Life Expectancy Analysis: Comprehensive Data Science Project Report

Table of Contents

Executive Summary

This comprehensive analysis examined global life expectancy patterns from 2000-2015 across 193 countries using advanced machine learning techniques and statistical analysis. The study successfully identified key factors influencing life expectancy and developed predictive models achieving 96.88% accuracy. The research provides actionable insights for policymakers and health organizations to improve global health outcomes.

Key Findings

- Global life expectancy improved by 4.9 years from 2000-2015
- HIV/AIDS deaths emerged as the strongest predictor (59.4% model importance)
- Education (schooling) shows the highest positive correlation (0.715)
- Significant gaps persist between developed (80.7 years) and developing countries (69.7 years)
- Random Forest model achieved superior performance with $R^2 = 0.9688$

Introduction and Objectives

Life expectancy serves as a fundamental indicator of population health and socioeconomic development. Despite significant global health improvements, substantial disparities persist between nations. This project addresses eight critical research questions:

1. Which predicting factors truly affect life expectancy?
2. Should countries with low life expectancy increase healthcare expenditure?
3. How do infant and adult mortality rates impact life expectancy?
4. What is the relationship between lifestyle factors and life expectancy?
5. What is the impact of schooling on human lifespan?

6. Does alcohol consumption correlate with life expectancy?

7. Do densely populated countries have lower life expectancy?

8. What is the impact of immunization coverage on life expectancy?

Dataset Overview

The analysis utilized WHO Global Health Observatory data covering:

- **Countries**: 193 nations
- **Time Period**: 2000-2015 (16 years)
- **Total Observations**: 2,938 country-year records
- **Variables**: 22 indicators including health, economic, and social factors
- **Country Status**: 161 developing countries, 32 developed countries

Data Quality Assessment

Initial data examination revealed missing values in key variables:

- Population: 652 missing values (22.2%)
- Hepatitis B: 553 missing values (18.8%)
- GDP: 448 missing values (15.3%)
- Total expenditure: 226 missing values (7.7%)

Missing values were addressed using mean imputation for numerical variables, ensuring data integrity while preserving sample size.

Exploratory Data Analysis

Global Life Expectancy Trends

[^4]

The analysis revealed consistent improvement in global life expectancy over the study period:

- **2000**: 66.75 years (global average)
- **2015**: 71.62 years (global average)
- **Total Improvement**: 4.9 years over 15 years

Development Status Disparities

[^8]

Significant disparities exist between country development classifications:

- **Developed Countries (2015)**: 80.71 years average
- **Developing Countries (2015)**: 69.69 years average
- **Gap**: 11.02 years (narrowed from 12.18 years in 2000)

Factor Correlation Analysis

[^6]

Comprehensive correlation analysis identified the strongest predictors of life expectancy:

Positive Correlations:

- Schooling: 0.715 (strongest positive relationship)
- Income composition of resources: 0.692
- BMI: 0.559
- GDP: 0.430

Negative Correlations:

- Adult Mortality: -0.696 (strongest negative relationship)
- HIV/AIDS deaths: -0.556
- Thinness (malnutrition indicators): -0.467 to -0.472

Machine Learning Analysis

Model Development and Performance

Two primary models were developed and evaluated:

Random Forest Regressor:

- R² Score: 0.9688
- RMSE: 1.6447
- MAE: 1.0581

Linear Regression:

- R² Score: 0.8241
- RMSE: 3.9042
- MAE: 2.8583

The Random Forest model demonstrated superior performance, capturing complex non-linear relationships between predictors and life expectancy.

Feature Importance Analysis

[^11]

Machine learning analysis revealed the relative importance of factors:

1. **HIV/AIDS deaths**: 59.40% (dominant factor)
2. **Adult Mortality**: 15.97%
3. **Income composition of resources**: 14.31%
4. **Schooling**: 1.87%
5. **BMI**: 1.40%

The overwhelming importance of HIV/AIDS deaths highlights the devastating impact of this epidemic on global life expectancy, particularly in sub-Saharan Africa.

Country-Specific Analysis

Top Performing Countries (2015)

[^13]

Highest Life Expectancy:

1. Slovenia: 88.0 years

2. Denmark: 86.0 years

3. Chile: 85.0 years

4. Cyprus: 85.0 years

5. Japan: 83.7 years

Lowest Performing Countries (2015)

Lowest Life Expectancy:

1. Sierra Leone: 51.0 years

2. Angola: 52.4 years

3. Central African Republic: 52.5 years

4. Chad: 53.1 years

5. Côte d'Ivoire: 53.3 years

The 37-year gap between highest and lowest performers underscores the urgent need for targeted interventions in low-performing regions.

Research Question Analysis

1. Key Predicting Factors

The machine learning analysis definitively established HIV/AIDS deaths as the most critical factor, followed by adult mortality and income composition. This emphasizes the need for comprehensive HIV prevention and treatment programs.

2. Healthcare Expenditure Impact

Countries with life expectancy below 65 years spent an average of 72.75% on healthcare expenditure, while those above 65 years spent 1,009.62%. This suggests that strategic healthcare investment can significantly improve population health outcomes.

3. Mortality Rates Impact

- **Adult Mortality**: Strong negative correlation (-0.696)
- **Infant Deaths**: Moderate negative correlation (-0.197)

Adult mortality showed the stronger relationship, indicating that improvements in adult healthcare and prevention programs yield substantial life expectancy gains.

4. Lifestyle Factors

- **BMI**: Moderate positive correlation (0.559)
- **Alcohol**: Weak positive correlation (0.392)

Proper nutrition (reflected in healthy BMI) shows stronger association with longevity than alcohol consumption patterns.

5. Education Impact

Schooling demonstrated the strongest positive correlation (0.715), reinforcing education's role as a fundamental determinant of health outcomes through improved health literacy and economic opportunities.

6. Alcohol Consumption

Alcohol showed a weak positive correlation (0.392), likely reflecting the complex relationship between moderate consumption in developed countries versus harmful use patterns globally.

7. Population Density

Population size showed minimal correlation (-0.020), indicating that country size is not a significant predictor of life expectancy outcomes.

8. Immunization Impact

Immunization programs showed moderate positive correlations:

- Diphtheria: 0.475
- Polio: 0.462
- Hepatitis B: 0.204

Global Health Disparities

The analysis reveals persistent and significant health disparities:

- **Geographic Concentration**: Lowest-performing countries are predominantly in sub-Saharan Africa
- **Development Gap**: 11-year difference between developed and developing nations
- **HIV/AIDS Impact**: Disproportionate effect on African nations with high HIV prevalence
- **Economic Factors**: Strong correlation between income composition and life expectancy

Model Validation and Robustness

Cross-validation analysis confirmed model reliability:

- 20-fold cross-validation mean R^2: 0.962
- Standard deviation: 0.0096
- Consistent performance across different data splits

Policy Implications and Recommendations

Priority Interventions

1. HIV/AIDS Combat Programs

- Implement comprehensive prevention strategies
- Expand access to antiretroviral therapy
- Strengthen testing and treatment infrastructure
- Target high-prevalence regions with intensive interventions

2. Healthcare System Strengthening

- Increase healthcare expenditure in low-performing countries
- Develop primary healthcare access
- Implement preventive care programs
- Strengthen maternal and child health services

3. Education Investment

- Expand educational access and quality
- Develop health literacy programs
- Support girls' education initiatives
- Integrate health education in curricula

4. Economic Development

- Address income inequality
- Develop social safety nets
- Support economic diversification
- Promote sustainable development

Monitoring and Evaluation Framework

Key Performance Indicators:

- Annual life expectancy improvements
- Reduction in development status gap
- HIV/AIDS mortality trends
- Educational attainment rates
- Healthcare expenditure efficiency

Target Setting:

- Reduce developed-developing gap to 8 years by 2030
- Achieve minimum 60-year life expectancy globally
- Eliminate countries below 55-year life expectancy
- Improve bottom decile countries by 5 years

Study Limitations and Future Research

Limitations

- Data availability constraints for some countries
- Time period limited to 2000-2015
- Potential reporting variations between countries
- Missing values addressed through imputation

Future Research Directions

- Extend analysis to post-2015 data including COVID-19 impact
- Investigate regional and cultural factors
- Develop country-specific prediction models
- Analyze intervention effectiveness through natural experiments

Conclusions

This comprehensive analysis successfully identified the key determinants of global life expectancy and developed highly accurate predictive models. The findings provide clear evidence for policy priorities:

Critical Success Factors:

1. HIV/AIDS prevention and treatment (highest impact)
2. Adult mortality reduction through healthcare improvement
3. Educational investment for long-term health gains
4. Economic development and income equality

Global Progress:

- Significant improvements achieved (4.9 years globally)
- Developing countries showing faster improvement rates
- Persistent gaps requiring targeted interventions

Model Performance:

- Random Forest achieved 96.88% accuracy
- Robust cross-validation results
- Clear feature importance rankings

The study demonstrates that while progress has been made, substantial opportunities exist for further improvements through evidence-based policy interventions. Countries can use these findings to prioritize resource allocation and develop targeted strategies for improving population health outcomes.

The machine learning approach successfully captured complex relationships between health determinants, providing a robust framework for understanding and predicting life expectancy patterns. This methodology can be extended to support ongoing health policy development and international cooperation efforts.

Data Sources and References

1. World Health Organization Global Health Observatory Data Repository

2. United Nations Development Programme Human Development Reports

3. World Bank Development Indicators

4. GitHub Repository: kiranshahi/Life-expectancy-prediction

5. Kaggle Dataset: Life Expectancy (WHO) Fixed

This report represents a comprehensive analysis of global life expectancy determinants using advanced data science techniques. The findings provide evidence-based recommendations for improving global health outcomes through targeted policy interventions.
[1]

⁂

1. Life-Expectancy-Analysis-Analysis-Machine-Learning-Project-Data-Analyst.pdf