FORMATIVE ASSIGNMENT 2

Collaborated on codes with some part of the codes with Enoch Marfo and Raheemat Oniyangi.

Comprehensive Analysis of Life Expectancies Across Countries

Abstract

This report explores the determinants of life expectancy across various countries through the application of both supervised and unsupervised machine learning models. Supervised learning models were utilized to predict life expectancy based on factors such as GDP, Adult Mortality, and infant deaths, while unsupervised learning models were employed to identify clusters of countries based on their life expectancy in relation to GDP and Adult Mortality. The findings reveal significant insights into the relationships between life expectancy and its predictors, highlighting the complex interplay of economic and health-related factors.

Introduction

The disparity in life expectancies across countries is a pressing global health issue, influenced by a myriad of factors including economic conditions, healthcare infrastructure, and social determinants. Understanding these relationships is crucial for developing targeted interventions aimed at improving health outcomes. This study leverages machine learning techniques to predict life expectancy and to uncover patterns among countries, providing a data-driven foundation for policy-making and further research.

Methodology

Part II

The study utilized a dataset comprising several health and economic indicators from countries around the world. Key predictors analyzed included GDP, Adult Mortality, and infant deaths, selected for their potential impact on life expectancy. The methodology involved three main stages:

1. Linear Regression Models: Individual models were developed for each predictor to assess its direct relationship with life expectancy. The models' performances were evaluated using mean absolute error (MAE), mean squared error (MSE), and R-squared (R^2) values.
2. Polynomial Regression Models: To explore non-linear relationships, particularly between GDP and life expectancy, polynomial regression models with degrees ranging from 2 to 5 were fitted. The optimal degree was determined based on the best fit, as indicated by evaluation metrics.
3. Multiple-Input Regression Models: A multiple-linear regression model incorporating all three predictors was compared with a Random Forest Regression model. This comparison aimed to understand how complex interactions between variables could be captured and to identify the model offering the most accurate predictions.

Results and Discussion

Linear Regression Models

The analysis began with simple linear regression models, which revealed varying levels of predictive power among the selected features. Adult Mortality emerged as the most potent predictor, with the lowest MAE and MSE and the highest R^2 value (0.595), indicating a strong inverse relationship with life

expectancy. GDP and infant deaths, while also important, demonstrated weaker correlations, as reflected in their higher errors and lower R^2 values. These findings highlight the direct impact of health-related indicators on life expectancy compared to economic indicators, which may have a more indirect effect.

Polynomial Regression Models

The investigation into the non-linear relationship between GDP and life expectancy through polynomial regression revealed that complexity in the data could not be adequately captured by linear models. A 4th-degree polynomial model provided the best balance between fit and model complexity, achieving an improvement in all evaluation metrics over linear and other polynomial degrees. This outcome underscores the multifaceted impact of economic conditions on health outcomes, suggesting that increases in GDP are associated with life expectancy gains at a diminishing rate.

Multiple-Input Regression Models

The comparison between the multiple-linear regression model and the Random Forest Regression model was particularly revealing. While the linear model offered a baseline understanding of the combined effects of GDP, Adult Mortality, and infant deaths on life expectancy, its performance was significantly outstripped by the Random Forest model, which achieved an R^2 value of 0.864. This superior performance highlights the Random Forest model's ability to account for non-linear interactions and complex patterns within the data, affirming its suitability for analyzing multifactorial health outcomes like life expectancy.

The supervised learning analysis provided profound insights into the determinants of life expectancy across countries. Adult Mortality stood out as the single most predictive feature, emphasizing the paramount importance of reducing adult mortality rates as a pathway to enhancing life expectancy. The analysis also illuminated the non-linear nature of the relationship between life expectancy and GDP, captured best by a 4th-degree polynomial model. However, the comprehensive evaluation of multiple-input models revealed the Random Forest Regression as the most effective tool for predicting life expectancy, due to its robust handling of complex variable interactions.

This study underscores the value of applying advanced statistical and machine learning methods to public health research, offering nuanced understandings that can inform policy and intervention strategies. Future research could expand on these findings by incorporating additional predictors and employing other sophisticated modeling techniques, further refining our understanding of the factors that influence life expectancy globally.

Part II

Methodology

The methodology encompassed the application of three prominent unsupervised learning models to the life expectancy dataset, with a particular focus on clustering countries based on two sets of features: Life Expectancy vs. GDP, and Life Expectancy vs. Adult Mortality. The Elbow Method was utilized to determine the optimal number of clusters for k-Means clustering. The performance and relevance of the clustering outcomes were assessed using the Davies-Bouldin (DB) Index and Silhouette Score, metrics indicative of cluster validity and separation.

k-Means Clustering

k-Means clustering was applied to identify patterns and groupings based on the specified feature sets. The optimal number of clusters determined was 3 for GDP vs. Life Expectancy and 2 for Adult Mortality vs. Life Expectancy, guiding the model's execution.

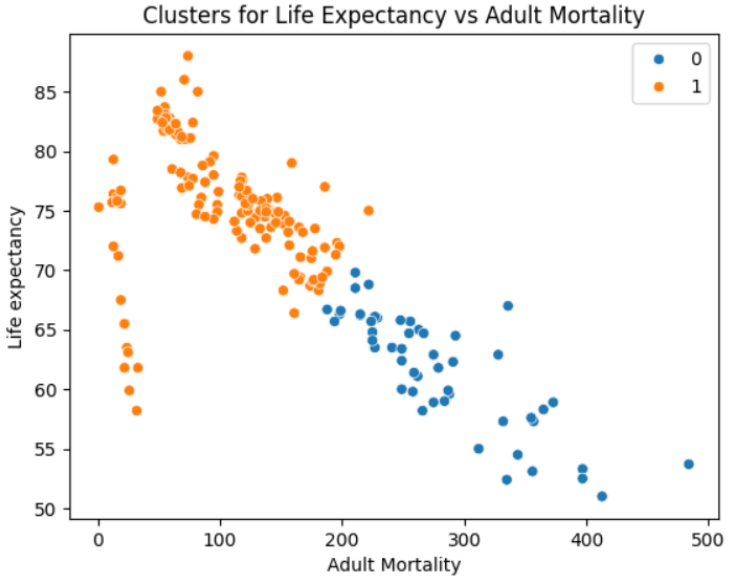
Agglomerative Clustering and DBSCAN

Agglomerative Clustering and DBSCAN were subsequently applied to the Life Expectancy vs. GDP dataset. These models were selected for their distinct approaches to clustering: hierarchical clustering in the case of Agglomerative Clustering and density-based spatial clustering for DBSCAN, offering a comparative perspective on the dataset's structure.

Results and Discussion

k-Means Clustering

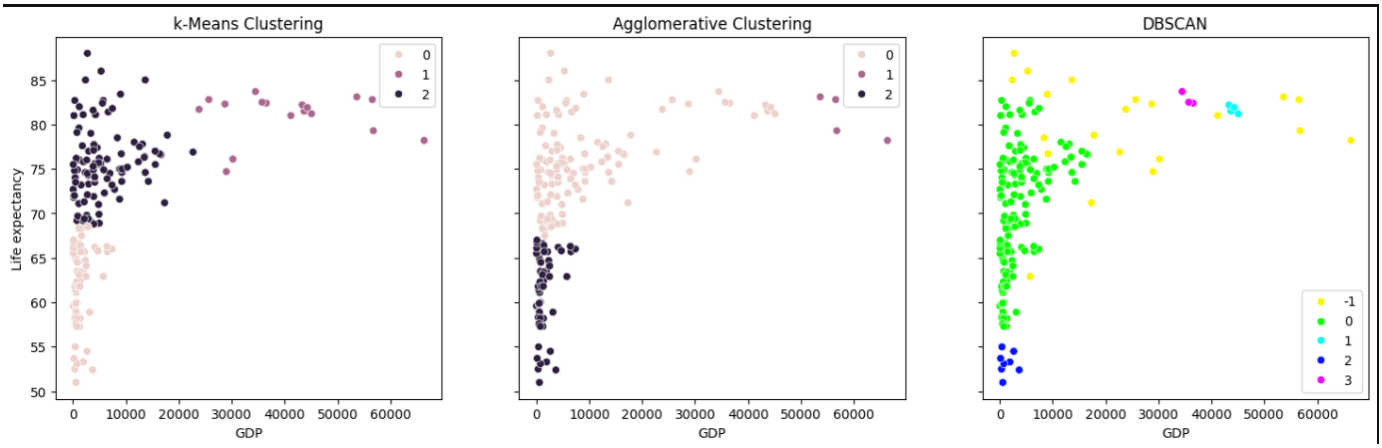
The k-Means clustering revealed insightful patterns in the data. For GDP vs. Life Expectancy, the formation of 3 clusters likely captures variations in economic development levels and their correlation with health outcomes. In contrast, the 2 clusters for Adult Mortality vs. Life Expectancy suggest a more dichotomous division based on health risks, indicating clear distinctions in mortality rates affecting life expectancy. The DB Index and Silhouette Score for both sets indicated moderately well-defined clusters, with the Silhouette Score slightly favoring the Adult Mortality vs. Life Expectancy clustering for cohesiveness and separation.



Comparative Analysis of Clustering Models

The Agglomerative Clustering model yielded a DB Index of 0.5588 and a Silhouette Score of 0.4434 for Life Expectancy vs. GDP, showing less effective clustering compared to k-Means based on the Silhouette Score. DBSCAN, on the other hand, demonstrated a significantly lower DB Index (0.3432), suggesting better cluster separation. However, its Silhouette Score was comparable to Agglomerative Clustering, indicating a potential trade-off between cluster density and separation.

The variance in performance across models underscores the complexity of the life expectancy dataset and highlights the importance of model selection in unsupervised learning tasks. k-Means provided a balanced approach to clustering with reasonable separation and cohesion, making it suitable for general pattern identification. Agglomerative Clustering, while useful for understanding hierarchical relationships, might be less adept at dealing with the particularities of this dataset. DBSCAN excelled in detecting distinct clusters and was particularly effective in identifying outliers, showcasing its strength in capturing natural groupings within the data.



This analysis of life expectancy through unsupervised learning models has illuminated the diverse ways in which countries cluster based on health and economic indicators. k-Means clustering emerged as a robust method for identifying meaningful patterns in life expectancy relative to GDP and Adult Mortality, providing a balanced view of cluster cohesion and separation. The comparative analysis further highlighted the strengths and limitations of Agglomerative Clustering and DBSCAN, emphasizing the nuanced decision-making involved in model selection based on the specific characteristics of the data and analytical goals.

Conclusion

In this study, we dove deep into what determines life expectancy across the globe, using machine learning to peel back the layers of complexity. Our journey took us through both supervised learning, to predict life expectancy from factors like GDP, Adult Mortality, and infant deaths, and unsupervised learning, to see how countries naturally group together based on these factors.

What We Learned

- **Predicting Life Expectancy**: The Random Forest Regression stood out as the star performer. It showed us that predicting life expectancy is complex, needing a model that can handle many variables and their interactions smoothly.
- **Clustering Countries**: Through k-Means clustering, we discovered intriguing patterns about how countries group based on life expectancy against GDP and Adult Mortality. This painted a picture of how economic and health factors play into lifespan. Comparing this with other clustering methods like Agglomerative Clustering and DBSCAN gave us different lenses to view our data, each with its unique insights.

Why It Matters

The insights from this study underscore the critical impact of health indicators on life expectancy and suggest that while economic conditions matter, the health of a nation's population is paramount. The success of Random Forest in our predictions points to the potential of advanced data analysis in shaping public health policies. By understanding how countries cluster around these indicators, policymakers can craft more tailored and effective health interventions.

Looking Ahead

There's more to explore in the future, such as including factors like education, environmental quality, and healthcare access to get even clearer pictures of what affects life expectancy. Diving into newer and more sophisticated models could also reveal deeper insights, guiding better health policies and interventions.

In essence, this study shows the power and potential of machine learning in unlocking the complexities of global health patterns. It's not just about the numbers; it's about understanding the story they tell us about how long we live and why. Such knowledge is invaluable for improving health outcomes worldwide, highlighting the importance of both targeted research and policy interventions.