PROJECT REPORT

A Predictive Analysis of Life Expectancy Based on Global Socioeconomic Data

1. INTRODUCTION

Life expectancy is considered to be one of the critical indicators of a nation's overall well-being, as this reflects the overall health, economic, and social conditions of its general population. If analysed and understood effectively, it can help policymakers, healthcare professionals, and researchers develop strategies to improve public health and economic stability.

This project aims to develop a predictive model to estimate life expectancy in different countries based on key socioeconomic factors such as **GDP, homicide rates, urban population, population growth, import/export, fertility rates, education levels, and more**. By leveraging machine learning techniques, we analyze how these variables interact and contribute to variations in life expectancy across various nations. With efficient predictions, organizations can make informed decisions to enhance quality of life, allocate resources efficiently, and implement policies that promote healthier societies.

---

2. DATA SOURCE

This dataset comprises 204 entries and 38 attributes, providing a comprehensive analysis of key economic and social indicators across various countries. It includes a diverse range of metrics, allowing for in-depth exploration of global trends related to GDP, education, health, and environmental factors. Here are the key features:

GDP: Gross Domestic Product (in current US dollars), representing the total economic output of a country.
Sex Ratio: The ratio of males to females in the population, highlighting demographic trends.
Life Expectancy: Average lifespan for males and females, an essential indicator of healthcare quality.
Education Enrollment Rates: Data on primary, secondary, and post-secondary education enrollment for males and females, reflecting educational attainment.
Unemployment Rate: Percentage of the labor force that is unemployed, indicating economic health.
Homicide Rate: Number of homicides per 100,000 population, providing insight into safety and crime levels.
Urban Population Growth: Rate of growth in urban populations, illustrating migration trends.

CO2 Emissions: Carbon dioxide emissions per capita, an important measure of environmental impact.
Forested Area: Percentage of land covered by forests, indicating biodiversity and environmental health.
Tourist Numbers: Total number of international visitors, which can reflect a country's tourism potential.

Dataset link:
https://www.kaggle.com/datasets/arslaan5/global-data-gdp-life-expectancy-and-more

---

3. DATA WRANGLING

Data types: There are a total of 38 columns in the dataset, out of them 5 are non-numeric and the rest are numeric types. The non-numeric type includes the country names, the country code, the capital, the currency and the region. The other 33 columns are numeric and seem perfect from the modelling standpoint.
Missing data: There seem to be quite a few missing values in the dataset. The feature 'CO2-emissions' has the highest value at 59 (29%). Dropping every row that has a missing value is not an ideal way to handle the missing data. Further analysis suggested dropping 1 row, which has most of the values missing.
Duplicate Values: Grouping by the country code, it was found that 12 rows have been duplicated, so they can be dropped as well.
Fill Values: The numerical columns with missing values were decidedly filled with the median values based on the regions the countries belong to.
Outliers: There seem to be a lot of the columns that tend to have outliers in large numbers but the outliers do hold information for countries ranging in all sorts of ways. I believe these values cannot be removed/deleted as they hold important information, and are not mistakes. So the outliers have been included in the final dataset.

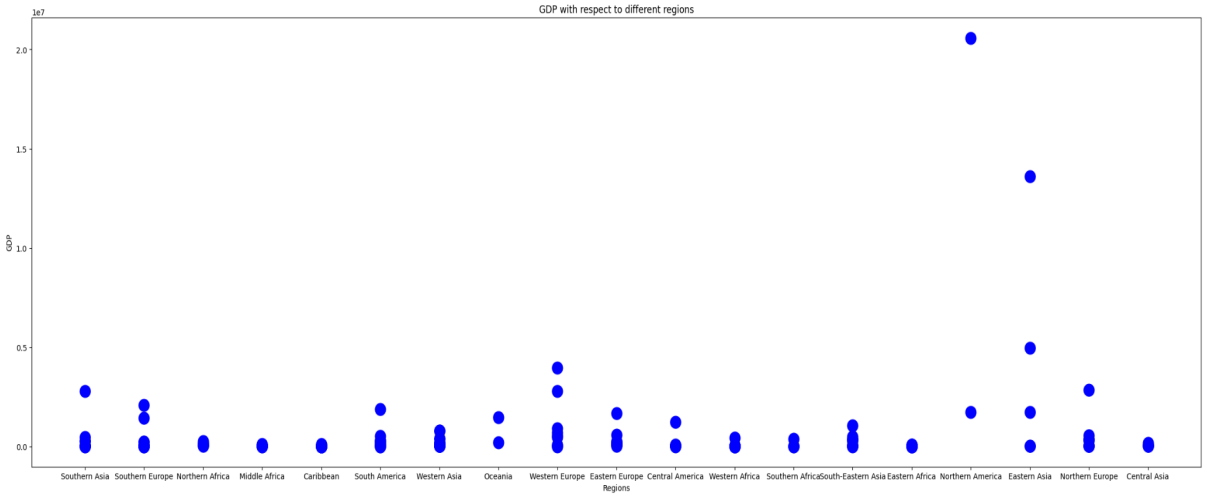
---

4. Exploratory Data Analysis (EDA)

EDA is an essential step for exploring the data to visually understand the patterns, and uncover relationships and potential data quality issues. After performing the necessary data wrangling steps and cleaning the data of any duplicates and missing values, we are left with a dataset having 38 features and 179 rows of information to further explore and understand.

1. <u>GDP distribution across different regions:</u>

   Northern America and Eastern Asia have outliers with extremely high GDP (e.g., the U.S. and China). Regions like Middle Africa, Western Africa, and the Caribbean have lower GDP values, showing economic disparity.
   The distribution suggests a **significant economic gap** between different world regions.

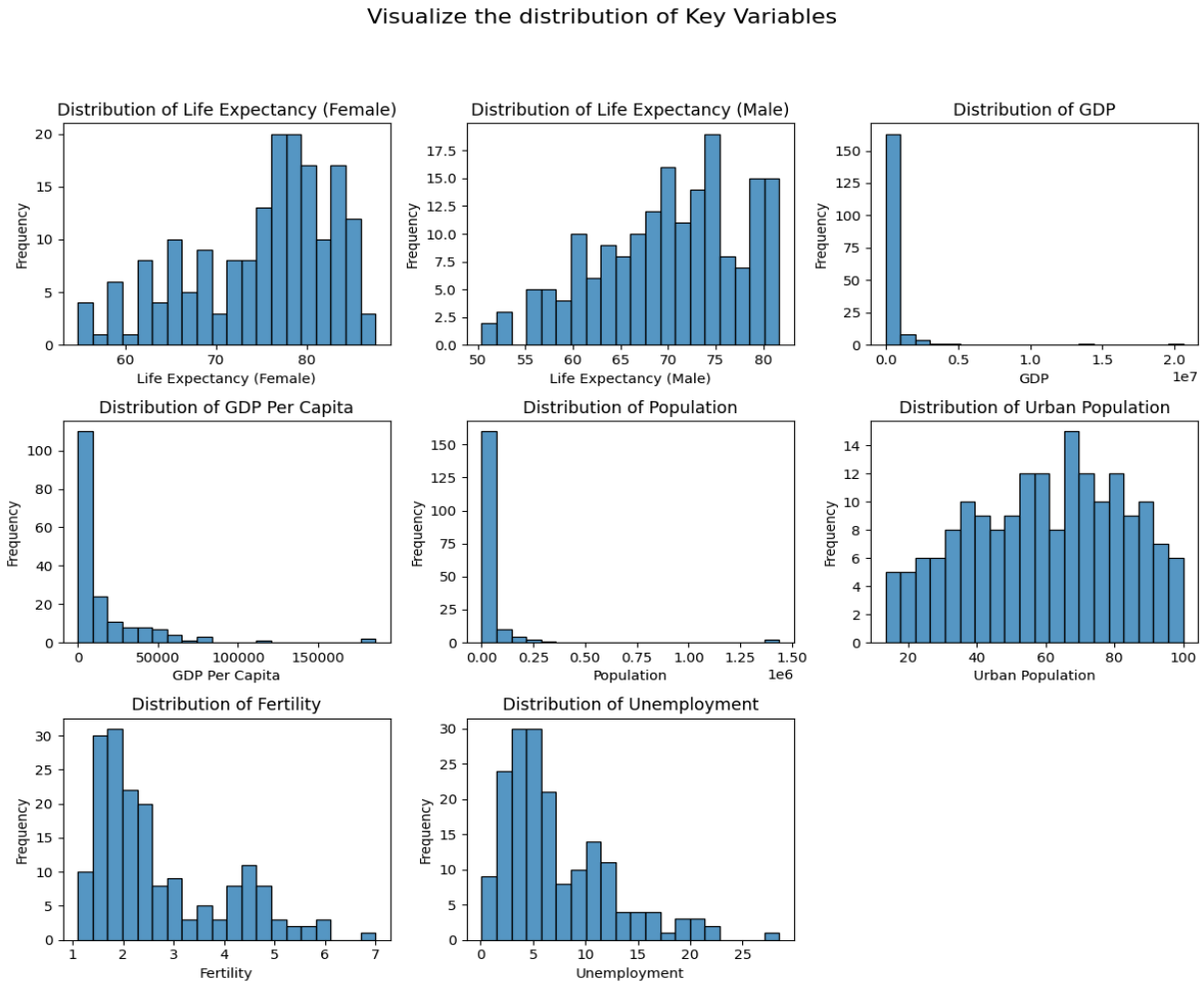


2. <u>Histograms for key economic variables related to Life Expectancy:</u>

   The distributions reveal a few key socioeconomic trends affecting life expectancy.

   Life Expectancy: Both male and female life expectancy are generally high, with most countries having values above 65 years. This indicates overall good health outcomes, although disparities exist, with some countries having lower life expectancy.
   Economic Disparities: The skewed distributions of GDP and GDP per capita highlight substantial economic inequalities between countries. A few countries dominate the higher end of the spectrum, while the majority have much lower economic output and per capita income.
   Urbanization: The urban population distribution shows a diverse range of urbanization levels, indicating varying degrees of urban development across countries. Some countries have high urban populations, while others remain largely rural.
   Fertility Rates: The left-skewed fertility distribution suggests that many countries have low fertility rates, potentially indicating advanced stages of demographic transition with lower birth rates.
   Unemployment: The right-skewed unemployment distribution shows that while most countries have relatively low unemployment rates, there are significant outliers with much higher rates, indicating areas with potential economic distress. These insights can help in understanding the socio-economic landscape of different countries and identifying areas for targeted interventions or further research.

   The visualizations of the data distributions provide essential insights into global trends. Life expectancy is generally high, especially among females, while GDP and GDP per capita show significant economic disparities. Countries with lower fertility rates often have higher life expectancy. Having a clear understanding of these distributions helps in building a predictive model for life expectancy more effectively.



3. <u>Comparing countries with the highest and lowest GDPs:</u>

   Massive GDP Disparity: The United States ($20.58 trillion) and China ($13.6 trillion) dominate the global economy, while São Tomé and Príncipe have a GDP of just $411 million—50,000 times smaller than the United States of America.

   CO_2 Emissions Correlation: Higher GDP countries tend to have significantly higher CO_2 emissions (e.g., China: 9,257.9, U.S.A: 4,761.3), reflecting higher industrialization levels. Lower GDP countries, like São Tomé and Príncipe (3.40), contribute minimally to global emissions.

   Trade Activity Difference: The top economies have billions in trade activity, with China exporting $2.49 trillion and importing $2.07 trillion. The bottom economies trade in millions, with São Tomé and Príncipe exporting just $14 million, which is a negligible fraction of global trade.

   These insights highlight the economic gap between developed and developing nations, which can impact factors like life expectancy, infrastructure, and public services.

5. PREPROCESSING AND FEATURE ENGINEERING

This step involves **transforming raw data into a structured, meaningful format** by normalizing and scaling features, encoding categorical variables, and creating new informative variables from the already existing ones to improve model performance and interpretability. A few steps were taken for this dataset as well before starting with the modelling process.

A few columns related to life expectancy and various education levels have data separately for the male and female populations. Since the correlation between them is very high (0.97 for life expectancy_male/female), they can interfere with the modelling process. Upon statistical inference, we can decide to replace the two separate columns for male and female, with a single column taking the average value for each entry.
For the desired analysis, columns like country code, capital and currency are not required, so we can drop them before going further into the modelling. This leaves us with only two non-numeric columns, the name of the country and the region it belongs to.

Creating dummy variables and scaling the data:
We separate the dataset into two, one with only numeric columns and the other with a categorical column which is the region in this case. Since the dataset varies significantly, it requires scaling the numeric values.
Also, for the purpose of modelling, we require the data to be completely numeric. So we create dummies for the region column. After the necessary transformations, the data is again put together into one completely processed dataset.

Train and Test Split:
The dataset needs to be split into two datasets at random sets to first train the model using the training dataset and then to evaluate the model performance using the test dataset. For this project, the dataset has been split with 80% of the data used for training and 20% for evaluating the performance of the trained model.

---

6. MODELLING

The target variable here is life expectancy, which is a continuous numeric variable. Hence we need to go for regression models. Regression is appropriate for life expectancy prediction because it deals with continuous, real-valued predictions, and it allows you to model relationships between life expectancy and various influencing factors. Three modelling techniques have been used to evaluate and compare the performance and select the model that best fits our data.

A. Linear Regression as the Baseline Model

The modelling process was started with **Linear Regression**, a simple and interpretable method for predicting continuous outcomes such as life expectancy. This model was

selected as a baseline due to its straightforward nature and the assumption of a linear relationship between the predictors and the target variable.
To evaluate the model's performance and compare with other models, we used three key metrics:

1. **Mean Squared Error (MSE)**: Measures the average of the squares of the errors, providing insight into the model's accuracy.
2. **Root Mean Squared Error (RMSE)**: The square root of MSE, which gives a more interpretable error metric in the same unit as the target variable (life expectancy).
3. **R-squared (R^2)**: Indicates the proportion of variance in the target variable explained by the model. A higher R^2 means a better model fit.

After training the model, the results were:

- **MSE**: 0.2009
- **RMSE**: 0.4483
- **R^2**: 0.8543

The R^2 of 0.854 indicated that the model could explain 85% of the variance in life expectancy, which was a strong start, but there was room for improvement.

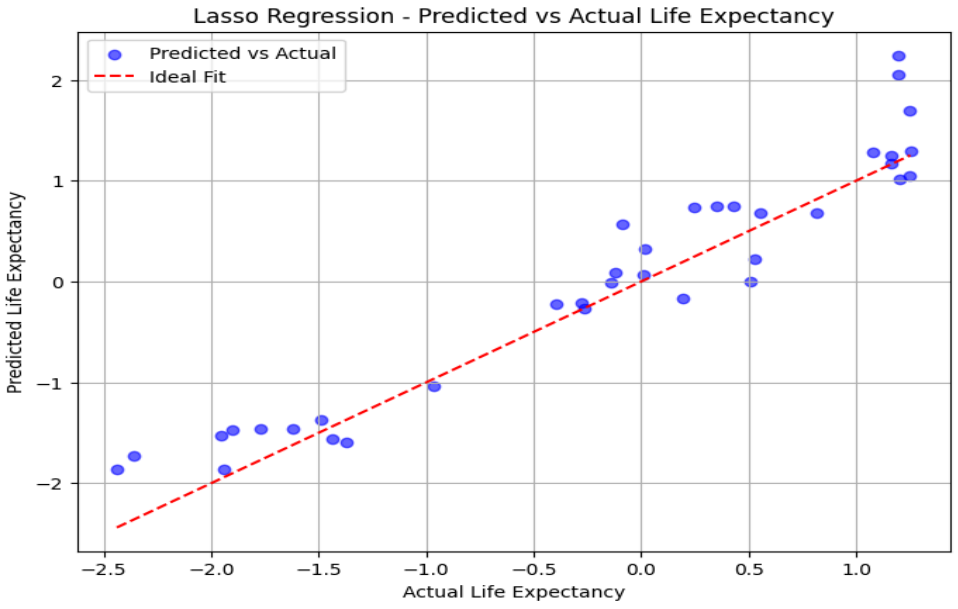
Hyperparameter Tuning with Lasso Regression

To improve the model's performance, we turned to **Lasso Regression**. Lasso, an extension of linear regression with L1 regularization, helps prevent overfitting by penalizing large coefficients, effectively performing feature selection.

GridSearchCV was used for hyperparameter tuning to find the best value for the **alpha** parameter (regularization strength). This was done to improve the model's generalization ability and reduce overfitting.

The hyperparameter tuning process resulted in:

- **Best alpha**: 0.01
- **MSE**: 0.1424
- **RMSE**: 0.3774
- **R^2**: 0.8967



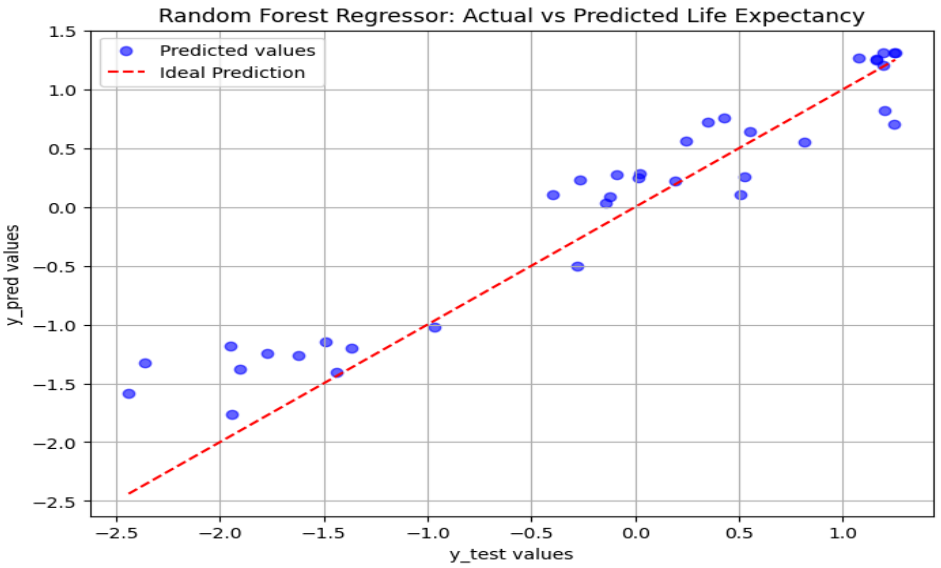
Lasso Regression with an optimized alpha performed better than Linear Regression, showing a decrease in both MSE and RMSE, and a significant increase in R^2, indicating a stronger fit to the data.

B. Random Forest Regressor

Random Forest Regressor was the second model of choice, a more complex model known for handling non-linear relationships and capturing feature interactions. Initially, it was trained with default parameters. Although Random Forest had strong predictive power, it did not outperform Lasso Regression significantly. Hyperparameter tuning via **RandomizedSearchCV** was then conducted, but it slightly reduced performance.

The final results for Random Forest were:

- **MSE**: 0.1536
- **RMSE**: 0.3919
- **R^2**: 0.8886



C. XGBoost Regressor

Lastly, I implemented **XGBoost**, a popular algorithm for regression tasks. Despite its ability to handle complex relationships, it did not outperform Lasso Regression. We applied **GridSearchCV** to tune the hyperparameters, but the results were similar to the default model.

The final results for XGBoost were:

- **MSE**: 0.1465
- **RMSE**: 0.3828
- **R^2**: 0.8938

---

7. CONCLUSION

	Model	MSE	RMSE	R-squared
0	Lasso Regression	0.142431	0.377400	0.896742
1	Random Forest	0.153603	0.391922	0.888643
2	XGBoost	0.146533	0.382796	0.893768

The results showed that **Lasso Regression** with optimized hyperparameters provided the best performance in terms of MSE, RMSE, and R^2. The model explained nearly 90% of the variance in life expectancy and offered the most accurate predictions, making it the most suitable model for this task.

The **Random Forest** and **XGBoost** models showed competitive performance but did not surpass **Lasso Regression**. This suggests that while more complex models can capture intricate relationships, a simpler, regularized model like **Lasso Regression** can often yield better results, especially with the appropriate hyperparameter tuning.

---

8. Future Work

To further improve the model's performance or have further work done with respect to life expectancy analysis worldwide, the following routes could be considered:

1. **Feature Engineering**: Identifying new features or transforming existing ones could enhance the model's ability to capture underlying patterns.
2. **Advanced Models**: Exploring other advanced models such as **Gradient Boosting Machines (GBM)** or **Neural Networks** might yield better results.
3. Understanding the life expectancy classifications based on regions could uncover geographic trends or clusters of nations with similar health characteristics, allowing for targeted interventions.
4. **Life Expectancy Analysis Based on Gender**: The gender gap in life expectancy can be studied and explored, particularly focusing on factors that contribute to differences between men and women, such as healthcare access, lifestyle choices, and societal roles.