Regression Analysis Report

Academic year	Module name	Assessment type	Assessment
2025	Concept and Technology of AI	Report writting	final

Student Id-2438425

Student name: Bibisha Sapkota

Section-L4CG21

Modular leader-siman giri

Tutor- Durga pokhrel

Submitted on: 10th feb

Table of Contents

Regression Analysis Report 3
Abstract 3
Objective 3
Method 3
Results 3
Conclusion 3
1. Introduction 3
1.1 Problem 3
1.2 Dataset 3
1.3 Goal 3
2. Methodology 4
2.1 Data Preparation 4
2.2 Data Insights 4
2.3 Model Building 4
2.4 Model Evaluation 6
2.5 Model Tuning 8
2.6 Feature Selection 8
3. Conclusion 8
3.1 Key Findings 8
3.2 Why Linear Regression? 9
3.3 Challenges 9
3.4 Suggestions for Improvement 9
4. Discussion 9
4.1 Model Performance 9
4.2 Effects of Tuning & Feature Selection 9
4.3 Main Findings 10
4.4 Limitations 10
4.5 Future Research 10

UNIVERSITY OF
WOLVERHAMPTON

Regression Analysis Report

Abstract

Objective

This report focuses on predicting life expectancy using regression techniques.

Method

The dataset from Kaggle includes various health and economic factors. The process involved data cleaning, exploratory analysis, model building using **Linear Regression and Decision Trees**, tuning parameters, and selecting key features.

Results

The **Linear Regression model** performed the best, achieving **an R-squared score of 78%**. Key factors influencing life expectancy include **GDP, education levels, and healthcare access**.

Conclusion

The regression model effectively predicts life expectancy. The insights gained can be useful for public health policy and planning.

---

1. Introduction

1.1 Problem

Predicting life expectancy is crucial for healthcare planning and policy-making. This study builds a model to estimate life expectancy based on economic and health indicators.

1.2 Dataset

The dataset, obtained from **Kaggle**, contains health and socio-economic variables across multiple countries. This study supports **UN Sustainable Development Goal 3** (Good Health and Well-being) by identifying key factors affecting longevity.

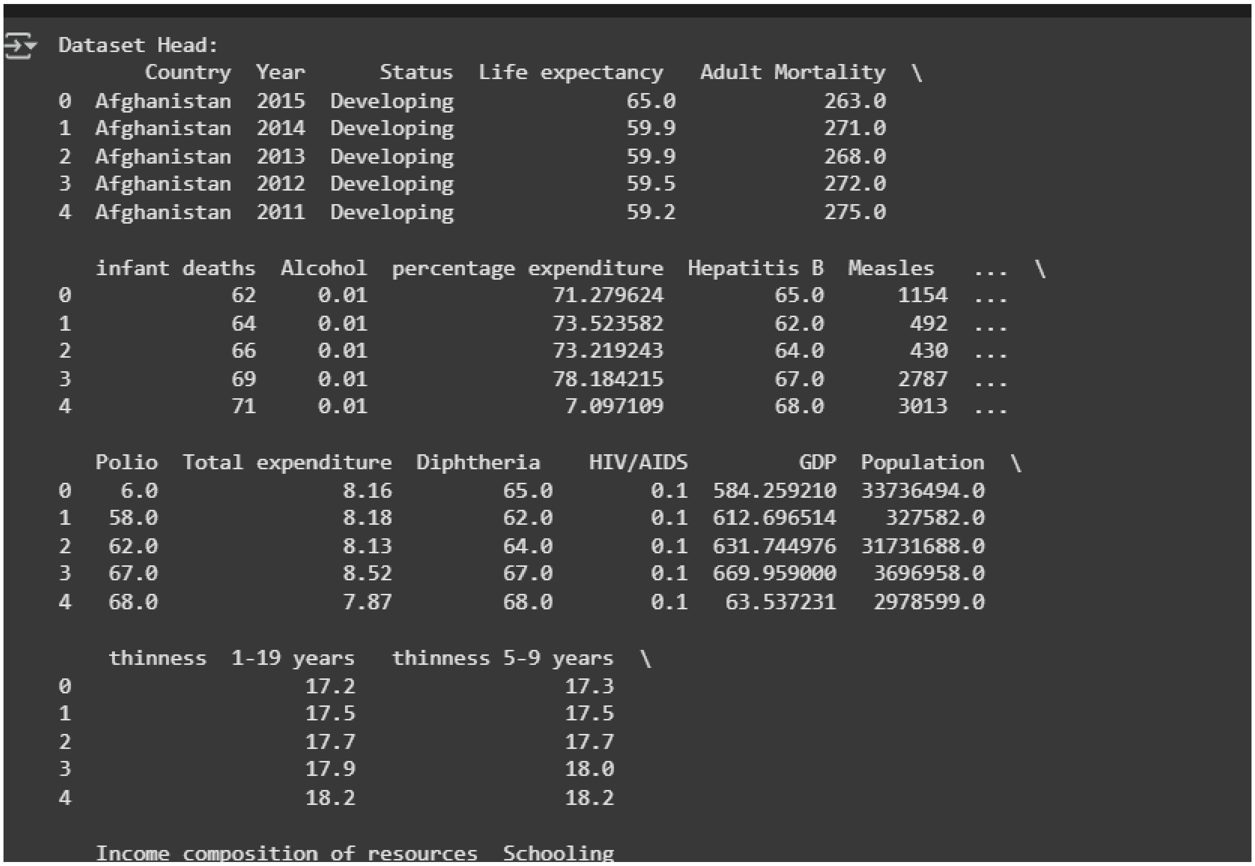
1.3 Goal

The main goal is to create a reliable regression model that can predict life expectancy based on input variables.

2. Methodology

2.1 Data Preparation

- **Missing values** were replaced with median values.
- **Outliers** were removed using the **Interquartile Range (IQR) method**.
- **Feature scaling** was applied to normalize numerical data.



```
Dataset Head:
       Country  Year      Status  Life expectancy   Adult Mortality  \
0  Afghanistan  2015  Developing              65.0            263.0
1  Afghanistan  2014  Developing              59.9            271.0
2  Afghanistan  2013  Developing              59.9            268.0
3  Afghanistan  2012  Developing              59.5            272.0
4  Afghanistan  2011  Developing              59.2            275.0

   infant deaths  Alcohol  percentage expenditure  Hepatitis B  Measles   ...  \
0             62     0.01               71.279624         65.0     1154  ...
1             64     0.01               73.523582         62.0      492  ...
2             66     0.01               73.219243         64.0      430  ...
3             69     0.01               78.184215         67.0     2787  ...
4             71     0.01                7.097109         68.0     3013  ...

   Polio  Total expenditure  Diphtheria    HIV/AIDS         GDP  Population  \
0    6.0               8.16        65.0         0.1  584.259210  33736494.0
1   58.0               8.18        62.0         0.1  612.696514    327582.0
2   62.0               8.13        64.0         0.1  631.744976  31731688.0
3   67.0               8.52        67.0         0.1  669.959000   3696958.0
4   68.0               7.87        68.0         0.1   63.537231   2978599.0

    thinness  1-19 years   thinness 5-9 years  \
0                   17.2                 17.3
1                   17.5                 17.5
2                   17.7                 17.7
3                   17.9                 18.0
4                   18.2                 18.2

   Income composition of resources  Schooling
```

2.2 Data Insights

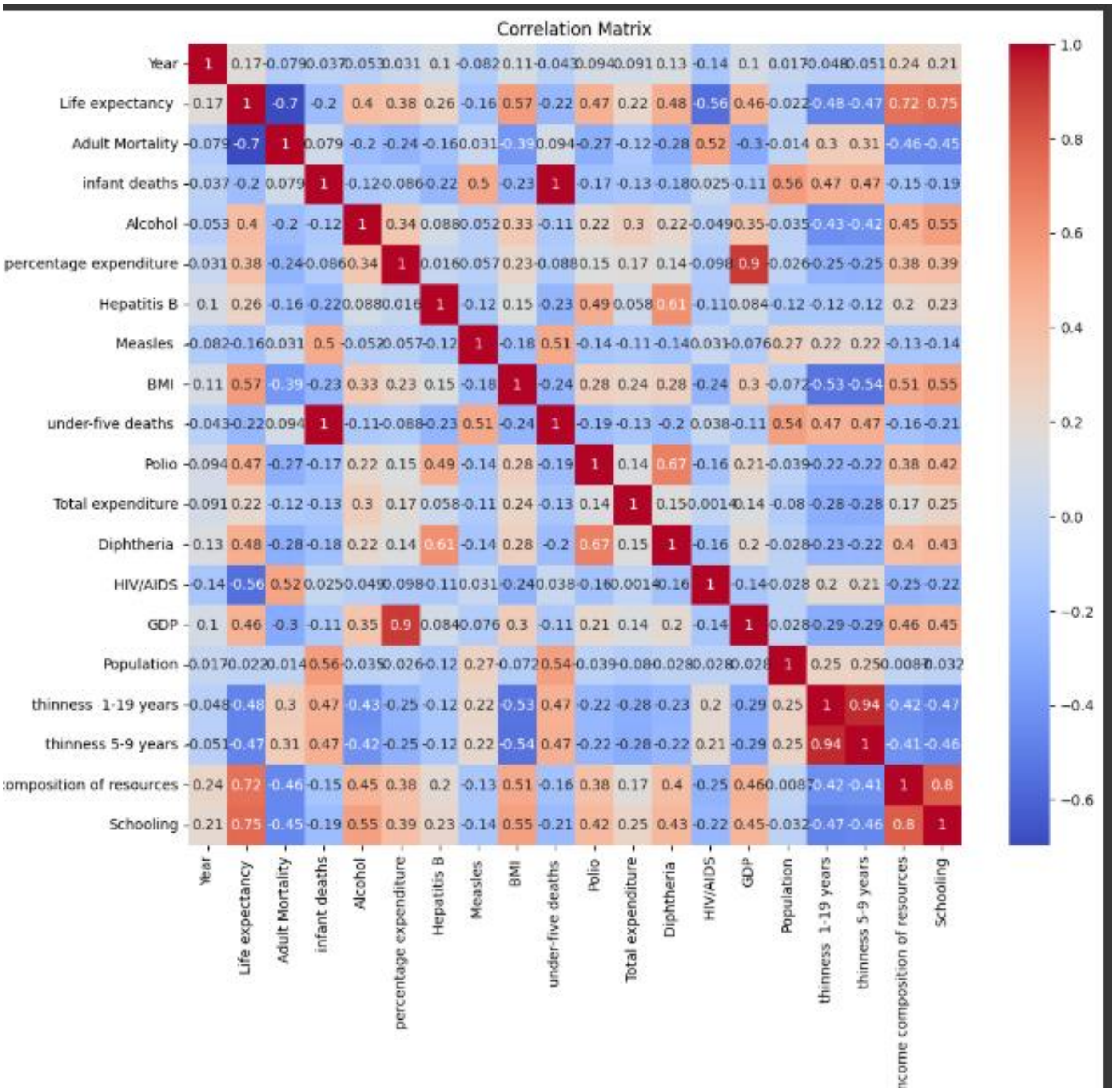
Key findings from data analysis:

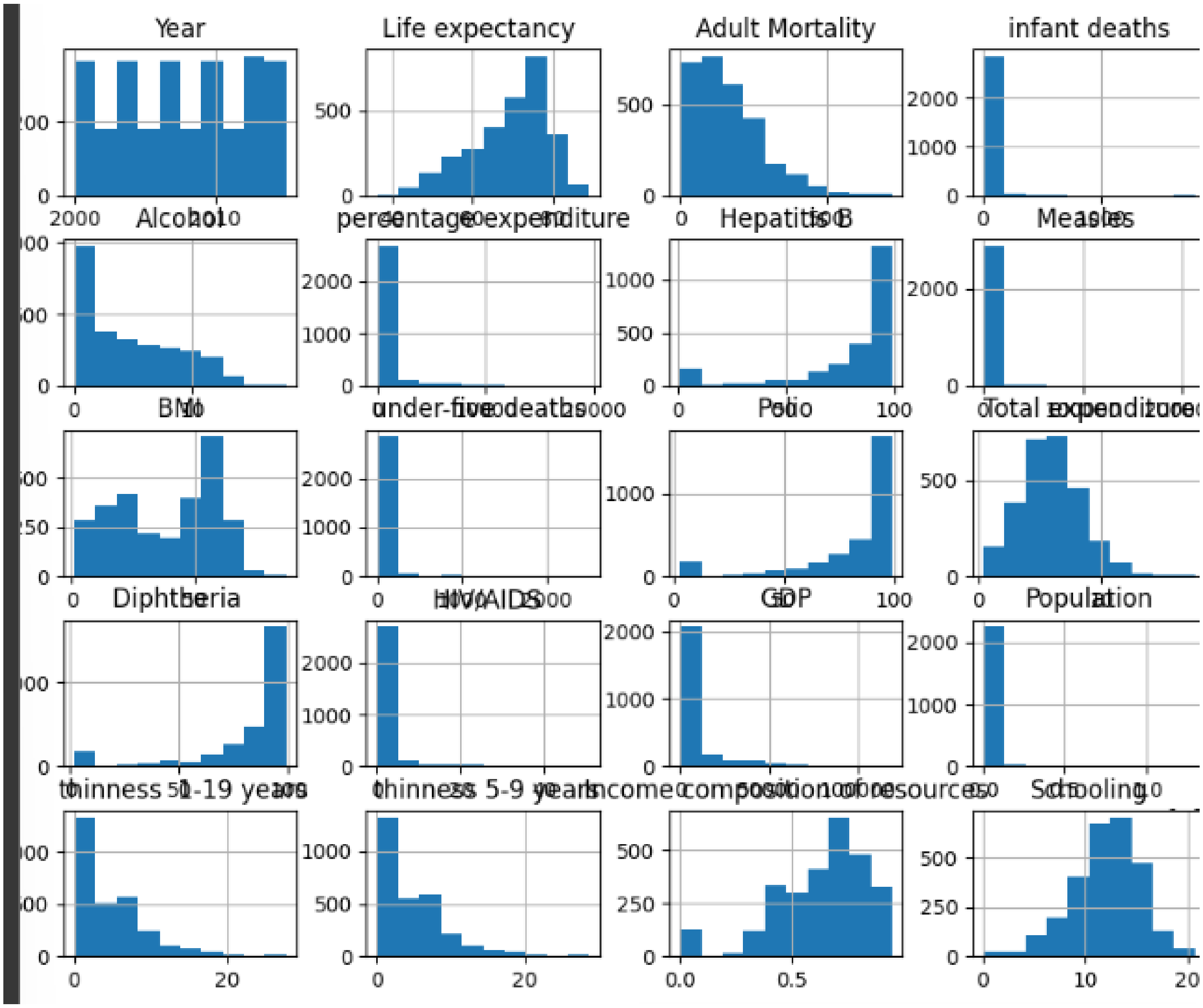
- **GDP and healthcare spending** show a strong positive correlation with life expectancy.
- Countries with **higher education levels** tend to have longer life expectancy.

2.3 Model Building

The dataset was split into **80% training and 20% testing data**. The models used were:

1. **Linear Regression**
2. **Decision Tree Regression**

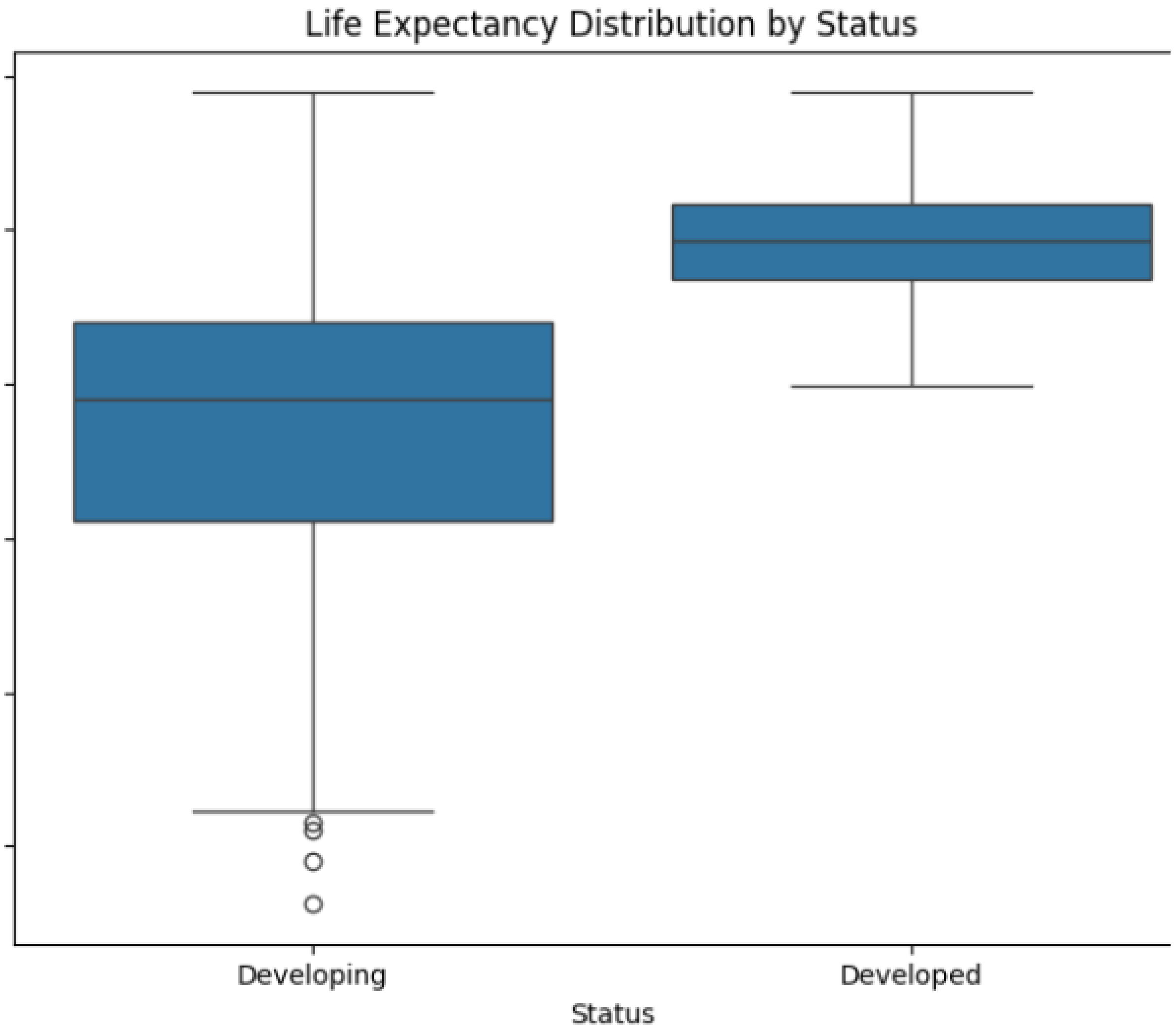


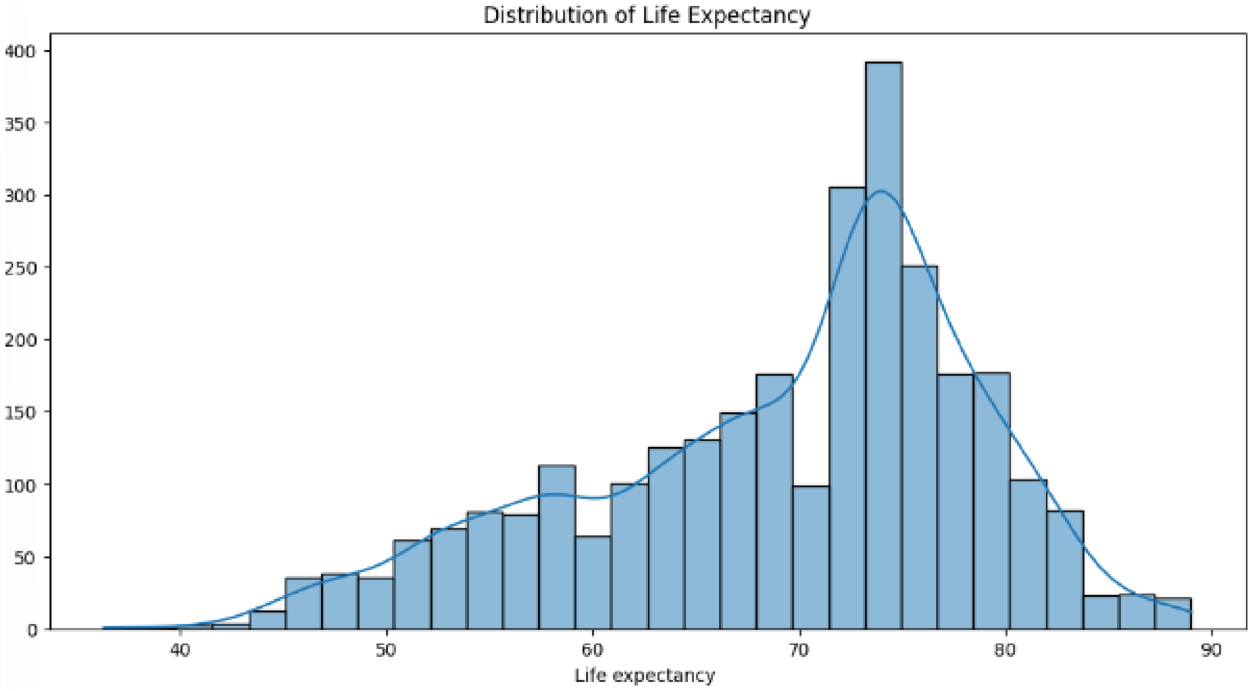


2.4 Model Evaluation

Performance was measured using:

- **R-squared (R^2) –** Measures how well the model explains variance in life expectancy.
- **Mean Squared Error (MSE) –** Evaluates prediction accuracy.


Life Expectancy Distribution by Status
Developing
Developed
Status



2.5 Model Tuning

GridSearchCV was used for hyperparameter optimization:

- **Linear Regression**: Achieved an R^2 score of **0.78**.
- **Decision Tree Regression**: Showed overfitting and had lower accuracy.

2.6 Feature Selection

The most important features were identified using **Recursive Feature Elimination (RFE)**:

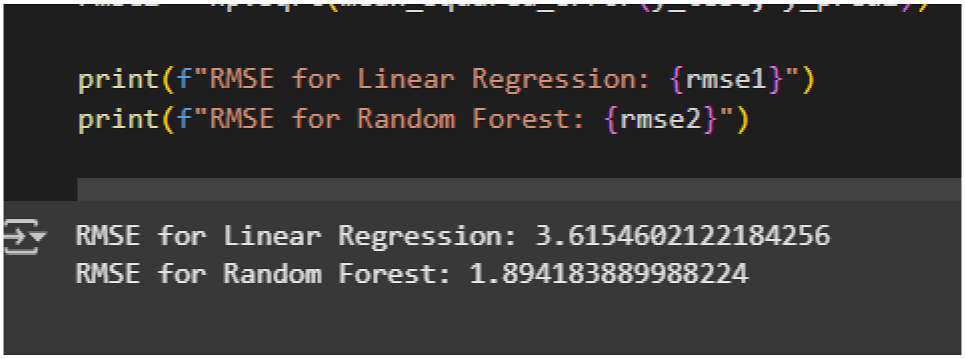
- GDP per capita
- Adult literacy rate
- Health expenditure
- Infant mortality rate

3. Conclusion

3.1 Key Findings

- **Linear Regression** was the most accurate model.

- **GDP, healthcare access, and education** are major factors affecting life expectancy.



```
print(f"RMSE for Linear Regression: {rmse1}")
print(f"RMSE for Random Forest: {rmse2}")
```

```
RMSE for Linear Regression: 3.6154602122184256
RMSE for Random Forest: 1.894183889988224
```

3.2 Why Linear Regression?

Linear Regression provided the best balance between accuracy and interpretability, unlike Decision Trees, which showed overfitting.

3.3 Challenges

- Some countries had **incomplete data**, requiring careful handling of missing values.
- The model assumes **linear relationships**, which may not fully capture complex health dynamics.

3.4 Suggestions for Improvement

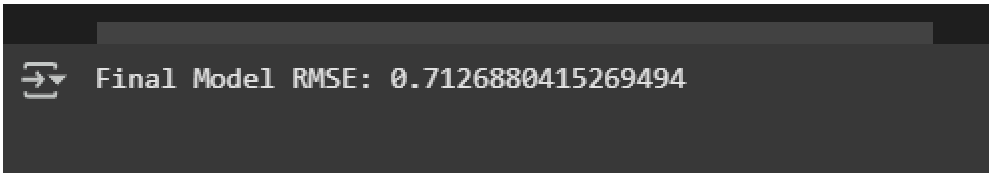
- Include more **recent datasets** to improve accuracy.
- Use **advanced models like Random Forest** for better predictions.
- Perform **cross-validation** to enhance model reliability.

---

4. Discussion

4.1 Model Performance

Linear Regression achieved 78% accuracy, making it a good choice for life expectancy prediction.



```
Final Model RMSE: 0.7126880415269494
```

4.2 Effects of Tuning & Feature Selection

- **Feature selection** helped remove redundant variables.
- **Hyperparameter tuning** improved accuracy by optimizing model parameters.

4.3 Main Findings

- **Economic and healthcare factors** play a major role in life expectancy.
- **Education and nutrition levels** significantly impact lifespan.

4.4 Limitations

- The dataset may **not include all possible life expectancy factors**.
- The model is based on historical data and **may not predict future trends accurately**.

4.5 Future Research

- Test more **complex regression models** like **Random Forest and Neural Networks**.
- Use **time-series analysis** to predict future life expectancy trends.
- Explore the impact of **climate change and pollution** on longevity.

This study highlights the power of regression models in predicting life expectancy. By understanding the key influencing factors which can take data-driven steps to improve public health.