Economic Development and Life Expectancy

Factors affecting life expectancy: Application to 120 countries

Students: Lia Poghosyan, Elen Sahakyan,
Tamara Sedrakyan, Anna Tantushyan, Meri Torosyan

Outline

- Data Description
- Linear Regression Model
- Naive Bayes Model
- k-nearest neighbors
- Decision Tree
- Logistic Regression Model
- Random Forest

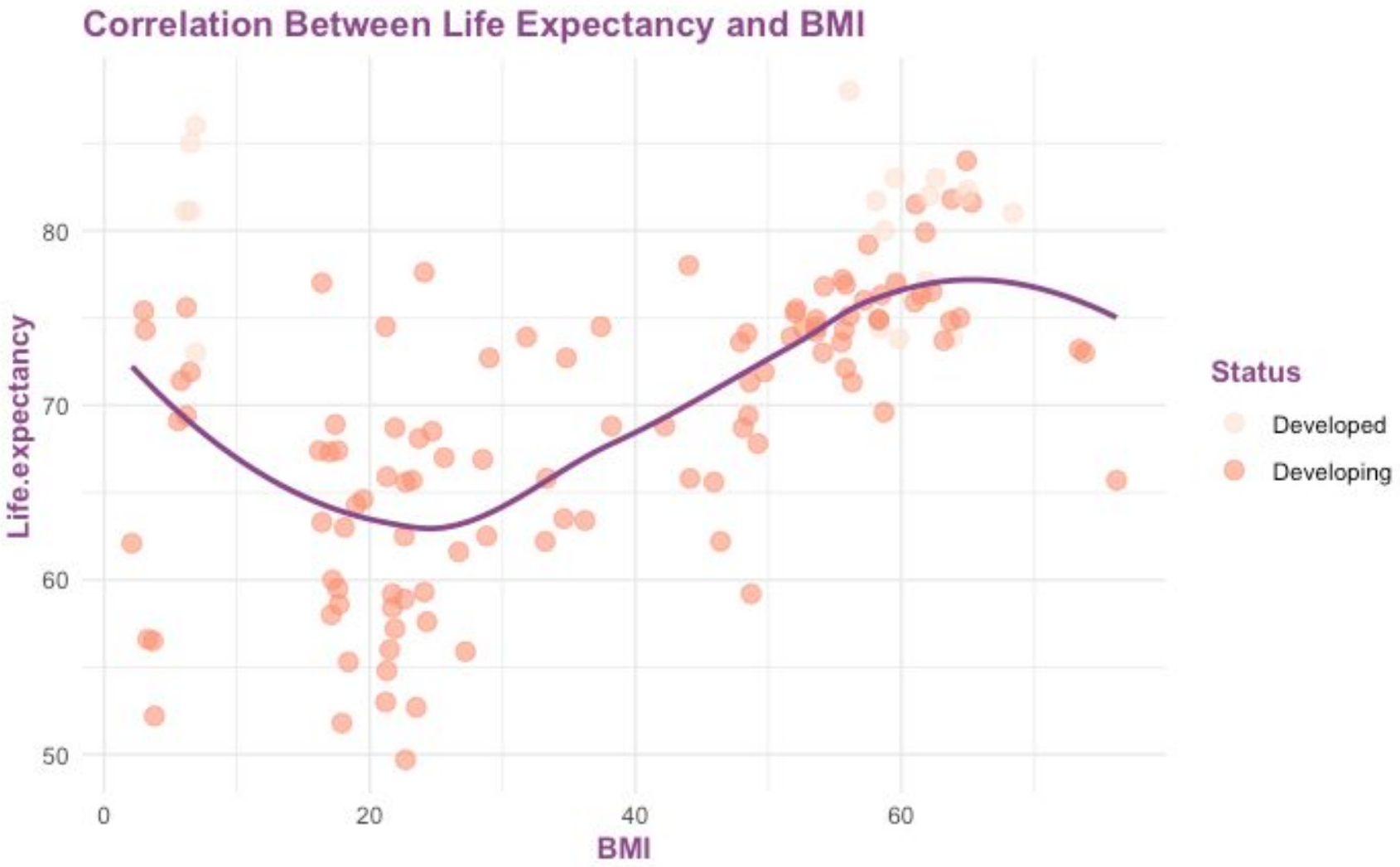
Life Expectancy

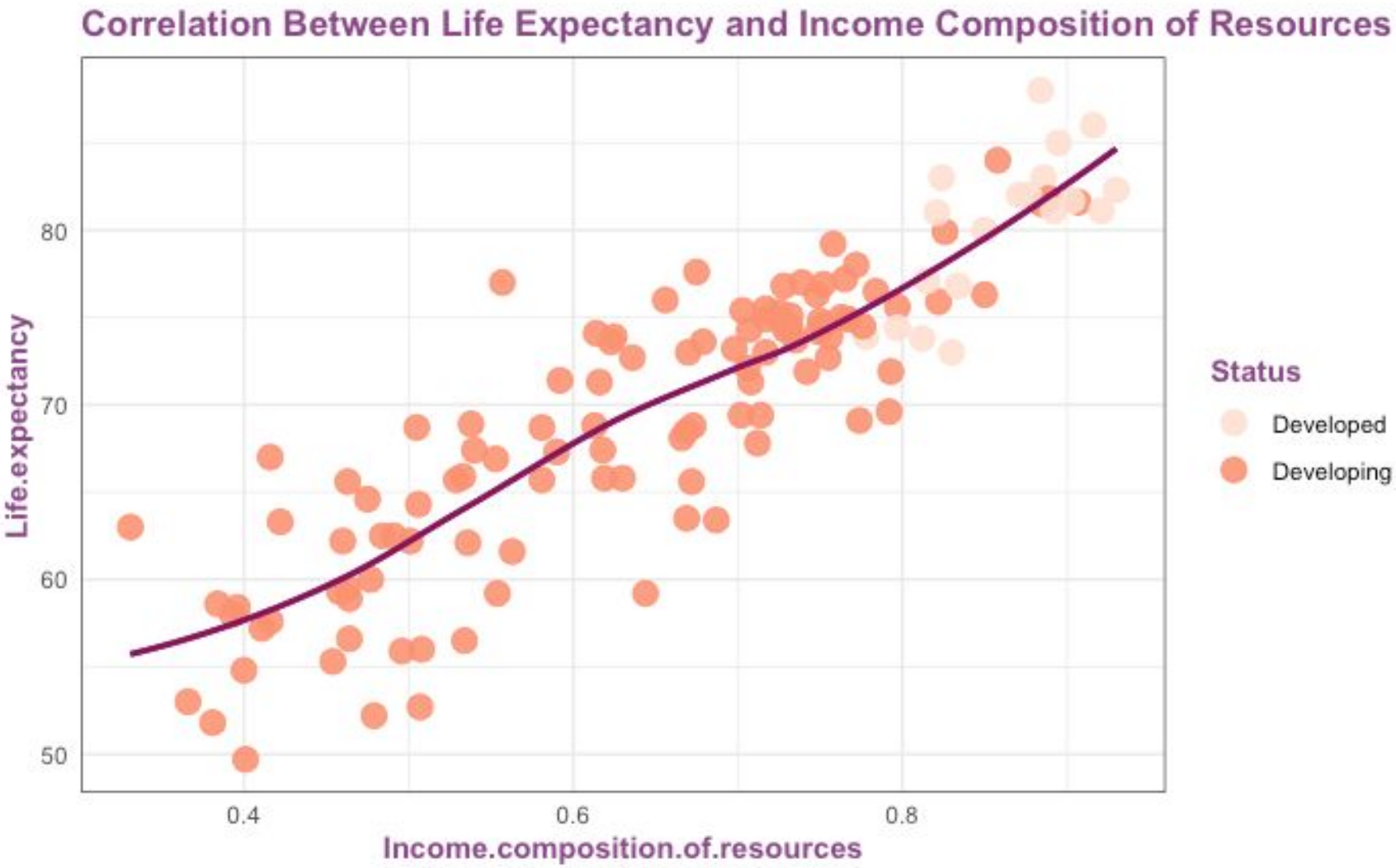
Data collected from World Health Organization (WHO)

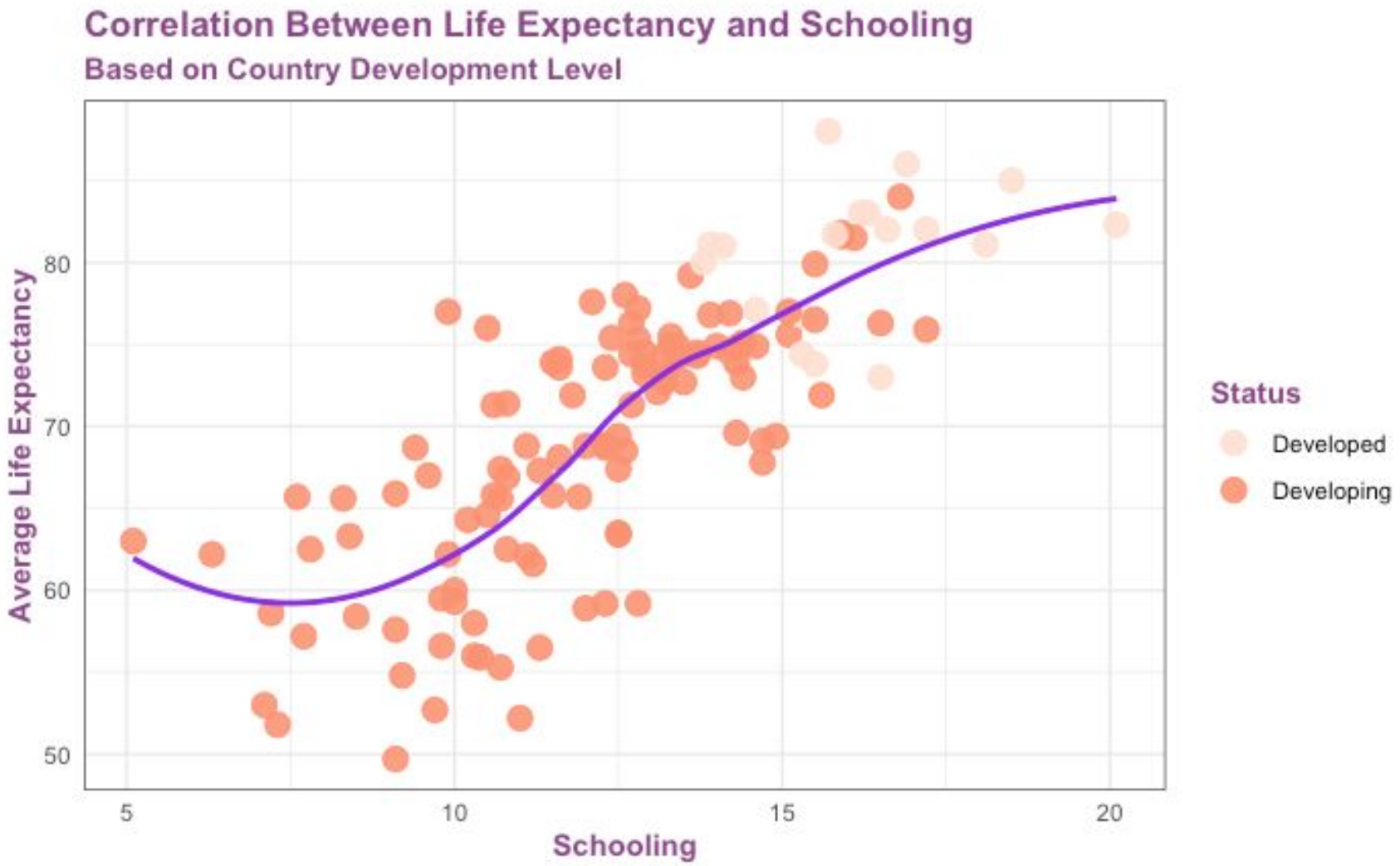
Dataset: Data Structure

```
'data.frame':   2938 obs. of  22 variables:
 $ Country                        : Factor w/ 193 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Year                           : int  2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 ...
 $ Status                         : Factor w/ 2 levels "Developed","Developing": 2 2 2 2 2 2 2 2 2 2 ...
 $ Life.expectancy                : num  65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...
 $ Adult.Mortality                : int  263 271 268 272 275 279 281 287 295 295 ...
 $ infant.deaths                  : int  62 64 66 69 71 74 77 80 82 84 ...
 $ Alcohol                        : num  0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 ...
 $ percentage.expenditure         : num  71.3 73.5 73.2 78.2 7.1 ...
 $ Hepatitis.B                    : int  65 62 64 67 68 66 63 64 63 64 ...
 $ Measles                        : int  1154 492 430 2787 3013 1989 2861 1599 1141 1990 ...
 $ BMI                            : num  19.1 18.6 18.1 17.6 17.2 16.7 16.2 15.7 15.2 14.7 ...
 $ under.five.deaths              : int  83 86 89 93 97 102 106 110 113 116 ...
 $ Polio                          : int  6 58 62 67 68 66 63 64 63 58 ...
 $ Total.expenditure              : num  8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ...
 $ Diphtheria                     : int  65 62 64 67 68 66 63 64 63 58 ...
 $ HIV.AIDS                       : num  0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...
 $ GDP                            : num  584.3 612.7 631.7 670 63.5 ...
 $ Population                     : num  33736494 327582 31731688 3696958 2978599 ...
 $ thinness..1.19.years           : num  17.2 17.5 17.7 17.9 18.2 18.4 18.6 18.8 19 19.2 ...
 $ thinness.5.9.years             : num  17.3 17.5 17.7 18 18.2 18.4 18.7 18.9 19.1 19.3 ...
 $ Income.composition.of.resources: num  0.479 0.476 0.47 0.463 0.454 0.448 0.434 0.433 0.415 0.405 ...
 $ Schooling                      : num  10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4 8.1 ...
```

Testing Correlations Between Variables

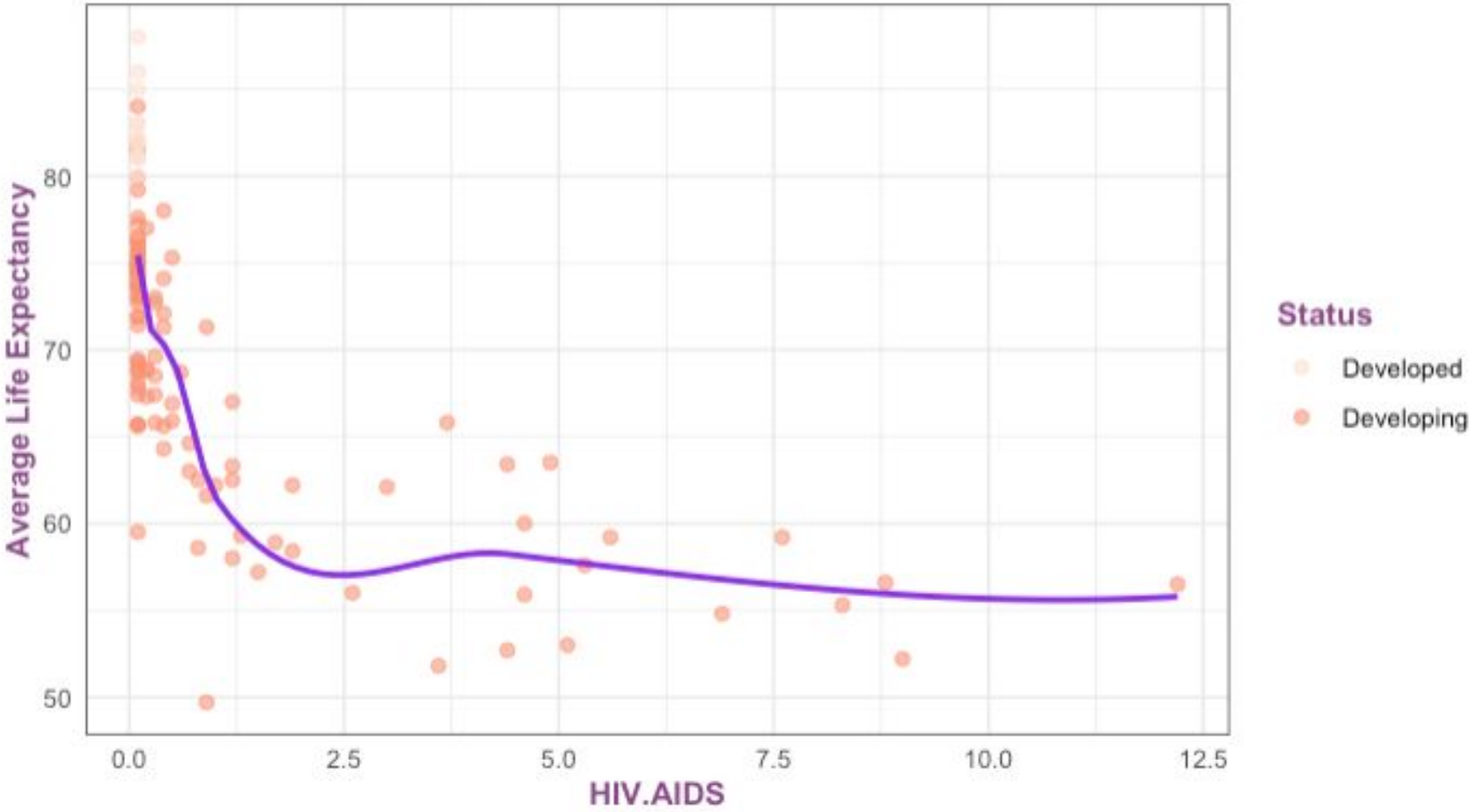

Correlation Between Life Expectancy and BMI
Life.expectancy
80
70
60
50
0
20
40
60
BMI
Status
Developed
Developing


Correlation Between Life Expectancy and Income Composition of Resources
Life.expectancy
80
70
60
50
0.4
0.6
0.8
Income.composition.of.resources
Status
Developed
Developing


Correlation Between Life Expectancy and Schooling
Based on Country Development Level
Average Life Expectancy
80
70
60
50
5
10
15
20
Schooling
Status
Developed
Developing

Correlation Between Life Expectancy and HIV Viruses

Based on Country Development Level



Objective of the Study

Economic Growth and Life Expectancy – Do Wealthier Countries Live Longer?

Objective 1*: Which microeconomic factors affect the life expectancy level of the country.

Objective 2*: Does higher life expectancy lead to higher development level?

*As determined on the WHO cross-country data of 2012.
*Objectives are used as a general guideline to capture the reader. The actual regression results may deviate in an acceptable range from the title.

Linear Regression Model

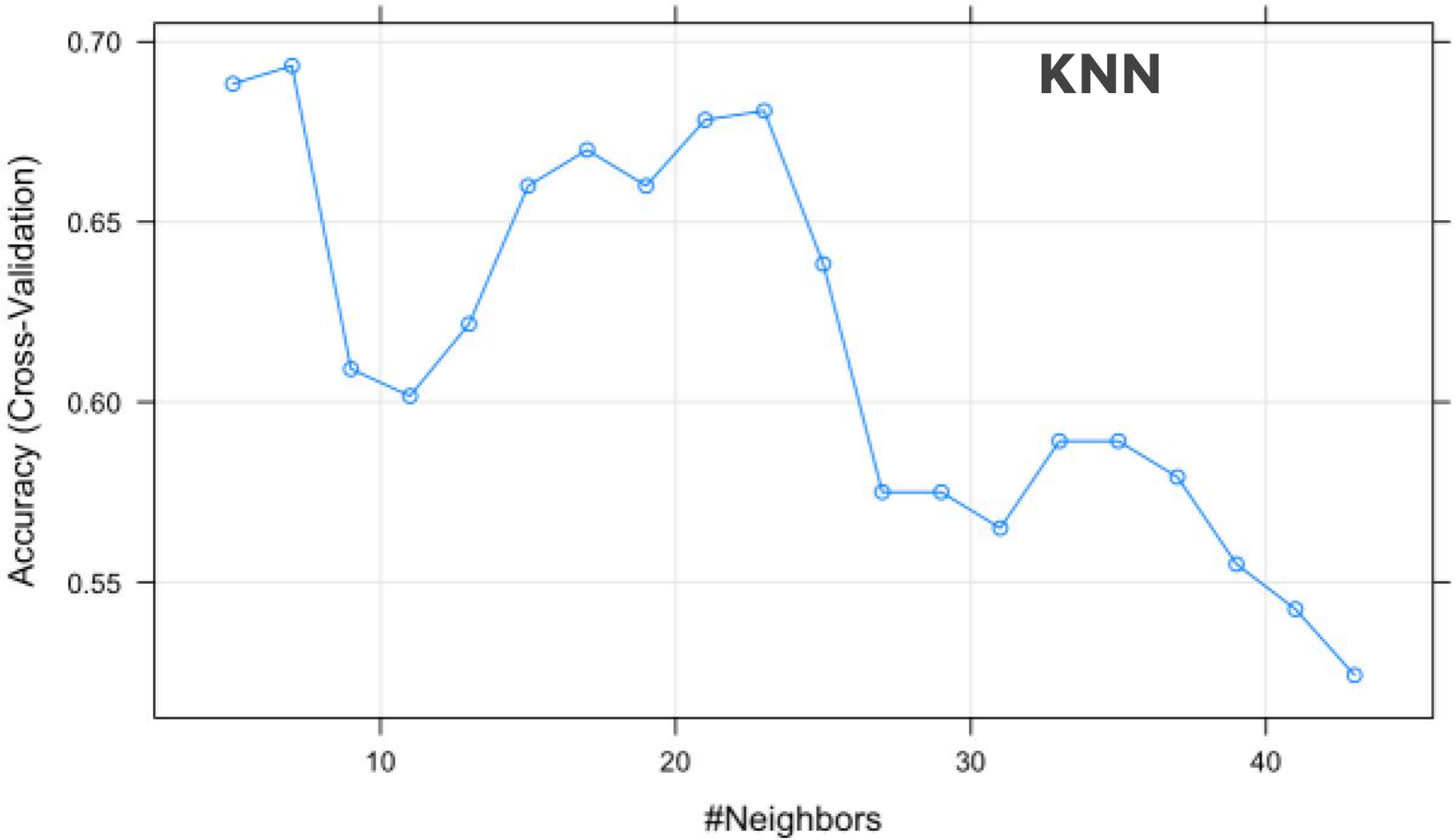
Linear regression model is used to predict the possible relationship between the observed variables and life expectancy.

Dependent variable: Life Expectancy

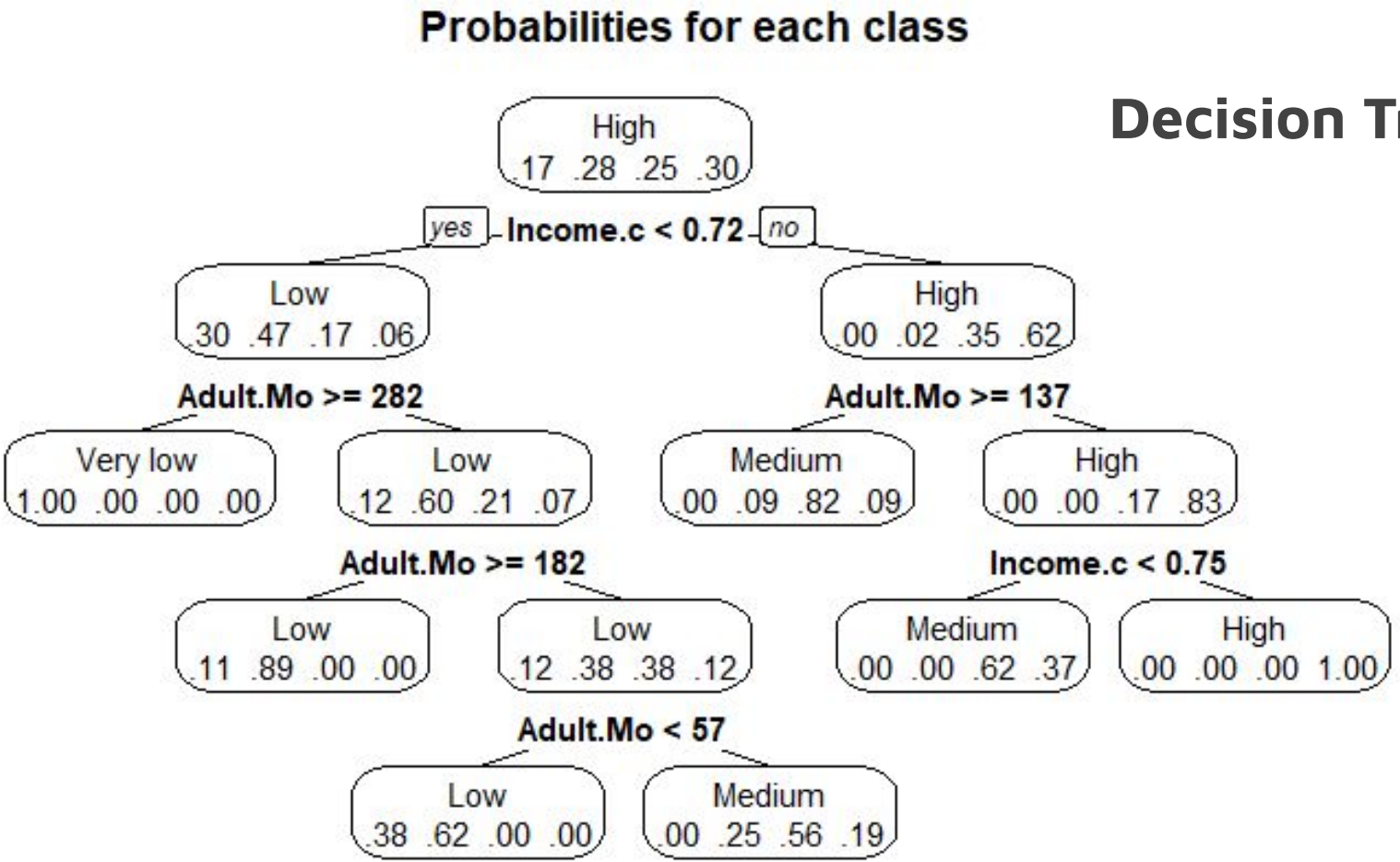
```
lm(formula = Life.expectancy ~ Income.composition.of.resources +
    Adult.Mortality + HIV.AIDS, data = life_expect)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8579 -2.0067 -0.1016  1.8318  9.4897

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      48.344289   1.882858  25.676  < 2e-16 ***
Income.composition.of.resources 37.543898   2.312278  16.237  < 2e-16 ***
Adult.Mortality                  -0.012010   0.003324  -3.613 0.000437 ***
HIV.AIDS                         -1.006020   0.151516  -6.640 8.62e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.09 on 125 degrees of freedom
Multiple R-squared:  0.8714,    Adjusted R-squared:  0.8683
F-statistic: 282.3 on 3 and 125 DF,  p-value: < 2.2e-16
```

Naive Bayes Model

```
          Reference
Prediction Very low Low Medium High
  Very low        6   1      0    0
  Low             0   6      0    0
  Medium          0   2      5    2
  High            0   1      4    9

Overall Statistics

               Accuracy : 0.7222
                 95% CI : (0.5481, 0.858)
    No Information Rate : 0.3056
    P-Value [Acc > NIR] : 3.229e-07

                  Kappa : 0.6246

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Very low Class: Low Class: Medium Class: High
Sensitivity                    1.0000     0.6000        0.5556      0.8182
Specificity                    0.9667     1.0000        0.8519      0.8000
Pos Pred Value                 0.8571     1.0000        0.5556      0.6429
Neg Pred Value                 1.0000     0.8667        0.8519      0.9091
Prevalence                     0.1667     0.2778        0.2500      0.3056
Detection Rate                 0.1667     0.1667        0.1389      0.2500
Detection Prevalence           0.1944     0.1667        0.2500      0.3889
Balanced Accuracy              0.9833     0.8000        0.7037      0.8091
```


KNN
0.70
0.65
0.60
0.55
Accuracy (Cross-Validation)
10
20
30
40
#Neighbors

Decision Tree



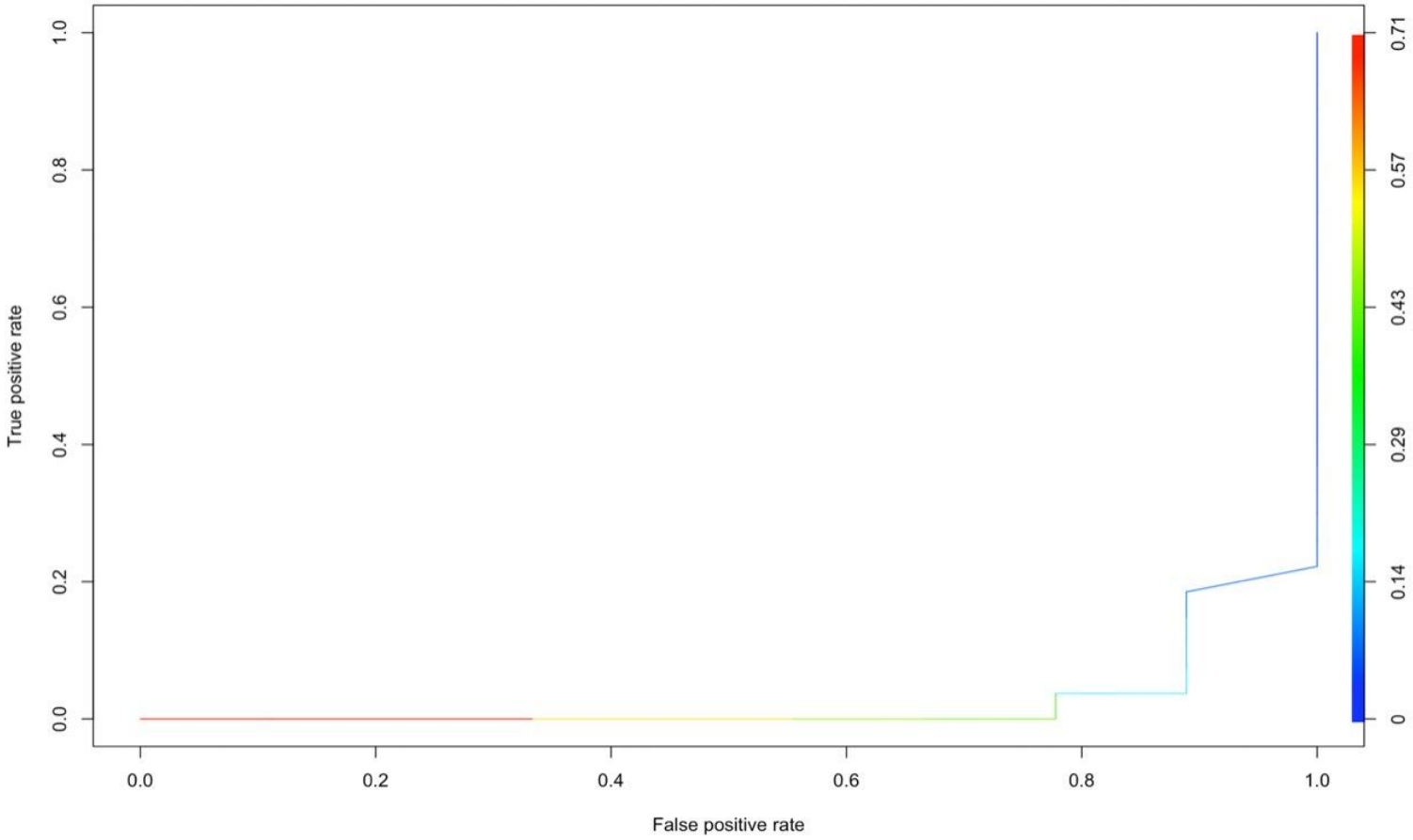
Decision Tree Outcome

```
Confusion Matrix and Statistics

          Reference
Prediction Very low Low Medium High
  Very low        4   1      0    0
  Low             2   8      1    0
  Medium          0   1      6    4
  High            0   0      2    7

Overall Statistics

               Accuracy : 0.6944
                 95% CI : (0.5189, 0.8365)
    No Information Rate : 0.3056
    P-Value [Acc > NIR] : 1.782e-06

                  Kappa : 0.5866

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Very low Class: Low Class: Medium Class: High
Sensitivity                    0.6667     0.8000        0.6667      0.6364
Specificity                    0.9667     0.8846        0.8148      0.9200
Pos Pred Value                 0.8000     0.7273        0.5455      0.7778
Neg Pred Value                 0.9355     0.9200        0.8800      0.8519
Prevalence                     0.1667     0.2778        0.2500      0.3056
Detection Rate                 0.1111     0.2222        0.1667      0.1944
Detection Prevalence           0.1389     0.3056        0.3056      0.2500
Balanced Accuracy              0.8167     0.8423        0.7407      0.7782
```

Logistic Regression Outcome

```
           pr_class_LR_st
            Developed Developing
  Developed          6          3
  Developing        27          0
Confusion Matrix and Statistics

           Reference
Prediction   Developed Developing
  Developed          6         27
  Developing         3          0

               Accuracy : 0.1667
                 95% CI : (0.0637, 0.3281)
    No Information Rate : 0.75
    P-Value [Acc > NIR] : 1

                  Kappa : -0.1765

 Mcnemar's Test P-Value : 2.679e-05

            Sensitivity : 0.6667
            Specificity : 0.0000
         Pos Pred Value : 0.1818
         Neg Pred Value : 0.0000
             Prevalence : 0.2500
         Detection Rate : 0.1667
   Detection Prevalence : 0.9167
      Balanced Accuracy : 0.3333

       'Positive' Class : Developed
```

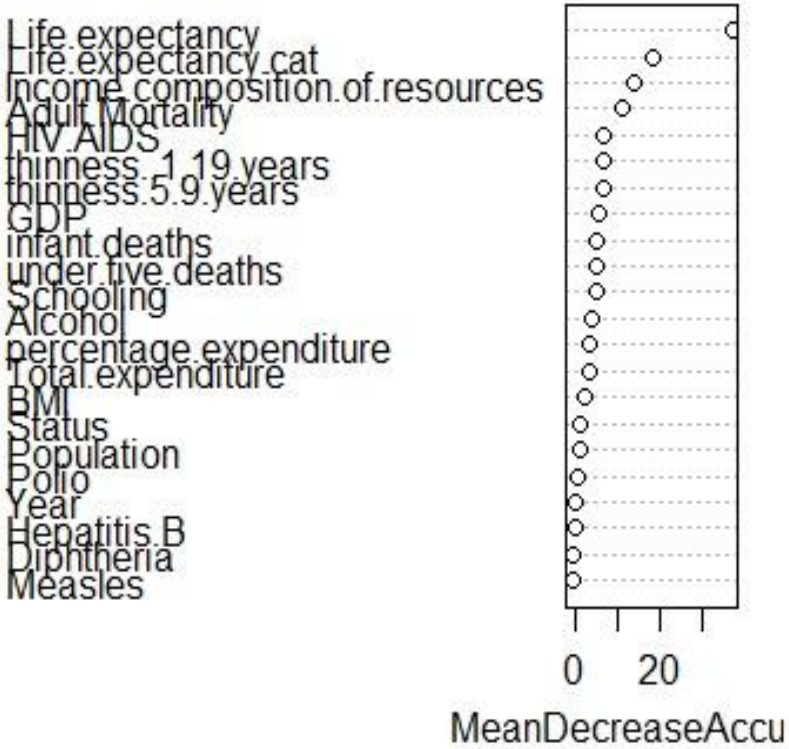
Logistic Regression: Area Under the Curve

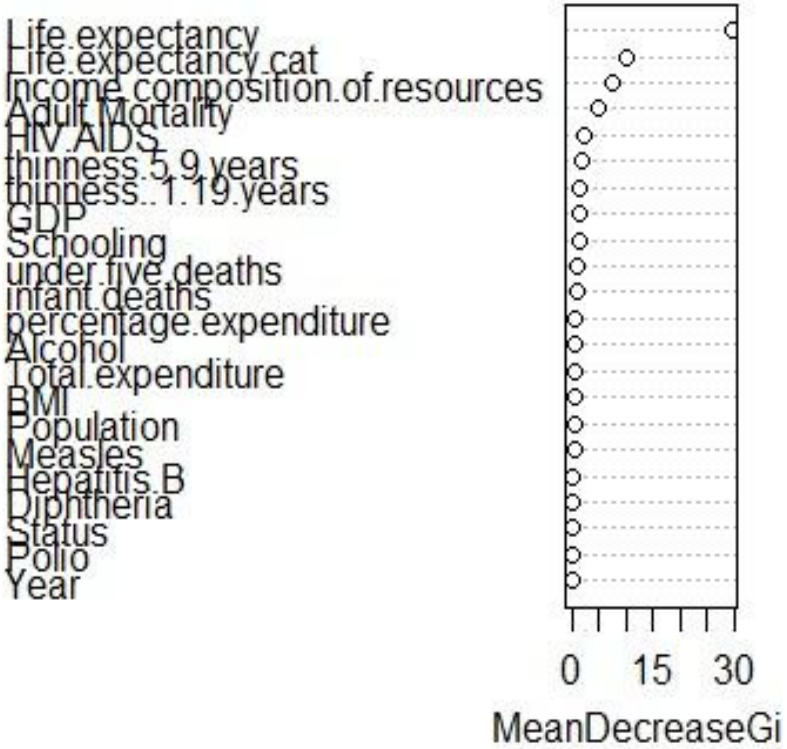


Random Forest

```
Confusion Matrix and Statistics

          Reference
Prediction Very low Low Medium High
  Very low        6   0      0    0
  Low             0  10      0    0
  Medium          0   0     15    0
  High            0   0      1    5

Overall Statistics

               Accuracy : 0.973
                 95% CI : (0.8584, 0.9993)
    No Information Rate : 0.4324
    P-Value [Acc > NIR] : 1.675e-12

                  Kappa : 0.9616

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Very low Class: Low Class: Medium Class: High
Sensitivity                    1.0000     1.0000        0.9375      1.0000
Specificity                    1.0000     1.0000        1.0000      0.9688
Pos Pred Value                 1.0000     1.0000        1.0000      0.8333
Neg Pred Value                 1.0000     1.0000        0.9545      1.0000
Prevalence                     0.1622     0.2703        0.4324      0.1351
Detection Rate                 0.1622     0.2703        0.4054      0.1351
Detection Prevalence           0.1622     0.2703        0.4054      0.1622
Balanced Accuracy              1.0000     1.0000        0.9688      0.9844
```

Importance of variables





Conclusion

- The life expectancy is mostly dependent on HIV aids,Income contribution of resources and Adult Mortality
- The best model was created with Random Forest with an accuracy of 97.3%.
- Life Expectancy was analysed with countries Status. The low accuracy of 16% showed that the country's development status is only 16% varied based on variance of Life Expectancy. Other factors may have higher impact on the country's status.

Torture the data and it will confess to anything!

- Ronald Coase