Thanya Nguyen

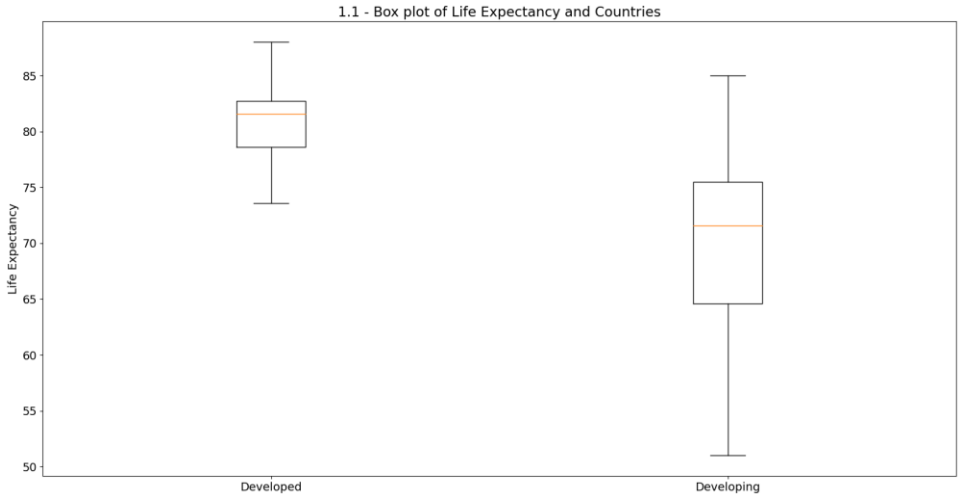
30 August 2023

Intro to Machine Learning

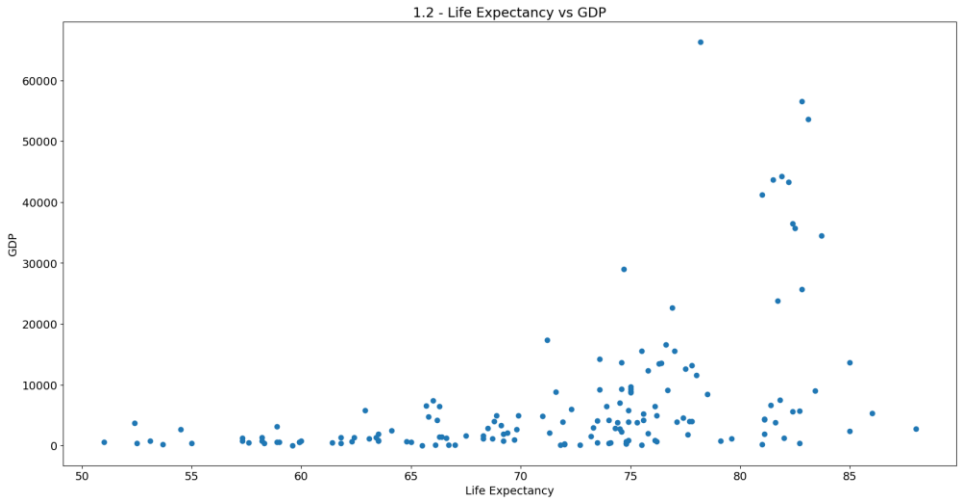
Professor Catherine Schuman

Lab 1: Python Data Manipulation and Visualization and Decision Trees Report

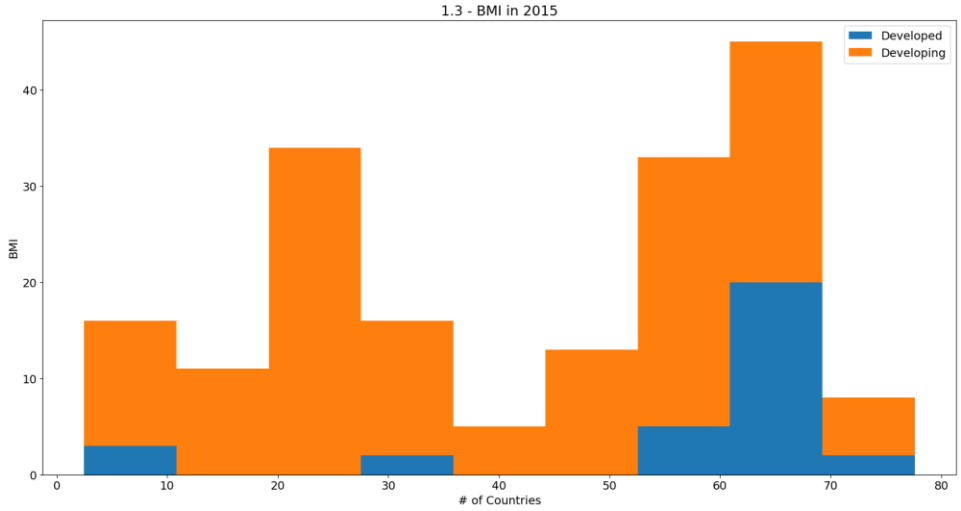
1.1 -



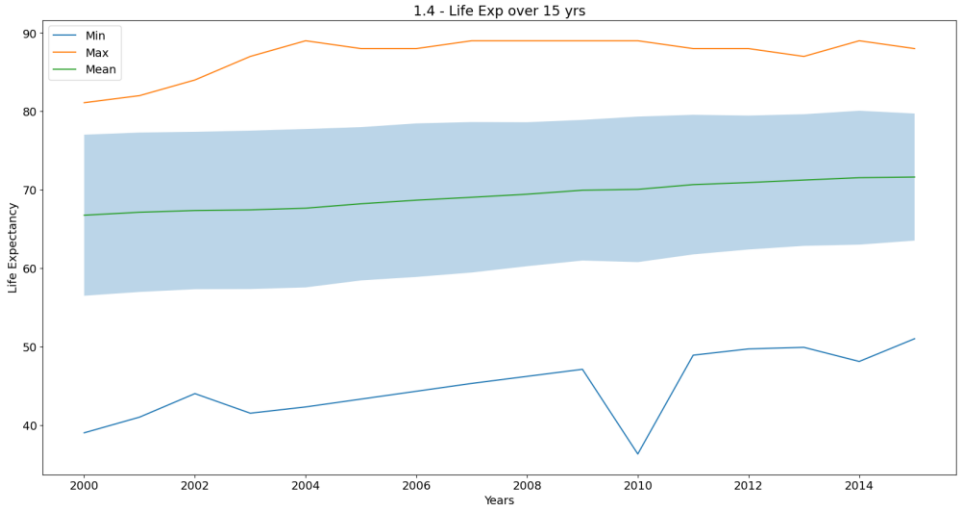
1.2 -



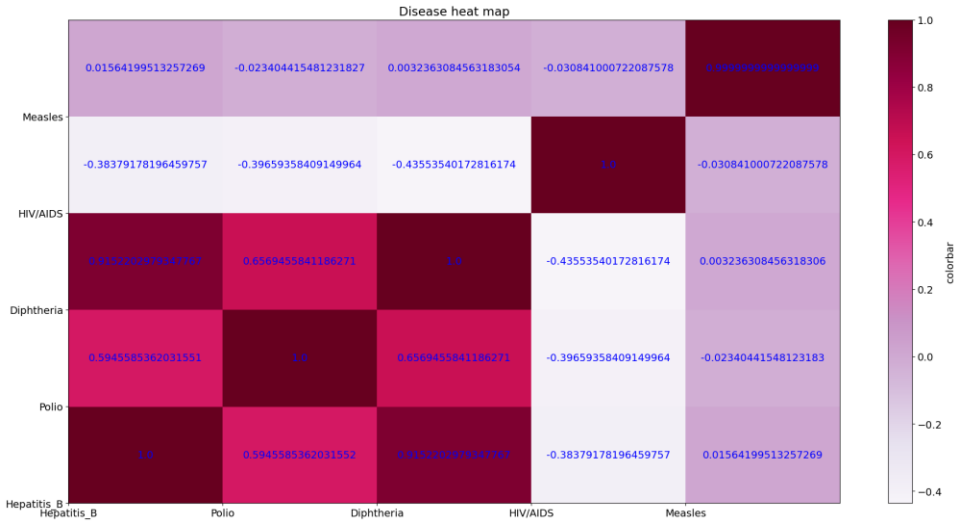
1.3 -


1.3 - BMI in 2015
Developed
Developing
BMI
# of Countries
0
10
20
30
40
50
60
70
80

1.4 -

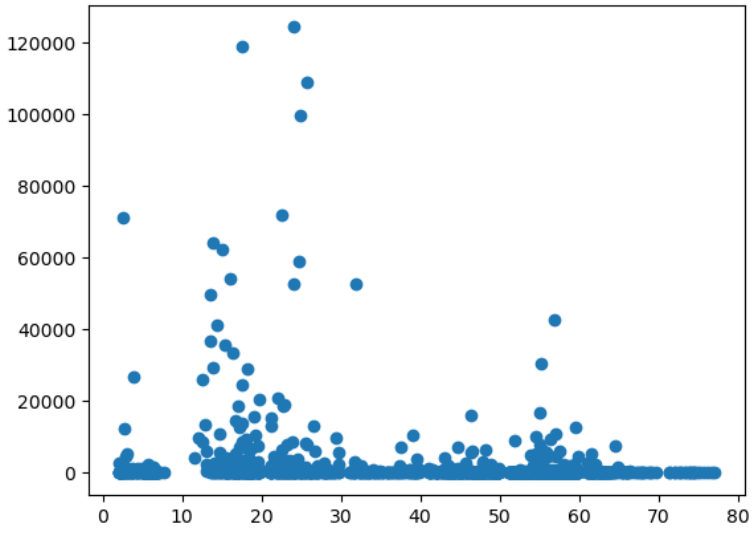

1.4 - Life Exp over 15 yrs
Min
Max
Mean
Life Expectancy
Years
40
50
60
70
80
90
2000
2002
2004
2006
2008
2010
2012
2014

1.5 - I think the most heavily correlated is the Diphtheria because it is closest to the dark pink color.



Lab 1 part 2

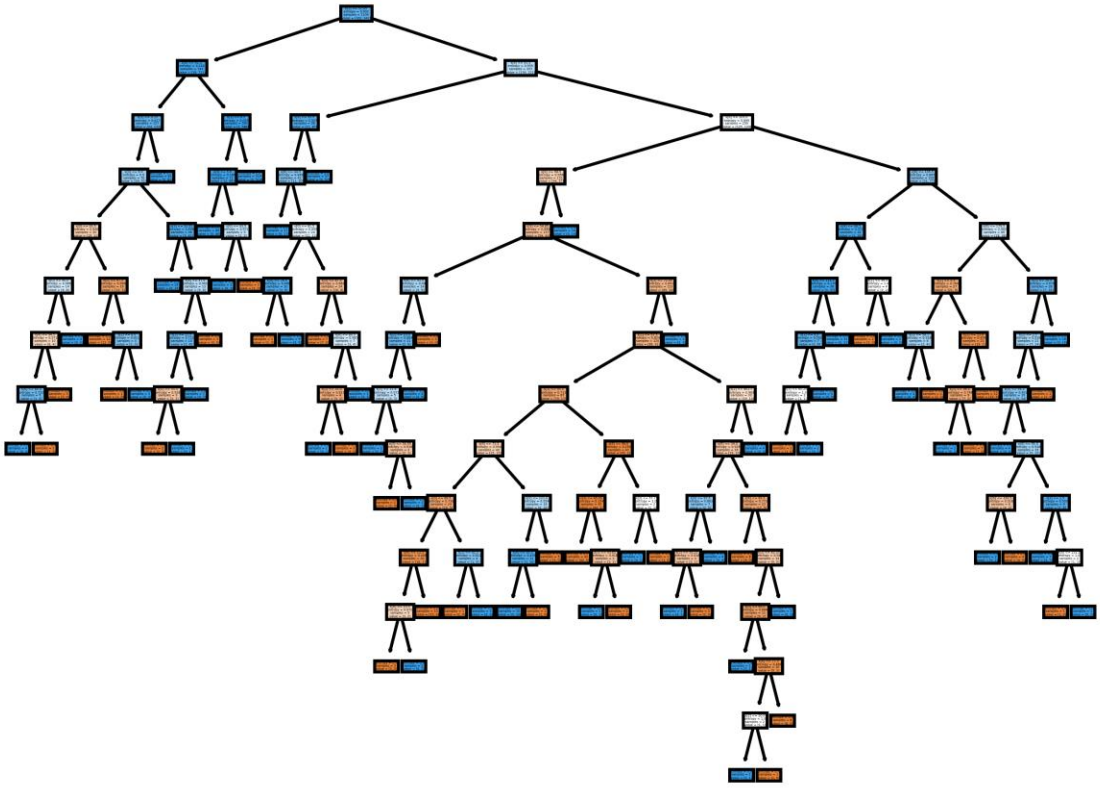
2.1 - The information gain if the split is BMI < 50 is 0.1373.



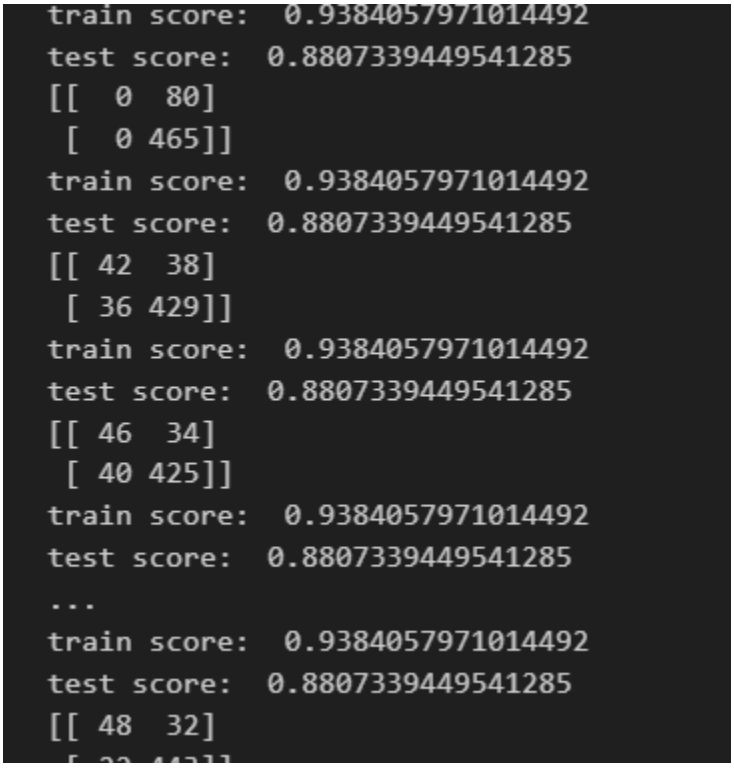
BMI < 50 graph

2.2 -

The accuracy score is 0.89724. The first decision it used to split is the BMI of 52.85. This uses feature 3 and it splits the value from 162 and 942.

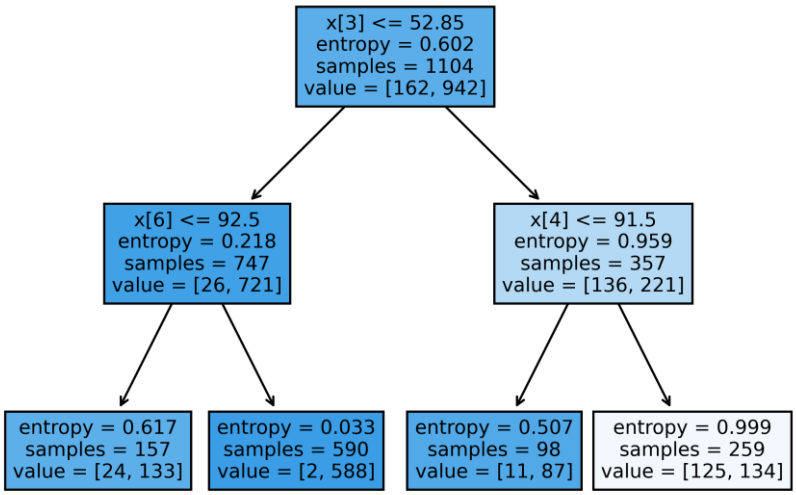


2.3 -



The picture above is a snippet of the full output. If I were to pick between the best score being training or testing, I would pick training score. The training score is consistently higher for every iteration I print out. The higher the information gained, the better the decision made. Thus, the higher scores are the better scores to choose from

2.4 -


x[3] <= 52.85
entropy = 0.602
samples = 1104
value = [162, 942]
x[6] <= 92.5
entropy = 0.218
samples = 747
value = [26, 721]
x[4] <= 91.5
entropy = 0.959
samples = 357
value = [136, 221]
entropy = 0.617
samples = 157
value = [24, 133]
entropy = 0.033
samples = 590
value = [2, 588]
entropy = 0.507
samples = 98
value = [11, 87]
entropy = 0.999
samples = 259
value = [125, 134]

2.5 -

```
[[  0  80]
 [  0 465]]
[[ 42  38]
 [ 36 429]]
[[ 46  34]
 [ 40 425]]
[[ 55  25]
 [ 41 424]]
[[ 54  26]
 [ 41 424]]
[[ 56  24]
 [ 41 424]]
[[ 56  24]
 [ 40 425]]
[[ 50  30]
 [ 32 433]]
[[ 53  27]
 [ 32 433]]
[[ 53  27]
 [ 32 433]]
[[ 49  31]
 [ 22 443]]
[[ 47  33]
 [ 21 444]]
[[ 51  29]
 [ 29 436]]
[[ 44  36]
 [ 22 443]]
[[ 49  31]
 [ 21 444]]
[[ 48  32]
 [ 27 438]]
[[ 49  31]
 [ 23 442]]
[[ 48  32]
 [ 22 443]]
[[ 49  31]
 [ 20 445]]
```

2.6 -

```
decision tree 1

train score: 0.9384057971014492

test score: 0.8807339449541285

decision tree 2

train score: 0.9329710144927537

test score: 0.8770642201834863
```

Looking at the output given above, Tree 1 performed better on Training score and Test score. Although they are close numbers, the split difference played a factor into Tree 1's performance being better than Tree 2 because the calculations are higher. Since the split for Tree 1 is 2 and the other is 30, the higher splits may cause overfitting for the model. The training score for Tree 1 is 0.9384 while Tree 2 is 0.9329. The test score for Tree 1 is 0.8807 while Tree 2 is 0.87706.