In [2]:

```
import pandas as pd
import numpy as np
```

In [3]:

```
df = pd.read_csv("train_data.csv")
```

In [4]:

```
df.shape
```

Out[4]:

```
(2644, 23)
```

In [5]:

```
df.isna().sum()
```

Out[5]:

```
id                                  0
Country                             0
Year                                0
Status                              0
Life expectancy                     9
Adult Mortality                     9
infant deaths                       0
Alcohol                           174
percentage expenditure              0
Hepatitis B                       487
Measles                             0
 BMI                               28
under-five deaths                   0
Polio                              15
Total expenditure                 204
Diphtheria                         15
 HIV/AIDS                           0
GDP                               401
Population                        595
 thinness  1-19 years              28
 thinness 5-9 years                28
Income composition of resources   149
Schooling                         146
dtype: int64
```

In [8]:

```
df.dtypes
```

Out[8]:

```
id                                  int64
Country                            object
Year                                int64
Status                             object
Life expectancy                   float64
Adult Mortality                   float64
infant deaths                       int64
Alcohol                           float64
percentage expenditure            float64
Hepatitis B                       float64
Measles                             int64
 BMI                              float64
under-five deaths                   int64
Polio                             float64
Total expenditure                 float64
Diphtheria                        float64
 HIV/AIDS                         float64
GDP                               float64
Population                        float64
 thinness  1-19 years             float64
 thinness 5-9 years               float64
Income composition of resources   float64
Schooling                         float64
dtype: object
```

In [ ]:

```

```

In [12]:

```
df = df.fillna(df.mean(axis=0))
```

In [13]:

```
df.isna().sum()
```

Out[13]:

```
id                                   0
Country                              0
Year                                 0
Status                               0
Life expectancy                      0
Adult Mortality                      0
infant deaths                        0
Alcohol                              0
percentage expenditure               0
Hepatitis B                          0
Measles                              0
 BMI                                 0
under-five deaths                    0
Polio                                0
Total expenditure                    0
Diphtheria                           0
 HIV/AIDS                            0
GDP                                  0
Population                           0
 thinness  1-19 years                0
 thinness 5-9 years                  0
Income composition of resources      0
Schooling                            0
dtype: int64
```

In [17]:

```
df['Status']= df['Status'].astype('category').cat.codes
```

In [19]:

```
df.drop(columns=['Country'], inplace=True)
```

In [20]:

```
df.head()
```

Out[20]:

	id	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles
0	2752	2009	1	76.0	84.0	1	1.73	292.402267	93.000000	0
1	2486	2002	1	67.9	221.0	0	4.41	250.711237	81.197033	0
2	642	2014	0	77.8	97.0	0	12.14	1884.098811	95.000000	0
3	1229	2004	1	71.8	139.0	28	0.01	0.000000	95.000000	3
4	1583	2002	1	44.0	67.0	46	1.10	3.885395	64.000000	92

5 rows × 22 columns

In [23]:

```
df.columns
```

Out[23]:

```
Index(['id', 'Year', 'Status', 'Life expectancy ', 'Adult Mortality',
       'infant deaths', 'Alcohol', 'percentage expenditure', 'Hepatiti
s B',
       'Measles ', ' BMI ', 'under-five deaths ', 'Polio', 'Total expe
nditure',
       'Diphtheria ', ' HIV/AIDS', 'GDP', 'Population',
       ' thinness  1-19 years', ' thinness 5-9 years',
       'Income composition of resources', 'Schooling'],
      dtype='object')
```

In [21]:

```
from sklearn.model_selection import train_test_split
```

In [27]:

```
X_train, X_test, y_train, y_test = train_test_split(df.drop(columns=['Life expectanc
                                                    df['Life expectancy '], test_siz
```

In [28]:

```
X_train.shape, y_train.shape
```

Out[28]:

```
((1771, 21), (1771,))
```

In [29]:

```
X_test.shape, y_test.shape
```

Out[29]:

```
((873, 21), (873,))
```

In [ ]:

```

```

In [30]:

```
from sklearn.linear_model import LinearRegression
```

In [31]:

```
model = LinearRegression()
```

In [32]:

```
model.fit(X_train, y_train)
```

Out[32]:

```
LinearRegression()
```

In [33]:

```
y_pred = model.predict(X_test)
```

In [34]:

```
model.score(X_train, y_train)
```

Out[34]:

```
0.8198646717133773
```

In [35]:

```
model.score(X_test, y_test)
```

Out[35]:

```
0.8113515851735386
```

In [ ]:

```

```