**"Global Life Expectancy Modeling Project Final Report"**

**TEAM 5**

**Date: "April 2025"**

**1. Introduction**

**a. Project Motivation**

We chose to study **global life expectancy** due to its importance as a summary measure of population health and socioeconomic development. Understanding what influences life expectancy can guide better policy decisions and resource allocation across nations.

**b. Background Research**

Previous research has shown that life expectancy is influenced by economic, demographic, and health-related factors. However, global-scale analysis with recent machine learning tools remains limited. This project applies modern regression models to investigate patterns and determinants of global life expectancy.

**c. Dataset Description**

We used a cleaned dataset derived from the **World Bank** database. It contains country-level data from 1990 to 2023, including indicators such as GDP, death rates, population structure, and rural demographics.

**d. Key EDA Findings**

* Life expectancy increased steadily from 1990 to 2019, peaking around 2019 before dropping sharply during the COVID-19 pandemic (2020–2021).
* Some countries displayed stagnation or decrease due to conflict, poor healthcare, or external shocks.
* Variables like GDP per capita, crude death rate, and child population percentage showed strong correlations with life expectancy.
* There appeared to be 3 main clusters of data: smaller countries with weaker economies and lower life expectancies, larger countries with more robust economies and higher life expectancies, and those that fell in between for both of those categories. Most of the countries fell into the low to mid ranges.

**2. SMART Research Question**

**SMART Question:** *What are the most important socio-economic and demographic factors influencing life expectancy across countries from 1990 to 2023, and how accurately can we predict life expectancy using machine learning models?*

This question guided our analysis and model building. We aimed to identify top predictors and compare model performance using standard evaluation metrics.

**3. Modeling Approach**

**a. Feature Engineering**

* Dropped gender-specific life expectancy to avoid multicollinearity.
* Label encoded country names.
* Sorted data by year and split into training (<=2022) and test (>2022) sets.

**b. Algorithms Used**

We applied and compared three regression models:

* **Random Forest Regressor**
* **Decision Tree Regressor**
* **Bayesian Ridge Regressor**

**c. Evaluation Metrics**

* **Mean Squared Error (MSE)**
* **R-squared (R²)**

**d. Results Summary**

|  |  |  |
| --- | --- | --- |
| **Model** | **MSE** | **R²** |
| Random Forest Regressor | 8.62 | 0.884 |
| Decision Tree Regressor | 9.54 | 0.872 |
| Bayesian Ridge Regressor | 15.38 | 0.794 |

**e. Feature Importance (from Random Forest)**

Top predictors:

* Population ages 0-14 (% of total population)
* Death rate, crude (per 1,000 people)
* GDP per capita (current US$)

**4. Interpretation of Results**

* The **Random Forest model** showed the best predictive performance, with high R² and the lowest MSE.
* The **most influential features** were demographic (youth population), economic (GDP per capita), and health-related (death rate).
* Surprisingly, other commonly assumed predictors like total population or urban/rural breakdown had minimal predictive power.

**5. Time Series Forecasting**

**a. Trend Analysis**

We created a global average time series of life expectancy from 1990 to 2023. The trend showed a consistent rise until 2019, followed by a dip during the COVID-19 pandemic and a partial rebound.

**b. ARIMA and SARIMAX Models**

We forecasted the next 5 years of life expectancy:

* **ARIMA** predicted a continued modest increase.
* **SARIMAX**, which includes seasonal components, produced a less reliable result with higher error.

**6. Predictions and Implications**

Our models suggest that:

* Countries with **high youth population** and **low death rates** will likely continue experiencing **higher life expectancies**.
* **Policy interventions** that reduce mortality and support child and maternal health may have the most impact.
* Economic growth (especially GDP per capita) still plays a vital role in improving life expectancy.

**7. Conclusion**

By combining machine learning and time series models, we found that life expectancy can be **accurately predicted** using a small number of demographic and economic variables.

Our findings answer the SMART question by highlighting:

* Top influencing factors.
* The comparative performance of different models.
* Reliable future predictions with ARIMA.

The analysis offers insights for global health organizations and governments aiming to enhance life expectancy through targeted data-driven strategies.

**8. References**

* World Bank Data. (n.d.). *World Development Indicators*. Retrieved from https://databank.worldbank.org
* Himmelfarb Health Sciences Library. (n.d.). *APA Citation Style, 7th Edition*. GWU.
* Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research.
* Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3rd ed.). OTexts.