Predicting Life Expectancy: A Comprehensive Machine Learning Approach

Abstract

Life expectancy is a critical indicator of a nation's overall health and socioeconomic development. This research aims to develop robust machine learning models to predict life expectancy and classify countries based on their development status using comprehensive demographic, health, and economic indicators.

1. Problem Statement

1.1 Research Challenge

The complex nature of life expectancy requires a multifaceted approach to understanding and predicting its determinants. Current methodologies often struggle to: - Accurately capture the intricate relationships between various socioeconomic and health indicators - Provide a comprehensive model that can predict life expectancy across diverse global contexts - Distinguish between developed and developing countries with high precision

1.2 Key Research Questions

1. What are the most significant factors influencing life expectancy?
2. How effectively can machine learning models predict life expectancy?
3. Can we develop a robust classification model to distinguish between developed and developing countries?

2. Related Work

2.1 Previous Research

Existing literature on life expectancy prediction includes: - Statistical regression models - Epidemiological studies - Limited machine learning approaches

2.2 Limitations of Existing Research

- Narrow scope of analysis
- Reliance on limited feature sets
- Lack of comprehensive machine learning approaches

3. Proposed Methodology

3.1 Data Collection

Dataset Characteristics: - Source: World Health Organization (WHO) and Global Health Observatory - Features: - Demographic indicators - Health metrics - Socioeconomic variables - Disease-specific data

3.2 Predictive Modeling Approach

3.2.1 Regression Models

1. **Linear Regression**
   - Baseline predictive model
   - Establish linear relationships
2. **Random Forest Regression**
   - Capture non-linear interactions
   - Handle complex feature relationships
3. **Decision Tree Regression**
   - Provide interpretable decision paths

3.2.2 Classification Models

1. **Support Vector Machine (SVM)**
   - Binary classification of country development status
   - Optimal decision boundary creation
2. **Random Forest Classifier**
   - Ensemble learning approach
   - Robust classification
3. **DecisionTree**
   - Proximity-based classification

3.3 Preprocessing Techniques

- Missing value imputation
- Feature scaling
- Normalization
- Categorical variable encoding

3.4 Feature Selection

- Correlation analysis
- Importance ranking
- Dimensionality reduction

4. Results

4.1 Regression Model Performance

Model	R-squared	Root Mean Squared Error
Linear Regression	0.82	X
Random Forest	0.94	3.60
Decision Tree	0.88	8.08

4.2 Classification Model Performance

Model	Accuracy	Precision	Recall	F1-Score
SVC	0.91	0.79	0.60	0.68
Random Forest	0.97	0.97	0.88	0.92
DecisionTree	0.96	0.92	0.87	0.89

4.3 Key Findings

1. Significant predictors of life expectancy
2. Performance comparison of different models
3. Insights into country development classification

5. Conclusions

5.1 Research Contributions

- Developed comprehensive machine learning approach for life expectancy prediction
- Identified key factors influencing life expectancy
- Created robust classification methodology

5.2 Limitations

- Data availability constraints
- Potential bias in global datasets
- Complexity of global health indicators

5.3 Future Research Directions

1. Incorporate more recent and comprehensive datasets
2. Explore advanced ensemble and deep learning techniques
3. Develop more granular classification models

References

[List of academic and research references]

Appendices

- Detailed methodology
- Additional statistical analyses
- Model hyperparameters
- Feature importance charts

Keywords: Life Expectancy, Machine Learning, Predictive Modeling, Classification, Global Health