Assignment: End-to-End Machine Learning Pipeline

Project Summary: Life Expectancy Classification Analysis

1. Data Overview

This project utilizes a dataset from the Global Health Observatory (GHO) by the World Health Organization (WHO), focusing on life expectancy and key social, economic, and health indicators across 193 countries from 2000 to 2015. After merging WHO and UN sources, the dataset contained 2,938 records and 22 features.

To maintain data integrity, rows with missing values in critical columns like Hepatitis B, GDP, and Population were dropped. These were mostly from smaller or less documented countries such as Vanuatu, Tonga, and Cabo Verde.

Since the task was classification, the continuous target variable Life Expectancy was converted into three categorical classes:

- Low
- Medium
- High

These classes were then label-encoded into numerical values. Predictor variables included metrics such as GDP, Schooling, Adult Mortality, and immunization coverage, ensuring the dataset was well-structured for model training.

2. Key Insights from Data Visualization

- A bar chart highlighted that developed nations (e.g., Japan, Switzerland) had the highest life expectancy, while developing countries scored significantly lower.
- A correlation heatmap revealed:
    - Positive correlations between life expectancy and factors like GDP and Schooling.
    - Negative correlations with Adult Mortality and Infant Deaths.

3. Model Evaluation & Comparison

Three machine learning models were trained and evaluated on the preprocessed data:

- K-Nearest Neighbors (KNN)
- Decision Tree
- Random Forest

Observations:

- KNN showed poor performance due to sensitivity to feature scales and noise in the data.
- Decision Tree achieved better accuracy but had a higher chance of overfitting.
- Random Forest provided the best results by reducing overfitting and handling complex patterns more effectively.

Performance Summary:

Model	Accuracy	Precision	Recall	F1 Score	Testing Accuracy
KNN	0.59	0.58	0.59	0.58	0.59
Decision Tree	0.90	0.90	0.90	0.90	0.90
Random Forest	0.92	0.93	0.92	0.92	0.92

4. Conclusions & Insights

- The Random Forest model achieved the highest accuracy (92%), making it the most effective and reliable for this classification task.
- Important features that influenced life expectancy predictions included:
    - Income Composition of Resources
    - Adult Mortality
    - HIV/AIDS
    - Schooling
    - GDP
- Hyperparameter tuning, done using RandomizedSearchCV, significantly boosted model performance, especially for Random Forest by finding optimal settings like tree depth and number of estimators.