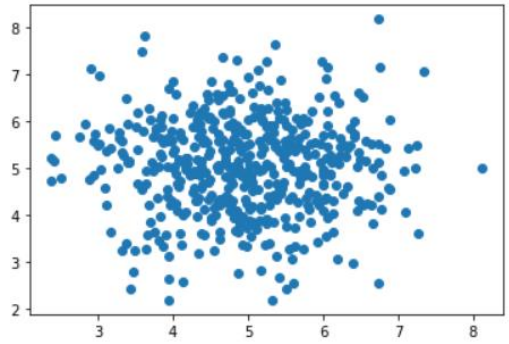
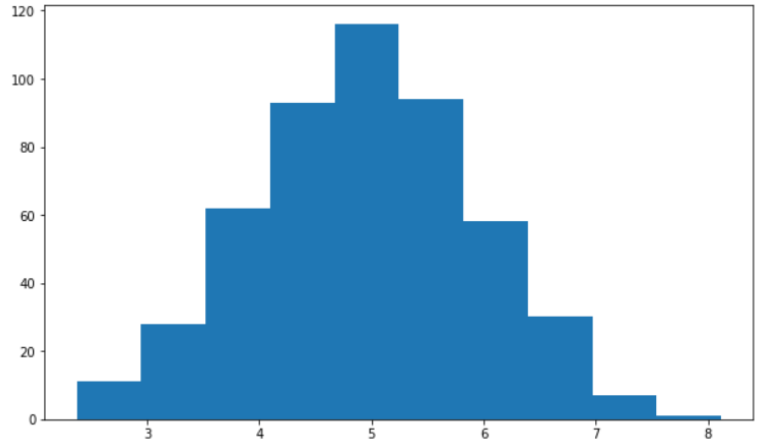
CLL788 Assignment 1

Q1)

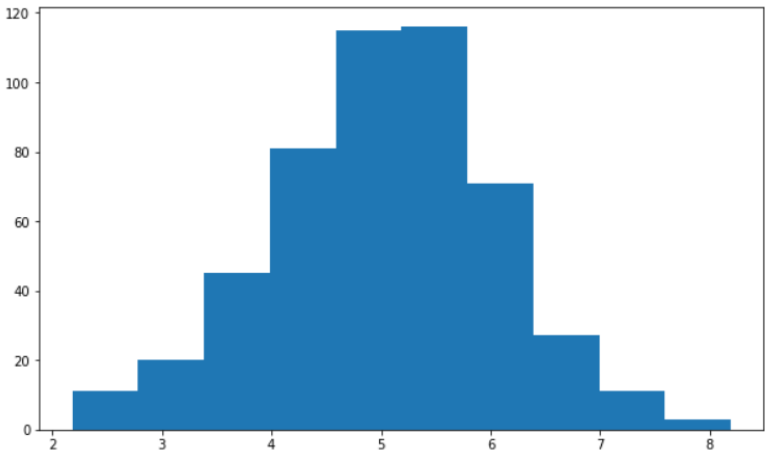
a) Scatter Plot:



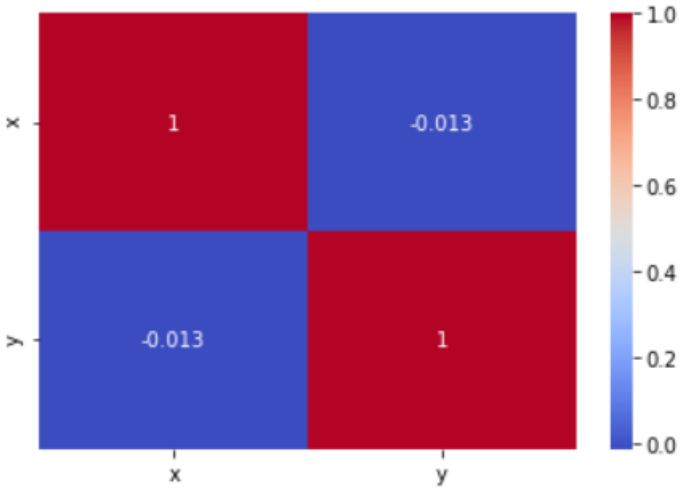
Histogram(X):



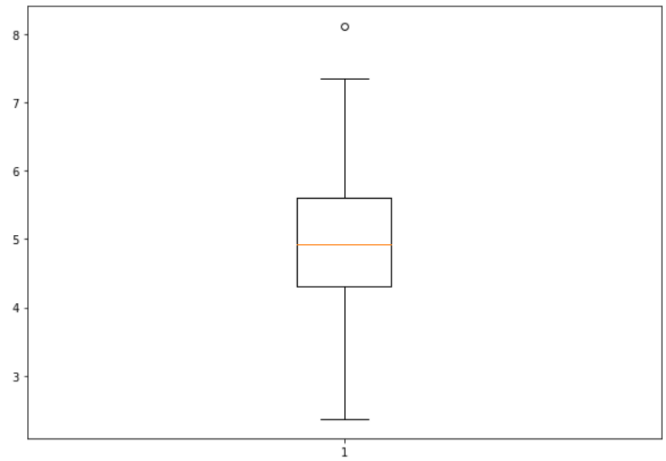
Histogram (Y):



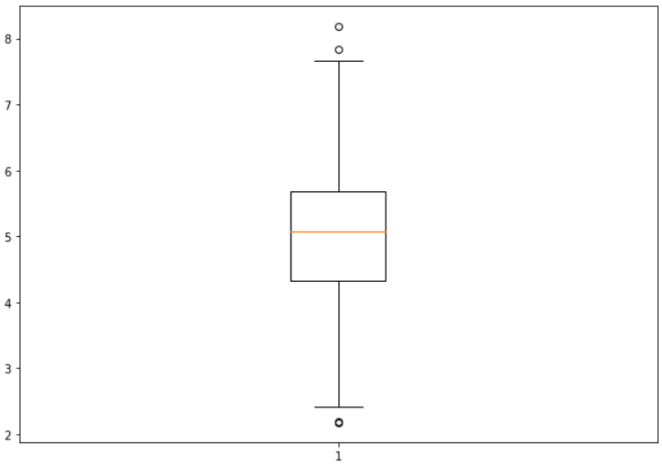
Heat Map:



Box Plot(X):

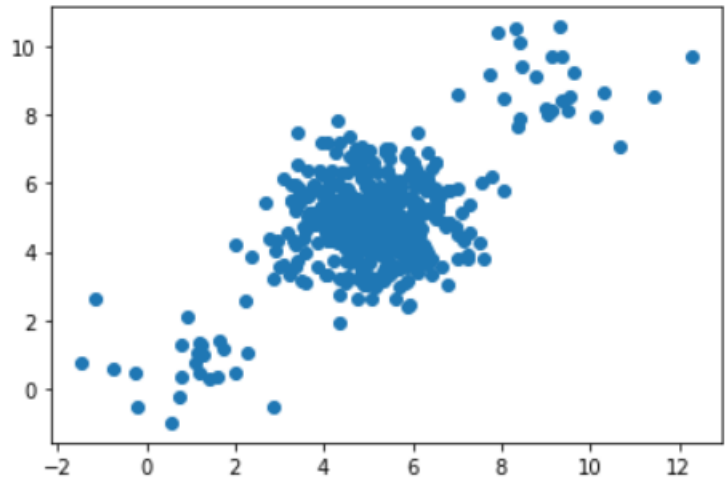


Box Plot(Y):

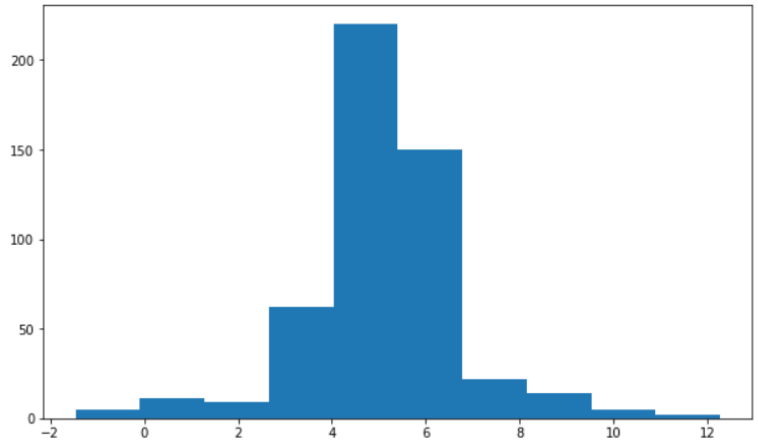


b)

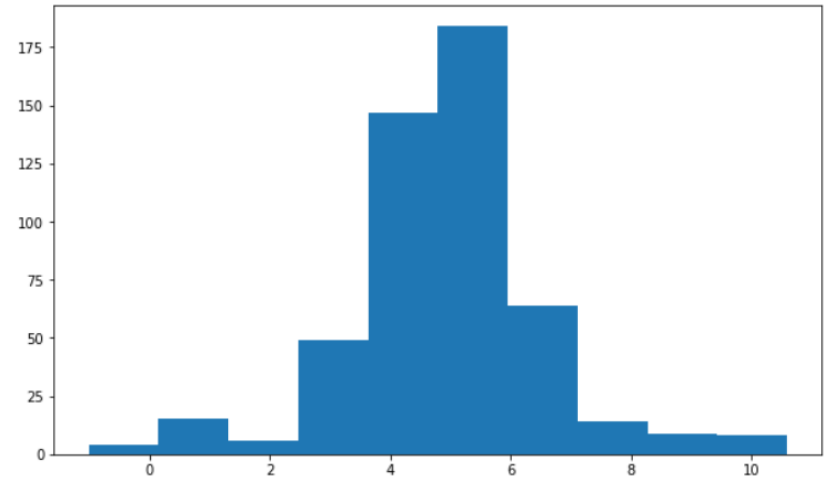
Scatter Plot:



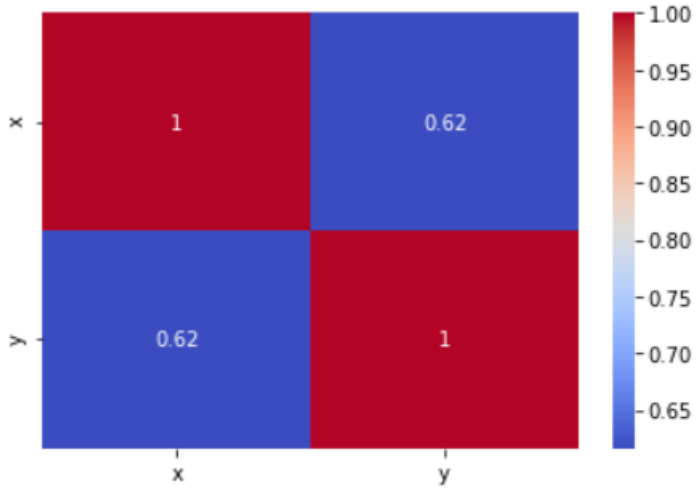
Histogram(X):



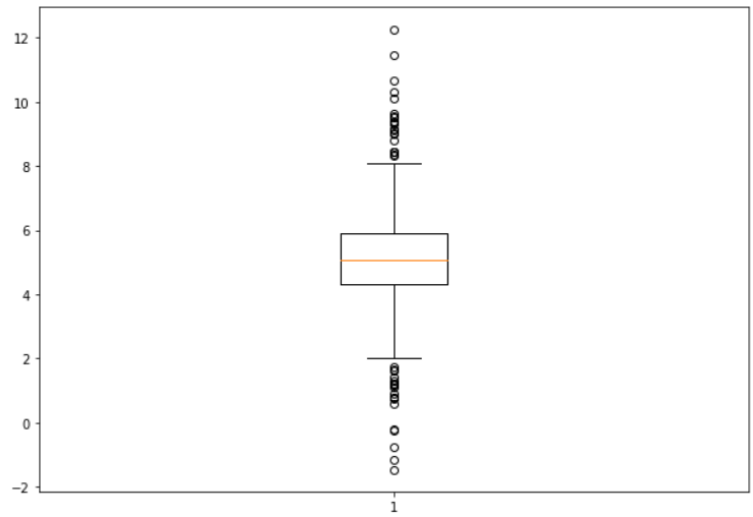
Histogram(Y):



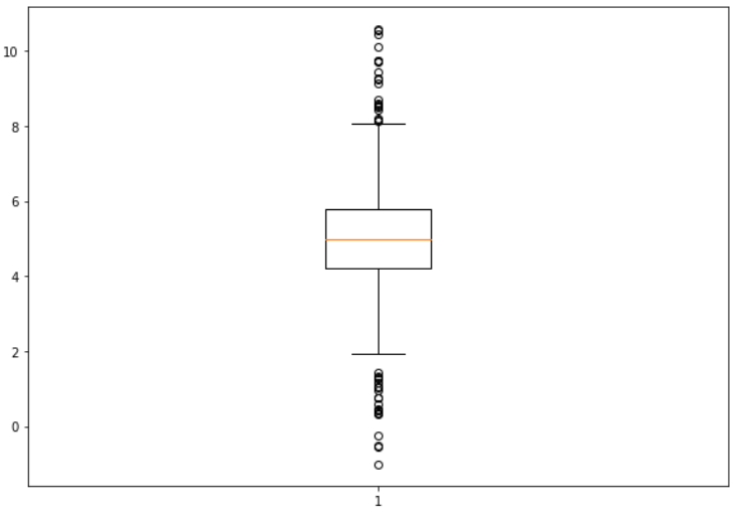
Heat Map:



Box Plot(X):



Box Plot(Y):



c) Data 1:

```
Statistics
Mean of X:  4.939742918488824
Median of X:  4.924278397765319
Mode of X:  5.621808487464226
SD of X:  0.9858156848198057

Mean of Y:  5.042983531876115
Median of Y:  5.074768390017326
Mode of Y:  3.739319183222699
SD of Y:  1.0071882136675947
```

Data 3:

```
Statistics
Mean of X:  5.082467729231954
Median of X:  5.05487549200244
Mode of X:  4.792464301861226
SD of X:  1.6301021662394792

Mean of Y:  4.9528964245044955
Median of Y:  4.978847078474789
Mode of Y:  5.465293862739062
SD of Y:  1.6215698004456818
```

c) Outliers in X using Standard Deviation(k=3):

10.10393747643514, 10.67798532964827, 12.2670245277286, 11.44965456902049, 10.28782093813039, -1.458402520742969, -1.171853375119153, -0.763132811291513, -0.2198563833130491, -0.2439263484483771

Outliers in Y using Standard Deviation(k=3):

10.43902165955905, 10.55061243676815, 10.58925243952102, 10.11551641326384, -0.5334956955315495, -0.2557338130229054, -1.004100360593079, -0.5528195925039532

Outliers in X using MAD:

10.10393747643514, 9.14841306110033, 9.630313376450168, 9.109233422770801, 9.37144942459267, 9.005201225564582, 9.039507424436444, 10.67798532964827, 9.510497893324324, 12.2670245277286, 9.5456448855292, 9.375609831291062, 8.780482242073463, 11.44965456902049, 9.321302250247287, 10.28782093813039

Outliers in Y using MAD:

9.745989964192137, 9.265264240840457, 10.43902165955905, 9.219071559753377, 9.717401494534483, 8.5540601585995, 10.55061243676815, 9.749970745029252, 9.127858533076473, 8.581245337617629, 8.529730294461498, 10.58925243952102, 8.693157145943976, 9.431652444355333, 10.11551641326384

Q2)

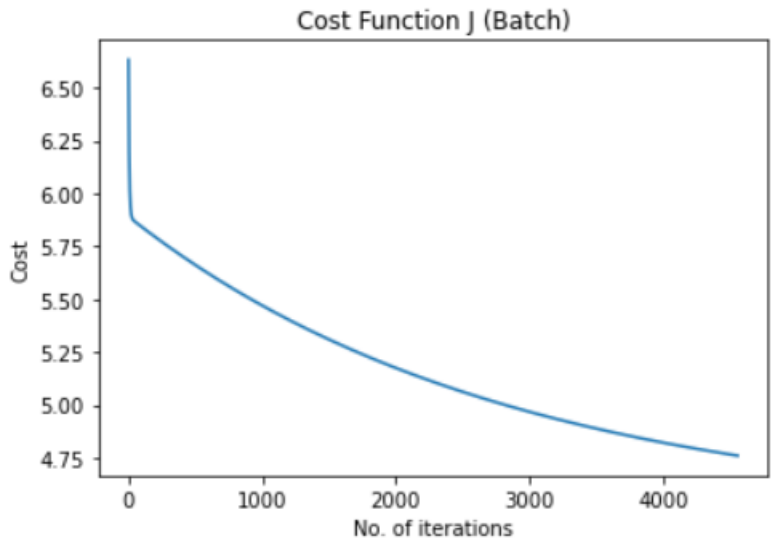
Batch LMS:

Convergence Time: 0.09677481651306152 sec

Theta: [-2.13481414, 1.01558921]

No. of iterations to converge: 4586

Mean Square Error: 9.519459125616603



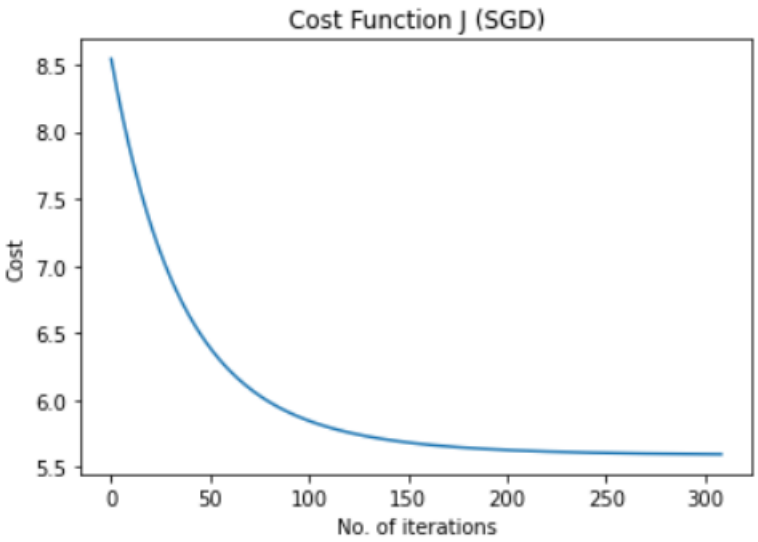
Stochastic LMS:

Convergence Time: 0.09822201728820801 sec

Theta: [-3.80030889, 1.01762938]

No. of iterations to converge: 309

Mean Square Error: 11.189636832591258



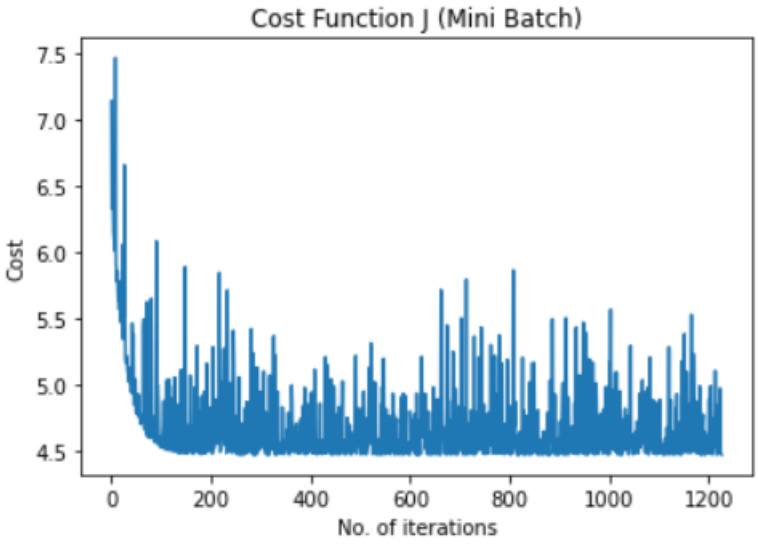
Mini-batch LMS:

Convergence Time: 0.18363738059997559 sec

Theta: [-3.93246316, 1.19570523]

No. of iterations to converge: 1227

Mean Square Error: 8.954070515787446



Least Square Closed Form:

Theta: [-3.91508424, 1.19303364]

Mean Square Error: 8.953942751950358

b) Theta: [0.71349679 0.41111413]

Cost= [7.362915897411117, 6.759007559107173, 6.204754994354164, 5.696070153918476]

Weights: [0.0239600931, 0.000504599365, 0.113718462, 0.639492951]

Prediction for query point: 3.828097461610978

(Computed using code)

c)

Ridge Regression: Mean Square Error= 51.116680052500485

Lasso Regression: Mean Square Error= 9.021411820230476

Elastic Regression: Mean Square Error= 9.029855945191517

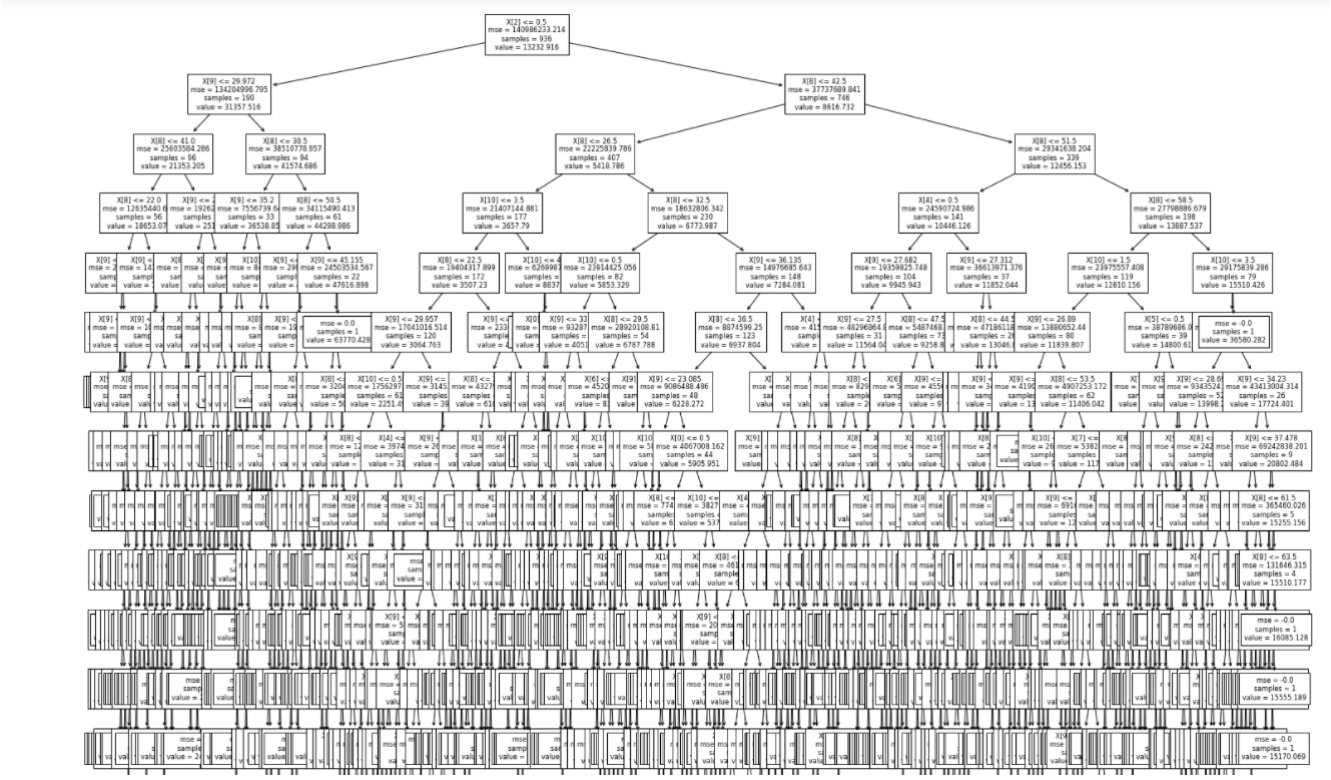
Best Model is **Lasso** as it has least error.

Q3) Theta: [0.0466342, 0.02988982, -0.02543302]

Q4) Using sum-of-the-squares as a criterion to select tree root: Root Node comes as **Smoker** with MSE equals to 140986233.214. With it's left child is **bmi<29.972** and right child as **Age<42.5**

Mean Square Error in train set: 5119427.156394816

Mean Square Error in test set: 36518391.76351633



Q5) For this we must compute cross entropy (amount of uncertainty) and information gain (Cross entropy of a set-Average Entropy Information for that attribute)

The dataset consists of 14 samples in which 9 says ‘Yes’ and 5 says ‘No’

Computing the complete cross entropy of dataset:

H(S) = - p(yes) * log2(p(yes)) - p(no) * log2(p(no))

= - (9/14) * log2(9/14) - (5/14) * log2(5/14)

= - (-0.41) - (-0.53)

= 0.94

Now let us compute the Information gain of various features i,e Outlook, Temperature, Humidity and Wind.

For Outlook:

H(Outlook=sunny) = -(2/5) *log (2/5) -(3/5) *log (3/5) =0.971

H(Outlook=rain) = -(3/5) *log (3/5) -(2/5) *log (2/5) =0.971

H(Outlook=overcast) = -(4/4) *log (4/4)-0 = 0

Average Entropy Information for Outlook -

I(Outlook) = p(sunny) * H(Outlook=sunny) + p(rain) * H(Outlook=rain) + p(overcast) * H(Outlook=overcast)

= (5/14) *0.971 + (5/14) *0.971 + (4/14)*0

= 0.693

Information Gain = H(S) - I(Outlook)

= 0.94 - 0.693

= 0.247

For Temperature:

H(Temperature=hot) = -(2/4) *log(2/4) -(2/4)*log(2/4) = 1

H(Temperature=cool) = -(3/4) *log (3/4) -(1/4) *log(1/4) = 0.811

H(Temperature=mild) = -(4/6) *log (4/6) -(2/6)*log(2/6) = 0.9179

Average Entropy Information for Temperature -

I(Temperature) = p(hot)*H(Temperature=hot) + p(mild)*H(Temperature=mild) + p(cool)*H(Temperature=cool)

= (4/14) *1 + (6/14) *0.9179 + (4/14)*0.811

= 0.9108

Information Gain = H(S) - I(Temperature)

= 0.94 - 0.9108

= 0.0292

For Humidity:

H(Humidity=high) = -(3/7) *log(3/7) -(4/7)*log(4/7) = 0.983

H(Humidity=normal) = -(6/7) *log(6/7) -(1/7)*log(1/7) = 0.591

Average Entropy Information for Humidity -

I(Humidity) = p(high)*H(Humidity=high) + p(normal)*H(Humidity=normal)

= (7/14) *0.983 + (7/14)*0.591

= 0.787

Information Gain = H(S) - I(Humidity)

= 0.94 - 0.787

= 0.153

For Wind:

H(Wind=weak) = -(6/8) *log(6/8)-(2/8)*log(2/8) = 0.811

H(Wind=strong) = -(3/6) *log(3/6)-(3/6)*log(3/6) = 1

Average Entropy Information for Wind -

I(Wind) = p(weak)*H(Wind=weak) + p(strong)*H(Wind=strong)

= (8/14) *0.811 + (6/14)*1

= 0.892

Information Gain = H(S) - I(Wind)

= 0.94 - 0.892

= 0.048

We can see that maximum entropy gain was found in **Outlook**. So, it becomes the root node.

We can also see that when Outlook=Overcast, our target is always ‘Yes’. So, it is a pure class. Now, we will remove the rows with Outlook as Overcast and repeat the same procedure with the rows contains Sunny and Rain.

Complete entropy of Sunny:

H(S) = - p(yes) * log2(p(yes)) - p(no) * log2(p(no))

= - (2/5) * log2(2/5) - (3/5) * log2(3/5)

= 0.971

Now we will go for other attributes:

For temperature:

H (Sunny, Temperature=hot) = -0-(2/2) *log (2/2) = 0

H (Sunny, Temperature=cool) = -(1) *log (1)- 0 = 0

H (Sunny, Temperature=mild) = -(1/2) *log (1/2) -(1/2) *log (1/2) = 1

Average Entropy Information for Temperature:

I(Sunny, Temperature) = p(Sunny, hot)*H(Sunny, Temperature=hot) + p(Sunny, mild)*H(Sunny, Temperature=mild) + p(Sunny, cool)*H(Sunny, Temperature=cool)

= (2/5) *0 + (1/5) *0 + (2/5) *1

= 0.4

Information Gain = H(Sunny) - I(Sunny, Temperature)

= 0.971 - 0.4

= 0.571

For Humidity:

H (Sunny, Humidity=high) = - 0 - (3/3)*log(3/3) = 0

H (Sunny, Humidity=normal) = -(2/2)*log(2/2)-0 = 0

Average Entropy Information for Humidity -

I(Sunny, Humidity) = p(Sunny, high)*H(Sunny, Humidity=high) + p(Sunny, normal)*H(Sunny, Humidity=normal)

= (3/5)*0 + (2/5)*0

= 0

Information Gain = H(Sunny) - I(Sunny, Humidity)

= 0.971 - 0

= 0.971

For wind:

H(Sunny, Wind=weak) = -(1/3)*log(1/3)-(2/3)*log(2/3) = 0.918

H(Sunny, Wind=strong) = -(1/2)*log(1/2)-(1/2)*log(1/2) = 1

Average Entropy Information for Wind:

I(Sunny, Wind) = p(Sunny, weak)*H(Sunny, Wind=weak) + p(Sunny, strong)*H(Sunny, Wind=strong)

= (3/5) *0.918 + (2/5)*1

= 0.9508

Information Gain = H(Sunny) - I(Sunny, Wind)

= 0.971 - 0.9508

= 0.0202

We can see that humidity has maximum information gain, so it comes under sunny. Now when outlook is sunny and humidity is high, target is no and when humidity is normal, target is normal. From here we get the leaf nodes.

Now, we will go towards rain attribute:

Complete entropy of Rain is -

H(S) = - p(yes) * log2(p(yes)) - p(no) * log2(p(no))

= - (3/5) * log (3/5) - (2/5) * log (2/5)

= 0.971

Let's see its attributes:

For temperature:

H (Rain, Temperature=cool) = -(1/2)*log(1/2)- (1/2)*log(1/2) = 1

H (Rain, Temperature=mild) = -(2/3)*log(2/3)-(1/3)*log(1/3) = 0.918

Average Entropy Information for Temperature:

I(Rain, Temperature) = p(Rain, mild)*H(Rain, Temperature=mild) + p(Rain, cool)*H(Rain, Temperature=cool)

= (2/5)*1 + (3/5)*0.918

= 0.9508

Information Gain = H(Rain) - I(Rain, Temperature)

= 0.971 - 0.9508

= 0.0202

For wind:

H(Wind=weak) = -(3/3) *log(3/3)-0 = 0

H(Wind=strong) = 0-(2/2) *log(2/2) = 0

Average Entropy Information for Wind:

I(Wind) = p (Rain, weak) *H(Rain, Wind=weak) + p(Rain, strong)*H(Rain, Wind=strong)

= (3/5) *0 + (2/5)*0

= 0

Information Gain = H(Rain) - I(Rain, Wind)

= 0.971 - 0

= 0.971

Here the max info gain is from **Wind,** so this becomes the node. Now when Outlook is Rain and Wind is Strong, target is No and when wind is weak target is Yes. So, these becomes the leaf nodes.

The tree looks as following: