Life Expectancy Prediction Using ML

MINI PROJECT

SUBMITTED BY :
SOUMYA DANTRE
02302102025
MTECH- CSE(AI)

Predicting Life Expectancy using different Regression Algorithms and comparing their performance on the Life Expectancy dataset

Dataset Used :

https://www.kaggle.com/datasets/saurabhbadole/life-expectancy-based-on-geographic-locations

The dataset contains country-wise health, economic, and social factors affecting life expectancy.

The dataset has ~22 features, collected from WHO, United Nations, and World Bank.

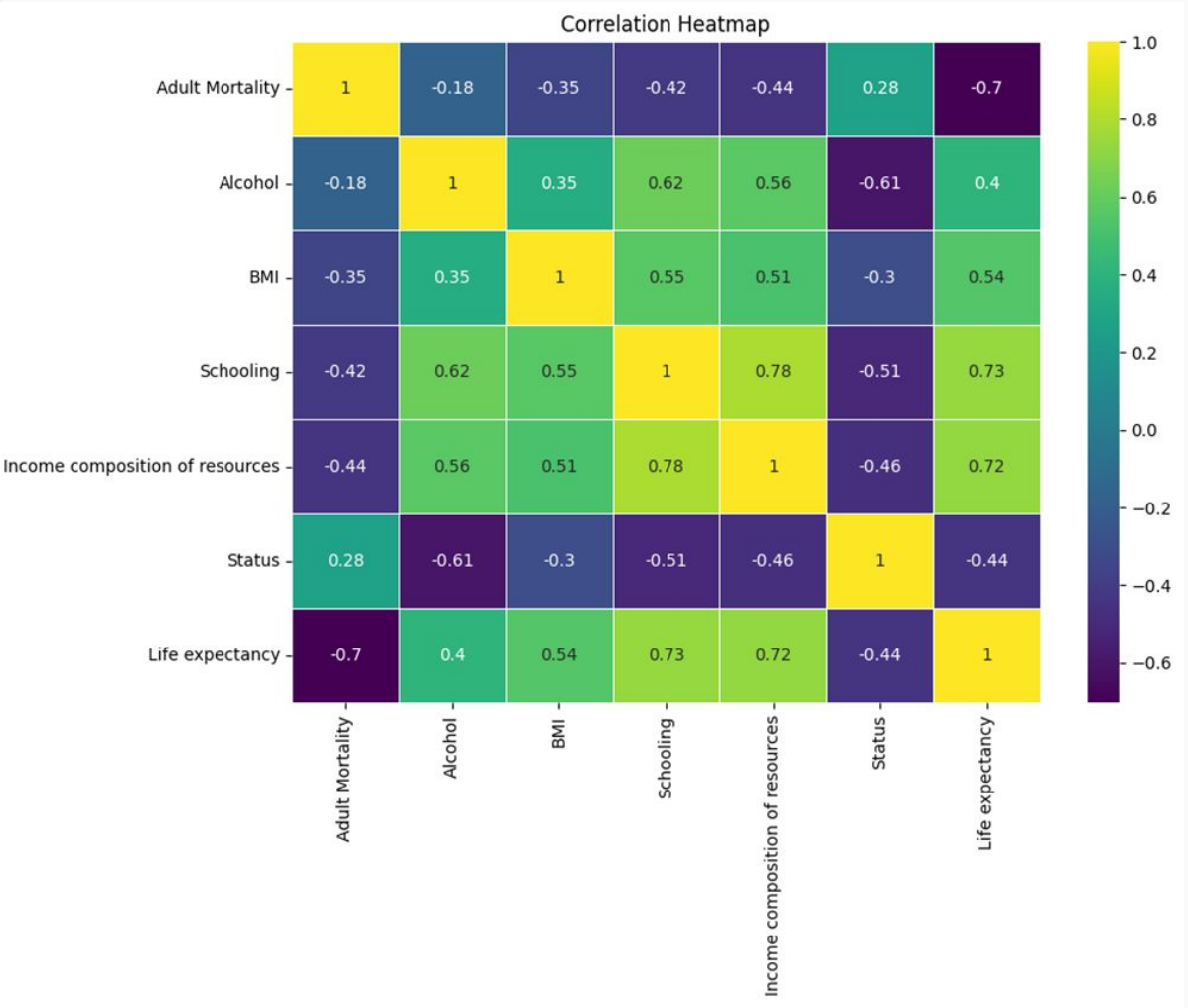
Features: Adult Mortality, Alcohol, BMI, Schooling, Income Index, Immunization Data, HIV/AIDS, GDP, Thinness Indicators, Population.

Correlation Heatmap

Heatmap displays the correlation between the selected features and the target variable Life Expectancy. Correlation shows how strongly two variables move together.

Value close to +1 → Strong positive relationship
Value close to −1 → Strong negative relationship
Value near 0 → No strong relationship

Colors in the heatmap represent the strength:
Yellow / Light Green → High positive correlation
Green → Moderate correlation
Blue / Purple → Negative or weak correlation

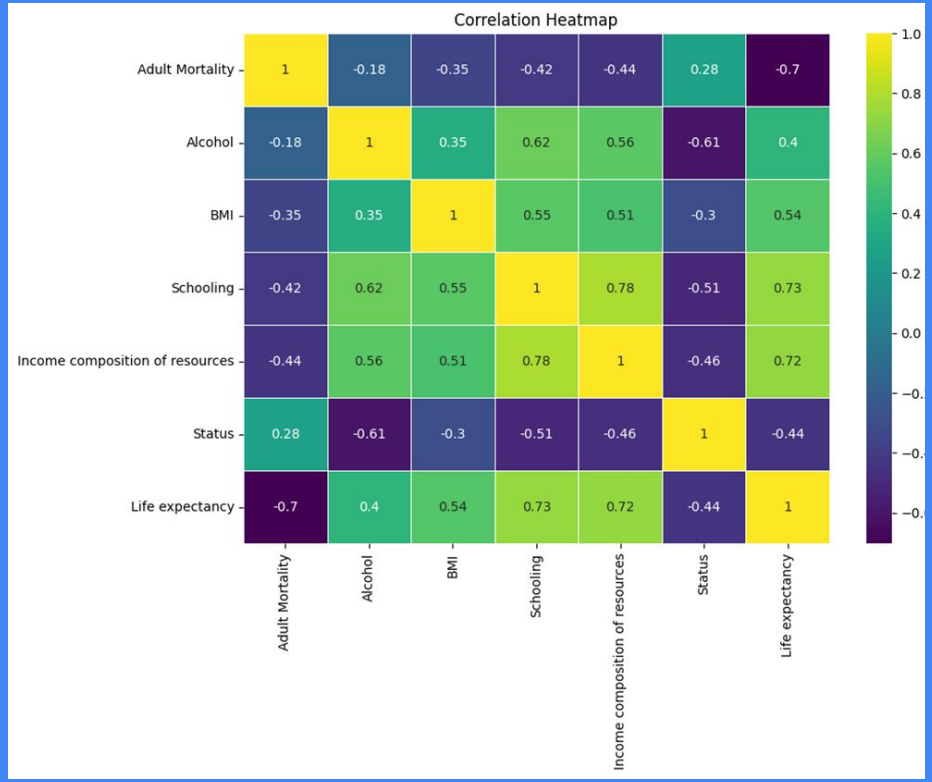


Observation

1. Relationship Between Features and Life Expectancy

The heatmap tells you which features influence life expectancy:
Schooling → Strong positive correlation
More years of education → Higher life expectancy
Income Composition → Strong positive correlation
Higher income index → Better living conditions → Higher life expectancy
BMI → Moderate positive correlation
Better nutrition → Longer lifespan
Adult Mortality → Strong negative correlation
High mortality rate → Lower life expectancy

2. How Features Relate to Each Other
Some features are naturally related:
Schooling ↔ Income index
Adult Mortality ↔ Life expectancy
BMI ↔ Life expectancy



Algorithms Used

1. Linear Regression

Linear Regression is a supervised learning algorithm used to predict a continuous output (here: Life Expectancy).

It models the relationship between multiple input features (Adult Mortality, BMI, Schooling, Income Index, etc.) and the target value.

The model fits a straight-line equation of the form:

$$LifeExpectancy = \beta_0 + \beta_1(BMI) + \beta_2(Alcohol) + \dots$$

It is simple, interpretable, and useful as a baseline model.

Algorithms Used

2. Random Forest Regression

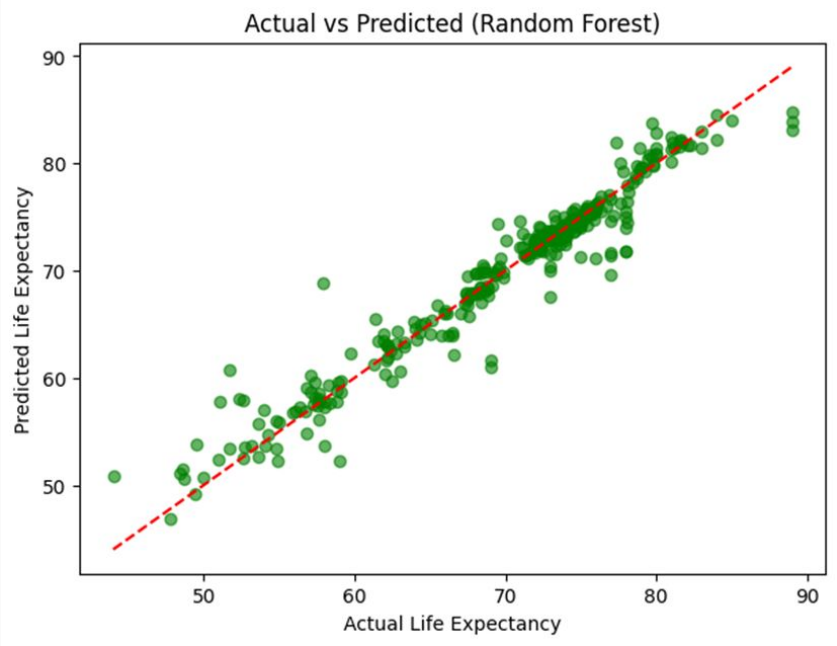
Random Forest is an ensemble algorithm that combines many decision trees.

Each tree makes a prediction, and the final result is the average of all trees.

It captures non-linear patterns, handles large datasets, and gives higher prediction accuracy.

Works better than a single tree and reduces overfitting.


Regression Line - Linear Regression
Perfect Fit
Predicted Life Expectancy
Actual Life Expectancy
50
60
70
80
90


Actual vs Predicted (Random Forest)
Predicted Life Expectancy
Actual Life Expectancy
50
60
70
80
90

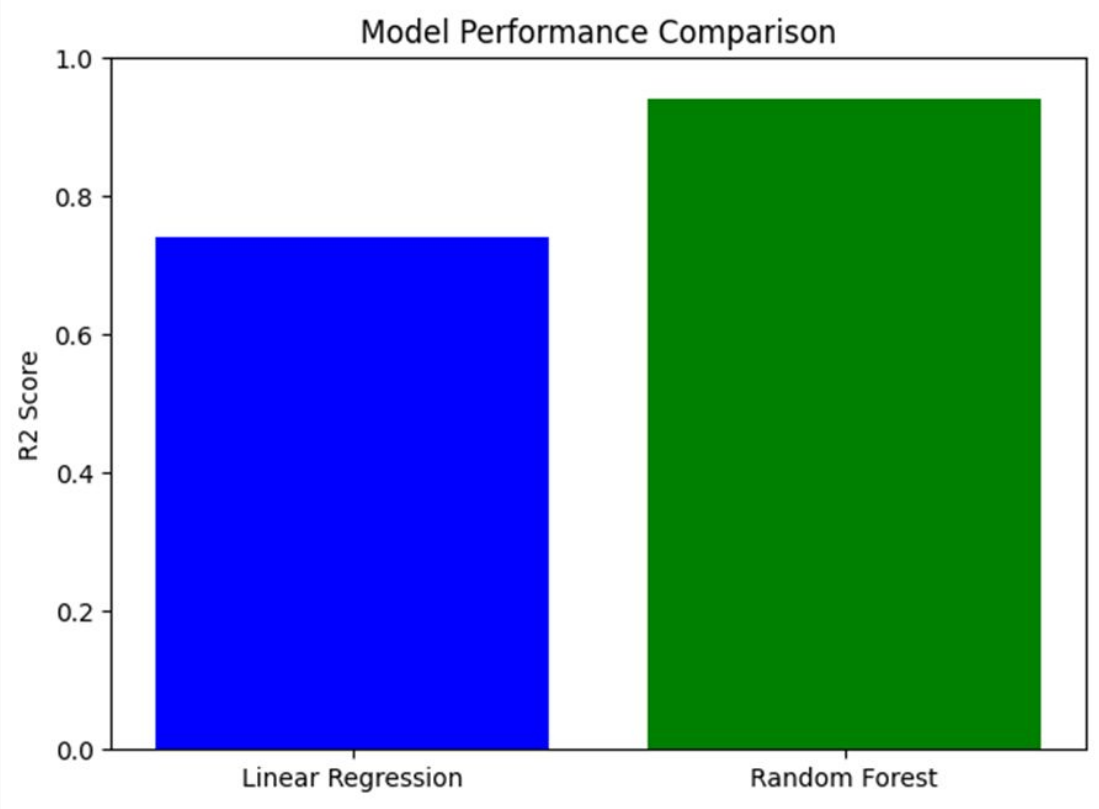
Results

```
...
    ---- Linear Regression ----
    R2 Score: 0.7412292204451125
    MSE: 18.378543675118657

    ---- Random Forest ----
    R2 Score: 0.9407460887255386
    MSE: 4.208359993939385
```

Random Forest performs best.

Important factors: Adult Mortality, Schooling, BMI, Income Index.



THANKYOU