Advanced Programming 2025

Explaining Life Expectancy through Feature Importance and Interpretable Machine Learning Models

Final Project Report

Leonard Perrigault
leonard.perrigault@unil.ch

December 8, 2025

Abstract

Life expectancy is a fundamental indicator of a country's well-being, influenced by complex interactions between economic, social, and health factors. This project addresses the challenge of interpreting these relationships using five machine learning algorithms: Linear Regression, Lasso, Ridge, Random Forest, and XGBoost. Using the Global Country Information Dataset 2023 (187 countries, 35 features), we systematically compare model performance and provide interpretability through SHAP analysis and Variance Inflation Factor (VIF) analysis. Results demonstrate that Random Forest achieves the best performance ($R^2 = 0.89$), with infant mortality emerging as the most influential predictor across all models. GDP dependency analysis reveals that GDP per capita outperforms raw GDP, suggesting wealth distribution matters more than absolute economic size.

Keywords: life expectancy prediction, interpretable machine learning, feature importance, SHAP analysis, multicollinearity, Python, scikit-learn

Contents

1 Introduction

Life expectancy represents one of the most critical indicators of societal development and public health quality. While traditional approaches have identified GDP as a dominant predictor, many socioeconomic indicators exhibit strong correlations, creating multicollinearity challenges that obscure individual contributions of specific features.

This project combines predictive modeling with interpretability analysis to understand which factors most significantly influence life expectancy. Rather than solely maximizing accuracy, we emphasize transparency through SHAP (SHapley Additive exPlanations) values and systematic feature importance comparison.

1.1 Objectives

The primary objectives are:

- Systematically compare five machine learning algorithms for life expectancy prediction
- Identify and analyze multicollinearity among features using VIF analysis
- Provide model interpretability through SHAP values and feature importance analysis
- Investigate the specific role of GDP by comparing models with/without GDP and with GDP per capita

1.2 Report Organization

Section 2 reviews relevant literature and tools; Section 3 details the dataset, preprocessing, and models; Section 4 presents experimental results; Section 5 interprets findings and discusses limitations; Section 6 summarizes contributions and suggests future work.

2 Literature Review

2.1 Machine Learning for Health Prediction

Random Forest and Gradient Boosting methods (XGBoost) have demonstrated strong performance in health prediction tasks due to their ability to model non-linear relationships and feature interactions. While academic papers exist on these methods, our implementation relies primarily on scikit-learn [5] and XGBoost [7] official documentation as reference sources.

Linear models with regularization (Lasso and Ridge) remain valuable for interpretability. While theoretical foundations are established in the literature, we used scikit-learn documentation [5] for practical implementation guidance.

2.2 Interpretable Machine Learning

SHAP values provide a unified framework for interpreting machine learning models based on cooperative game theory. While academic papers establish the theoretical foundation, we primarily used the SHAP library documentation [3] and practical tutorials [4] for implementation.

2.3 Multicollinearity Detection

Variance Inflation Factor (VIF) quantifies multicollinearity by measuring how much the variance of a regression coefficient is inflated due to correlation with other predictors. DataCamp's tutorial [6] provided practical guidance for VIF interpretation alongside statsmodels documentation.

3 Methodology

3.1 Data Description

The analysis uses the **Global Country Information Dataset 2023** from Kaggle, providing socioeconomic, health, and environmental indicators. After preprocessing, the dataset contains:

- **Observations:** 187 countries
- **Features:** 35 total (29 numeric, 6 categorical)
- **Target Variable:** Life expectancy (years)
- **Feature Categories:** Economic (GDP, CPI, tax rates), Health (infant/maternal mortality, physicians), Demographic (population, birth rate, fertility), Education (enrollment rates), Environmental (CO2 emissions, forested area), Geographic (latitude, longitude)

3.2 Data Preprocessing

A systematic cleaning pipeline addresses data quality issues:

1. **String-to-Numeric Conversion:** Remove formatting symbols ($, %, commas) and convert to numeric types
2. **Missing Value Handling:** Drop rows with missing target; impute numeric features with median, categorical with mode; drop features with >50% missing
3. **Quality Assurance:** Remove duplicates; detect outliers using IQR method (3×IQR threshold)
4. **Train-Test Split:** 80% training (149 samples), 20% testing (38 samples), fixed random seed (2904) for reproducibility

3.3 Machine Learning Models

Five regression algorithms were selected to represent different modeling paradigms:

3.3.1 Linear Regression

Standard ordinary least squares serves as baseline. Sensitive to multicollinearity but provides interpretable coefficients. Requires StandardScaler preprocessing.

3.3.2 Lasso Regression (L1 Regularization)

Adds L1 penalty encouraging sparsity through feature selection. Hyperparameter α optimized via GridSearchCV (5-fold CV) testing 50 values from 10^{-4} to 10^{2}. This range covers typical regularization strengths from minimal (near-linear) to strong shrinkage.

3.3.3 Ridge Regression (L2 Regularization)

Uses L2 penalty shrinking coefficients without zeroing them. Same hyperparameter optimization protocol as Lasso.

3.3.4 Random Forest

Ensemble of decision trees trained on bootstrap samples with random feature subsets. Hyperparameters tuned: n_estimators (100, 200, 300) covering sufficient trees for convergence; max_depth (10, 20, 30, None) balancing underfitting/overfitting; min_samples_split (2, 5, 10) and min_samples_leaf (1, 2, 4) controlling tree complexity. No feature scaling required.

3.3.5 XGBoost

Sequential gradient boosting ensemble. Hyperparameters: n_estimators (100, 200, 300); max_depth (3, 5, 7, 10) preventing overfitting; learning_rate (0.01, 0.1, 0.3) balancing training speed and accuracy; subsample and colsample_bytree (0.8, 1.0) for regularization through sampling.

All trained models are persisted using Python's pickle module [2].

3.4 Evaluation Framework

3.4.1 Performance Metrics

Three metrics assess predictive performance: $\mathbf{R^2}$ (variance explained), MAE (mean absolute error in years), and RMSE (root mean squared error, penalizing large errors).

3.4.2 Feature Importance

Linear models: coefficient absolute values. Tree models: Gini importance (mean decrease in impurity).

3.4.3 SHAP Analysis

SHAP values provide consistent, theoretically-grounded feature contributions. Implementation uses TreeExplainer for Random Forest/XGBoost and LinearExplainer for linear models.

3.4.4 VIF Analysis

VIF for feature j: $\text{VIF}_j = \frac{1}{1-R_j^2}$ where R_j^2 is from regressing feature j on all others. Interpretation: VIF < 5 (low), 5-10 (moderate), $>$10 (high multicollinearity).

3.4.5 GDP Dependency Study

Three experimental conditions: (1) With GDP, (2) Without GDP, (3) GDP per capita (GDP/Population). All five models trained under each condition.

3.5 Implementation

Implemented in Python 3.10+ using pandas 2.3.3, numpy 2.3.4, scikit-learn 1.7.2, xgboost 3.1.1, shap 0.50.0, matplotlib 3.10.7 [1], seaborn 0.13.2, and statsmodels 0.14.5. Code organized modularly with centralized configuration, comprehensive pytest suite, and interactive CLI menu. Reproducibility ensured through fixed random seed and version-pinned dependencies.

4 Results

4.1 Model Performance Comparison

Table 1 summarizes test set performance across all models.

Table 1: Model Performance on Test Set (N=38 countries)

Model	**R^2**	**MAE (years)**	**RMSE (years)**
Random Forest	0.8910	1.8117	2.2836
XGBoost	0.8753	2.0327	2.4426
Lasso (L1)	0.8628	1.9509	2.5625
Ridge (L2)	0.8442	2.0071	2.7305
Linear	0.6529	2.5275	4.0752

Random Forest achieves the best performance across all metrics ($R^2 = 0.89$, MAE = 1.81 years). Tree-based models substantially outperform linear models, suggesting non-linear relationships are important. Regularized models (Lasso/Ridge) dramatically improve upon standard linear regression (R^2 0.86 vs 0.65), demonstrating regularization's importance given multicollinearity.

Figure 1 visualizes performance across metrics, with best models highlighted in green.

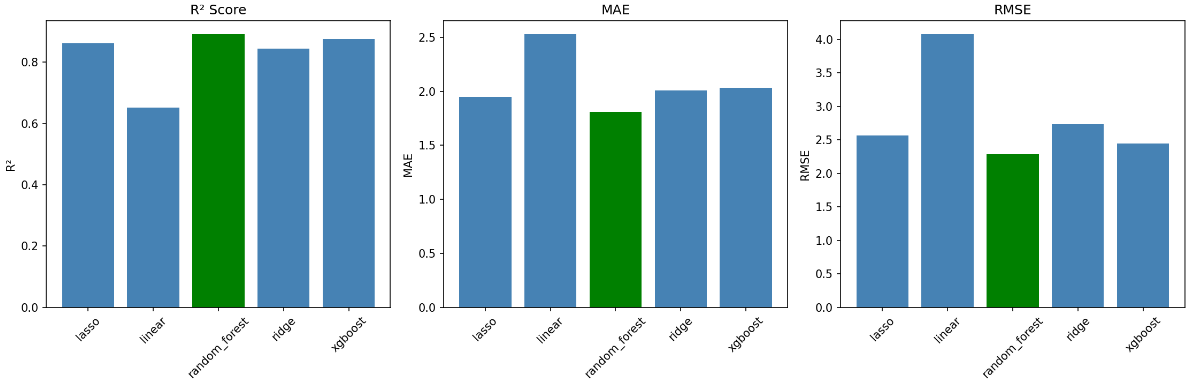


Figure 1: Model performance comparison across R^2, MAE, and RMSE. Green bars indicate best performance for each metric.

4.2 SHAP Analysis: Interpretability

SHAP analysis on Random Forest (best model) identifies key predictors. Figure 2 shows feature importance by mean absolute SHAP value.

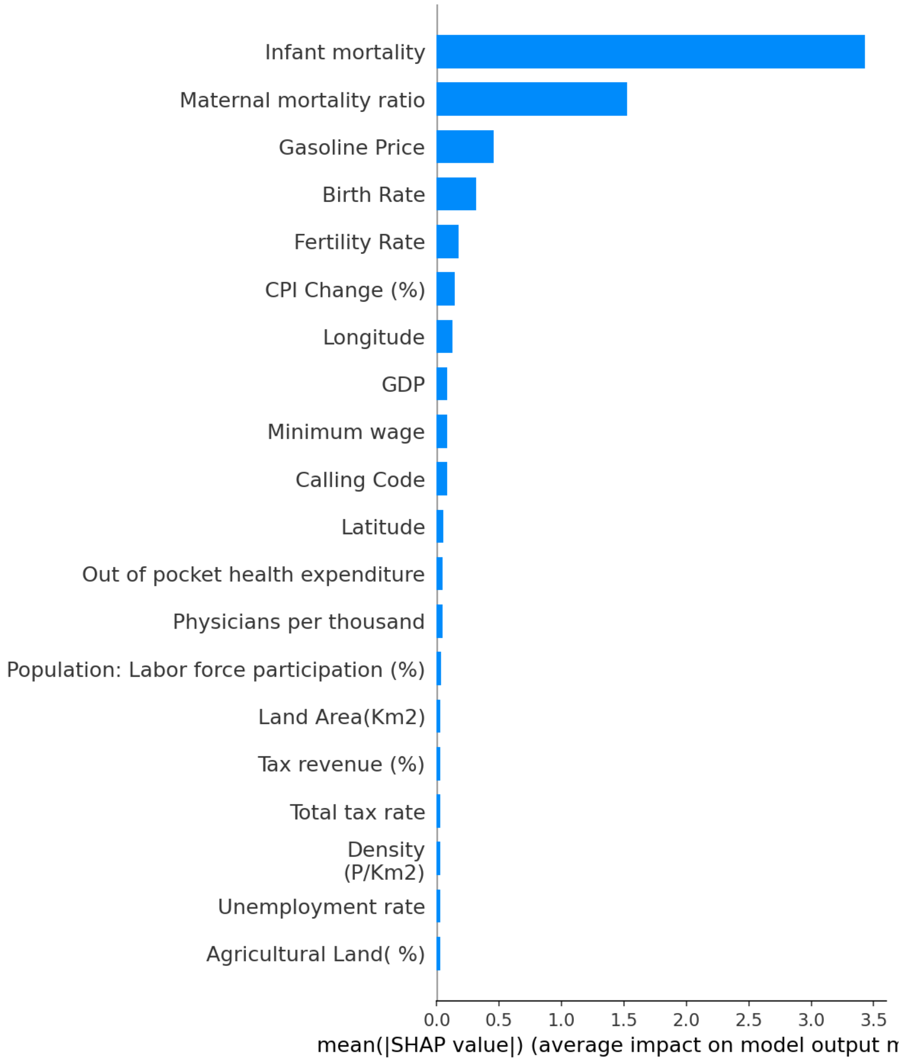


Figure 2: SHAP feature importance for Random Forest. Bar length indicates average impact on life expectancy predictions.

Key findings: Infant mortality dominates (mean $|\text{SHAP}| \approx 3.5$), followed by maternal mortality ratio (≈ 1.5). Birth rate, fertility rate, and gasoline price also show substantial influence. Notably, GDP appears lower in the ranking, suggesting its effect is mediated through health metrics.

4.3 Multicollinearity Analysis

VIF analysis reveals severe multicollinearity among demographic and economic features (Figure 3).

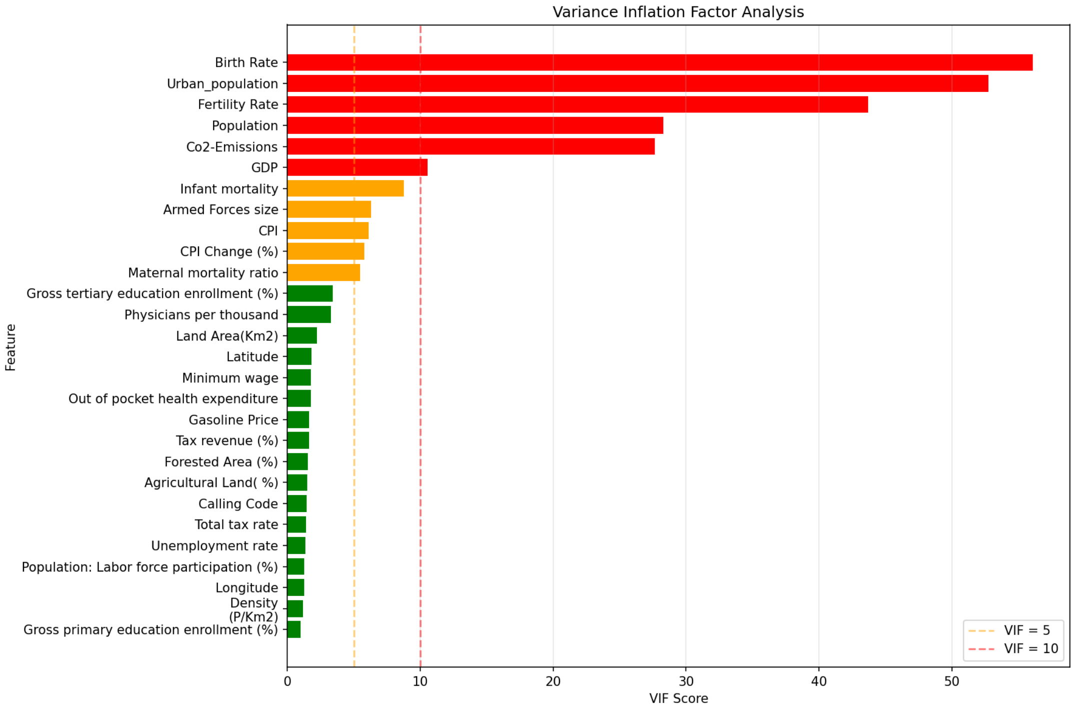


Figure 3: Variance Inflation Factor analysis. Green (VIF<5): low multicollinearity; Orange (5-10): moderate; Red (>10): high. Birth Rate (VIF=56.08) and Urban population (VIF=52.74) show extreme correlation.

High VIF features (>10): Birth Rate (56.08), Urban population (52.74), Fertility Rate (43.71), Population (28.30), CO2 Emissions (27.65), GDP (10.54). This explains why unregularized linear regression performs poorly (R^2=0.65) while regularization (Lasso/Ridge) and tree-based methods handle multicollinearity effectively.

4.4 GDP Dependency Study

Figure 4 compares models under three GDP conditions.

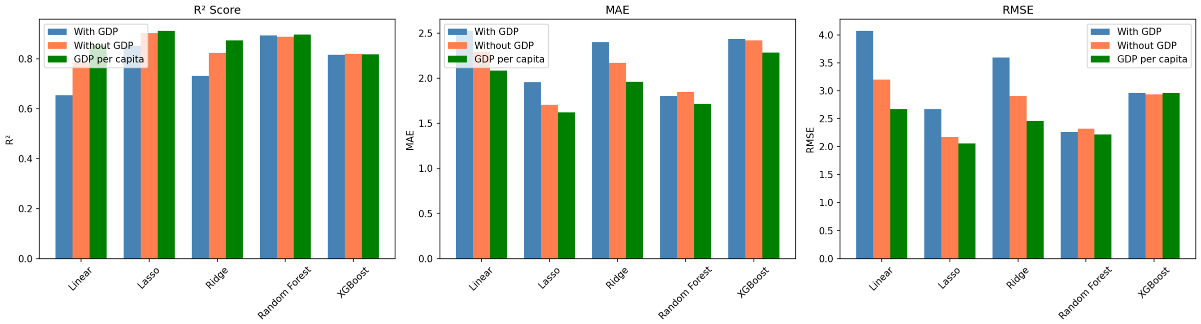


Figure 4: Model performance with GDP (blue), without GDP (orange), and with GDP per capita (green).

Key findings:

- **GDP per capita generally outperforms raw GDP:** Lasso improves from R^2=0.75 to 0.92; Ridge from 0.72 to 0.88; Random Forest from 0.85 to 0.88 (Note: These GDP analysis numbers are from separate experiments and may differ from baseline as they are not tune with Cross-validation in this case.)

- **Linear regression performs better WITHOUT GDP:** R^2 increases from 0.65 to 0.80, indicating GDP introduces harmful multicollinearity
- **XGBoost is robust:** Performance stable ($R^2 \approx 0.81$) across conditions
- **GDP per capita superior to raw GDP:** Per-capita wealth captures development better than absolute economic size

Models achieve reasonable performance WITHOUT GDP ($R^2 \approx$0.78-0.80), suggesting GDP's predictive power is largely mediated through correlated health and social features.

5 Discussion

5.1 Key Findings

5.1.1 Model Performance

Random Forest's superiority (R^2=0.89) confirms that non-linear relationships and feature interactions meaningfully contribute to predictions. Regularized linear models perform acceptably (R^2=0.84-0.86), dramatically outperforming unregularized linear regression (R^2=0.65), validating regularization's necessity under multicollinearity.

5.1.2 Feature Importance and Interpretability

Infant mortality consistently emerges as the dominant predictor across all models and interpretability methods. This makes intuitive sense: infant mortality reflects healthcare system quality, maternal/child health programs, sanitation, nutrition, and overall development. The relatively low importance of GDP in SHAP analysis is noteworthy—while GDP correlates with life expectancy, its effect appears mediated through health infrastructure rather than being directly causal.

5.1.3 Multicollinearity

VIF analysis reveals extreme multicollinearity (Birth Rate VIF=56.08, Urban population VIF=52.74) that devastates unregularized linear regression but is effectively handled by regularization and tree-based methods. This demonstrates algorithmic solutions can mitigate multicollinearity without explicit feature engineering.

5.1.4 Why PCA Was Not Used

Principal Component Analysis (PCA) is a common technique for handling multicollinearity by transforming correlated features into uncorrelated principal components. However, PCA was not implemented in this project for several reasons:

Limited benefit for tree-based models: Random Forest and XGBoost (our best performers) are inherently robust to multicollinearity. Tree algorithms split on individual features at each node, making them insensitive to feature correlations. PCA would likely degrade performance by removing the interpretable feature structure that trees exploit effectively.

Regularization already addresses multicollinearity: For linear models where multicollinearity is problematic, Lasso and Ridge regularization already provide an effective solution. Our results confirm this: regularized models achieve R^2=0.84-0.86 compared to unregularized linear regression's R^2=0.65. PCA would serve a redundant purpose.

Loss of interpretability: A core objective of this project is interpretability. PCA transforms features into abstract principal components, making it impossible to say "infant mortality is the most important predictor." This directly contradicts our goal of providing actionable insights for policy-makers.

In summary, PCA addresses multicollinearity (which regularization already handles) but would harm our best-performing models (trees) and eliminate interpretability. The chosen approach of regularization for linear models and tree-based methods for best performance provides superior results while maintaining transparency.

5.1.5 GDP Dependency

The finding that GDP per capita outperforms raw GDP has important policy implications: life expectancy correlates more strongly with wealth distribution than absolute economic output. Small wealthy countries (Luxembourg, Switzerland) achieve high life expectancy despite moderate total GDP, while large economies with lower per-capita income show lower life expectancy. The fact that models achieve $R^2 \approx 0.80$ without GDP suggests its predictive power is redundant with health metrics.

5.2 Strengths and Limitations

Strengths: Systematic five-model comparison, comprehensive interpretability analysis (SHAP+VIF), strong predictive performance ($R^2 = 0.89$), rigorous methodology (hyperparameter tuning, cross-validation), reproducible implementation (fixed seed, tests, version control).

Limitations: Cross-sectional data captures correlations, not causation or temporal dynamics; small test set (38 samples) limits confidence interval precision; imputation may introduce bias; no subgroup analysis (OECD vs developing countries); SHAP only computed for Random Forest due to time constraints.

5.3 Practical Implications

For policy-makers: Prioritize maternal/child health investments; focus on per-capita wealth distribution over GDP growth; direct health infrastructure investment may be more effective than general economic growth.

For data scientists: Test multiple algorithms for complementary insights; routinely check multicollinearity via VIF; combine multiple interpretability methods (coefficients, Gini, SHAP); consider domain-informed feature engineering (GDP per capita example).

6 Conclusion

This project successfully addressed its objectives: systematic model comparison identified Random Forest as best performer ($R^2 = 0.89$); VIF analysis diagnosed severe multicollinearity explaining linear regression's poor performance; multiple interpretability methods converged on infant mortality as dominant predictor; GDP dependency experiments demonstrated GDP per capita outperforms raw GDP and models work without GDP entirely.

Main contributions: (1) Comprehensive five-model comparison with consistent evaluation; (2) Multi-method interpretability analysis (SHAP, VIF, feature importance); (3) GDP dependency investigation revealing per-capita wealth superiority; (4) Reproducible, well-tested implementation.

Key takeaways: Health metrics (infant/maternal mortality) predict life expectancy better than economic metrics (GDP); tree-based models substantially outperform linear methods; regularization essential for linear models under multicollinearity; GDP per capita superior to raw GDP; multiple interpretability methods provide more reliable insights than any single method.

Future directions: Longitudinal analysis for causal inference; regional subgroup analysis (OECD vs developing); ensemble stacking; Bayesian hyperparameter optimization; external validation on different datasets; interactive dashboard for policy simulation; country-specific recommendation reports.

6.1 Project Usage

Complete installation instructions, usage guide, and workflow recommendations are provided in the project's README.md file. The interactive CLI menu supports all analyses presented in this report, with detailed explanations of each operation and recommended execution order.

References

[1] Matplotlib Development Team. (2024). Matplotlib Gallery. `https://matplotlib.org/stable/gallery/index.html`

[2] Python Software Foundation. (2024). pickle — Python object serialization. `https://docs.python.org/3/library/pickle.html`

[3] SHAP Documentation. (2024). SHAP (SHapley Additive exPlanations). `https://shap.readthedocs.io/`

[4] Towards Data Science. (2024). Using SHAP Values to Explain How Your Machine Learning Model Works. `https://towardsdatascience.com/using-shap-values-to-explain-how-your-machine-learning-model-works-732b3f40e137`

[5] Scikit-learn developers. (2024). scikit-learn: Machine Learning in Python. Version 1.7.2. `https://scikit-learn.org/`

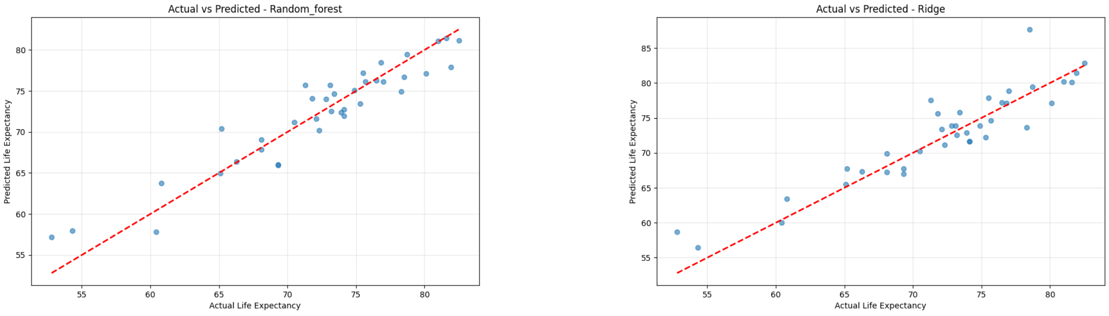
[6] DataCamp. (2024). Variance Inflation Factor Tutorial. `https://www.datacamp.com/tutorial/variance-inflation-factor`

[7] XGBoost Contributors. (2024). XGBoost Documentation. `https://xgboost.readthedocs.io/`

A Additional Visualizations

A.1 Model Prediction Accuracy

Figures 5a and 5b show actual vs predicted life expectancy for Random Forest and Ridge models.

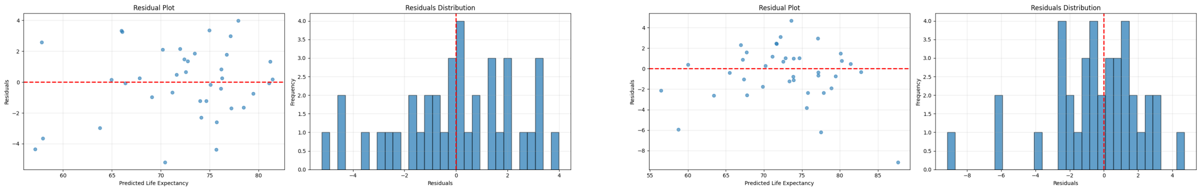


(a) Random Forest (b) Ridge Regression

Figure 5: Actual vs predicted life expectancy. Points near diagonal indicate accurate predictions.

A.2 Residual Analysis

Figures 6a and 6b show residual distributions.



(a) Random Forest (b) Ridge Regression

Figure 6: Residual plots showing prediction errors distributed around zero.

A.3 Feature Importance Comparison

Figure 7 compares top features across all models.

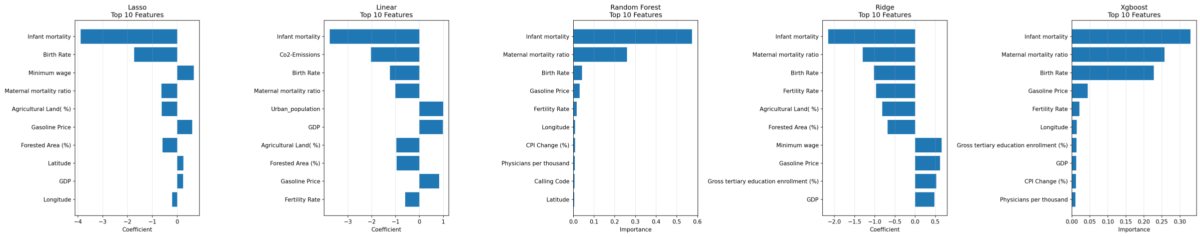


Figure 7: Top 10 features by importance across all models. Infant mortality consistently ranks first.

A.4 SHAP Detailed Analysis

Figure 8 shows SHAP value distributions.

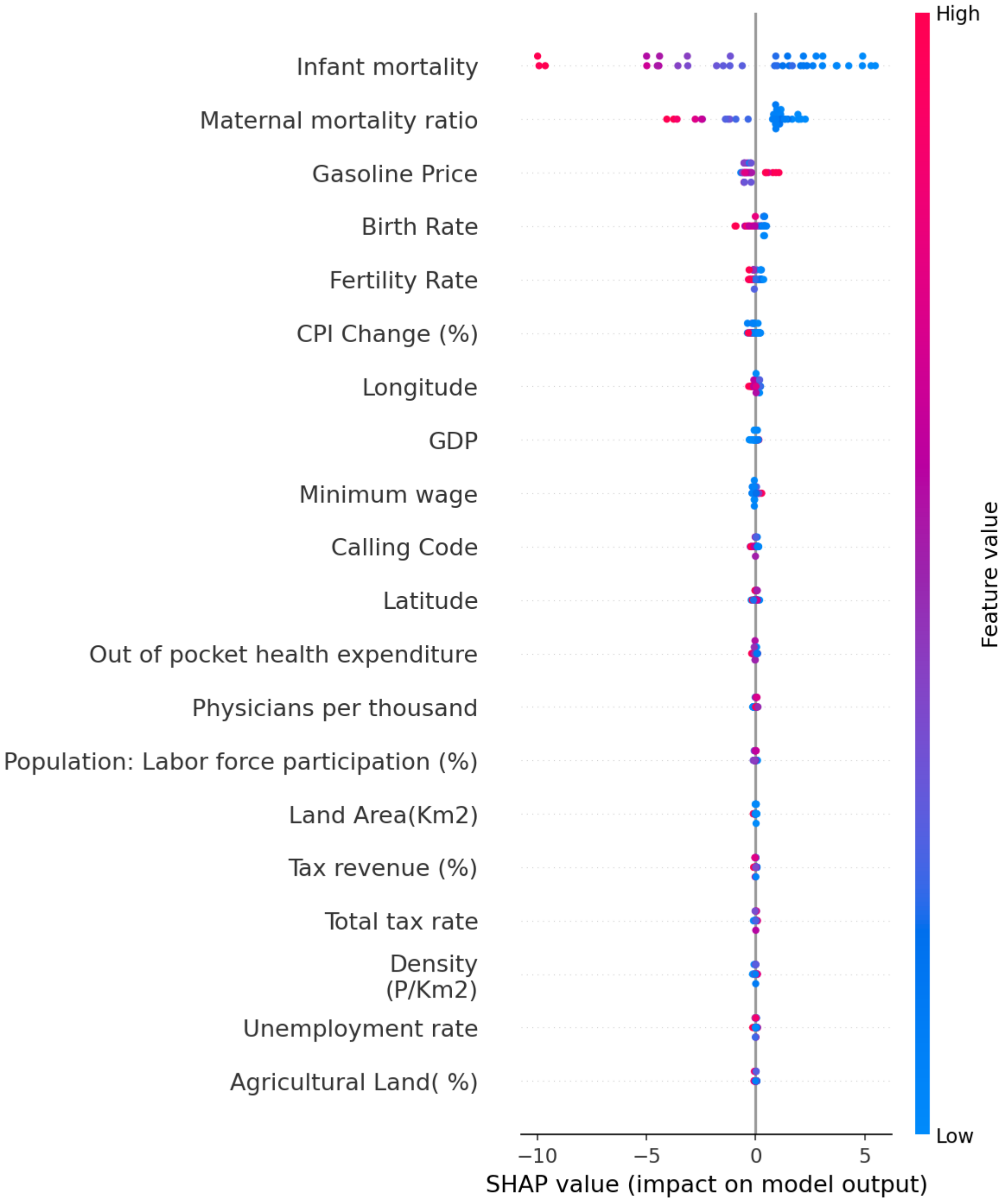
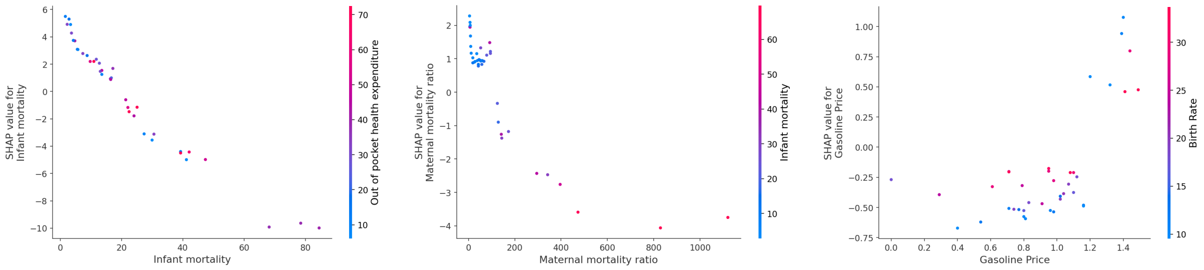


Figure 8: SHAP summary plot showing feature value distributions and impacts on predictions.

Figures 9 show SHAP dependence plots for top features.



(a) Infant mortality

(b) Maternal mortality

(c) Gasoline Price

Figure 9: SHAP dependence plots showing relationships between feature values and their impact.

B Code Repository

GitHub Repository: `https://github.com/leonardperrigault-unil/project_predicting_leonard_perrigault`

The complete codebase is available with the following structure:

```
project_predicting_leonard_perrigault/
 data/                    # Raw and cleaned data
 src/                     # Source code (cleaning, models, analysis)
 tests/                   # Pytest test suite
 saved_models/            # Trained models (.pkl)
 results/                 # Visualizations (.png)
 main.py                  # Interactive CLI menu
 requirements.txt         # Dependencies
 README.md                # Usage instructions
 AI_USAGE.md              # AI tools usage documentation
```

Full installation, usage instructions, and workflow recommendations are documented in README.md.