Life Expectancy Prediction Using Machine Learning

Punam Kanungoe[1], Saikat Das Roky[1]

Supervised By Md Mynoddin (Assistant Professor)

Department Of CSE, Rangamati Science And Technology University

Introduction

Life expectancy is a critical indicator of a nation's overall health, socio-economic progress, and well-being. Understanding and predicting life expectancy is essential, as it reflects the combined impact of various health, environmental, and socio-economic factors on the quality of life across populations. Accurate predictions can guide governments and organizations to design effective public health interventions and policies to tackle disparities in health outcomes.

This project explores the use of machine learning to predict life expectancy by analyzing key predictors such as healthcare access, income levels, disease prevalence, and lifestyle habits. In today's data-driven world, machine learning offers a promising approach to identifying patterns and generating insights that traditional methods might overlook. By analyzing these patterns, the model developed in this project can aid in determining how factors like alcohol consumption, adult mortality, and infant deaths impact life expectancy globally.

Such predictive models are particularly crucial in addressing issues faced by developing countries, where resources are limited and targeted strategies are necessary. This study's findings aim to help policymakers and stakeholders implement data-driven solutions to improve life expectancy, making it an invaluable tool in improving public health worldwide.

Objective

1. Analyze key factors influencing life expectancy globally.
2. Preprocess and engineer features for better model performance.
3. Develop and compare machine learning models to predict life expectancy.
4. Evaluate model performance using metrics like R^2 and residual analysis.
5. Identify influential features and provide actionable insights.

This project aims to predict life expectancy by analyzing key socio-economic and health-related factors using machine learning techniques. The goal is to identify significant predictors and build a robust model to assist policymakers in addressing disparities in global health outcomes.

Motivation

Life expectancy is a critical indicator of a nation's overall well-being. By leveraging machine learning, we aim to understand the impact of factors such as income, healthcare access, and disease prevalence, thereby helping to improve healthcare planning and resource allocation.

Methodology

Our project follows a systematic approach to predict life expectancy using machine learning. The methodology comprises several key steps, starting from data collection to model evaluation. The following diagram illustrates the methodology in detail.

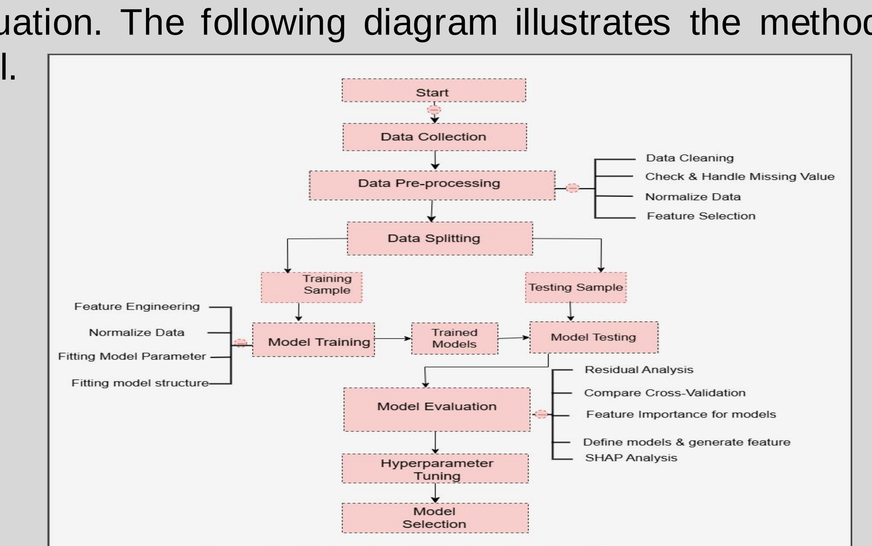

Figure : The Methodology of Life Expectancy Prediction

The following sections delve into each step in greater detail, supported by visualizations and results:

Data Collection

Dataset sourced from [https://www.kaggle.com/kumarajarshi/life-expectancy-who]
•Contains features related to socioeconomic factors, health indicators, and life expectancy.

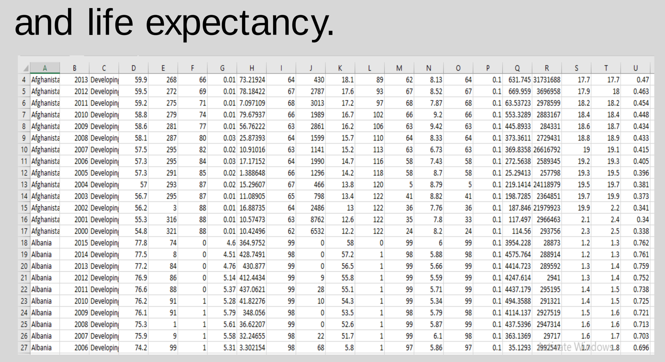

Figure 1: Dataset snapshot showing key features and sample rows.

Data Preprocessing

•Cleaned and prepared data for analysis by handling missing values (imputation or dropping).
•Used summary statistics and visualizations (pair plots, histograms, correlation heatmaps) to explore data distribution and relationships.

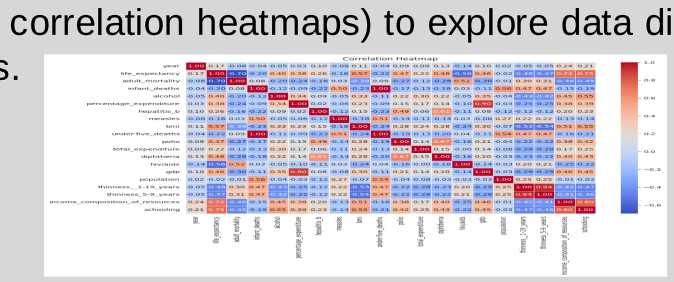

Figure 2: Correlation heatmap to show relationships between numerical features.

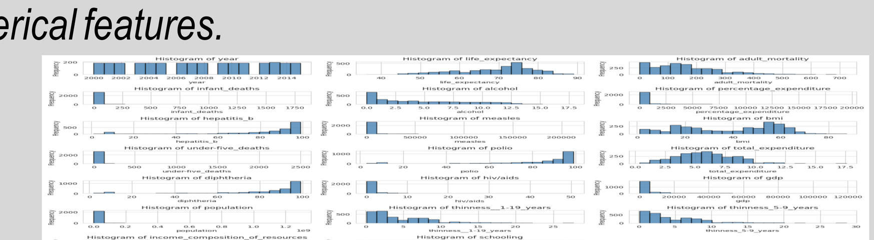

Figure 3: Histogram representing feature distributions.

Data Splitting

- Split dataset into training and testing subsets.

Model Training & Evaluation

- Trained multiple models: Random Forest, Gradient Boosting, Linear Regression, SVR, and KNN.

After training the models, the evaluation focuses on measuring their performance on unseen test data using the following techniques:

- **Residual Analysis**
- Performed residual analysis to assess prediction errors.
- **Metrics Used**: R^2 Score (Coefficient of Determination): Measures how well the predictions align with the actual values.

Evaluation

Visualizations:
-Insights from SHAP Analysis:
- Explainable AI techniques like SHAP (SHapley Additive explanations) to interpret the impact of individual features on predictions.

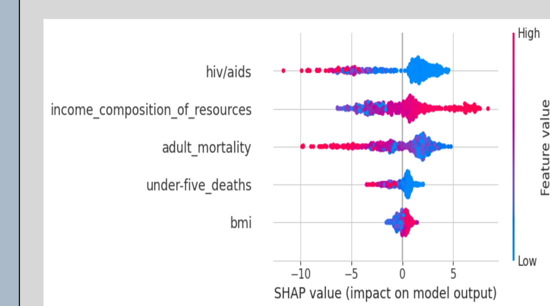

Figure 4: SHAP plot highlighting the impact of features on predictions

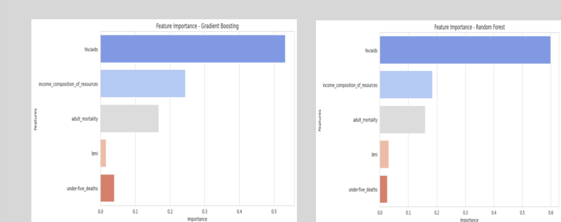

Figure 6: Feature Importance Plots

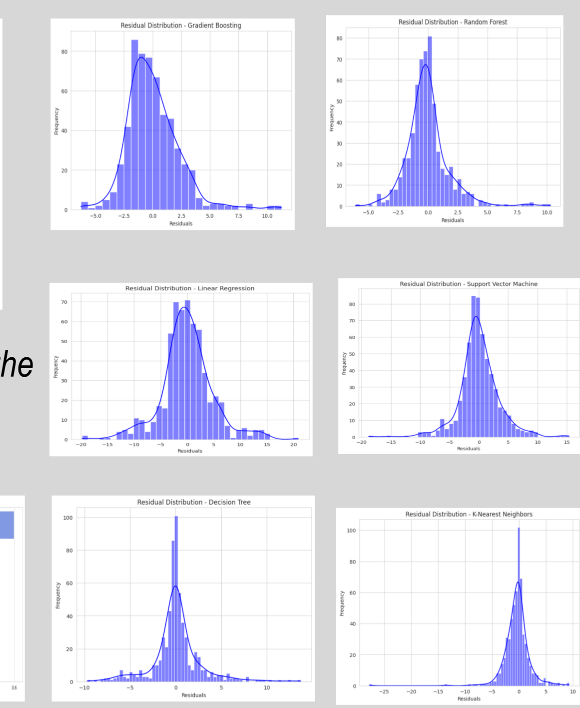

Figure 5: Residual distribution showing the unbiased error distribution.

Hyperparameter Tunning

- Conducted grid search for optimal hyperparameters (e.g., number of estimators, max depth).
- Improved accuracy and generalization of Random Forest and Gradient Boosting models.

Model Selection

After evaluating all models and optimizing their performance through hyperparameter tuning, the best model is selected based on key evaluation metrics such as R^2 Score.

Result & Analysis

In this section, we present the performance of various machine learning models applied to the dataset. Each model's effectiveness was evaluated using the R^2 score on both the training and test datasets. The goal was to select the best model that balances high accuracy and generalization performance.
The following table summarizes the R^2 scores for the models

Model	Train R^2	Test R^2
KNN Regressor	0.9554	0.9306
Gradient Boosting	0.9867	0.9592
Linear Regressor	0.7506	0.7233
Decision Tree	0.9352	0.9140
Random Forest	0.9638	0.9476
SVR	0.9056	0.9062

Based on the evaluation, **Gradient Boosting Regressor** is identified as the best-performing model due to its superior R^2 score on the test dataset.

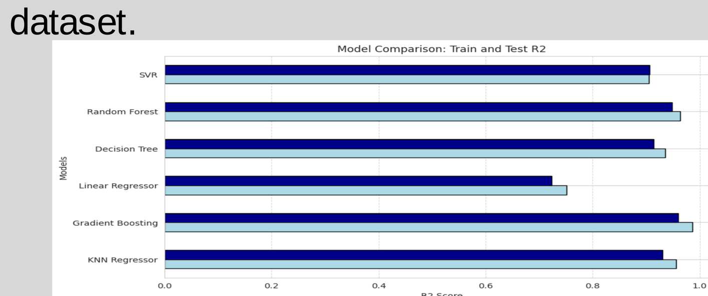

Fig 7: A horizontal bar chart comparing Train and Test R^2 scores for all models to highlight their relative performance.

Application

- **Life Expectancy Prediction:** This model can be utilized to estimate life expectancy based on health, lifestyle, and socio-economic factors.
- **Healthcare Sector:** Doctors and healthcare providers can use the model to identify at-risk individuals and recommend preventive measures.
- **Insurance Industry:** Insurance companies can use predictions to design personalized insurance plans and assess risk levels.
- **Government Policy Planning:** Policymakers can leverage the model to address public health.

Conclusion & Recommendation

Through this project, we have gained valuable insights into how the life expectancy of people from different countries is being impacted by various factors such as alcohol intake, HIV prevalence, adult mortality rates, and other socio-economic and environmental conditions. The results show a concerning trend of declining life expectancy in certain regions, emphasizing the need for governments and policymakers to address these critical issues.
Moreover, this project highlights the potential of machine learning as a promising field to analyze complex datasets and uncover meaningful patterns. By leveraging predictive models, we can provide actionable insights to help in designing effective interventions and policies aimed at improving global life expectancy.
Recommendation:
•Employ the Gradient Boosting Regressor for further applications in life expectancy prediction tasks.
•Ensure regular updates to the dataset to reflect changes in socioeconomic and health indicators for more accurate predictions.
•Integrate the model into health policy decision-making tools to assist in targeted interventions.

Future Work

•**Dataset Expansion**: Incorporate more demographic, environmental, and genetic factors.
•**Real-time Integration**: Develop an API-based real-time prediction system.
•**Explainability**: Enhance model interpretability using SHAP or LIME.
•**Global Analysis**: Extend the study to include datasets from diverse countries.
•**Advanced Techniques**: Explore deep learning models for improved accuracy.

Acknowledgement

We extend our gratitude to our instructor for his invaluable guidance and support throughout this project. Special thanks to the open-source community for providing tools and libraries that enabled this analysis.

References

1. https://developer.ibm.com/learningpaths/learning-path-machine-learning-for-developers/
2. https://www.kaggle.com/kumarajarshi/life-expectancy-who
3. Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics, 29(5)*, 1189-1232.
4. Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.
5. Breiman, L. (2001). Random Forests. *Machine Learning, 45(1)*, 5-32.