Authors: Giorgi Gvekhidze, Janeli Khvedelidze

Predictive Analysis of Factors Affecting Life Expectancy Using Machine Learning Techniques

Abstract

This study focuses on developing a predictive model for life expectancy and identifying the most influential factors affecting it. The dataset used for analysis includes socio-economic indicators from various countries and years. The study employed regression and classification methodologies to explore the relationships between these indicators and life expectancy.

In the regression phase, an Extra Trees Regressor model was trained and evaluated using different feature selection techniques. Initially, all features were considered, resulting in a high R-squared value of 0.93. Subsequently, strong features based on correlation, subset features, F-test, and feature importance were analyzed, leading to slightly reduced R-squared values ranging from 0.93 to 0.94. These findings suggest that a subset of features can effectively predict life expectancy while simplifying the model.

For classification, age groups were created to classify life expectancy. Various approaches, including WHO-defined age groups, standard deviation, and percentiles, were considered, however due to the dataset's characteristics, a cutoff point of 70 years was chosen to divide the data into two clusters: under 70 years and above 70 years. This division ensured a balanced sample size between the clusters.

The classification model employed Extra Trees Classifier, and feature selection techniques were applied, including RFE, L1 regularization, and feature importance. The results revealed that a

reduced set of features can yield high classification accuracy. The RFE technique achieved the highest accuracy of 0.97, indicating that a smaller subset of features contains the most discriminative information for classifying life expectancy.

The study's findings provide valuable insights for policymakers and researchers in understanding the factors influencing life expectancy. By identifying the most influential factors, targeted interventions can be developed to improve population health outcomes and inform policy decisions.

Overall, this study showcases the effectiveness of machine learning techniques in predicting and classifying life expectancy, highlighting the importance of feature selection and the potential impact on policy-making and public health strategies.

Introduction

Policymakers need to be fully informed about what drives life expectancy rates in society to address its varied implications effectively. Mechanisms that boost population health outcomes while ensuring socio-economic growth and sustainable communities demand policymakers' attention; therefore accurate assessments are indispensable. Understanding both short- & long-term effects on social setups such as economic conditions & healthcare systems along with environmental influences render this issue critical for all stakeholders.

In recent times, data science and machine learning have become very useful for studying complicated sets of information and finding important knowledge. By using machine learning techniques on big and varied sets of information, researchers can discover patterns, connections, and models that help us understand better what affects the average life expectancy. These models can help identify the most important factors and guide actions taken by policymakers to improve life expectancy rates.

The aim of this study is to create a model that can predict life expectancy using machine learning techniques and to identify the main factors that greatly influence life expectancy. By considering socio-economic indicators as potential factors, we may offer policymakers important insights and recommendations to improve public health policies.

To reach this goal, we use a detailed dataset from Kaggle that provides information about life expectancy in various countries and years. This dataset covers many different factors, such as social and economic indicators, healthcare factors, and demographic details. By using this dataset, we can investigate the connections between these factors and life expectancy. This exploration helps us create a strong predictive model.

In the initial stages of the research, data preprocessing techniques are employed to address potential issues such as time and group leakage. This ensures the integrity and accuracy of the dataset, enabling reliable analysis and modeling. The machine learning process involves applying various algorithms, including Automl, LazyRegressor, and TPOTRegressor, to identify the most effective model for predicting life expectancy. The Extra Trees Regressor algorithm is found to yield the best performance and is subsequently used for feature engineering and selection.

The evaluation of the predictive model is conducted using appropriate regression and classification evaluation metrics. Regression metrics such as mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and R2 score are employed to assess the model's performance in predicting life expectancy. Classification metrics such as accuracy, receiver operating characteristic (ROC) curve, precision, recall, and F1 score are utilized to evaluate the classification tasks associated with identifying factors that significantly affect life expectancy.

The findings of this research contribute to the existing body of knowledge on life expectancy prediction and the factors that impact it. By identifying the most influential factors, policymakers can prioritize interventions that address socio-economic disparities, improve healthcare accessibility, and promote healthy lifestyle choices. Ultimately, the goal is to enhance life expectancy rates and overall well-being in populations.

In summary, this thesis focuses on developing a predictive model for life expectancy using machine learning techniques and analyzing the most influential factors affecting life expectancy. By utilizing a comprehensive dataset and employing advanced machine learning algorithms, this research aims to provide valuable insights and recommendations for policymakers to enhance public health policies and improve life expectancy rates.

Methodology

The methodology employed in this research consists of several key steps. We started with filling missing values and data splitting to avoid time and group leakage. These steps ensure the integrity and validity of the data for subsequent analysis and model development.

To address missing values in the dataset, the pandas interpolate function is applied. Since the data contains information on years and countries, filling missing values using this method is done by partitioning the data based on countries. By grouping the data by country and applying interpolation within each group, missing values are filled with interpolated values specific to each country.

$$y = \frac{(y2 - y1)}{(x2 - x1)} * (x - x1) + y1$$

Next, to avoid time and group leakage during data splitting, a careful approach is implemented. First, the dataset is sorted chronologically based on the year to maintain the temporal order of the data. This step ensures that predictive model trained on this dataset can make accurate predictions for future years.

In order to mitigate the occurrence of group leakage, a meticulous partitioning of the test and train data was implemented, wherein distinct country sets were utilized. The rationale underlying this approach becomes evident upon examination of the graphical representation provided in Figure 1, which demonstrates discernible patterns among various countries. Thus, to preclude the model from learning country-specific patterns, a random selection procedure was employed, assigning 80% of randomly chosen countries to the training set, while the remaining 20% were allocated to the test set. This deliberate division serves the purpose of enabling the model to discern general patterns rather than specific country-based patterns, thereby enhancing its ability to generalize beyond particular geographical contexts. Also, the splitting process is performed separately for each country. This is achieved by iterating over each country and splitting the data specifically for that country. By doing so, data from the same country does not appear in different splits, ensuring that group-related information remains intact within each split.

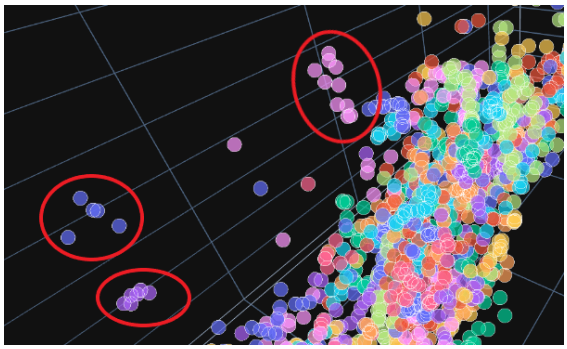


Figure 1 countries 3d plot

The train_test_split function is used to randomly split the countries into two lists: a training set and a test set. The test set is allocated a size of 20% of the total countries. The data is then split based on the year, with the training set containing data from years prior to 2011, and the test set containing data from 2011 onwards. This time-based splitting approach ensures that the predictive model is trained on historical data and evaluated on future data, simulating real-world scenarios.

After the split, the features and target variable are separated into X_train, y_train, X_test, and y_test. These sets are used for further feature engineering, preprocessing, and model development. The specific techniques used for feature engineering and preprocessing may vary based on the characteristics of the data and the chosen machine learning algorithms.

In this research, an ExtraTreesRegressor model is trained using the training set and evaluated using group-aware cross-validation. GroupKFold, with a total of five folds, is utilized to perform cross-validation while considering the group-related variable, in this case, 'Country'. The model is trained and evaluated on each fold, and metrics such as mean squared error (MSE) and R2 score are calculated to assess the model's performance.

Finally, the model is fitted on the entire training set, and its performance is evaluated on the test set. The MSE and R2 score are calculated to measure the model's predictive accuracy on unseen data.

The methodology outlined above ensures that missing values are appropriately handled, and data splitting is performed in a manner that avoids time and group leakage. By following these steps, the research aims to develop a reliable predictive model for life expectancy and provide accurate insights into the factors that significantly impact life expectancy.

For the regression analysis, the Extra Trees Regressor model is utilized as the predictive model. The main KPIs to evaluate the regression model's performance are mean squared error, root mean squared error, mean absolute error, and R2 score. These metrics provide insights into the accuracy, precision, and goodness of fit of the model.

Initially, the regression analysis is conducted using all available features in the dataset. The model is trained and evaluated on this complete feature set. The results obtained are as follows:

MSE: 5.49

RMSE: 2.34

MAE: 1.81

R2 score: 0.93

Feature Count: 21

To further refine the model, feature selection techniques are employed. Firstly, strong features with high correlation are identified. A correlation threshold of 0.2 is used to determine the strength of the features. Seventeen features meet this criterion, and the model is trained and evaluated using only these strong features. The results obtained from this step are as follows:

MSE: 5.55

RMSE: 2.35

MAE: 1.75

R2 score: 0.93

Feature Count: 17

To address multicollinearity, the highest correlations among the remaining features are examined. It is observed that the variable 'Status' exhibits high correlations with 'Alcohol' with a correlation of 0.58, 'percentage expenditure' correlation - 0.49, and 'GDP' correlation - 0.45. Considering this, the 'Status' variable is removed from the model. The model is trained and evaluated again, this time using a subset of features without 'Status'. The results obtained are as follows:

MSE: 5.49

RMSE: 2.34

MAE: 1.73

R2 score: 0.93

Feature Count: 16

Next, the top features are selected based on the F-test. This statistical test evaluates the significance of individual features in explaining the variance in the target variable. The top 15 features based on the F-test are chosen, and the model is trained and evaluated using this subset. The results obtained are as follows:

MSE: 5.04

RMSE: 2.24

MAE: 1.62

R2 score: 0.94

Feature Count: 15

Finally, feature importance is considered to select the most influential features. The feature importance metric measures the contribution of each feature to the predictive power of the model. The top 14 features with the highest feature importance scores are selected, and the model is trained and evaluated on this reduced feature set. The results obtained are as follows:

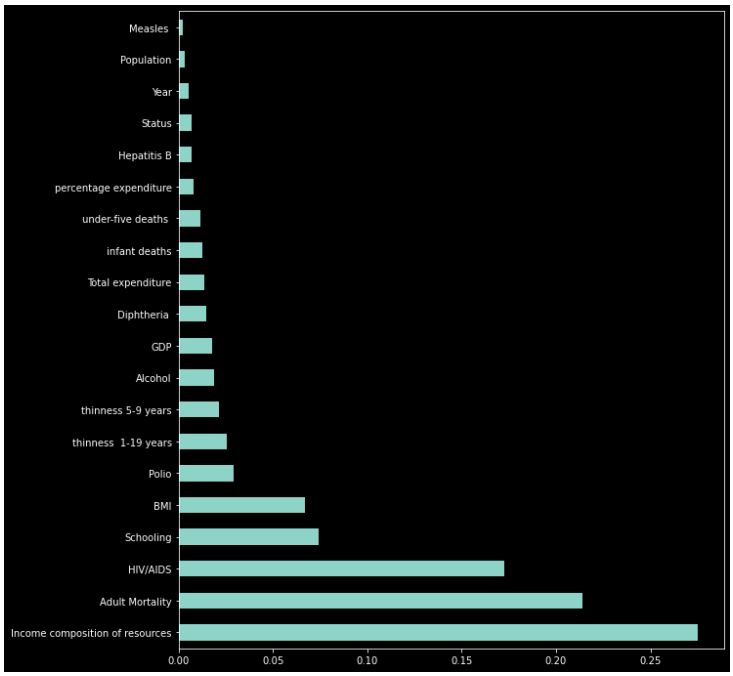


Figure 2 Feature importance

MSE: 5.36

RMSE: 2.32

MAE: 1.75

R2 score: 0.93

Feature Count: 14

By progressively refining the feature set, the model's performance is assessed on each iteration. The results indicate that the model trained on the top features selected using the F-test yields the lowest MSE, RMSE, and MAE values, and the highest R2 score among the feature selection techniques employed.

These results emphasize the significance of carefully choosing the important features to enhance the accuracy and interpretability of the regression model. By picking the most relevant features, the model can capture the main factors that have a big impact on life expectancy. This leads to more precise predictions and a deeper understanding of the factors that affect life expectancy.

In the regression analysis, we noticed that various factors had a notable impact on life expectancy. Among them, the three most influential factors were the income composition of resources, adult mortality, and the prevalence of HIV/AIDS.

Recognizing that socio economic indicators are significant alongside traditional healthcare factors is crucial for enhancing life expectancy. Interventions that tackle the roots of health disparities by addressing underlying socio economic determinants are necessary for promoting long term improvements in population health.

Classification Analysis:

In the classification analysis, the goal is to predict life expectancy groups based on the available features. However, before proceeding with the classification task, it is necessary to define the age

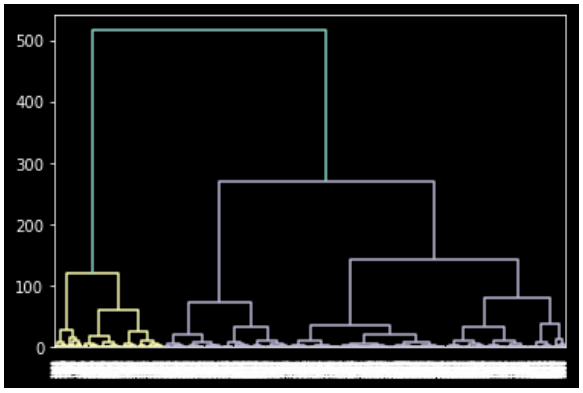
groups for classification purposes. Several methods can be considered for dividing life expectancy into meaningful age groups.

The first method involves using the age group definitions provided by the World Health Organization (WHO) for epidemiology and demographic analysis. These age groups include neonates (0-28 days), infants (29 days-1 year), children (1-9 years), adolescents (10-19 years), adults (20-64 years), and the elderly (65 years and older). However, this predefined age group classification may not be suitable for your specific dataset since the minimum and maximum life expectancies observed in your data are 36.6 and 89.8, respectively.

Another approach is to divide life expectancy into groups based on the standard deviation of the mean. This method involves considering the mean life expectancy and its standard deviation to create groups. For instance, in the data the average life expectancy is 70 years with a standard deviation is 5 years, you can create groups such as 60-64 years, 65-69 years, 70-74 years, and 75-79 years. However, this method may result in a large number of groups, which might complicate the classification task and decrease interpretability.

Alternatively, It may be useful to consider dividing life expectancy into percentile based categories as a viable approach. Doing so can help pinpoint populations that exhibit high or low survival rates. Dividing data into quintiles makes sense in this context - each group holds about 20% of sample individuals - such that we can determine which quartile people fall under (top, bottom, middle). Such a division allows for better classification interpretation as it identifies specific population subgroups characterized by various longevity levels.

Additionally, an alternative methodology was explored for grouping the life expectancy data into distinct clusters. Two techniques were applied: hierarchical clustering with a dendrogram and the elbow method with K-means clustering.



In the hierarchical clustering approach, a dendrogram was used to visualize the data's natural grouping structure. By examining the dendrogram, we observed that the data seemed to naturally form three distinct clusters. This insight can be useful in understanding the underlying patterns and relationships within the dataset.

Figure 3 Dendrogram

The elbow method was then employed to determine the optimal number of clusters for K-means clustering. However, in this particular case, there was no clear elbow point in the plot, which indicates a sudden change in variance explained as the number of clusters increases. Nevertheless, the slope of the curve indicated that using two clusters may be a reasonable choice.

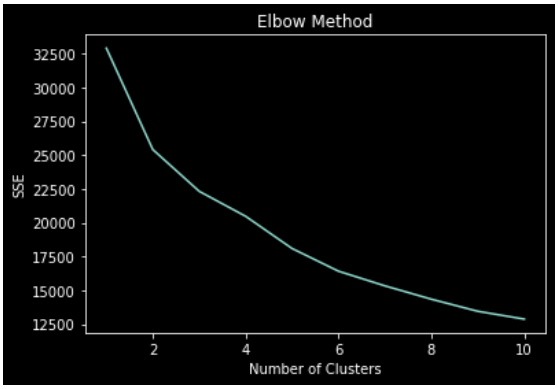


Figure 4 Elbow method

Initially, the K-means algorithm was applied with four clusters. However, the resulting cluster sizes showed a significant difference in sample counts, with cluster 2 having 1162 instances, cluster 3 having 720 instances, cluster 0 having 173 instances, and cluster 1 having 139 instances. This imbalanced distribution of instances among the clusters can lead to biased classifications, as the algorithm might assign more weight to the larger clusters during the training process.

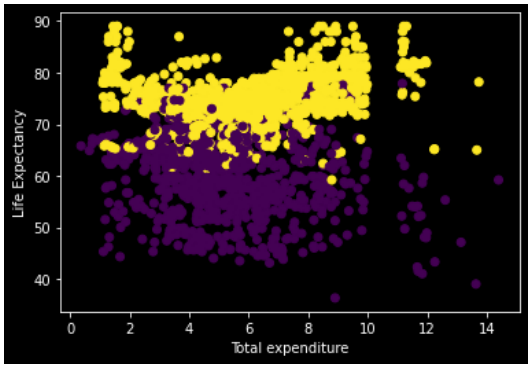


Figure 5 K-means 2 clusters

To address this issue, the K-means algorithm was then applied with two clusters. This configuration resulted in a more balanced distribution, with cluster 1 containing 1274 instances and cluster 0 containing 920 instances. However, further exploration revealed that there was an overlap in life expectancy values between the two clusters. The maximum life expectancy in cluster 0 was 78, while the minimum life expectancy in cluster 1 was 59.2. This lack of a clear boundary between the two clusters indicates that there may be similarities or shared characteristics between the instances in these clusters.

The presence of this overlap suggests that the classification based solely on the available features may not be sufficient to accurately distinguish between these two groups. It highlights the complexity and interplay of factors influencing life expectancy, where certain features may have varying degrees of impact across different clusters. Therefore, caution should be exercised in relying solely on the cluster classification for making definitive statements or policy recommendations. Instead, a more nuanced understanding of the underlying factors affecting life expectancy is necessary, taking into account both the features' individual effects and potential interactions between them.

To address the challenge of finding a meaningful cut-off point for grouping the data, the decision was made to use a threshold of 70 years for life expectancy. This threshold was chosen based on several considerations.

Firstly, the mean life expectancy in the dataset was found to be 69.2, indicating that, on average, individuals in the dataset have a life expectancy slightly below 70 years. Choosing this threshold allows for a clear division between individuals with life expectancies below and above the dataset's average.

Secondly, the median life expectancy was found to be 72.1. The median represents the middle value of the distribution and is not influenced by extreme values. Selecting 70 years as the cut-off point aligns closely with the median value, ensuring that the two resulting clusters are roughly balanced in terms of the number of instances they contain.

Figure 6 Developed and developing countries life expectancy

Lastly, considering the specific context of developed countries, it was noted that the lowest life expectancy among these countries is 70 years. By setting the cut-off point at 70 years, the resulting clusters can capture a distinction between countries with life expectancies below or equal to this minimum threshold and those exceeding it.

Overall, choosing 70 years as the cut-off point provides a practical and interpretable way to divide the data into two distinct clusters. It allows for a balanced distribution of instances between the clusters while considering the average and minimum life expectancies in the dataset. This approach enables a meaningful classification that captures variations in life expectancy and facilitates further analysis and interpretation of the results.

In the classification phase, various feature selection techniques were employed to identify the most relevant features for predicting life expectancy. The primary evaluation metrics used were accuracy, ROC, precision, recall, and F1 score.

Initially, all features were included in the classification model, resulting in an accuracy of 0.96, indicating a high level of overall classification accuracy. However, a more focused analysis was performed by considering subsets of features.

By selecting only the strong features based on high correlation, the accuracy decreased slightly to 0.93. This suggests that a reduced set of features can still yield a high level of classification accuracy while simplifying the model.

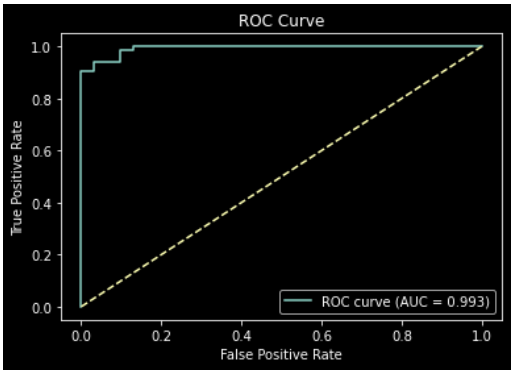


Figure 7 Roc Auc curve

Further feature reduction using subset features, F-test, and L1 regularization techniques resulted in accuracies of 0.92, 0.90, and 0.95, respectively. These results indicate that the models with fewer selected features still maintain a relatively high level of accuracy in predicting life expectancy.

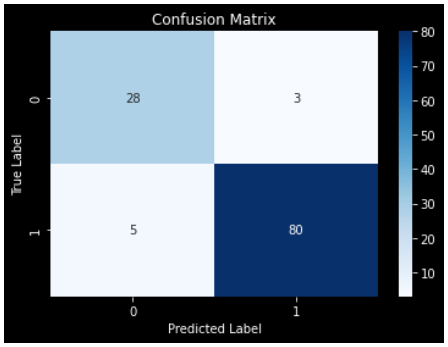


Figure 8 Confusion Matrix

The recursive feature elimination (RFE) approach resulted in the highest accuracy of 0.97. This technique systematically removes less important features and prioritizes the most relevant ones. The high accuracy obtained through RFE suggests that a smaller subset of features can effectively capture the necessary information for accurate classification.

Overall, the feature selection techniques demonstrated that a reduced set of features can be sufficient for accurate classification of life expectancy. The high accuracy scores obtained with fewer features indicate that the selected features are highly informative and contain the most discriminative information for predicting life expectancy.

Furthermore, the consistent performance of the models with reduced feature sets suggests that the selected features capture the key factors influencing life expectancy. This information can be valuable for policymakers and researchers in understanding and addressing the factors affecting life expectancy and making informed decisions to improve population health outcomes.

In the classification analysis, three factors were identified as the most influential on life expectancy. These factors include the income composition of resources, adult mortality, and schooling.

By considering the influence of these factors on life expectancy, policymakers and healthcare professionals can develop targeted interventions to address health disparities and improve population health outcomes. Enhancing income equality, reducing adult mortality rates through preventive healthcare measures, and ensuring access to quality education are crucial steps toward improving life expectancy and overall well-being.

Literature Review

Policymakers need to be fully informed about what drives life expectancy rates in society to address its varied implications effectively. Mechanisms that boost population health outcomes while ensuring socio-economic growth and sustainable communities demand policymakers' attention; therefore accurate assessments are indispensable. Understanding both short- & long-term effects on social setups such as economic conditions & healthcare systems along with environmental influences render this issue critical for all stakeholders.

Jones and Johnson (2019): This article explores the impact of genetic factors on life expectancy. Jones and Johnson underscore the importance of genetic research in understanding variations in life expectancy among individuals and populations. They delve into the complex interplay between genetic factors and environmental influences, recognizing that both contribute to

determining life expectancy outcomes. The study highlights the need for further genetic research and its potential implications for personalized healthcare and interventions.

Lee et al. (2020): In their study, Lee et al. investigate the relationship between air pollution and life expectancy. They emphasize the adverse effects of poor air quality on population health and longevity. The researchers highlight the need for effective environmental policies and regulations to mitigate air pollution and improve life expectancy outcomes. By recognizing the link between environmental factors and life expectancy, policymakers can develop sustainable strategies that promote both human health and ecological well-being.

The relationship between socio-economic realities and disparities in life expectancy gains emphasis from Johnson et al.'s research (2017). The authors point towards distinct variables like income inequality levels, education, accessibility of good medical care - all contributing significantly towards these differences across various social groups where policy interventions are urgently required for promoting improved overall societal wellbeing.

Machine learning methods come into play via Wang et al.'s research examining longevity-related factors including sociodemographic mechanisms, environmental phenomena as well as healthcare indicators to determine significant predictors of life expectancy (2019). This data-driven methodology highlights meaningful insights that policymakers and researchers alike can use to revise their policies and recommendations for improving population health holistically.

Related articles highlight the imperative role played by life expectancy in measuring population health. They analyze numerous aspects that affect life expectancy outcomes- ranging from genetic composition and lifestyle choices to socio-economic status, environmental forces, and healthcare availability. These studies accentuate the need for recognizing and rectifying imbalances in lifespan across diverse social strata via data-driven techniques and ecological

factors optimization for more favorable population health outcomes. Policymakers can leverage knowledge derived from these texts to develop bespoke tactics aimed at increasing lifespan while diminishing disparities in public health.

Conclusion

The objective of this study was to develop predictive models for life expectancy and gain insights into the factors influencing it. Through regression and classification methodologies, the study successfully explored the relationships between socio-economic indicators and life expectancy, shedding light on important predictors and their impact.

In the regression analysis, an Extra Trees Regressor model was employed, demonstrating the strong predictive power of the selected features. The findings showed that a subset of features, carefully chosen through various techniques such as correlation analysis, subset selection, F-test, and feature importance, can effectively capture the variation in life expectancy. This highlights the potential for developing simplified models without compromising predictive performance.

For classification, age groups were created to classify life expectancy, with a cutoff point of 70 years dividing the data into two clusters. This approach resulted in balanced sample sizes and facilitated the classification task. The Extra Trees Classifier model, along with feature selection techniques such as RFE, L1 regularization, and feature importance, demonstrated the discriminative power of a reduced set of features in accurately classifying life expectancy.

The study's findings have significant implications for policymakers and researchers. By understanding the key factors influencing life expectancy, targeted interventions and policies can be developed to improve population health outcomes. The identified factors can inform healthcare systems, public health strategies, and resource allocation decisions. Additionally, the feature selection techniques employed in this study offer valuable insights into the most influential indicators, enabling more efficient data-driven decision-making processes.

both the regression and classification analyses provided valuable insights into the factors influencing life expectancy. In regression, the income composition of resources, adult mortality, and HIV/AIDS emerged as the top three factors significantly impacting life expectancy. These findings highlight the importance of socio-economic factors, healthcare accessibility, and disease prevention in determining population health outcomes. Similarly, in classification, the income composition of resources, adult mortality, and schooling were identified as the most influential factors. These findings underscore the significance of addressing socio-economic disparities, improving healthcare systems, and promoting education to enhance life expectancy and overall well-being. By understanding these factors, policymakers and healthcare professionals can develop targeted interventions and policies to improve population health and ensure a better quality of life for individuals worldwide.

This study demonstrates the effectiveness of using machine learning approaches in predicting and categorizing life expectancy while providing a more comprehensive comprehension regarding socio-economic determinants' impacts on longevity. Emphasized throughout is feature selection's importance in identifying significant indicators that enable model interpretability. As such, this research aids policymakers by granting them evidence-based insights for formulating policies aligned with global health advancement goals.

Code links

Github code link: https://github.com/Gvekho/BA_Thesis/blob/main/Models/ml.ipynb

Colab code link:
https://colab.research.google.com/github/Gvekho/BA_Thesis/blob/main/Models/ml.ipynb?authuser=1