Project Overview

This project aims to accurately predict life expectancy at birth (Age 0) across various countries using historical data. Life expectancy is a crucial indicator of public health, socio-economic status, and overall quality of life, making it valuable for policy-making and planning.

Data Description

We utilize a comprehensive dataset that includes life expectancy values at different ages for numerous countries. For this project, we focus specifically on predicting life expectancy at birth (Age 0). The key variables include:

- **Year**: The year for each life expectancy measurement.
- **Entity (Country)**: The country for each life expectancy data point is encoded numerically for model compatibility.
- **Life Expectancy at Birth**: The target variable represents the life expectancy at birth for both sexes combined.

Methodology

Data Preparation

1. **Encoding**: We apply label encoding to the Entity column to convert country names into numerical format, ensuring compatibility with machine learning algorithms.
2. **Train-Test Split**: To evaluate the model's generalizability, we divide the dataset into training (80model'stesting (20%) sets.

Model Selection and Training

1. **Baseline Model**: To establish a baseline, we start with a linear regression model. However, the model's MAE and RMSE indicate that it needs to be more complex to capture relationships in the data accurately.
2. **Complex Model**: We implement a Random Forest Regressor due to its ability to handle non-linear relationships and its robustness in structured data.
3.

Hyperparameter Tuning

To maximize model accuracy, we conduct hyperparameter tuning using Grid Search Cross-Validation, testing different combinations of parameters:

- **n_estimators**: Number of trees in the forest.
- **max_depth**: Maximum depth of each tree.
- **min_samples_split**: Minimum samples required to split a node.
- **min_samples_leaf**: Minimum samples required to form a leaf node.

This process identifies the optimal parameters:

- max_depth: None
- min_samples_leaf: 1

- min_samples_split: 2
- n_estimators: 300

Results and Evaluation

The tuned Random Forest model achieves the following metrics:

- **Mean Absolute Error (MAE)**: 0.81 years
- **Root Mean Squared Error (RMSE)**: 1.56 years

These metrics indicate that the model makes accurate predictions, with an average deviation of approximately 0.81 years from the actual life expectancy values.

Visual Validation

The predicted vs. actual plot for the test data shows a strong alignment along the ideal prediction line, with points clustering closely around it. This alignment confirms the model's accuracy and consistency across life expectancy values.

Model Deployment

We save the tuned model for future applications as tuned_random_forest_model.joblib for easy reusability in production environments. The model is ready for deployment in systems requiring real-time life expectancy predictions, such as policy planning tools or public health applications.

Conclusion

This project successfully builds an accurate predictive model for life expectancy at birth, leveraging a Random Forest Regressor with optimized hyperparameters. The model demonstrates high accuracy, and its robust performance on test data confirms its suitability for practical applications. This model enables stakeholders to make informed healthcare, social services, and economic planning decisions based on reliable life expectancy predictions.

Recommendations for Future Work

1. **Feature Expansion**: To enhance model performance, incorporate additional socio-economic or health-related features (e.g., GDP, healthcare spending).
2. **Testing Advanced Models**: Experiment with other ensemble methods like XGBoost to explore potential accuracy improvements.
3. **Deployment and Integration**: Deploy the model within an API framework (e.g., Flask, FastAPI) for integration into systems requiring automated life expectancy predictions.

Appendix: Key Code and Hyperparameter Details

- **Data Preparation**: Encoding, train-test split
- **Model Initialization**: Random Forest Regressor with optimal hyperparameters
- **Evaluation Metrics**: MAE, RMSE, predicted vs. actual plot