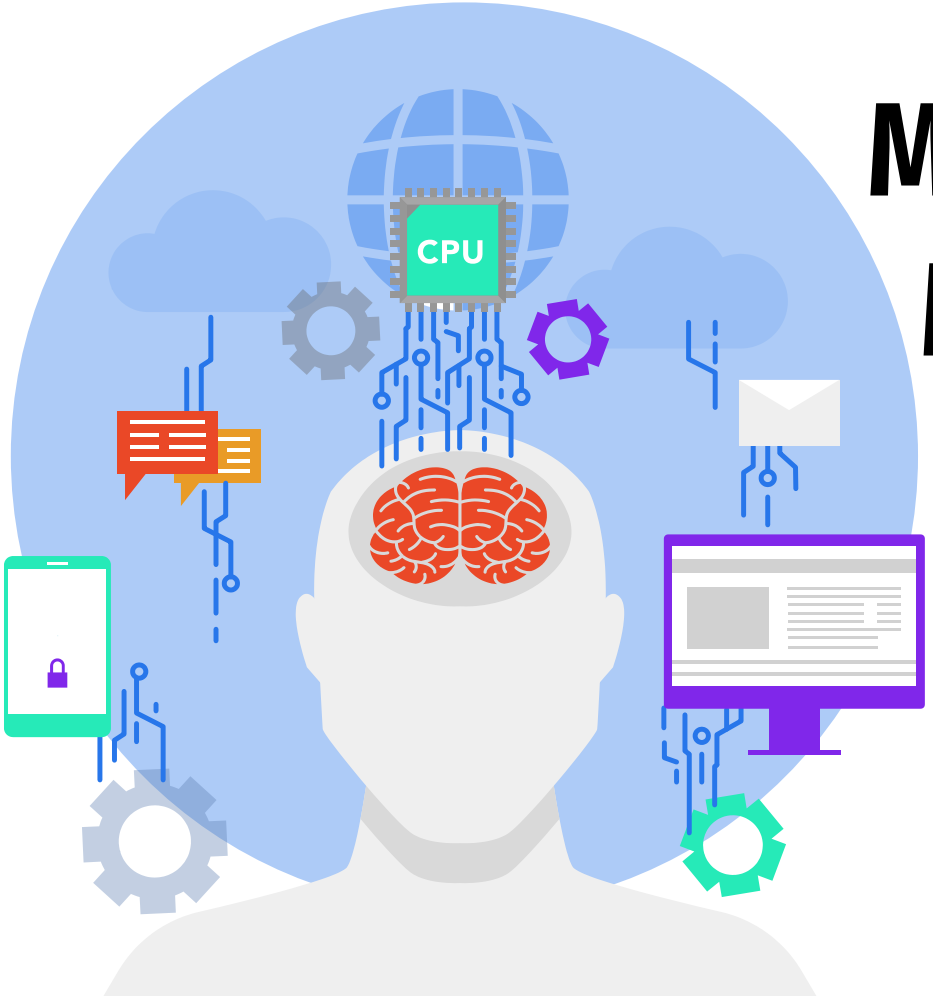


Machine Learning Model to predict life expectancy

By Alex Cortes

EDA and Data Preprocessing

Being aware about the impact of data preprocessing for the performance of the model

Inconsistencies
Handle extra spaces and special characters in column names

01

Correlation matrix

Drop features close to 0

02

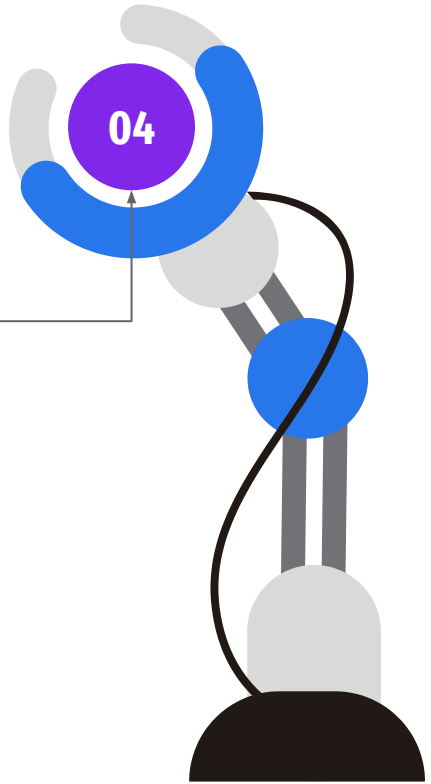
Visualization
Plot distribution of each variable

03

Imputation

.ffill().bfill() → .fillna(0)

04



Encoding

Data types

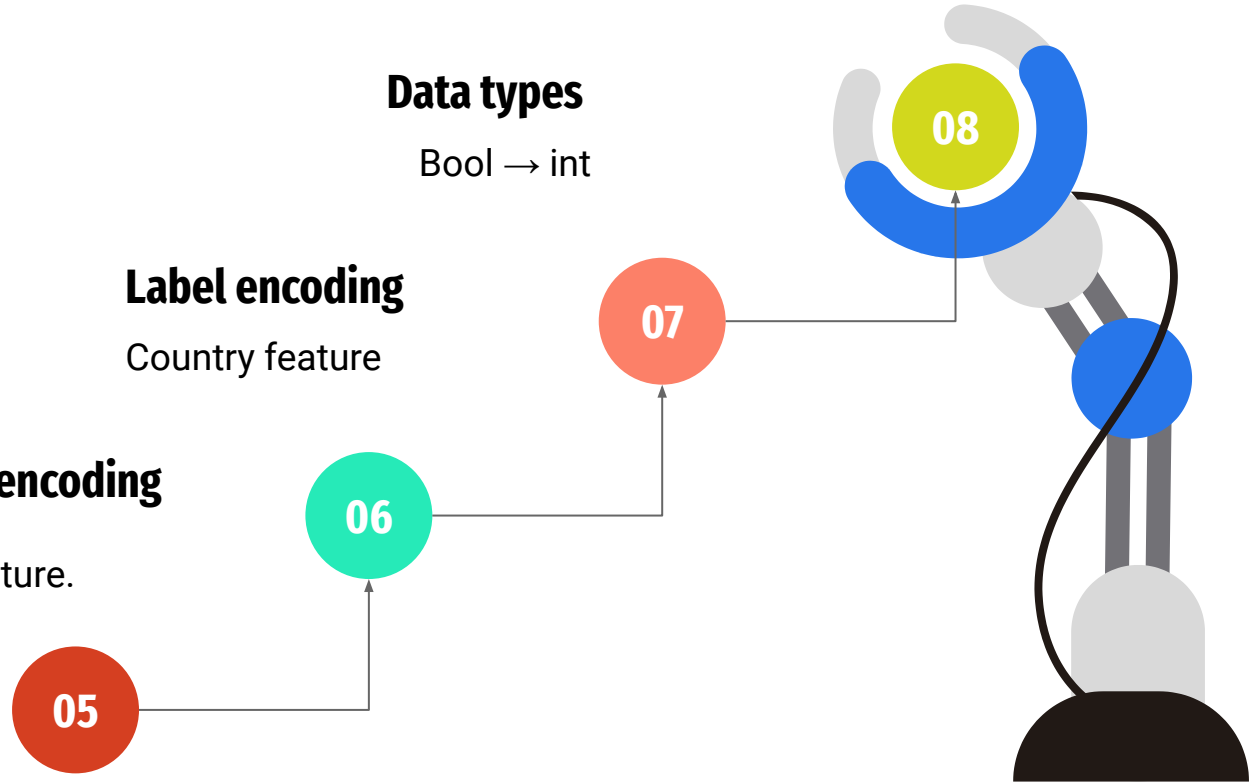
Bool → int

Label encoding

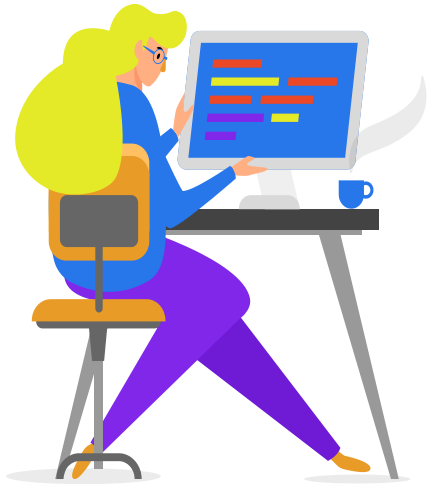
Country feature

One-hot encoding

Status feature.



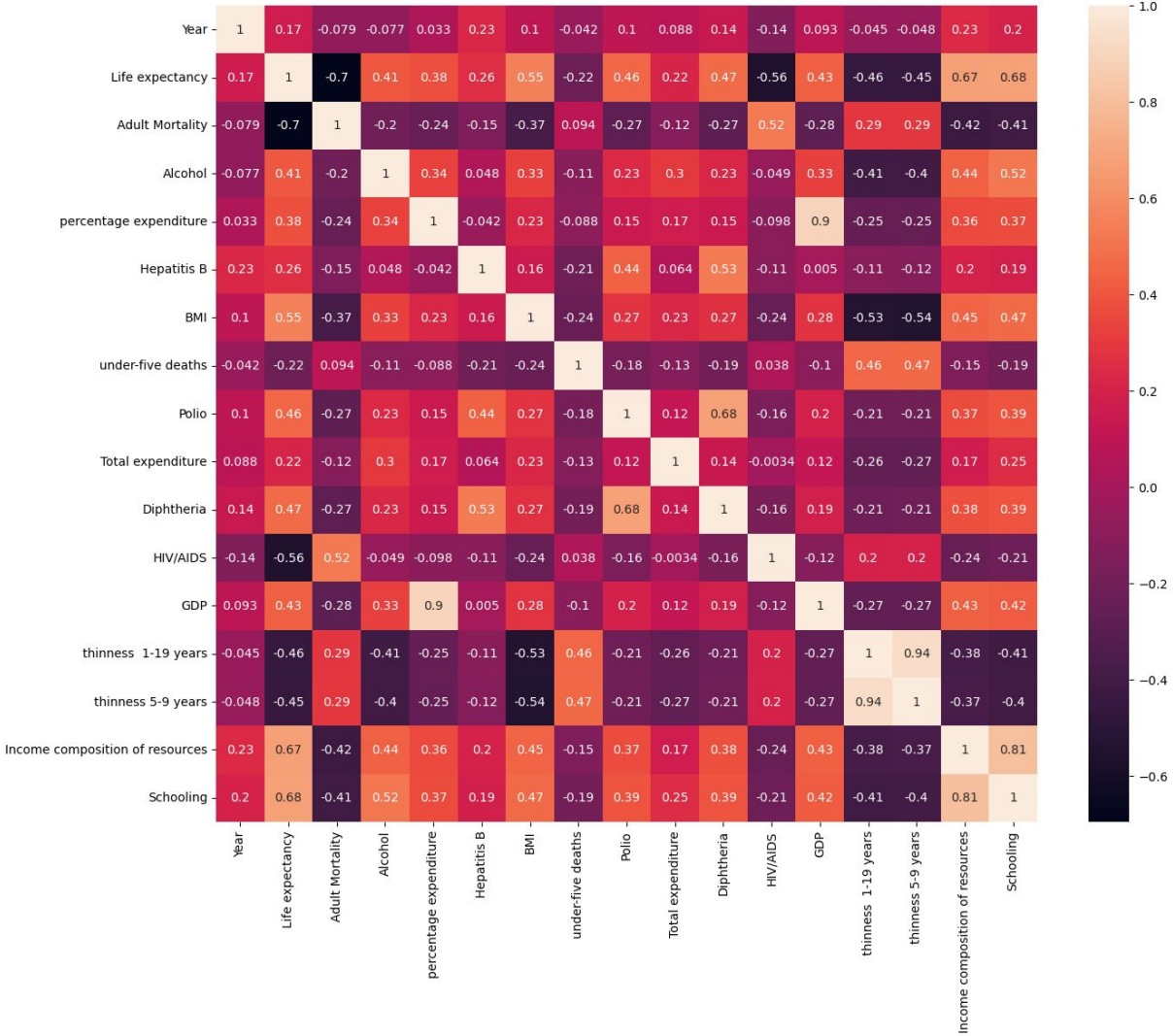
Feature Engineering



Variance Inflation Factor (VIF) method to test Multicollinearity.

	variable	VIF
0	Intercept	0.000000
1	Country	1.042866
2	Year	1.191255
3	Adult_Mortality	1.709408
4	Alcohol	1.943198
5	percentage_expenditure	5.485958
6	Hepatitis_B	1.546191
7	BMI	1.687098
8	under_five_deaths	1.367295
9	Polio	1.985652
10	Total_expenditure	1.221337
11	Diphtheria	2.238289
12	HIV_AIDS	1.447225
13	GDP	5.760995
14	thinness__1_19_years	8.860278
15	thinness_5_9_years	8.939633
16	Income_composition_of_resources	3.161793
17	Schooling	3.432671
18	Status_Developed	inf
19	Status_Developing	inf

	variable	VIF
0	Intercept	0.000000
1	Country	1.040230
2	Year	1.191250
3	Adult_Mortality	1.708502
4	Alcohol	1.942714
5	percentage_expenditure	5.485503
6	Hepatitis_B	1.543827
7	BMI	1.664113
8	under_five_deaths	1.350130
9	Polio	1.985433
10	Total_expenditure	1.217816
11	Diphtheria	2.233778
12	HIV_AIDS	1.446719
13	GDP	5.757905
14	thinness__1_19_years	1.928245
15	Income_composition_of_resources	3.161513
16	Schooling	3.431602
17	Status_Developed	inf
18	Status_Developing	inf

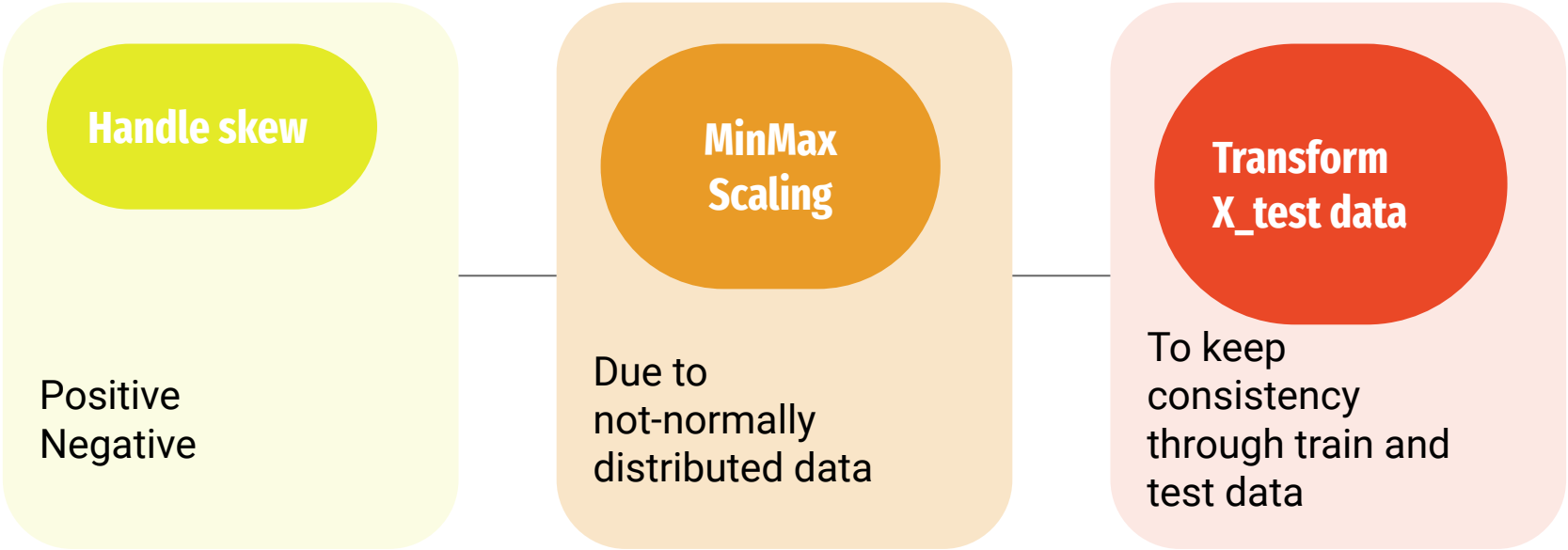

Year 1 0.17 -0.079 -0.077 0.033 0.23 0.1 -0.042 0.1 0.088 0.14 -0.14 0.093 -0.045 -0.048 0.23 0.2
Life expectancy 0.17 1 -0.7 0.41 0.38 0.26 0.55 -0.22 0.46 0.22 0.47 -0.56 0.43 -0.46 -0.45 0.67 0.68
Adult Mortality -0.079 -0.7 1 -0.2 -0.24 -0.15 -0.37 0.094 -0.27 -0.12 -0.27 0.52 -0.28 0.29 0.29 -0.42 -0.41
Alcohol -0.077 0.41 -0.2 1 0.34 0.048 0.33 -0.11 0.23 0.3 0.23 -0.049 0.33 -0.41 -0.4 0.44 0.52
percentage expenditure 0.033 0.38 -0.24 0.34 1 -0.042 0.23 -0.088 0.15 0.17 0.15 -0.098 0.9 -0.25 -0.25 0.36 0.37
Hepatitis B 0.23 0.26 -0.15 0.048 -0.042 1 0.16 -0.21 0.44 0.064 0.53 -0.11 0.005 -0.11 -0.12 0.2 0.19
BMI 0.1 0.55 -0.37 0.33 0.23 0.16 1 -0.24 0.27 0.23 0.27 -0.24 0.28 -0.53 -0.54 0.45 0.47
under-five deaths -0.042 -0.22 0.094 -0.11 -0.088 -0.21 -0.24 1 -0.18 -0.13 -0.19 0.038 -0.1 0.46 0.47 -0.15 -0.19
Polio 0.1 0.46 -0.27 0.23 0.15 0.44 0.27 -0.18 1 0.12 0.68 -0.16 0.2 -0.21 -0.21 0.37 0.39
Total expenditure 0.088 0.22 -0.12 0.3 0.17 0.064 0.23 -0.13 0.12 1 0.14 -0.0034 0.12 -0.26 -0.27 0.17 0.25
Diphtheria 0.14 0.47 -0.27 0.23 0.15 0.53 0.27 -0.19 0.68 0.14 1 -0.16 0.19 -0.21 -0.21 0.38 0.39
HIV/AIDS -0.14 -0.56 0.52 -0.049 -0.098 -0.11 -0.24 0.038 -0.16 -0.0034 -0.16 1 -0.12 0.2 0.2 -0.24 -0.21
GDP 0.093 0.43 -0.28 0.33 0.9 0.005 0.28 -0.1 0.2 0.12 0.19 -0.12 1 -0.27 -0.27 0.43 0.42
thinness 1-19 years -0.045 -0.46 0.29 -0.41 -0.25 -0.11 -0.53 0.46 -0.21 -0.26 -0.21 0.2 -0.27 1 0.94 -0.38 -0.41
thinness 5-9 years -0.048 -0.45 0.29 -0.4 -0.25 -0.12 -0.54 0.47 -0.21 -0.27 -0.21 0.2 -0.27 0.94 1 -0.37 -0.4
Income composition of resources 0.23 0.67 -0.42 0.44 0.36 0.2 0.45 -0.15 0.37 0.17 0.38 -0.24 0.43 -0.38 -0.37 1 0.81
Schooling 0.2 0.68 -0.41 0.52 0.37 0.19 0.47 -0.19 0.39 0.25 0.39 -0.21 0.42 -0.41 -0.4 0.81 1
Year
Life expectancy
Adult Mortality
Alcohol
percentage expenditure
Hepatitis B
BMI
under-five deaths
Polio
Total expenditure
Diphtheria
HIV/AIDS
GDP
thinness 1-19 years
thinness 5-9 years
Income composition of resources
Schooling
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Vs

Feature engineering

X and Y variables	
Data splitting 60/40	X_train, X_test, Y_train, Y_test
X_copy	To store 80/20 splited data
Histogram of each variable in X_train	verify if they have Gaussian-like distribution and to point to possible outliers presence

Feature engineering



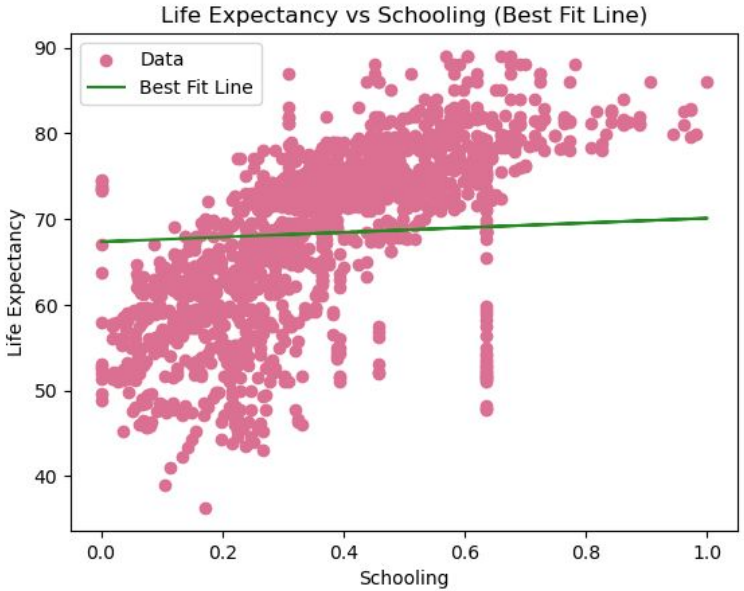
Model Building

Using the top 5 features to build a multiple linear regression model

	Adult_Mortality	Income_composition_of_resources	Schooling	HIV_AIDS	BMI
0	0.533021	0.400577	0.282387	0.153819	0.455746
1	0.428228	0.659728	0.575253	0.000000	0.062087
2	0.596082	0.259584	0.164671	0.000000	0.462351

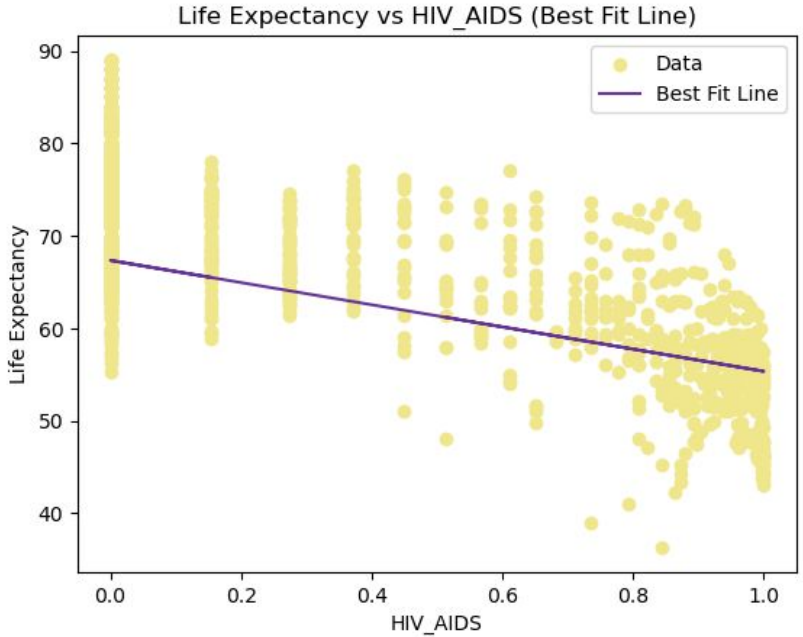
As Adult Mortality increases, Life Expectancy slightly decreases. Higher adult mortality typically suggests poorer health conditions. The best-fit line shows a weak negative trend while the spread in the data indicates Adult Mortality it's not the only factor impacting life expectancy

Even though life expectancy may increase slightly with schooling, it's not enough to create a noticeable slope in a linear regression



The downward slope of the best-fit line indicates a negative relationship, meaning that as HIV/AIDS prevalence increases, life expectancy tends to decrease.

At low HIV/AIDS rates, life expectancy becomes less predictable based on this feature alone



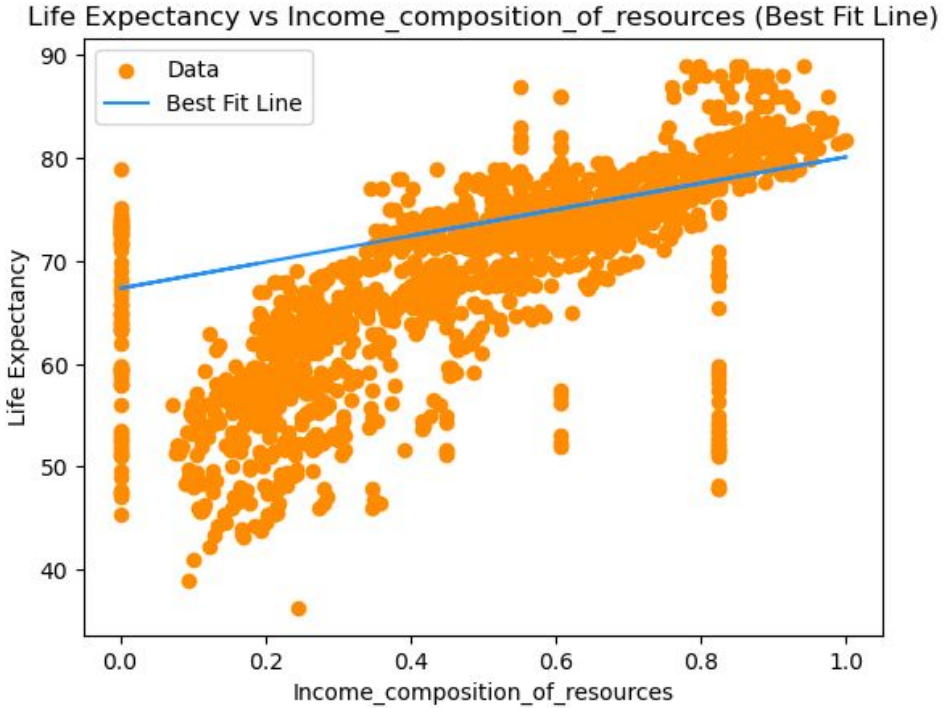
Since the fit line has a slight positive slope, it indicates a weak positive correlation between BMI and life expectancy. As BMI increases, life expectancy also increases slightly, The spread of the data means BMI is not a dominant predictor by itself.

Life expectancy vs income on multiple linear regression model 60/40

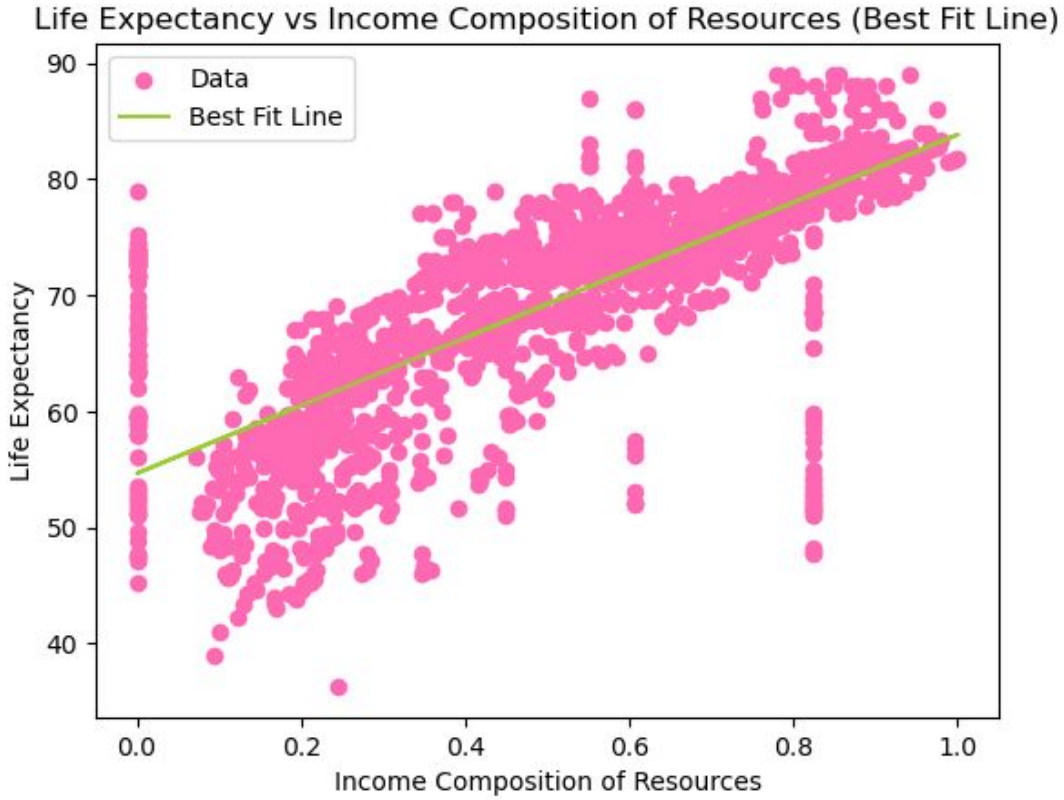
The positive slope of the best-fit line suggests a stronger relationship, as Income Composition of Resources increases, Life Expectancy generally increases.

Higher values in this feature are associated with better resource access, often translating to better health care, education, and overall living conditions, which positively affect life expectancy.

This feature might be a key predictor for this model



Life expectancy vs income on simple linear regression model 60/40



Income composition of resources was chosen as the key feature to predict life expectancy in a simple linear regression model.

Model Evaluation and Tuning:

Multilinear regression 60/40

Simple linear regression 60/40

	Predicted Value	Actual Value
2402	56.582810	54.0
203	67.749405	67.3
2325	76.232059	79.3
1744	65.510064	73.5
1093	54.467481	57.6
...	...	...
2544	71.033074	73.8
1408	76.418564	73.2
124	79.336922	83.0
2452	71.720805	69.1
1196	64.592463	64.4

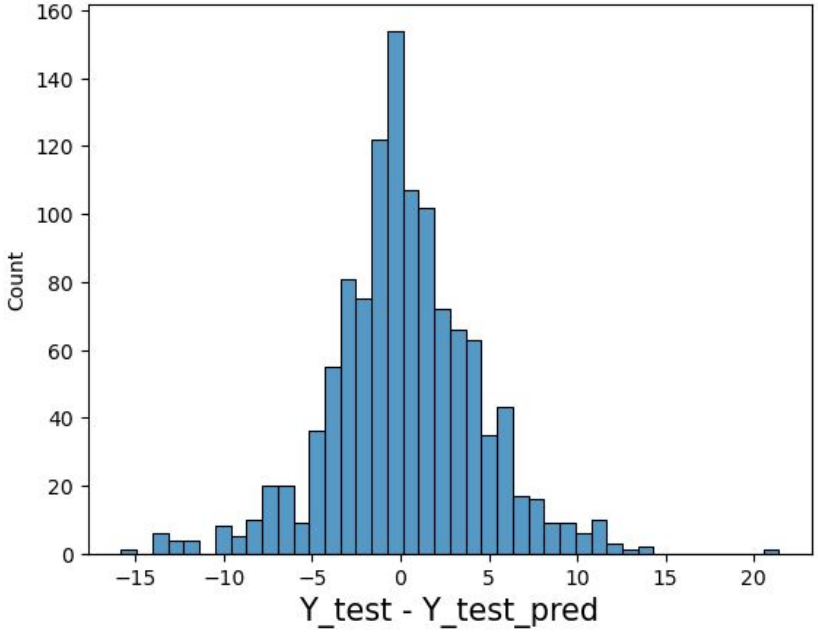
	Predicted Value	Actual Value
2402	66.687899	54.0
203	62.473818	67.3
2325	76.423791	79.3
1744	54.646892	73.5
1093	60.265321	57.6
...	...	...
2544	68.405997	73.8
1408	74.704231	73.2
124	81.237323	83.0
2452	70.783285	69.1
1196	63.629443	64.4

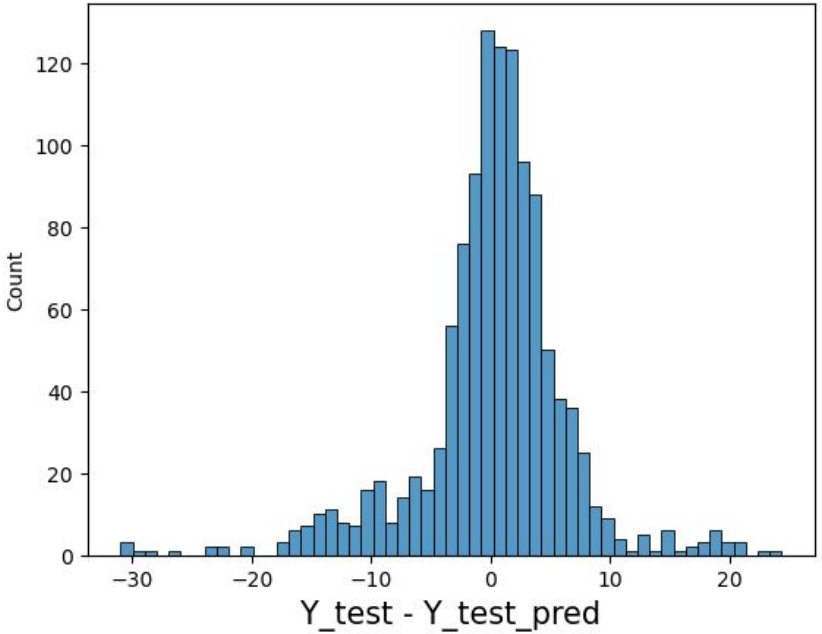
Model Evaluation and Tuning:

Multilinear regression 60/40

→

Simple linear regression 60/40





Model Evaluation and Tuning:

Multilinear regression 60/40

Simple linear regression 60/40

Overall, the model explains a significant proportion of the variance in life expectancy, which suggests that the selected features (Adult Mortality, Income Composition of Resources, Schooling, HIV/AIDS, and BMI) are relevant predictors of life expectancy.

```
Mean Absolute Error:  3.1009655823824445
Mean Squared Error:  17.707492360936232
R2 coefficient:  0.7982056906482318
```

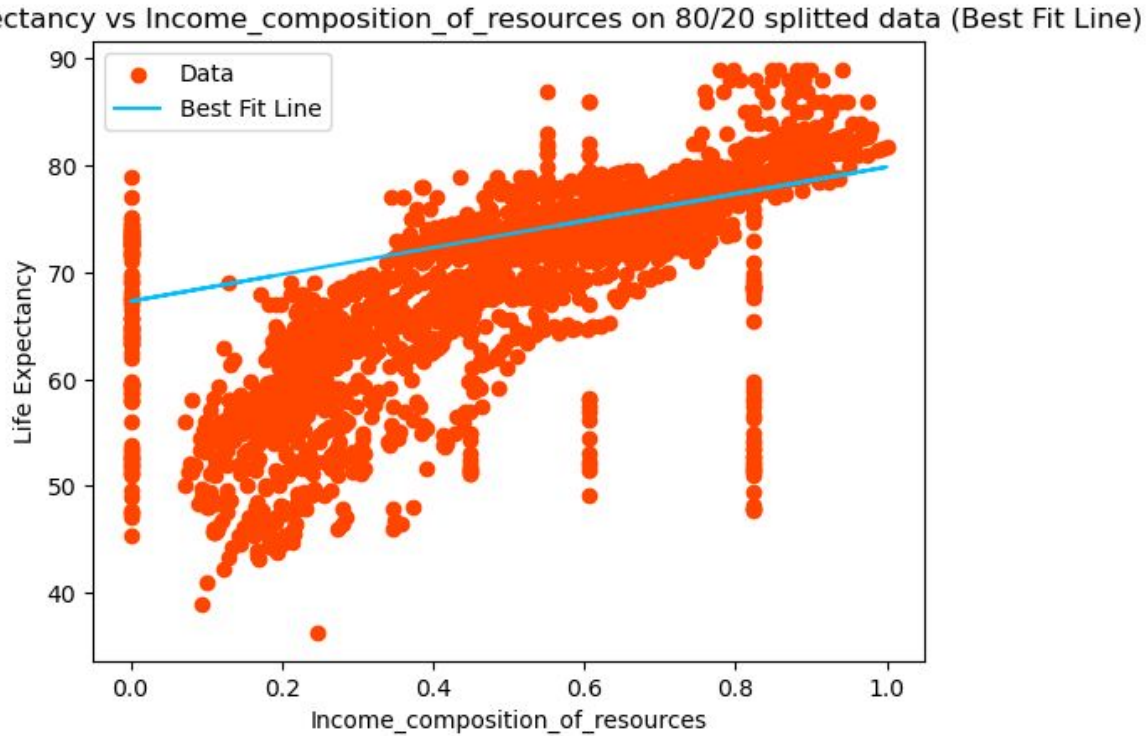
```
Mean Absolute Error:  4.2317631880222235
Mean Squared Error:  39.317630554639756
R2 coefficient:  0.5519368896849255
```

Multilinear regression 80/20

The most fitable model to capture life expectancy variations is multilinear regression

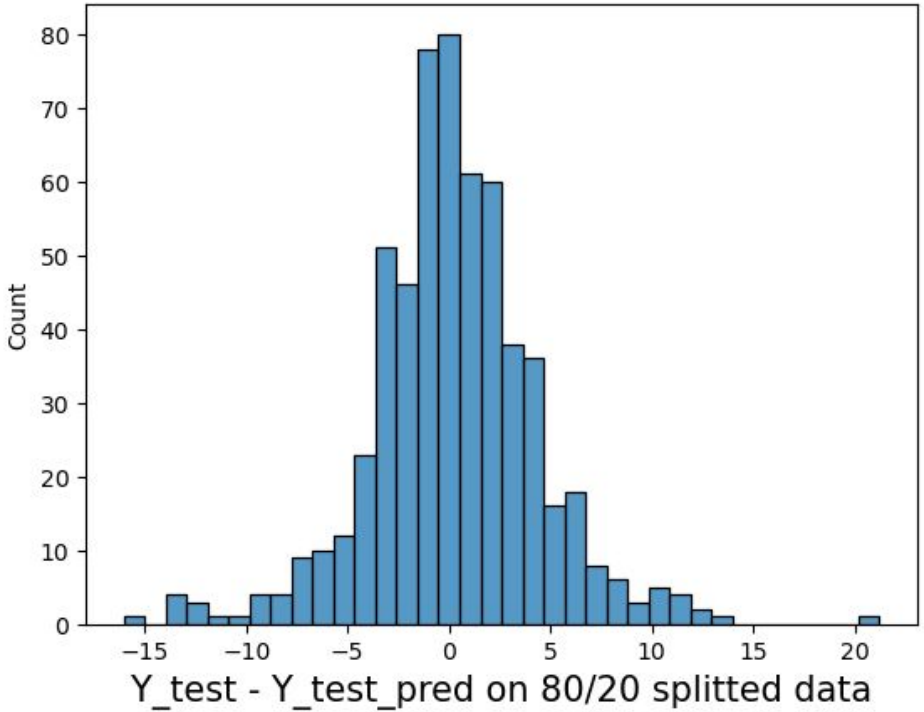
After increasing the train data by splitting 80/20, the multilinear model improved significantly

Life expectancy vs income on multiple linear regression model 80/20



Actual values vs Predicted value on multiple linear regression model 80/20

	Predicted Value	Actual Value
2402	56.682376	54.0
203	67.681383	67.3
2325	76.167788	79.3
1744	65.567808	73.5
1093	54.608398	57.6
...	...	...
2445	72.800379	74.5
370	77.670208	77.1
1791	61.341972	63.5
975	54.732963	56.6
1954	66.990155	63.5



Model Evaluation and Tuning:

Multilinear regression 80/20

After increasing the train data by splitting 80/20, the multilinear model improved significantly

```
Mean Absolute Error:  3.0214847725530407
Mean Squared Error:  17.31048110982294
R2 coefficient:  0.8054237858920115
```

Metric results are sign that the model fits the data well and explains a significant portion of the variability in life expectancy.

Cross Validation: Lasso Regression

- Since Multicollinearity was already handle by VIF method, the best option is Lasso(L1)-Regularization.

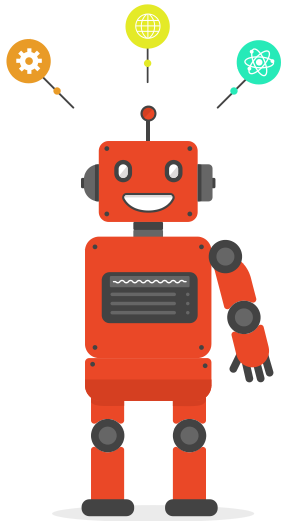
- K-fold validation will be performed.

```
# Best score is given by the scores list (i=0.25 = 0*0.25 = 0)
lm_best = Lasso(alpha=0, tol=0.0925) #alpha is customizable from 0 to 1
lm.fit(X_train_top_copy, Y_train)
```

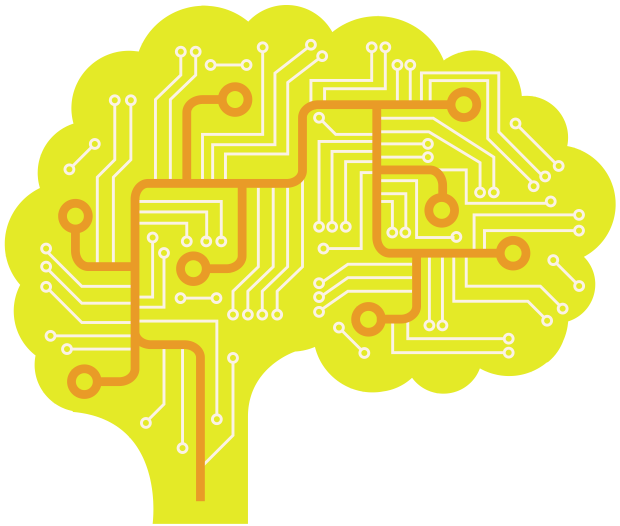
```
Lasso
Lasso(alpha=2.0, tol=0.0925)
```

```
# Return r2 score for test data after lasso regression
lr.score(X_test_top_copy, Y_test)
```

Before applying Lasso, the r2 score was 0.8054237858920115. Lasso improved 0.012% the r2 score. Even if its a small number, it is an improvement.



Challenges



- Not been able to understand the logic behind the workflow at the beginning
- Dealing with outliers
- Wrong interpretation of errors.