Assignment: End-to-End Machine Learning Pipeline

1. Data Insights:

The Global Health Observatory (GHO) of the World Health Organization (WHO) provides health-related datasets for research and analysis. This project uses life expectancy and related socio-economic and health factors for 193 countries, collected between 2000–2015. The dataset was merged from WHO and UN sources, consisting of 2938 rows and 22 columns.

Initially, the dataset had some missing values for features like Hepatitis B, GDP, and population, mostly from smaller countries (e.g., Vanuatu, Tonga, Cabo Verde). These rows were excluded to maintain data quality.

Since our task was classification, we transformed the continuous target variable (*Life Expectancy*) into categorical classes:

- Low
- Medium
- High

The categories were then label encoded into numerical values for training the models. Predictor variables included immunization rates, mortality factors, schooling, GDP, and social factors.

This preprocessing ensured the dataset was structured properly for classification.

2. Visualization Findings:

- The **bar chart** clearly showed that developed countries such as Japan and Switzerland had the **highest life expectancy**, while developing countries had much lower values.
- The **heatmap** revealed strong correlations, for example, **GDP and Schooling** were positively correlated with life expectancy, while **infant deaths and adult mortality** were negatively correlated.

3. Model Comparison Table

We tested three baseline models:

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

The following table summarizes the results:

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

Observation:

- KNN struggled due to sensitivity to scale and noisy features.
- Decision Tree performed much better but risked overfitting.
- Random Forest outperformed both, achieving the highest accuracy (92%).

4. Key Conclusions

- The **Random Forest model** performed best, achieving **92% accuracy**, because it reduces overfitting and captures complex feature interactions.
- The most important features influencing life expectancy were **Schooling, GDP, Adult Mortality, and Income Composition**, highlighting the role of both health and economic factors.
- **Hyperparameter tuning** improved model performance, especially for Random Forest, making it the most reliable model for life expectancy classification.