# README – Life Expectancy Prediction using XGBoost

This folder contains the code and the trained model (.pkl file) for predicting life expectancy using an XGBoost regression pipeline. The model was trained on a custom dataset constructed using MySQL and scientific references.

## 📦 Requirements

To run this code, install the following libraries:  
- pandas  
- numpy  
- scikit-learn  
- xgboost  
- joblib  
- scikit-optimize (for Bayesian optimization)  
  
You can install them using:  
pip install pandas numpy scikit-learn xgboost joblib scikit-optimize

## 📌 Steps Included in the Notebook

### 1. Loading the Data

• Imported and preprocessed the custom dataset from MySQL using Pandas.  
• Handled missing values, standardized formats, and validated data integrity.

### 2. Feature & Target Extraction

• Features (X): All lifestyle and health-related factors including:  
 - Age, Gender, Country  
 - Exercise Hours, Sleep Hours, BMI  
 - Diet Type, Medical History, Smoking, Alcohol, Stress, Social Life  
• Target (y): Life Expectancy (in years)

### 3. Splitting the Data

• Training Set: 80%  
• Testing Set: 20%

### 4. Feature Engineering

• Categorical Features (e.g., Gender, Country, Diet Type): One-Hot Encoded  
• Numerical Features (e.g., BMI, Sleep Hours, Exercise): Standard Scaled using StandardScaler

### 5. Pipeline Construction

• Step 1: One-Hot Encoding for categorical variables  
• Step 2: Standard Scaling for numerical variables  
• Step 3: XGBoost Regressor for predicting life expectancy  
  
Why use a pipeline?  
✓ Automates data preprocessing and modeling steps  
✓ Ensures reproducibility and robustness for deployment

### 6. Model Training

• Hyperparameter tuning was done using Bayesian Optimization via BayesSearchCV  
• Final model was trained using the best parameters on the full training set

### 7. Evaluation

• Model performance was assessed using R² score on the test set  
• Achieved an accuracy score (R²) of 0.89

### 8. Feature Importance

• XGBoost’s inbuilt feature importance was used to identify key contributors:  
 - Top factors included Smoking, Exercise Hours, BMI, and Stress Level

### 9. Saving the Model

• The final model pipeline was saved using joblib as 'life\_expectancy\_model.pkl'  
• This allows easy reuse for prediction and deployment

## 🗂 Included Files

• xgboost\_model\_code.ipynb – Jupyter Notebook with the entire model pipeline  
• life\_expectancy\_model.pkl – Trained model file for inference