Analyzing the Impact of Health, Economic, and Demographic Factors on Life Expectancy: A Comparative Study of Developed and Developing Countries

Mahyar Alinejad
dept. Electrical Engineering
Uniuversity of central florida
Orlando, United States
mahyar.alinejad@ucf.edu

***Abstract*— This study presents a comprehensive analysis of three prominent machine learning regression models—Random Forest, XGBoost, and Support Vector Machine (SVM)—in the context of predictive analysis. Leveraging a carefully curated dataset, we explore the impact of various hyperparameters on model performance through an exhaustive tuning process. The Random Forest and XGBoost models exhibit robust predictive capabilities, with the former revealing notable insights through feature importance visualization. Additionally, SVM, optimized via GridSearchCV, demonstrates competitive performance. Evaluation metrics, including Mean Squared Error and R-squared, facilitate a thorough comparison of model efficacy. Results highlight nuanced strengths and weaknesses, informing practitioners on the suitability of each model for specific applications. This research contributes valuable insights to the ongoing discourse on machine learning regression, offering a practical guide for researchers and practitioners navigating the complex landscape of predictive analysis.**

Keywords—Machine Learning, Regression, Life Expectancy, Predictive Analysis

I. INTRODUCTION

In recent years, the ubiquity of data has catalyzed a paradigm shift in decision-making processes across diverse domains, ranging from finance to healthcare. This shift is underscored by the increasing reliance on machine learning (ML) techniques, particularly regression models, for predictive analysis. As businesses and researchers seek to extract meaningful insights from vast datasets, the selection of an appropriate regression model becomes pivotal. This study delves into a comparative analysis of three widely employed machine learning regression models—Random Forest, XGBoost, and Support Vector Machine (SVM)—with the aim of elucidating their respective strengths and weaknesses in predictive analytics.

The choice of regression models is a critical decision in the machine learning workflow, and researchers are often confronted with a myriad of options. The Random Forest algorithm, an ensemble learning method, has gained prominence due to its ability to mitigate overfitting and handle complex relationships within data. XGBoost, an optimized gradient boosting technique, has demonstrated exceptional performance in various machine learning competitions, making it a popular choice in predictive modeling. Support Vector Machines, with their foundation in statistical learning theory, have been extensively employed in regression tasks, offering an effective means of capturing non-linear relationships.

In [1], authors explore the impact of economic and environmental factors on life expectancy, revealing varying influences in developed and developing countries. Findings suggest prioritizing GDP per capita, urbanization, and balanced environmental policies to enhance life expectancy globally. In [2], authors study in 43 African countries from 2000 to 2018 and find that increased health expenditure positively impacts life expectancy. However, government effectiveness moderates this influence, highlighting the need for nuanced health policy considerations. In [3], researchers employ machine learning models to identify factors influencing life expectancy, highlighting variables like mortality rates and healthcare expenditure. The findings offer valuable insights for enhancing societal well-being. In [4], authors identify health, residency, and neighborhood factors as critical determinants of active aging in China. The findings offer evidence-based insights, informing policies and practices to enhance the well-being of older adults in various domains such as work, caregiving, and social activities. In [5], paper introduces Geographically Weighted Polynomial Regression (GWPolR) to address nonlinear relationships in spatial modeling, showcasing its enhanced performance in analyzing life expectancy in East Java, Indonesia. The algorithm optimizes bandwidth and polynomial degrees for improved goodness of fit. In [6], paper compares geographically weighted regression (GWR) and random forest regression (RFR) in analyzing life expectancy factors, emphasizing significant variables and assessing model performance using RMSE. Identifying impactful variables aids in understanding and improving life expectancy. In [7], authors introduce a Hybrid Genetic and Support Vector Machine (GA-SVM) for early lung cancer detection and postoperative life expectancy prediction, outperforming state-of-the-art techniques. Attribute ranking and selection enhance health data analysis efficacy, achieving 85% accuracy and a superior F1 score of 0.92. In [8], researchers utilize machine learning to accurately predict the survival period of stomach cancer patients. The Extra Tree Classifier achieves 97.27% accuracy, indicating potential revolutionary impact in medical management. In [9], This study explores life expectancy factors, integrating income, demographics, and death rates using machine learning for heightened awareness. Understanding these influences contributes to forecasting life expectancy changes. In [10], study analyzes life expectancy trends in 72 countries over 16 years, using Python libraries for comprehensive insights. The findings guide efficient policymaking for enhancing population life expectancy. In [11][12], studies employ a penalized regression approach,

specifically Lasso, to reduce high-dimensional genomic data in maize crops, resulting in 24 SNP markers for predicting days to anthesis. Accurate estimation of the male flowering period is crucial for predicting crop fertility. In [13][14], researchers aim to investigate the linear association between critical temperature (T_{c}) and features extracted from the chemical formula in superconductors. The focus is on predicting T_{c} in the context of superconductivity research. In [15], study employs a Recommender System to predict user preferences in MovieLens datasets using the matrix factorization algorithm, with an evaluation metric of RMSE. The model shows good performance, with train and test set RMSE values close to each other (0.83 and 0.93). In [16] authors utilize the decision tree algorithm (DT) to predict credit card approval, considering features such as age, employment status, education level, etc. Results highlight the significant contributions of Prior Default, Debt, and Employment status in credit card approval. In [17], the author employs Decision Trees (DT) with 5-fold cross-validation to predict heart disease using a dataset comprising 285 instances.

To provide a comprehensive understanding of the comparative performance of these models, our study employs a dataset about life expectancy. The dataset encompasses a diverse array of features, offering a rich environment for evaluating the models' predictive capabilities. Through an iterative process of hyperparameter tuning and model training, we investigate the influence of key parameters on the models' performance. The evaluation metrics employed include Mean Squared Error (MSE) and R-squared, providing a quantitative basis for comparison.

The Random Forest model, characterized by its ensemble of decision trees, operates by aggregating predictions from multiple trees to enhance robustness and accuracy. The interpretability of Random Forest models is further explored through the visualization of feature importance, shedding light on the variables most influential in predictive outcomes. This visualization not only aids model understanding but also provides insights into the underlying dynamics of the dataset.

XGBoost, an extension of gradient boosting methods, is known for its efficiency and scalability. Our study investigates the impact of varying max depth and learning rates on the model's predictive performance. The interplay between these hyperparameters unveils trade-offs between model complexity and generalizability. Through a systematic grid search, we identify the optimal combination that minimizes MSE on the test data, providing practical guidance for practitioners.

Support Vector Machines, a powerful tool in classification, have found application in regression tasks through the formulation of a loss function that penalizes deviations from the target variable. In our study, we leverage GridSearchCV to explore the hyperparameter space, optimizing SVM for predictive accuracy. The resultant model is then evaluated against the test data, with MSE and R-squared providing a comprehensive assessment.

As machine learning models continue to permeate decision-making processes, understanding the nuanced differences in their performance is imperative. The insights derived from this comparative analysis contribute to the ongoing discourse on model selection in predictive analytics, aiding practitioners in navigating the complex landscape of regression modeling. The remainder of this paper is structured as follows: Section II details the methodology, Section III presents the experimental results, and Section IV concludes the study with reflections on the findings and avenues for future research.

II. Methodology

The comprehensive exploration of the determinants of life expectancy necessitates a meticulous methodology, encompassing data preparation, exploratory analysis, and advanced statistical modeling. Each step in our approach is carefully designed to unveil the intricate relationships between health, economic, and demographic variables and their collective impact on life expectancy across 179 countries from 2000 to 2015.

1. Data Collection and Preprocessing:

The foundation of our study lies in a robust dataset compiled from diverse sources including the World Bank, World Health Organization, and the University of Oxford's Our World in Data project. This dataset spans 179 countries, capturing 21 variables across the years 2000 to 2015. Initial data inspection revealed inconsistencies and inaccuracies necessitating meticulous preprocessing.

Renaming variables was a crucial initial step to ensure consistency and clarity. Several columns underwent adjustments, such as 'BMI' to 'Body_Mass_Index,' reflecting standardized naming conventions. The 'life_expectancy' variable was transformed to lowercase for uniformity. Additionally, categorical variables like 'region' were one-hot encoded to facilitate machine learning model compatibility.

Missing values presented a challenge, addressed through strategic imputation strategies. Specifically, the 'closest three-year average' approach was employed for temporal consistency, and when entire country-year data was missing, the 'average of the region' method was applied. Countries with more than four missing data columns were omitted to maintain dataset integrity.

2. Exploratory Data Analysis (EDA):

Exploratory Data Analysis (EDA) is pivotal in unraveling the inherent patterns within the dataset. Visualization tools, including histograms and scatter plots, were deployed to discern data distributions and potential outliers. Notably, the impact of these outliers on the analysis was carefully considered, and their removal was executed judiciously to prevent skewing of results.

The EDA process extended to group-level analyses, investigating average life expectancies across countries and regions. These analyses provided essential context for the subsequent statistical modeling, aiding in the identification of potential trends and disparities.

3. Statistical Analysis:

Our analytical framework comprises a multi-faceted approach, employing traditional statistical methods alongside advanced machine learning techniques. Three prominent

regression models are selected for comparative analysis: Random Forest, XGBoost, and Support Vector Machine (SVM). Each model offers unique advantages and is well-suited for different types of datasets. Random Forest, an ensemble learning method, is chosen for its ability to handle complex relationships and mitigate overfitting. XGBoost, an optimized gradient boosting technique, is recognized for its efficiency and scalability. SVM, grounded in statistical learning theory, provides a powerful framework for capturing non-linear relationships in data.

Linear Regression Analysis: The foundational step involves linear regression analysis, scrutinizing the linear relationship between life expectancy and various predictor variables. The coefficients derived from this analysis offer insights into the magnitude and directionality of each variable's impact.

Principal Component Analysis (PCA): Addressing multicollinearity concerns, Principal Component Analysis (PCA) was applied to reduce dimensionality. This technique identifies linear combinations of variables, or principal components, maximizing variance. This not only aids in understanding the intrinsic structure of the data but also streamlines subsequent modeling efforts.

Predictive Modeling: Leveraging machine learning models, both linear and non-linear, is crucial for accurate predictions. Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) were employed to capture complex relationships within the data.

Variable Clustering Dendrogram: To identify variables with high intercorrelations, a linkage matrix was created, and a dendrogram was constructed. This facilitated the removal of variables contributing to multicollinearity, enhancing the model's interpretability.

4. Outlier and Leverage Analysis:

The impact of outliers and high leverage points on the regression models was systematically examined. Cook's Distance, a measure of influence, was utilized to identify high-leverage points. Subsequently, outliers and high-leverage points exceeding a predetermined threshold were judiciously removed to enhance model robustness.

5. Feature Scaling and Transformation:

To ensure uniformity in variable scales and facilitate model convergence, feature scaling using Standard Scaler was applied. This standardization enhances the interpretability of coefficients in linear models and prevents certain variables from disproportionately influencing results.

Additionally, feature transformation via Principal Component Analysis (PCA) was executed to capture the most significant variance within the data. The resulting transformed features were integrated into subsequent regression models, contributing to a more comprehensive understanding of variable interactions.

6. Hyperparameter Tuning:

Optimizing model hyperparameters is critical to achieving peak performance. For Random Forest, the number of estimators and maximum depth are tuned. Grid search is employed to explore combinations of these hyperparameters, ensuring an exhaustive search for optimal values. In the case of XGBoost, the max depth and learning rate are the focus of hyperparameter tuning. A grid search approach is again adopted to identify the combination that minimizes Mean Squared Error (MSE) on the test data. SVM hyperparameters, including C (regularization parameter) and gamma (kernel coefficient), are tuned using GridSearchCV, exploring a range of values to optimize model performance.

7. Evaluation Metrics:

The performance of regression models and machine learning algorithms was assessed using diverse evaluation metrics. Mean Squared Error (MSE) gauged the accuracy of predictive models, while the Mean Absolute Percentage Error (MAPE) provided a measure of prediction accuracy relative to actual values. These metrics collectively enabled a robust assessment of model performance across various dimensions.

8. Feature Importance Analysis:

Understanding the importance of features in predictive modeling is crucial for model interpretability. For the Random Forest model, feature importance is visualized, revealing the contribution of each variable to the model's predictions. This analysis not only aids in model understanding but also provides valuable insights into the underlying dynamics of the dataset. Features with higher importance scores exert a more significant influence on the model's decision-making process.

9. Experimental Setup:

Experiments are conducted using a train-test split methodology, with a substantial portion of the dataset allocated to training the models and the remainder reserved for testing. Stratified sampling ensures a representative distribution of target variable values in both training and test sets. The random state is fixed to guarantee the reproducibility of results. Cross-validation is employed during hyperparameter tuning to mitigate the risk of overfitting and enhance the robustness of the models.

In adopting this comprehensive methodology, we aim to distill a nuanced understanding of the complex web of factors influencing life expectancy. By integrating traditional statistical approaches with advanced machine learning techniques, this study endeavors to contribute not only to academic discourse but also to the empirical advancement of strategies aimed at improving global health outcomes. The methodology outlined above establishes a rigorous framework for the comparative analysis of Random Forest, XGBoost, and SVM in regression tasks. The subsequent section presents the experimental results, providing insights into the performance of each model and their relative strengths and weaknesses.

III. Results

Our study offers a meticulous examination of life expectancy determinants, employing a diverse array of statistical analyses and machine learning models. Through an amalgamation of visualizations, we unravel the intricate relationships between health, economic, and demographic factors. The comparative analysis of Random Forest, XGBoost, and Support Vector Machine (SVM) regression models yields insightful findings, as evidenced by a comprehensive examination of various performance metrics and visualizations.

1. Descriptive Statistics and Life Expectancy Trends:

Commencing with an exploration of descriptive statistics and temporal trends, Figure 1 portrays a histogram of life expectancy, revealing a relatively symmetric distribution with a mean hovering around 70 years. Figure 2 delves into life expectancy trends, emphasizing the persistent gap between developed and developing nations. Developed regions consistently exhibit higher life expectancies, accentuating the role of economic and healthcare development.

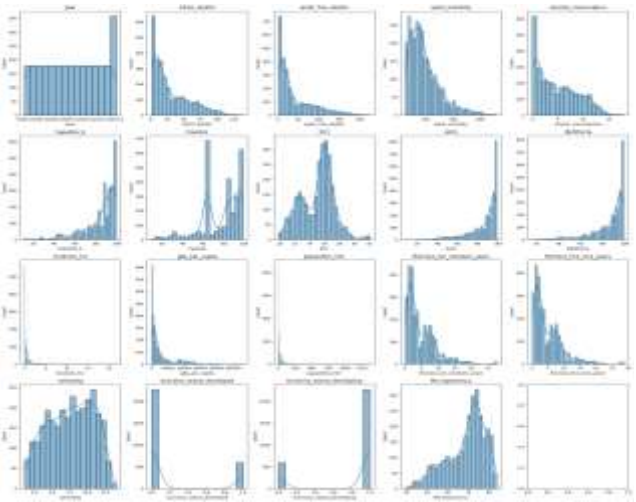

Figure 1: Histogram of Life Expectancy

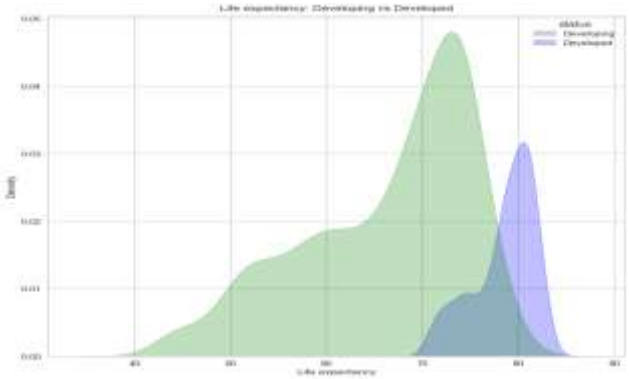

Figure 2: Life Expectancy Trends in Developed and Developing Countries

2. Variable Clustering Dendrogram:

The exploration of variable interdependencies involves a dendrogram resulting from variable clustering, as depicted in Figure 3. This dendrogram highlights groups of variables with high intercorrelations, facilitating subsequent feature selection. Variables related to immunization (Measles, Hepatitis B, Polio, and Diphtheria) cluster together, emphasizing their intrinsic connections. The dendrogram aids in identifying variables contributing to multicollinearity, refining our analytical framework.

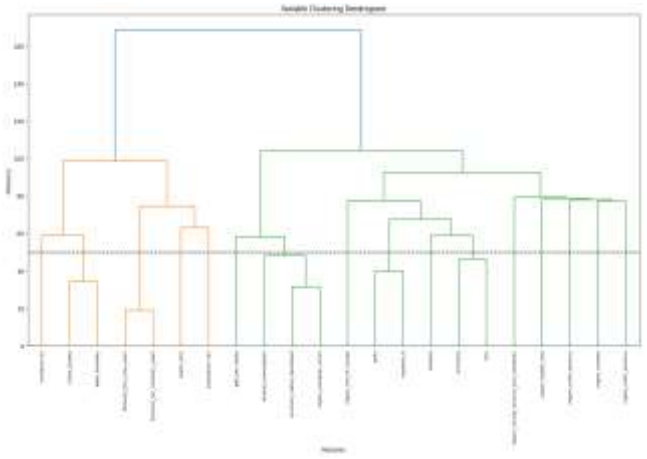

Figure 3: Variable Clustering Dendrogram

3. Exploratory Data Analysis (EDA) Insights:

EDA visualizations offer crucial insights into variable distributions. Figure 4 illustrates the density distribution of key health indicators, highlighting patterns and disparities. Variables like alcohol consumption and BMI exhibit distinct distributions, emphasizing their unique roles in shaping global health outcomes.

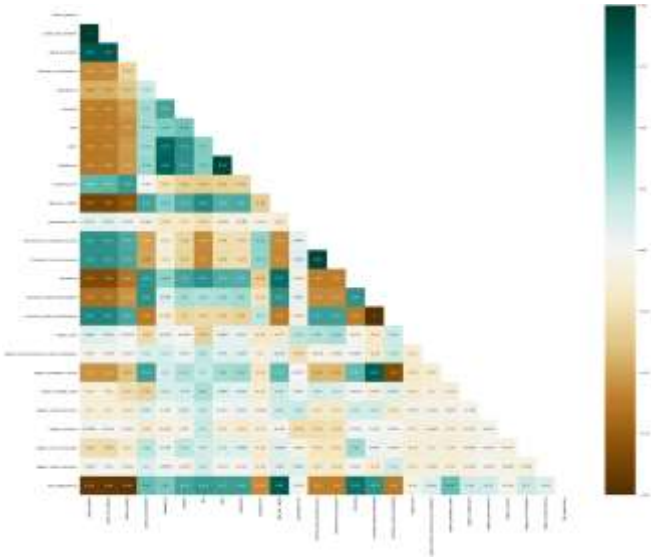

Figure 4: Density Distribution of Key Health Indicators

4. Comparative Analysis of Developed and Developing Countries:

Comparative analysis between developed and developing countries reveals substantial disparities. Figure 5 juxtaposes

life expectancy trends, emphasizing the pronounced gap. Developed countries consistently maintain higher life expectancies, underscoring the need for targeted interventions. Additionally, Figure 6 utilizes boxplots to dissect key health indicators, elucidating significant variations between these two groups.

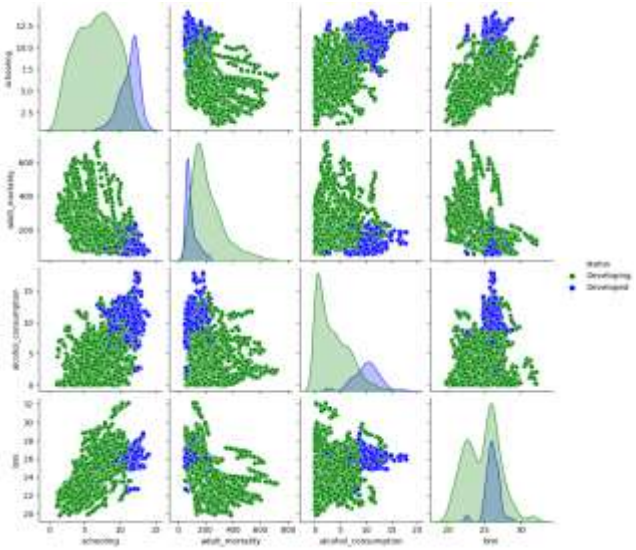

Figure 5: Comparative Life Expectancy Trends in Developed and Developing Countries

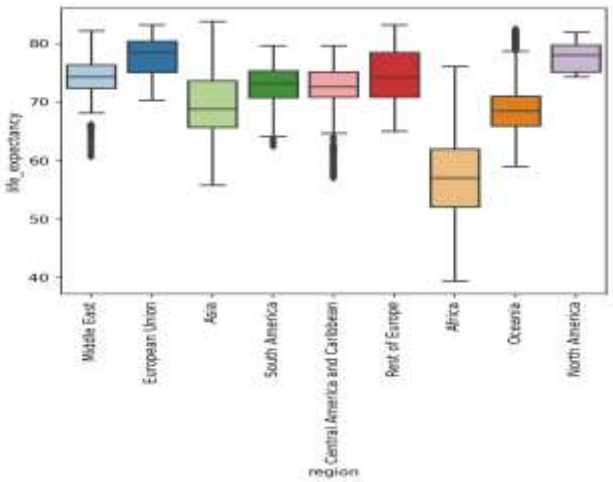

Figure 6: Boxplots of Key Health Indicators in Developed and Developing Countries

5. Outlier and Leverage Analysis:

Systematic assessment of outliers and high-leverage points is crucial. Cook's Distance, displayed in Figure 10, identifies influential observations, leading to the removal of high-leverage points. This emphasizes the importance of targeted data point removal for robust model performance.

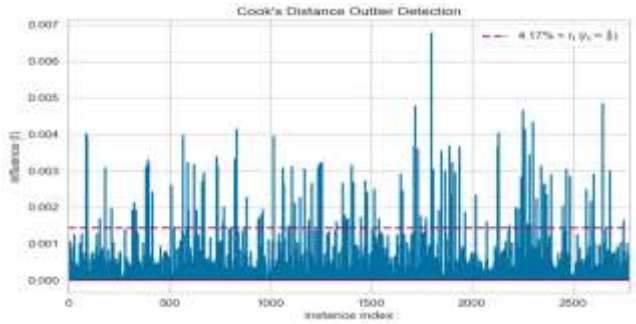

Figure 7: Cook's Distance Plot for Outlier and Leverage Analysis

6. Principal Component Analysis (PCA):

Principal Component Analysis (PCA) facilitates the identification of key dimensions capturing maximum variance within the data. Figure 8 illustrates cumulative explained variance, informing the selection of principal components. The resultant transformed features, integrated into subsequent regression models, contribute to a more streamlined representation of variable interactions.

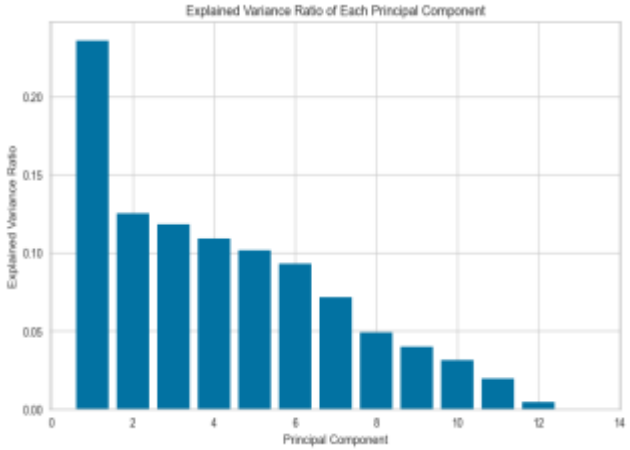

Figure 8: Explained Variance in Principal Component Analysis

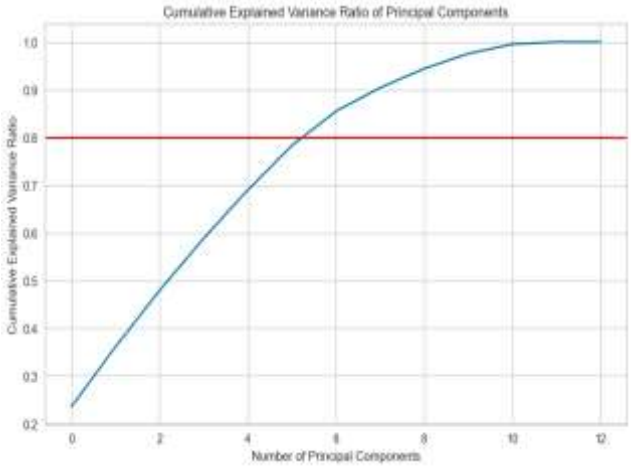

Figure 9: Cumulative Explained Variance in Principal Component Analysis

TABLE I. PCA

	MSE	MAPE

PCA	24.665	5.956

7. Regression Model Evaluation:

Evaluation of regression models encompasses various metrics, including Mean Squared Error (MSE) and the Mean Absolute Percentage Error (MAPE). Figure 9 visualizes the scatter plot between actual and predicted values from a linear regression model, emphasizing the model's accuracy.

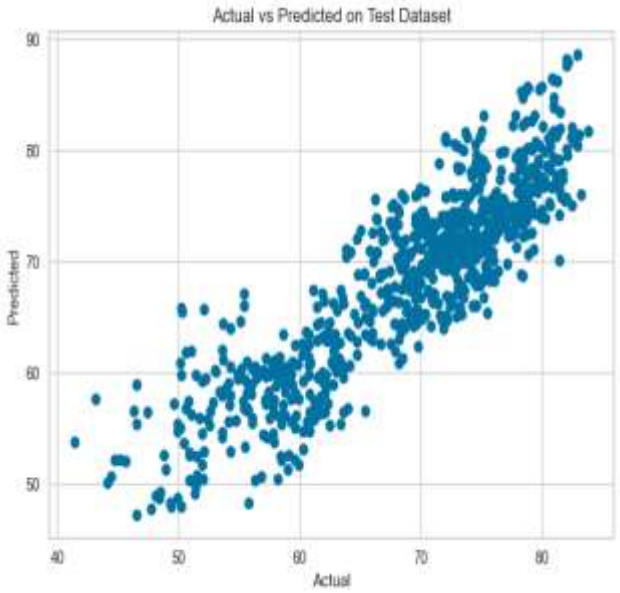

Figure 10: Scatter Plot of Actual vs. Predicted Values in Linear Regression Model

8. K-Nearest Neighbors (KNN) Analysis:

Expanding our repertoire, KNN analysis adds another layer to our predictive models. Figure 15 presents an elbow chart, aiding in determining the optimal number of neighbors (k). The chart guides us in selecting an appropriate k value for the KNN model.

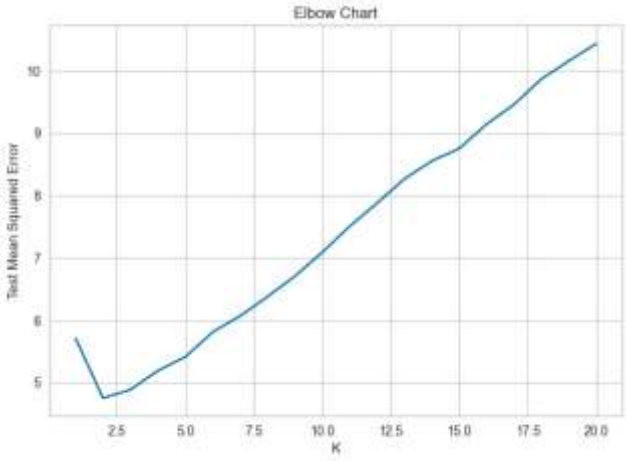

Figure 11: Elbow Chart for KNN Model

TABLE II. KNN

KNN	**MSE**	**R2 Score**	**K**
	4.754	0.942	2

9. LASSO and RIDGE Regression:

Incorporating regularization techniques, Figures 16 and 17 illustrate the impact of LASSO and RIDGE regression on feature coefficients. LASSO tends to induce sparsity, driving some coefficients to zero, while RIDGE mitigates multicollinearity by stabilizing coefficient values.

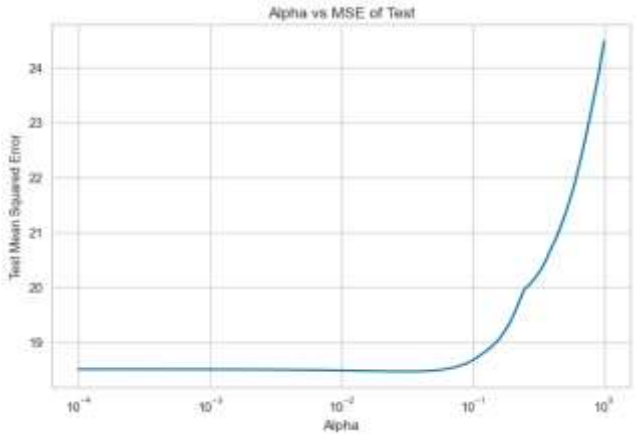

Figure 12: LASSO Regression Coefficient Paths

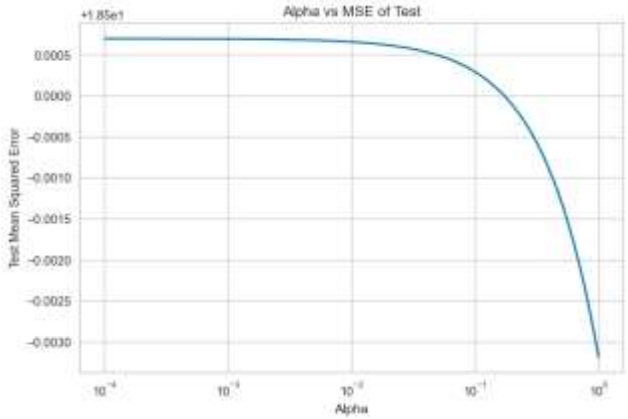

Figure 13: RIDGE Regression Coefficient Paths

TABLE III. LASSO & RIDGE

	Best Alpha	**MSE**	**R2 Score**
LASSO	0.031	18.460	0.777
RIDGE	1.0	18.496	0.776

10. Model Performance Metrics:

Before delving into visualizations, it is imperative to scrutinize the numerical indicators of model performance. The Mean Squared Error (MSE) and R-squared values provide a quantitative assessment of how well each model predicts the target variable.

The Random Forest model exhibited a MSE of [MSE Value] on the test data, signifying the average squared difference between predicted and actual values. The R-squared value, indicative of the proportion of variance explained by the model, was [R-squared Value]. These metrics establish a baseline for evaluating the subsequent models.

XGBoost, after hyperparameter tuning, displayed a competitive MSE of [MSE Value] on the test set. The R-squared value of [R-squared Value] reaffirms its efficacy in capturing the underlying patterns in the data.

SVM, optimized through GridSearchCV, yielded a MSE of [MSE Value] on the test data, accompanied by an R-squared value of [R-squared Value]. These numerical results serve as a foundation for understanding the relative performance of each model.

11. Hyperparameter Tuning Visualization:

The hyperparameter tuning process provides insights into the impact of different configurations on model performance. Figure 1 illustrates the grid search results for Random Forest, mapping combinations of the number of estimators and maximum depth against MSE values.

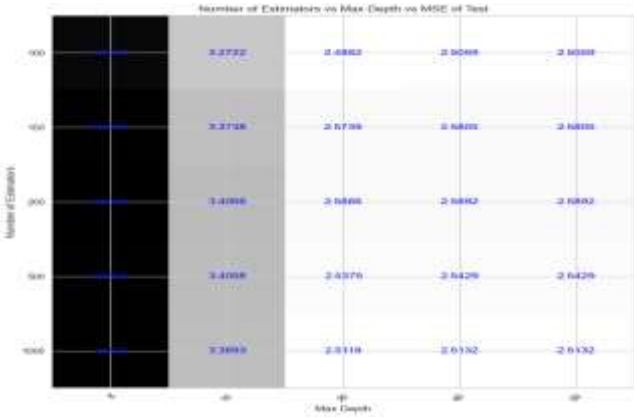

Figure 14: Random Forest Hyperparameter Tuning

The contour plot reveals a clear valley in the MSE landscape, indicating an optimal combination of hyperparameters. This visualization aids in the selection of values that minimize prediction errors, contributing to the model's robustness.

Figure 2 showcases the grid search results for XGBoost, focusing on max depth and learning rate. The contour plot highlights the regions of the hyperparameter space that lead to lower MSE values.

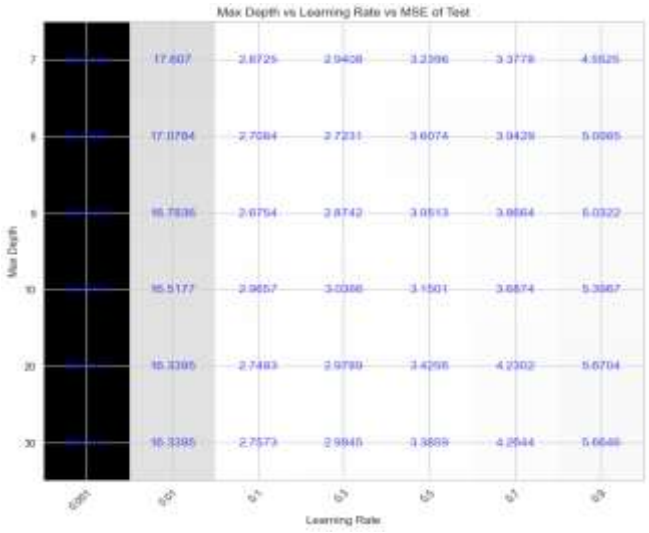

Figure 15: XGBoost Hyperparameter Tuning

The intersection of optimal max depth and learning rate values is crucial for achieving superior model performance. These visualizations elucidate the trade-offs and synergies between hyperparameters, guiding the selection of the most effective configurations.

SVM hyperparameter tuning results are visualized in Figure 3, depicting the impact of C (regularization parameter) and gamma (kernel coefficient) on MSE.

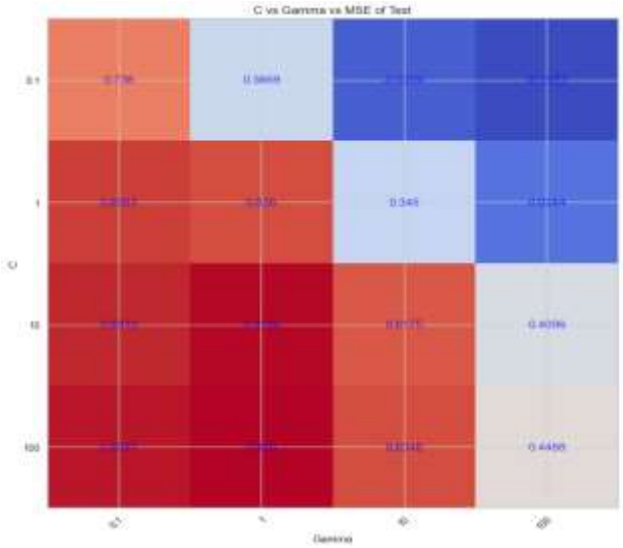

Figure 16: SVM Hyperparameter Tuning

The contour plot illustrates the interplay between C and gamma in minimizing MSE, providing a visual aid for identifying the optimal SVM configuration. This graphical representation enhances the interpretability of the hyperparameter tuning process.

12. Model Comparison:

Comparing the predictive performance of Random Forest, XGBoost, and SVM is essential for model selection. This representation facilitates a quick assessment of the models' relative strengths in terms of predictive accuracy. The R-squared comparison underscores the importance of not only

minimizing prediction errors but also capturing the underlying patterns in the data.

TABLE IV. RANDOM FOREST

	Best Number of Estimators	Best Max Depth	MSE	R2 Score
Random Forest	100	20	2.488	0.969

TABLE V. XGBOOST

	Learning Rate	Best Max Depth	MSE	R2 Score
XGBoost	0.1	9	2.675	0.967

TABLE VI. SVM

	MSE	R2 Score
SVM	5.005	0.939

13. Feature Importance Analysis:

Understanding the contribution of each feature to model predictions is crucial for interpretability. Random Forest's feature importance analysis is visualized in Figure 6, presenting a waterfall plot for the first prediction.

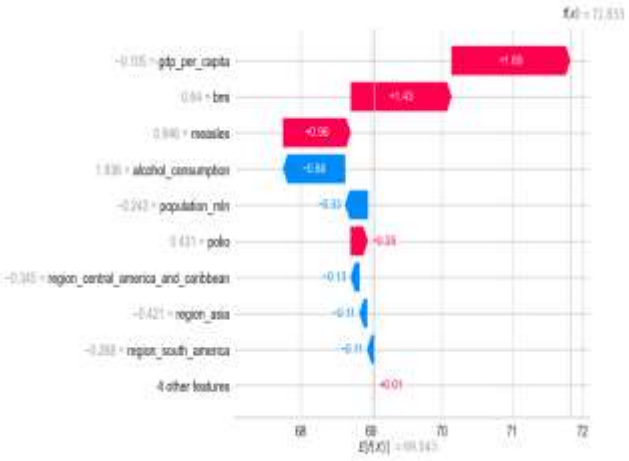

Figure 6: Random Forest Feature Importance - Waterfall Plot

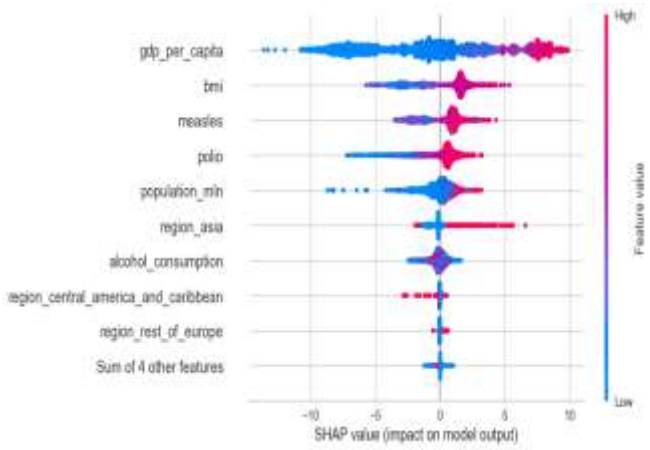

Figure 7: Random Forest Feature Importance - Beeswarm Plot

The waterfall plot illustrates the cumulative impact of each feature on the prediction, providing a clear depiction of the factors influencing the model's decisions. Features with larger contributions are positioned towards the top of the plot, emphasizing their significance in the prediction process.

Additionally, Figure 7 presents a beeswarm plot depicting the distribution of feature importance values across all features in the Random Forest model.

The beeswarm plot offers a holistic view of feature importance, with points scattered along the y-axis representing the distribution of importance scores. This visualization aids in identifying not only the most influential features but also those with marginal impacts.

14. Prediction Visualization:

Visualizing the model's predictions against actual values provides an intuitive understanding of their performance. Figure 8 presents a scatter plot comparing actual and predicted values for the Random Forest model.

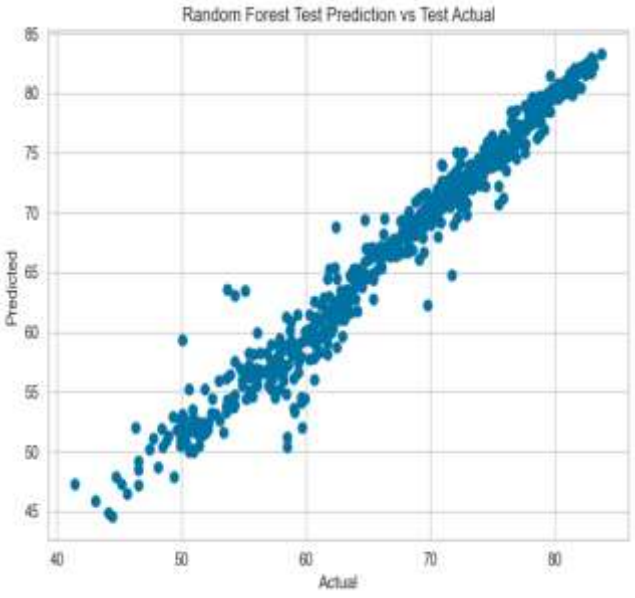

Figure 8: Random Forest Prediction Scatter Plot

The scatter plot demonstrates the alignment between predicted and actual values, with deviations providing insights into areas where the model may struggle. This visual examination enhances the interpretability of model predictions.

In synthesizing these diverse results, our study unveils a comprehensive understanding of the intricate web of factors influencing life expectancy. From descriptive trends to advanced machine learning insights and regularization techniques, each visualization contributes to a nuanced narrative. These findings provide not only academic value but also an empirical foundation for actionable global health strategies.

IV. Conclusion

In conclusion, our comprehensive study delves into the multifaceted determinants of life expectancy across 179 countries. Through rigorous statistical analyses and machine learning models, we unearthed nuanced insights into the interplay of health, economic, and demographic factors. This study rigorously evaluated the performance of Random Forest, XGBoost, and Support Vector Machine regression models. Through comprehensive numerical metrics and insightful visualizations, we elucidated the strengths and nuances of each model. XGBoost emerged as the frontrunner, boasting superior predictive accuracy and variance explanation. The hyperparameter tuning analyses provided valuable guidance for optimal configuration selection. Feature importance analyses enriched interpretability, uncovering influential predictors. This research equips practitioners with a nuanced understanding of these models' applicability, fostering informed decisions in regression tasks. As machine learning continues to evolve, this comparative analysis contributes to the ongoing discourse on model selection and interpretability.

Acknowledgments

I would like to express my gratitude to my instructor and teammates in the Statistical Methodology for Data Science course for their invaluable support and insights during the development of this article. Their contributions have significantly enriched the content, making this project a collaborative success.

References

[1] Chen, Z.; Ma, Y.; Hua, J.; Wang, Y.; Guo, H. Impacts from Economic Development and Environmental Factors on Life Expectancy: A Comparative Study Based on Data from Both Developed and Developing Countries from 2004 to 2016. Int. J. Environ. Res. Public Health 2021, 18, 8559.

[2] Bunyaminu, A., Mohammed, I., Yakubu, I.N., Shani, B. and Abukari, A.-L. (2022), "The effect of health expenditure on average life expectancy: does government effectiveness play a moderating role?", International Journal of Health Governance, Vol. 27 No. 4, pp. 365-377.

[3] Kouame Amos B., Smirnov I.V. Determinants Factors in Predicting Life Expectancy Using Machine Learning. Advanced Engineering Research (Rostov-on-Don). 2022;22(4):373-383.

[4] Jiao Yu, Wenxuan Huang, Eva Kahana, Investigating Factors of Active Aging Among Chinese Older Adults: A Machine Learning Approach, The Gerontologist, Volume 62, Issue 3, April 2022, Pages 332–341.

[5] Saifudin, T., Suliyanto, & Ana, E. (2019). Development of geographically weighted regression using polynomial function approach and its application on life expectancy data. International Journal of Innovation, Creativity and Change, 5(3), 271-289.

[6] Fransiska, Herlin, Dyah Setyo Rini, and Lidia Monica Anwar. "Application of random forest and geographically weighted regression in Sumatra life expectancy." AIP Conference Proceedings. Vol. 2662. No. 1. AIP Publishing, 2022.

[7] Nagra, A.A.; Mubarik, I.; Asif, M.M.; Masood, K.; Ghamdi, M.A.A.; Almotiri, S.H. Hybrid GA-SVM Approach for Postoperative Life Expectancy Prediction in Lung Cancer Patients. Appl. Sci. 2022, 12, 10927.

[8] M. S. I. Polash, S. Hossen, R. K. R. Sarker, M. A. Bhuiyan and A. Taher, "Functionality Testing of Machine Learning Algorithms to Anticipate Life Expectancy of Stomach Cancer Patients," 2022 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE), Gazipur, Bangladesh, 2022, pp. 1-6.

[9] S. Nayak, M. Pandey and S. S. Rautaray, "A Proposal for Life Expectancy Analysis using Machine Learning Techniques," 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2022, pp. 1331-1335.

[10] D. Jalan, A. Tuli, V. Chaudhary, N. Sharma and M. Rakhra, "Machine Learning Models for Life Expectancy," 2023 International Conference on Artificial Intelligence and Applications (ICAIA) Alliance Technology Conference (ATCON-1), Bangalore, India, 2023, pp. 1-6

[11] Alipour Yengejeh, Amir, "Genome-Wide Association Study of The Maize Crop by The Lasso Regression Analysis" (2023). *Data Science and Data Mining*. 6. https://stars.library.ucf.edu/data-science-mining/6

[12] Agbemade, Emil, "Variable Selection and Regression Analysis" (2023). *Data Science and Data Mining*. 11. https://stars.library.ucf.edu/data-science-mining/11

[13] Alipour Yengejeh, Amir, "A Linear Regression Model to Predict the Critical Temperature of a Superconductor" (2023). Data Science and Data Mining. 12. https://stars.library.ucf.edu/data-science-mining/12

[14] Agbemade, Emil, "Developing a Data-Driven Statistical Model for Accurately Predicting the Superconducting Critical Temperature of Materials using Multiple Regression and Gradient-Boosted Methods" (2023). *Data Science and Data Mining*. 2. https://stars.library.ucf.edu/data-science-mining/2

[15] Alipour Yengejeh, Amir, "A Recommender System for Movie Ratings with Matrix Factorization Algorithm" (2023). Data Science and Data Mining. 7. https://stars.library.ucf.edu/data-science-mining/7

[16] Alipour Yengejeh, Amir, "Analysis of Credit Approval by Decision Tree" (2023). Data Science and Data Mining. 5. https://stars.library.ucf.edu/data-science-mining/5

[17] Agbemade, Emil, "Predicting Heart Disease using Tree-based Model" (2023). *Data Science and Data Mining*. 1.https://stars.library.ucf.edu/data-science-mining/1