectives and Success Criteria

5.4.1 *Overview.* The performance of the predictive models developed in this project is evaluated against the business success criteria defined in the Business Understanding phase. These criteria focus on achieving strong predictive accuracy, robust generalizability across regions, and the provision of actionable insights for stakeholders in healthcare, insurance, and public policy.

5.4.2 *Comparison with Business Success Criteria.*

Accuracy of Models. The final Gradient Boosting Regressor model achieved strong performance on the test dataset, with an R^2 score of 0.8941, a Root Mean Squared Error (RMSE) of 2.7080, a Mean Absolute Error (MAE) of 2.0483, and a Mean Squared Error (MSE) of 7.3330. These results indicate a high level of predictive accuracy and reliability.

While the R^2 score narrowly missed the 90% target, the achieved value of 0.8941 demonstrates strong predictive accuracy, with RMSE and MAE remaining within acceptable practical limits. Regional analysis revealed areas for improvement, particularly in Europe and South America, where lower R^2 scores indicate the need for targeted refinements.

Feature importance analysis identified `HIV/AIDS (54.39%)`, `Income composition of resources (18.05%)`, and `Adult Mortality (14.54%)` as the most significant predictors, aligning with the goal of providing actionable insights. However, the overwhelming importance of `HIV/AIDS` raises concerns about potential biases or imbalances in the dataset, overshadowing other significant indicators such as `Schooling (0.94%)` and `under-five deaths (2.50%)`. This requires further investigation to ensure the model does not over-rely on a single variable.

Other predictors, such as `Diphtheria (1.63%)`, showed moderate influence, while features like `GDP (0.10%)` and `Population (0.10%)` contributed minimally. The distribution of feature importance highlights the need for future work on feature engineering or alternative models to balance significance across meaningful variables, improving model robustness and interpretability.

2. Customer Satisfaction.

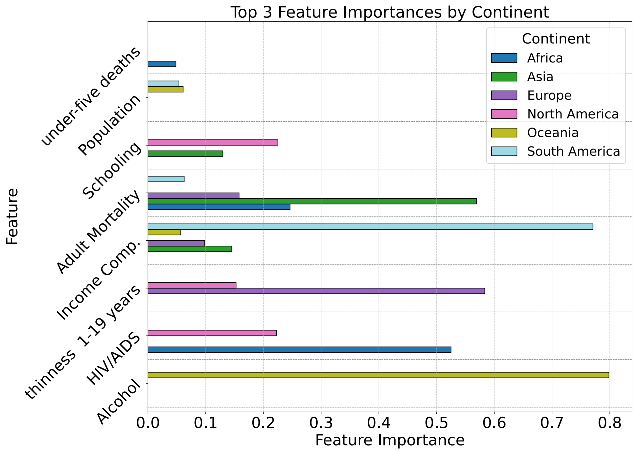

Figure 6: Visualization of the top 3 feature importances for each continent. The horizontal bars represent the relative importance of each feature within a specific continent, with distinct colors indicating the continents.

- The comprehensive analytical report and visualizations, such as feature importance plots and segmented performance metrics, provide actionable insights designed for healthcare providers, insurers, and policymakers. These insights are expected to facilitate better decision-making in aligning services with demographic needs.
- However, structured feedback mechanisms (e.g., surveys, workshops, or direct adoption rates) were not implemented in this project. Achieving this would have required engagement with real customers or stakeholders, which was beyond the scope of this exercise as it focused on a fiktive dataset and target audience.

5.4.3 *Comparison with Data Mining Success Criteria.*

1. Model Development.

- The final model met the success criteria in achieving strong predictive accuracy, with an R^2 Score close to the 90% target at 0.8941, while maintaining low error rates (RMSE: 2.7080).
- The model significantly outperformed baseline models, such as the mean baseline (R^2 = 0.0000, RMSE = 6.8508) and the group-based baseline (R^2 = 0.5815, RMSE = 4.2243). This demonstrates its ability to leverage health and socioeconomic indicators effectively for accurate life expectancy predictions.

2. Insight Provision.

- The feature importance analysis quantified the influence of key variables on life expectancy, meeting the goal of providing descriptive and visual analyses. The results are expected to be useful for stakeholders in healthcare, insurance, and public policy.
- Formal stakeholder feedback (e.g., surveys or workshops) was not collected, as the project lacked engagement with real-world customers. This limits the ability to assess the practical utility of the insights fully.

5.4.4 *Key Findings and Limitations.*

- **Achievement of Goals:** The project achieved strong predictive accuracy and provided actionable insights into life expectancy determinants. The final model demonstrated robust generalizability and outperformed baseline models.
- **Limitations:** The lack of engagement with real customers meant that certain business success criteria, such as structured feedback and direct adoption rates, could not be assessed. Future implementations involving real-world stakeholders would address this gap and validate the practical utility of the results.
- **Opportunities for Improvement:** Additional hyperparameter optimization (e.g., grid search cross-validation) and further model refinements could enhance performance, particularly in regions with lower R^2 scores.

5.5 Bias Analysis

The feature `Continent` was analyzed to identify potential model biases. Results revealed significant disparities in performance across regions. The model achieved the highest R^2 score of 0.8988 in `Oceania` and the lowest at 0.5832 in `Europe`, with RMSE ranging from 1.8853 (`Oceania`) to 3.1020 (`Africa`). These findings suggest better performance in smaller, less variable groups and lower accuracy for larger, heterogeneous populations.

Figure 7 shows the residual distribution across continents. `Africa` and `Europe` exhibit wider interquartile ranges and more outliers, indicating greater variability. Positive residuals suggest underestimation, while negative residuals point to overestimation. Notably, `Europe` shows a bias towards overestimation, with a higher proportion of negative residuals.

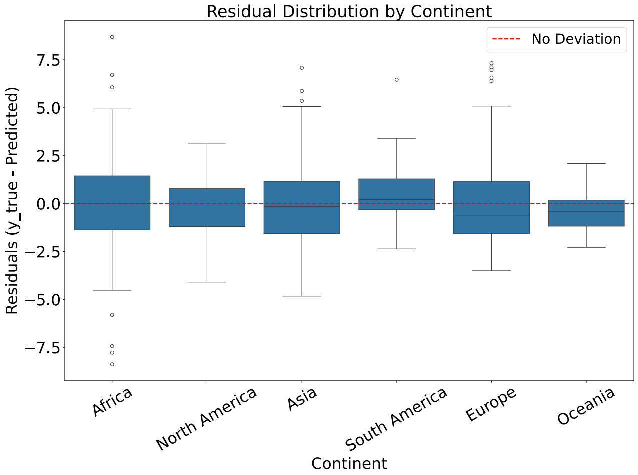

Figure 7: Residual Distribution by Continent. The boxplot shows the residuals for each continent, where positive values indicate underestimation and negative values indicate overestimation by the model. Wider interquartile ranges and outliers in some regions, such as `Africa` and `Europe`, highlight greater variability in model performance.

Group sizes provide additional context: `Africa` (29.51%) and `Asia` (25.68%) account for the largest shares of the population, while smaller groups like `Oceania` (5.46%) and `South America` (6.56%) may benefit from less complex, more uniform data, resulting in higher model performance. However, the limited representation of smaller regions raises concerns about potential overfitting.

Incorporating the `Status_Developed` feature revealed further biases. Regions with only Developing countries, such as `Africa` (0.9134) and `South America` (0.8910), demonstrated high total correlations. In contrast, regions with both Developed and Developing subgroups showed stark disparities. For instance, North America displayed a strong correlation for Developing countries (0.9129) but a negative correlation for Developed ones (-1.0000). Similarly, Oceania exhibited a strong correlation for Developing groups (0.9611) but a negative correlation for Developed groups (-0.7786). These results indicate the model struggles to generalize for smaller, underrepresented subgroups.

While Africa demonstrated balanced over- and underestimation rates (50%-50%), its wider residual spread contributed to a relatively low R^2 score (0.7779). In contrast, Oceania's high R^2 score corresponds to a narrower residual distribution, potentially due to smaller group size and less variability.

These findings emphasize the need for a finer-grained analysis at the country level to identify localized underperformance or bias. Future work should address dataset imbalances, introduce fairness constraints to equalize performance across regions, and re-examine dominant features such as HIV/AIDS and Alcohol to determine whether their prominence reflects data biases or actual predictive relevance. Such measures will improve the model's fairness, robustness, and generalizability across diverse subgroups.

6 Deployment

This chapter evaluates the model's performance against business objectives, discusses ethical aspects, potential risks, and deployment strategies, and identifies key monitoring aspects with intervention triggers. Reproducibility is revisited to highlight well-documented areas and remaining challenges.

6.1 Relating Results to Business Objectives

The Gradient Boosting Regressor achieved an R^2 of 0.8941 on the test set when trained using the training set, which can be considered a sufficient performance to meet the first business objective when using the model in production. Even though the result is just short of the 90% which was the target value for R^2, the small difference can be justified with the stakeholders. The provided insights into key predictors, such as healthcare expenditure, schooling, and immunization rates, align with the second business objective of aiding policymakers, healthcare providers and insurers in making informed decisions.

To further improve the usefulness of the analysis, the obtained insights could be mapped to policy recommendations to provide governments and other interested groups with a clear path to improving health factors. To achieve this, the results should be validated by experts and long-term studies should be performed that verify if long term effects on health outcomes resulting from recommended policies are within expectations. Implementing structured feedback mechanisms of provided analyses in collaboration with target users, as outlined in the business and data mining success criteria, was out of the scope of this project and will be carried out at a later date.

Since predictions after deployment might form a basis for WHO recommendations regarding health related factors, the impact of predictions and analyses accompanying them might have a considerable impact. Therefore, instead of opting for a fully automated deployment, a hybrid deployment is suggested where predicitons are complemented by expert analyses before decsions are formed on top of them. It is also recommended that deployment is performed and verified for regions with high predictive performance first, such as the continent of South America.

Regarding recommendations for future analyses, it is recommended to expand the space of recommended features. Firstly, the feature selection mechanism should be discussed with the data providers to address potential confirmation biases. Furthermore, we recommend adding features more closely associated with health care levels of a country, e.g. metrics on physician numbers and availability of hospital beds.

Lastly, confirmation bias is a concern, as features might be selected based on prior beliefs about what affects life expectancy. This can lead to the risk of overlooking other potentially relevant factors.

6.2 Model Performance and Recommendations

The Gradient Boosting Regressor achieved an R^2 score of 0.8941 on the test set, demonstrating sufficient performance to meet the primary business objective for production use. While the result falls slightly short of the 90% target, the small discrepancy can be justified to stakeholders. Key predictors, such as healthcare expenditure, schooling, and immunization rates, align well with the second business objective of supporting policymakers, healthcare providers, and insurers in making informed decisions.

To enhance the analysis, the derived insights could be translated into actionable policy recommendations. Validation by experts and longitudinal studies are necessary to confirm the long-term impact of suggested policies. Structured feedback mechanisms, as outlined in the success criteria, were beyond this project's scope and are planned for future iterations.

Given the significant impact of predictions on health-related decisions, such as WHO recommendations, a hybrid deployment is advised. Predictions should be complemented by expert analysis before guiding decisions. Deployment should initially target regions with high predictive performance, such as South America.

For future analyses, expanding the feature space is recommended. Collaboration with data providers can address potential confirmation bias in feature selection, ensuring relevant health metrics like physician availability and hospital bed density are included. Addressing confirmation bias is essential to avoid overlooking factors that may significantly influence life expectancy.

6.3 Ethical Aspects and Impact Assessment/Risks in Deployment

The deployment of the model raises key ethical concerns and potential risks that must be addressed to ensure fairness, transparency, and responsible use. One critical aspect is **bias and fairness**, as the model demonstrates variable performance across regions, particularly in Europe and Africa, potentially due to data imbalances. Ensuring **transparency** in model operations is essential to build stakeholder trust, while strict adherence to **privacy** standards is required to comply with data protection laws.

The deployment also introduces risks such as **unintended consequences**, where misuse could lead to discriminatory outcomes, and **security vulnerabilities**, such as data breaches compromising sensitive information. Additionally, socioeconomic impacts, like influencing healthcare decisions and diminishing local expertise, must be considered. Ensuring compliance with legal frameworks and minimizing environmental impacts associated with large-scale model use are further priorities.

Mitigation strategies include conducting regular **bias audits** to ensure equitable performance, promoting **transparency** through clear explanations of predictions, and reinforcing **data security** to protect user privacy. Further steps involve assessing **social impacts** in collaboration with stakeholders, ensuring **legal compliance**, and adopting sustainable practices to minimize resource use. These measures will help facilitate ethical, secure, and socially responsible deployment of the model.

6.4 Deployment Monitoring and Intervention Triggers

This section outlines the continuous monitoring and intervention mechanisms required to ensure the model's reliability and ethical integrity post-deployment. Key aspects include tracking performance metrics such as RMSE, MSE, MAE, and R^2, particularly for variable regions like `Africa` and `Europe`. A focus on retraining or data augmentation is necessary if performance degrades significantly. Residual analysis will identify systematic errors or bias amplification, while data drift monitoring addresses shifts in feature distributions or input quality. Fairness metrics will be continuously assessed to detect and address worsening disparities, and user feedback will serve as a critical indicator for identifying usability issues.

Intervention triggers include performance drops, such as a 10% decrease in R^2 or significant increases in RMSE, as well as the detection of biases exceeding fairness thresholds. Data anomalies, including spikes in missing or invalid inputs, will prompt immediate data quality checks. User-reported issues and potential ethical violations will trigger audits or temporary model suspension if necessary.

Mitigation strategies encompass automated alerts for performance monitoring, periodic fairness audits, and robust protocols for addressing user-reported concerns. Fallback models or manual oversight will be employed during critical adjustments to maintain operational stability. These measures ensure the model performs effectively, mitigates risks, and upholds fairness and ethical standards.

6.5 Reproducibility

The report has been produced mindful to reproducibility. The measures taken to this end include detailed descriptions of the training process, such as documenting the exact procedure and methods for hyperparameter optimization, the exact modeling configuration used to obtain results, using seeds to produce the same outputs every time the code is executed, and providing clear explanations of the feature engineering steps such as scaling techniques, outlier handling, missing values handling, redundancy cleaning and encoding processes applied to the dataset. Additionally, the inclusion of performance metrics for baseline models and segmented error analyses ensures transparency in model evaluation.

Some potential problems for reproducibility remain. Firstly, the exact hyperparameter values produced for optimization are not included in the report. Secondly, due to the extensiveness of the dataset, not all visualizations used for gaining insights are included. Furthermore, subgroup-specific performance results, such as segmented analysis across continents or grouped correlation metrics, require access to the full code and raw outputs for verification. Lastly, specific model evaluations, like baseline comparisons and equalized odds analysis, rely on extensive intermediate calculations not exhaustively documented here. These details are, however, made available via the Jupyter Notebooks provided alongside the report.

7 Summary

The findings of this project highlight the successful implementation and evaluation of a predictive model for life expectancy, achieving strong accuracy with an R^2 score of 0.8941 and RMSE of 2.708 on the test data. This reflects the Gradient Boosting Regressor's robustness in leveraging health and socioeconomic indicators. Key predictors such as HIV/AIDS prevalence, income composition of resources, and adult mortality rates were identified, providing actionable insights for policymakers and healthcare providers. However, regional disparities in performance, particularly in Europe and South America, stress the need for addressing potential biases and improving data balance.

Lessons learned revolve around the importance of following every step of the CRISP-DM framework while paying attention to several factors throughout the whole process, such as ensuring reproducibility and transparancy and the consideration of risks and biases. By doing so, it is not only ensured that useful results are produced, but also that they can be justified and explained to any interested group. Ethical considerations and fairness, particularly regarding regional representation, remain areas for improvement. Future enhancements should focus on engaging stakeholders for feedback, refining features, and improving model fairness and generalizability.

References

[1] Abdelrahman Abozarifa. 2023. Life Expectancy Prediction Using Advanced Machine Learning Models. https://github.com/abdelrahman-abozarifa04/Life-Expectancy-with-R2-Scorer-96.5- and https://www.kaggle.com/code/abdelrahmansami04/life-expectancy-with-r2-score-96-5. Accessed: 2024-12-11.

[2] Sercan Ö. Arik and Tomas Pfister. 2021. TabNet: Attentive Interpretable Tabular Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 35 (05 2021), 6679–6687. https://doi.org/10.1609/aaai.v35i8.16826

[3] P. Chandirasekeran, S. Saravanan, S. Kannan, and et al. 2022. Analyzing Implications of Various Social Factors on Life Expectancy. *National Academy Science Letters* 45 (2022), 311–316. https://doi.org/10.1007/s40009-022-01118-6

[4] B.A. Lipesa, E. Okango, B.O. Omolo, and et al. 2023. An application of a supervised machine learning model for predicting life expectancy. *SN Applied Sciences* 5 (2023), 189. https://doi.org/10.1007/s42452-023-05404-w

[5] World Health Organization. 2024. *WHO Methods and Data Sources for Life Tables 2000-2021.* Technical Report. World Health Organization, Department of Data and Analytics, Geneva, Geneva. https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-life-expectancy-and-healthy-life-expectancy Global Health Estimates Technical Paper WHO/DDI/DNA/GHE/2024.1.

[6] Scikit-learn. 2024. GradientBoostingRegressor - Scikit-learn Documentation. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html Accessed: 2024-12-12.

[7] Ravid Shwartz-Ziv and Amitai Armon. 2022. Tabular data: Deep learning is not all you need. *Information Fusion* 81 (05 2022), 84–90. https://doi.org/10.1016/j.inffus.2021.11.011

[8] Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C. Bayan Bruss, and Tom Goldstein. 2021. SAINT: Improved Neural Networks for Tabular Data via Row Attention and Contrastive Pre-Training. *arXiv:2106.01342 [cs, stat]* (06 2021). https://arxiv.org/abs/2106.01342

A Tables

	Real-world interpretation and context	**Value interpretation**	**Data type**	**Unit of measurement**
Country	Country of interest	String values of country names	Nominal	-
Year	Year of interest (several years for each country)	Integer values of year from 2000 to 2015	Time/Discrete	Years
Status	The country's development status	String values of Developed/Developing	Nominal	-
Life expectancy	Life expectancy of the country and year	Float values of life expectancy in years	Continuous	Years
Adult mortality	Adult mortality rate over both sexes	Deaths of ages 15 to 60 per 1000 inhabitants (integer)	Discrete	Count per 1000 inhabitants
Infant deaths	Infant death rate	Infant deaths per 1000 inhabitants (integer)	Discrete	Count per 1000 inhabitants
Alcohol	Consumption per capita	Float values of alcohol consumption	Continuous	Litres of pure alcohol per capita (15+) per year
Percentage expenditure	Country expenditure on health	Expenditure for health as ratio of GDP per capita (not a percentage)	Continuous	Health expenditure / GDP per capita
Hepatitis B	HepB immunization coverage	Integer values of Hepatitis immunization coverage	Discrete	Percentage
Measles	Number of measles cases	Integer count of cases	Discrete	Count of cases per 1000 inhabitants
BMI	Average BMI of population	Float values of BMI	Continuous	Average of entire population
Under-five deaths	No. of deaths under 5 years	Integer values of deaths under 5 years	Discrete	Count per 1000 inhabitants
Polio	Polio immunization coverage	Integer values of Polio immunization coverage	Discrete	Percentage
Total expenditure	Health expenditure / total government expenditure	Float values of ratio health / total expenditure	Continuous	Percentage
Diphteria	Immunization coverage of DTP3 (vaccine against diphtheria, tetanus, and pertussis)	Integer values of DPT3 immunization coverage	Discrete	Percentage
HIV/AIDS	Deaths per 1000 (0-4 years)	Float values of deaths per 1000	Continuous	Deaths from 0 to 4 years per 1000 live births
GDP	Gross domestic product per capita	Float values of GDP per capita	Continuous	USD per capita
Population	Country population	Integer values for country population	Discrete	Total number
Thinness 1-19 years	Portion of "very thin" (1-19 years)	Float values for percentages of "very thin"	Continuous	Percentage
Thinness 5-9 years	Portion of "very thin" (5-9 years)	Float values for percentages of "very thin"	Continuous	Percentage
Income composition of resources	Human development index based on income and resource availability	Float values from 0 to 1	Continuous	Index (interpretable as percentage)
Schooling	Average number of schooling years	Float values of average schooling years	Continuous	Years

Table 4: Table for semantics and data types of all attributes. "Very thin" refers to weights two standard deviations below the median.