1. Data Cleaning
2. Feature Identification
3. Visualization
4. Preliminary Analysis

# Data Cleaning

The dataset has many missing values in several columns, including critical values which may influence life expectancy, such as Alcohol Consumption, Hepatitis B Immunization, BMI of the population, Polio Cases, Total Expenditure on Healthcare, etc. The dataset even has missing values for life expectancy.

Multiple approaches to cleaning this data can be used:

1. Removing Rows with missing values- While this is the most straightforward way to deal with the missing values, it would result in large amounts of loss in data.
2. Imputing missing values- Replacing the missing values with a statistical measure like mean, median or mode, as appropriate.

Since it is of utmost importance that the amount of data is maximized to improve the performance of any model being implemented, imputing the missing values is the best approach.

# Feature Identification

The RandomForest is an ensemble learning method, combining predictions from multiple decision trees to enhance prediction accuracy over any single model. The ensemble consists of numerous decision trees trained via the bagging method.

**Decision Trees**

Understanding RandomForests begins with comprehension of decision trees, specifically regression trees for regression tasks. A decision tree's structure is akin to a flowchart where each internal node denotes a test on an attribute, branches represent test outcomes, and leaf nodes hold the target variable's value.

For regression trees, each leaf's value typically represents the mean of the target values of samples within that leaf. The objective is to segment the feature space to minimize the target variable's variance within each leaf.

Mathematically, for feature space *X* and target space *Y*, a decision tree approximates the true function *f*\* from *x*∈*X* to *y*∈*Y* with a piecewise constant function *f*, where each leaf represents a partition *Xi*​⊂*X* with a constant prediction value *ci*​:

*f(x)=∑i​ci​****1****(x∈Xi​)​*

# Visualization

# Preliminary Analysis