Data Description
About this file
The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries The datasets are made available to the public for the purpose of health data analysis. The dataset related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from the United Nation website. Among all categories of health-related factors only those critical factors were chosen which are more representative. It has been observed that in the past 15 years , there has been a huge development in the health sector resulting in improvement of human mortality rates especially in the developing nations in comparison to the past 30 years. Therefore, in this project we have considered data from the years 2000-2015 for 193 countries for further analysis. The final merged file(final dataset) consists of 10 Columns and 2938 rows which means 9 predicting variables.

Columns Descriptions:

1. **Life expectancy:** Life Expectancy in age
2. **Status:** Developed or Developing status
3. **Adult Mortality:** Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
4. **BMI:** Average Body Mass Index of entire population
5. **HIV/AIDS:** Deaths per 1000 live births HIV/AIDS (0-4 years)
6. **Diphtheria:**Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
7. **Polio:** Polio (Pol3) immunization coverage among 1-year-olds (%)
8. **GDP:**Gross Domestic Product per capita (in USD)
9. **Income composition of resources:** Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
10. **Schooling:** Number of years of Schooling(years)

Questions:
EDA

1. How many independent variables/features are there in the data set?
2. How many numerical features are there?
3. How many categorical features are there?
4. Which is the third most important feature based on correlation for regression?
5. Which is the least important feature based on correlation for regression?

Multiple Linear Regression:
Convert Status into binary boolean variable which will be true if country is developed and answer the following post multiple linear regression(Note: Take default parameters of sklearn module and answer the following):

1. What is mean_squared_error for the test dataset?
2. What is r2_score for the test dataset?
3. What is the mean absolute error for the test dataset?
4. What is the value of cost function on the train and test dataset for the model previously trained?

5. What is mean_squared_error for the test dataset after we take top 5 features based on correlation?

Code email

6. What is r2_score for the test dataset after we take the top 5 features based on correlation?

What is the predicted Life expectancy for the following countries?

a. Case 1
    i. Status=Developing
    ii. Adult Mortality = 730
    iii. BMI = 27
    iv. HIV/AIDS = 33
    v. Diphtheria=70
    vi. Polio =60
    vii. GDP=450
    viii. Income composition of resources = 0.35
    ix. Schooling=10
b. Case 2
    i. Status=Developed
    ii. Adult Mortality = 100
    iii. BMI = 20
    iv. HIV/AIDS = 1
    v. Diphtheria=7
    vi. Polio =6
    vii. GDP=4500
    viii. Income composition of resources = 0.5
    ix. Schooling=12

Logistic regression:

Convert Life expectancy into binary variable >=65 which is true if life expectancy is greater than or equal to 65
Take default parameters of Logistics regression in sklearn module

1. What is the accuracy of the logistic regression model on the test dataset?
2. Find the number of false positives & False negatives on the test data set.
3. Calculate the precision and recall on the test dataset.
4. What is the value of the cost function on the train and test dataset?
5. If we keep the threshold for positive class to be 70% then what is the accuracy of the model on the test dataset?

Decision Tree:

Convert Life expectancy into binary variable >=65 which is true if life expectancy is greater than or equal to 65
Take default parameters of Logistics regression in sklearn module

1 What is the accuracy of the Decision Tree model on the test dataset?
2 Find the number of false positives & False negatives on the test data set.
3 Calculate the precision and recall on the test dataset.
4 What is the Gini Impurity of the train dataset?
5 What is the entropy of the train dataset?
6. What is the Gini Impurity of the test dataset?
7. What is the entropy of the test dataset?

Random Forest:

Convert Life expectancy into binary variable >=65 which is true if life expectancy is greater than or equal to 65
Take default parameters of Logistics regression in sklearn module

1 What is the accuracy of the model on the test dataset?
2 Find the number of false positives & False negatives on the test data set.
3 Calculate the precision and recall on the test dataset.