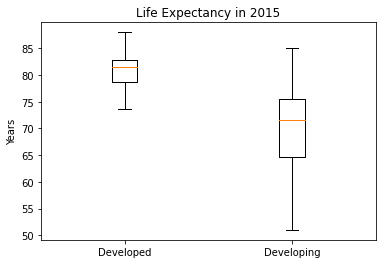
Name: Sarah Huang

Date: 9/8/2023

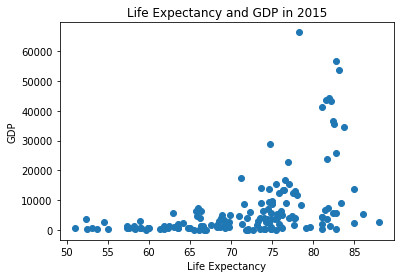
# Lab 1: Python Data Manipulation and Visualization and Decision Trees

**Question 1.1 (5 points): Plot the boxplot for life expectancy in developing and developed countries in 2015. Label y-axis and change xticks to appropriate labels for the boxes (Developing and Developed). You should have two boxes in the plot: one for developing and one for developed.**

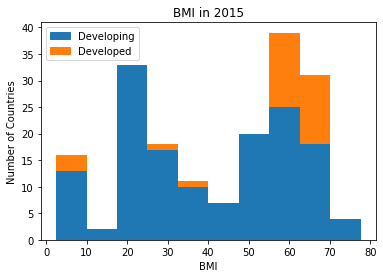


**Question 1.2 (5 points): Create a scatter plot between life expectancy and GDP in 2015.**

**Label axes appropriately**

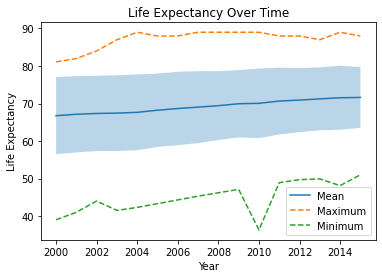


**Question 1.3 (10 points): Create a stacked histogram for BMI in 2015, where the stacked bars are on status (developed and developing). Don’t forget to include a legend.**



**Question 1.4 (10 points): Create a solid line plot showing the average (across all nations)**

**life expectancy changing over time, with the standard deviation above and below shown with filled between plot (use fill between plotting function with alpha = 0.3). Additionally, plot the maximum life expectancy for each year and the minimum life expectancy for each year as dotted line plots over time. Include legend for mean, minimum, and maximum, and don’t forget axes labels.**

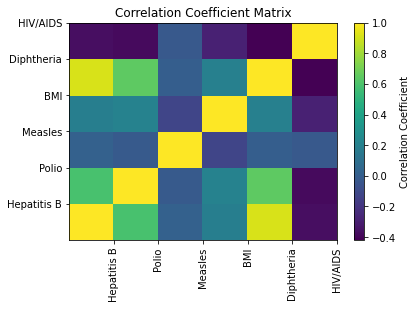


**Question 1.5 (15 points): Extract Hepatitis B, Polio, Measles, BMI, Diphtheria, HIV/AIDS**

**for 2015 and remove the rows with NaN elements (use dropna()). Create the correlation**

**coefficient matrix (use np.corrcoef()) and create a heatmap showing the correlations, using pcolor() and matplotlib. Label the rows and columns with the appropriate diseases. Include a colorbar and a label on the colorbar. Which of these two are the most heavily correlated?**

The two most heavily correlated are Polio and Diphtheria (the light green squares).



**Question 2.1 (10 points): What is the information gain if the split is on BMI < 50? Look**

**at slides from lecture on calculating information gain to use the correct formula.**

Total Count: 1104

Parent Developed Count: 162

Parent Developing Count: 942

Entropy of Parent: 0.6016194553265831

Child 1 Total Count: 701

Child 1 Developed Count: 21

Child 1 Developing Count: 680

Entropy of Parent: 0.6016194553265831

Child 2 Total Count: 403

Child 2 Developed Count: 141

Child 2 Developing Count: 262

Weighted Average Entropy: 0.46422369963331955

Information Gain: 0.13739575569326357

**Question 2.2 (10 points): Create a decision tree classifier with the entropy criterion and fit**

**to the training data. What is the accuracy score of this classifier on the testing set? What**

**is the first decision that is used to split the data (which feature does it use and what value does it split on)?**

The accuracy score of the classifier on the testing set is 0.8990825688073395.

The first decision that is used to split the data is feature\_3 with value 52.85.

Accuracy score on training set: 1.0

Accuracy score on testing set: 0.8990825688073395

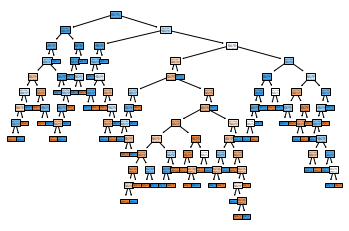
|--- feature\_3 <= 52.85

| |--- feature\_6 <= 92.50

| | |--- feature\_5 <= 0.15

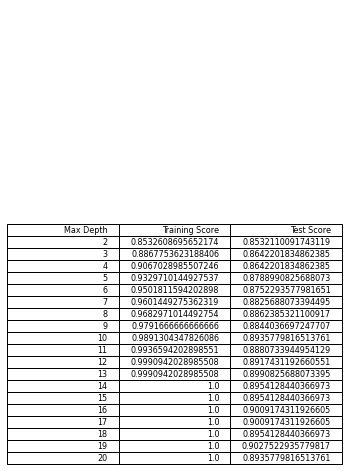
| | | |--- feature\_3 <= 9.90

…

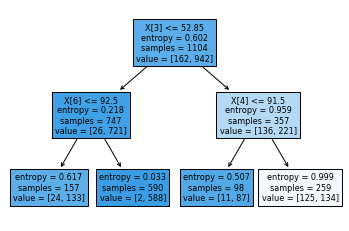


**Question 2.3 (10 points): Create decision trees with max depths from 2 to 20 and calculate the training and testing accuracy for each decision tree and display in a table. Which one is best? Provide a short (1-2 sentence) explanation for why that depth is the best performing.**

The decision tree with a max depth of 19 had the best accuracy test score because it is more complex (more accurate) and doesn’t start overfitting like with the max depth of 20.



**Question 2.4 (5 points): Show the diagram of the decision three with max depth of 2.**

****

**Question 2.5 (5 points): For the best performing tree in Question 2.3 (in terms of testing**

**results), show the confusion matrices that are produced for the training set and testing set.**

Training Confusion Matrix:

[[162 0]

[ 0 942]]

Test Confusion Matrix:

[[ 47 33]

[ 22 443]]

**Question 2.6 (15 points): Create two additional decision trees, where min samples leaf set is set to 10 and min samples split set to 2 and a second where min samples leaf set is set to 10 and min samples split set to 30. Calculate the training and testing scores for each. Which of these performs better? Provide a short explanation (2-3 sentences) for why one performs better than the other.**

The 2 min sample split performed better because there is a lot of data, so splitting is necessary to make the decision tree more complex, cover more patterns, and ultimately become more accurate. A 30 min sample split means there must be 30 samples at an internal node for it to split, but that would make the tree more simple and not capture all the patterns in the data.

2 min sample split

Accuracy score on training set: 0.9384057971014492

Accuracy score on testing set: 0.8807339449541285

30 min sample split

Accuracy score on training set: 0.9329710144927537

Accuracy score on testing set: 0.8770642201834863