**A Step By Step Regression Tree Example**

Decision trees are a powerful way to classify problems. On the other hand, they can be adapted into regression problems, too. Decision trees which built for a data set where the target column could be real number are called **regression trees**. In this case, approaches we’ve applied such as information gain for ID3, gain ratio for C4.5, or gini index for CART won’t work. Still, this is **CART** algorithm.

**Objective**

Decision rules will be found based on standard deviations.

**Data set**

The following data set might be familiar. We’ve used a similar data set in our previous experiments but that one denotes golf playing decisions based on some factors. In other words, the golf playing decision was a nominal target consisting of true or false values. Herein, the target column is the number of golf players and it stores real numbers. We have counted the number of instances for each class when the target was nominal. I mean that we can create branches based on the number of instances for true decisions and false decisions. Here, we cannot count the target values because it is continuous. Instead of counting, we can handle regression problems by switching the metric to standard deviation.

| **Day** | **Outlook** | **Temp.** | **Humidity** | **Wind** | **Golf Players**  **(Hours play)** |
| --- | --- | --- | --- | --- | --- |
| 1 | Sunny | Hot | High | Weak | 25 |
| 2 | Sunny | Hot | High | Strong | 30 |
| 3 | Overcast | Hot | High | Weak | 46 |
| 4 | Rain | Mild | High | Weak | 45 |
| 5 | Rain | Cool | Normal | Weak | 52 |
| 6 | Rain | Cool | Normal | Strong | 23 |
| 7 | Overcast | Cool | Normal | Strong | 43 |
| 8 | Sunny | Mild | High | Weak | 35 |
| 9 | Sunny | Cool | Normal | Weak | 38 |
| 10 | Rain | Mild | Normal | Weak | 46 |
| 11 | Sunny | Mild | Normal | Strong | 48 |
| 12 | Overcast | Mild | High | Strong | 52 |
| 13 | Overcast | Hot | Normal | Weak | 44 |
| 14 | Rain | Mild | High | Strong | 30 |

**Standard deviation**

Golf players = {25, 30, 46, 45, 52, 23, 43, 35, 38, 46, 48, 52, 44, 30}

Average of golf players =

(25 + 30 + 46 + 45 + 52 + 23 + 43 + 35 + 38 + 46 + 48 + 52 + 44 + 30)**/14** = 39.78

Standard deviation of golf players =

**√**[( (25 – 39.78)2 + (30 – 39.78)2 + (46 – 39.78)2 + … + (30 – 39.78)2 )/14] = **9.32**

**Outlook**

Outlook can be sunny, overcast and rain. We need to calculate standard deviation of golf players for all of these outlook candidates.

**Sunny outlook**

| Day | Outlook | Temp. | Humidity | Wind | Golf Players |
| --- | --- | --- | --- | --- | --- |
| 1 | Sunny | Hot | High | Weak | 25 |
| 2 | Sunny | Hot | High | Strong | 30 |
| 8 | Sunny | Mild | High | Weak | 35 |
| 9 | Sunny | Cool | Normal | Weak | 38 |
| 11 | Sunny | Mild | Normal | Strong | 48 |

Golf players for sunny outlook = {25, 30, 35, 38, 48}

Average of golf players for sunny outlook = (25+30+35+38+48)/5 = 35.2

Standard deviation of golf players for sunny outlook

= √(((25 – 35.2)2 + (30 – 35.2)2 + … )/5) = 7.78

**Overcast outlook**

| Day | Outlook | Temp. | Humidity | Wind | Golf Players |
| --- | --- | --- | --- | --- | --- |
| 3 | Overcast | Hot | High | Weak | 46 |
| 7 | Overcast | Cool | Normal | Strong | 43 |
| 12 | Overcast | Mild | High | Strong | 52 |
| 13 | Overcast | Hot | Normal | Weak | 44 |

Golf players for overcast outlook = {46, 43, 52, 44}

Average of golf players for overcast outlook = (46 + 43 + 52 + 44)/4 = 46.25

Standard deviation of golf players for overcast outlook = √(((46-46.25)2+(43-46.25)2+…)= 3.49

**Rainy outlook**

| Day | Outlook | Temp. | Humidity | Wind | Golf Players |
| --- | --- | --- | --- | --- | --- |
| 4 | Rain | Mild | High | Weak | 45 |
| 5 | Rain | Cool | Normal | Weak | 52 |
| 6 | Rain | Cool | Normal | Strong | 23 |
| 10 | Rain | Mild | Normal | Weak | 46 |
| 14 | Rain | Mild | High | Strong | 30 |

Golf players for overcast outlook = {45, 52, 23, 46, 30}

Average of golf players for overcast outlook = (45+52+23+46+30)/5 = 39.2

Standard deviation of golf players for rainy outlook

= √(((45 – 39.2)2+(52 – 39.2)2+…)/5)=10.87

**Summarizing standard deviations for the outlook feature**

| Outlook | Stdev of Golf Players | Instances |
| --- | --- | --- |
| Overcast | 3.49 | 4 |
| Rain | 10.87 | 5 |
| Sunny | 7.78 | 5 |

Weighted standard deviation for outlook

= (4/14)x3.49 + (5/14)x10.87 + (5/14)x7.78 = 7.66

You might remember that we have calculated the global standard deviation of golf players 9.32 in previous steps. Standard deviation reduction is the difference of the global standard deviation and standard deviation for current features. In this way, maximized standard deviation reduction will be the decision node.

Standard deviation reduction for outlook = 9.32 – 7.66 = 1.66

**Temperature**

Temperature can be hot, cool or mild. We will calculate standard deviations for those candidates.

**Hot temperature**

| Day | Outlook | Temp. | Humidity | Wind | Golf Players |
| --- | --- | --- | --- | --- | --- |
| 1 | Sunny | Hot | High | Weak | 25 |
| 2 | Sunny | Hot | High | Strong | 30 |
| 3 | Overcast | Hot | High | Weak | 46 |
| 13 | Overcast | Hot | Normal | Weak | 44 |

Golf players for hot temperature = {25, 30, 46, 44}

Standard deviation of golf players for hot temperature = 8.95

**Cool temperature**

| Day | Outlook | Temp. | Humidity | Wind | Golf Players |
| --- | --- | --- | --- | --- | --- |
| 5 | Rain | Cool | Normal | Weak | 52 |
| 6 | Rain | Cool | Normal | Strong | 23 |
| 7 | Overcast | Cool | Normal | Strong | 43 |
| 9 | Sunny | Cool | Normal | Weak | 38 |

Golf players for cool temperature = {52, 23, 43, 38}

Standard deviation of golf players for cool temperature = 10.51

**Mild temperature**

| Day | Outlook | Temp. | Humidity | Wind | Golf Players |
| --- | --- | --- | --- | --- | --- |
| 4 | Rain | Mild | High | Weak | 45 |
| 8 | Sunny | Mild | High | Weak | 35 |
| 10 | Rain | Mild | Normal | Weak | 46 |
| 11 | Sunny | Mild | Normal | Strong | 48 |
| 12 | Overcast | Mild | High | Strong | 52 |
| 14 | Rain | Mild | High | Strong | 30 |

Golf players for mild temperature = {45, 35, 46, 48, 52, 30}

Standard deviation of golf players for mild temperature = 7.65

**Summarizing standard deviations for temperature feature**

| Temperature | Stdev of Golf Players | Instances |
| --- | --- | --- |
| Hot | 8.95 | 4 |
| Cool | 10.51 | 4 |
| Mild | 7.65 | 6 |

Weighted standard deviation for temperature = (4/14)x8.95 + (4/14)x10.51 + (6/14)x7.65 = 8.84

Standard deviation reduction for temperature = 9.32 – 8.84 = 0.47

**Humidity**

Humidity is a binary class. It can either be normal or high.

**High humidity**

| Day | Outlook | Temp. | Humidity | Wind | Golf Players |
| --- | --- | --- | --- | --- | --- |
| 1 | Sunny | Hot | High | Weak | 25 |
| 2 | Sunny | Hot | High | Strong | 30 |
| 3 | Overcast | Hot | High | Weak | 46 |
| 4 | Rain | Mild | High | Weak | 45 |
| 8 | Sunny | Mild | High | Weak | 35 |
| 12 | Overcast | Mild | High | Strong | 52 |
| 14 | Rain | Mild | High | Strong | 30 |

Golf players for high humidity = {25, 30, 46, 45, 35, 52, 30}

Standard deviation for golf players for high humidity = 9.36

**Normal humidity**

| Day | Outlook | Temp. | Humidity | Wind | Golf Players |
| --- | --- | --- | --- | --- | --- |
| 5 | Rain | Cool | Normal | Weak | 52 |
| 6 | Rain | Cool | Normal | Strong | 23 |
| 7 | Overcast | Cool | Normal | Strong | 43 |
| 9 | Sunny | Cool | Normal | Weak | 38 |
| 10 | Rain | Mild | Normal | Weak | 46 |
| 11 | Sunny | Mild | Normal | Strong | 48 |
| 13 | Overcast | Hot | Normal | Weak | 44 |

Golf players for normal humidity = {52, 23, 43, 38, 46, 48, 44}

Standard deviation for golf players for normal humidity = 8.73

**Summarizing standard deviations for humidity feature**

| Humidity | Stdev of Golf Player | Instances |
| --- | --- | --- |
| High | 9.36 | 7 |
| Normal | 8.73 | 7 |

Weighted standard deviation for humidity = (7/14)x9.36 + (7/14)x8.73 = 9.04

Standard deviation reduction for humidity = 9.32 – 9.04 = 0.27

**Wind**

Wind is a binary class, too. It can either be Strong or Weak.

**Strong Wind**

| Day | Outlook | Temp. | Humidity | Wind | Golf Players |
| --- | --- | --- | --- | --- | --- |
| 2 | Sunny | Hot | High | Strong | 30 |
| 6 | Rain | Cool | Normal | Strong | 23 |
| 7 | Overcast | Cool | Normal | Strong | 43 |
| 11 | Sunny | Mild | Normal | Strong | 48 |
| 12 | Overcast | Mild | High | Strong | 52 |
| 14 | Rain | Mild | High | Strong | 30 |

Golf players for strong wind= {30, 23, 43, 48, 52, 30}

Standard deviation for golf players for strong wind = 10.59

**Weak Wind**

| 1 | Sunny | Hot | High | Weak | 25 |
| --- | --- | --- | --- | --- | --- |
| 3 | Overcast | Hot | High | Weak | 46 |
| 4 | Rain | Mild | High | Weak | 45 |
| 5 | Rain | Cool | Normal | Weak | 52 |
| 8 | Sunny | Mild | High | Weak | 35 |
| 9 | Sunny | Cool | Normal | Weak | 38 |
| 10 | Rain | Mild | Normal | Weak | 46 |
| 13 | Overcast | Hot | Normal | Weak | 44 |

Golf players for weak wind= {25, 46, 45, 52, 35, 38, 46, 44}

Standard deviation for golf players for weak wind = 7.87

**Summarizing standard deviations for wind feature**

| Wind | Stdev of Golf Player | Instances |
| --- | --- | --- |
| Strong | 10.59 | 6 |
| Weak | 7.87 | 8 |

Weighted standard deviation for wind = (6/14)x10.59 + (8/14)x7.87 = 9.03

Standard deviation reduction for wind = 9.32 – 9.03 = 0.29

So, we’ve calculated standard deviation reduction values for all features. The winner is outlook because it has the highest score.

| Feature | Standard Deviation Reduction |
| --- | --- |
| **Outlook** | **1.66** |
| Temperature | 0.47 |
| Humidity | 0.27 |
| Wind | 0.29 |

We’ll put the outlook decision at the top of the decision tree. Let’s monitor the new sub datasets for the candidate branches of outlook feature.

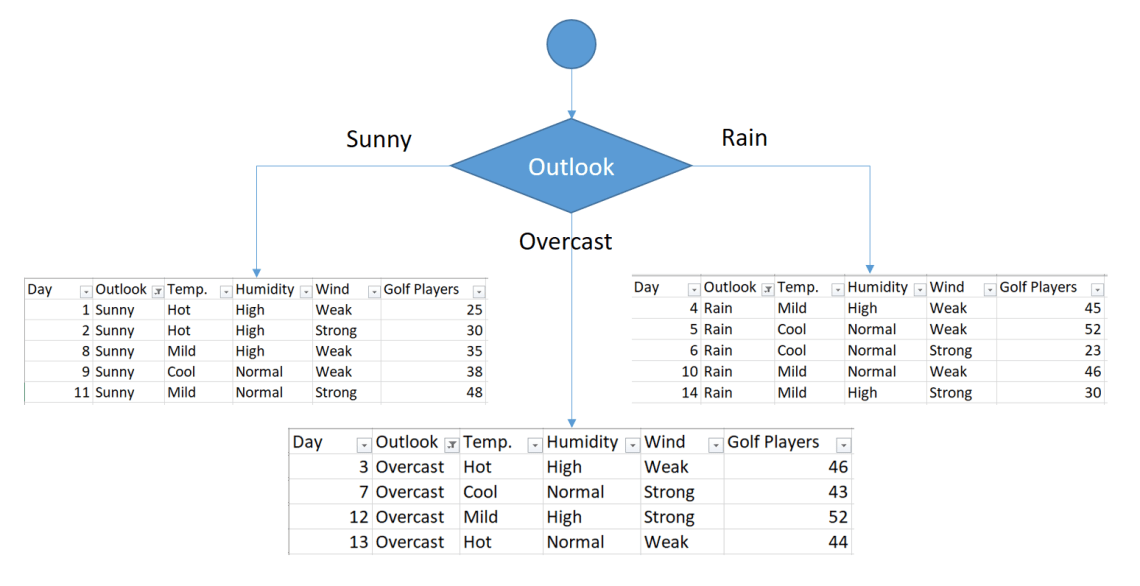


Fig. 1. Putting outlook at the top of the tree

**Sunny Outlook**

| Day | Outlook | Temp. | Humidity | Wind | Golf Players |
| --- | --- | --- | --- | --- | --- |
| 1 | Sunny | Hot | High | Weak | 25 |
| 2 | Sunny | Hot | High | Strong | 30 |
| 8 | Sunny | Mild | High | Weak | 35 |
| 9 | Sunny | Cool | Normal | Weak | 38 |
| 11 | Sunny | Mild | Normal | Strong | 48 |

Golf players for sunny outlook = {25, 30, 35, 38, 48}

Standard deviation for sunny outlook = 7.78

Notice that we will use this standard deviation value as global standard deviation for this sub data set.

**Sunny outlook and Hot Temperature**

| Day | Outlook | Temp. | Humidity | Wind | Golf Players |
| --- | --- | --- | --- | --- | --- |
| 1 | Sunny | Hot | High | Weak | 25 |
| 2 | Sunny | Hot | High | Strong | 30 |

Standard deviation for sunny outlook and hot temperature = 2.5

**Sunny outlook and Cool Temperature**

| Day | Outlook | Temp. | Humidity | Wind | Golf Players |
| --- | --- | --- | --- | --- | --- |
| 9 | Sunny | Cool | Normal | Weak | 38 |

Standard deviation for sunny outlook and cool temperature = 0

**Sunny outlook and Mild Temperature**

| Day | Outlook | Temp. | Humidity | Wind | Golf Players |
| --- | --- | --- | --- | --- | --- |
| 8 | Sunny | Mild | High | Weak | 35 |
| 11 | Sunny | Mild | Normal | Strong | 48 |

Standard deviation for sunny outlook and mild temperature = 6.5

**Summary of standard deviations for temperature feature when outlook is sunny**

| Temperature | Stdev for Golf Players | Instances |
| --- | --- | --- |
| Hot | 2.5 | 2 |
| Cool | 0 | 1 |
| Mild | 6.5 | 2 |

Weighted standard deviation for sunny outlook and temperature

= (2/5)x2.5 + (1/5)x0 + (2/5)x6.5 = 3.6

Standard deviation reduction for sunny outlook and temperature

= 7.78 – 3.6 = 4.18

**Sunny outlook and high humidity**

| Day | Outlook | Temp. | Humidity | Wind | Golf Players |
| --- | --- | --- | --- | --- | --- |
| 1 | Sunny | Hot | High | Weak | 25 |
| 2 | Sunny | Hot | High | Strong | 30 |
| 8 | Sunny | Mild | High | Weak | 35 |

Standard deviation for sunny outlook and high humidity = 4.08

**Sunny outlook and normal humidity**

| Day | Outlook | Temp. | Humidity | Wind | Golf Players |
| --- | --- | --- | --- | --- | --- |
| 9 | Sunny | Cool | Normal | Weak | 38 |
| 11 | Sunny | Mild | Normal | Strong | 48 |

Standard deviation for sunny outlook and normal humidity = 5

**Summarizing standard deviations for humidity feature when outlook is sunny**

| Humidity | Stdev for Golf Players | Instances |
| --- | --- | --- |
| High | 4.08 | 3 |
| Normal | 5.00 | 2 |

Weighted standard deviations for sunny outlook and humidity = (3/5)x4.08 + (2/5)x5 = 4.45

Standard deviation reduction for sunny outlook and humidity = 7.78 – 4.45 = 3.33

**Sunny outlook and Strong Wind**

| Day | Outlook | Temp. | Humidity | Wind | Golf Players |
| --- | --- | --- | --- | --- | --- |
| 2 | Sunny | Hot | High | Strong | 30 |
| 11 | Sunny | Mild | Normal | Strong | 48 |

Standard deviation for sunny outlook and strong wind = 9

**Sunny outlook and Weak Wind**

| Day | Outlook | Temp. | Humidity | Wind | Golf Players |
| --- | --- | --- | --- | --- | --- |
| 1 | Sunny | Hot | High | Weak | 25 |
| 8 | Sunny | Mild | High | Weak | 35 |
| 9 | Sunny | Cool | Normal | Weak | 38 |

Standard deviation for sunny outlook and weak wind = 5.56

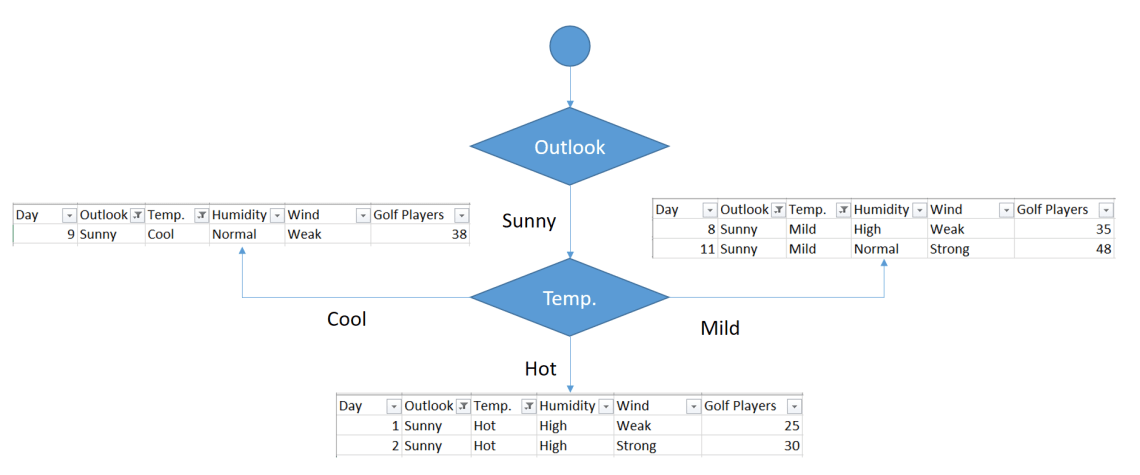
| Wind | Stdev for Golf Players | Instances |
| --- | --- | --- |
| Strong | 9 | 2 |
| Weak | 5.56 | 3 |

Weighted standard deviations for sunny outlook and wind = (2/5)x9 + (3/5)x5.56 = 6.93

Standard deviation reduction for sunny outlook and wind = 7.78 – 6.93 = 0.85

We’ve calculated standard deviation reductions for sunny outlook. The winner is temperature.

| Feature | Standard Deviation Reduction |
| --- | --- |
| Temperature | 4.18 |
| Humidity | 3.33 |
| Wind | 0.85 |

Putting temperature decision at the bottom of sunny outlook

**Pruning**

Cool branch has one instance in its sub data set. We can say that if the outlook is sunny and the temperature is cool, then there would be 38 golf players.

But what about hot branches? There are still 2 instances. Should we add another branch for weak wind and strong wind? No, we should not.

Because this causes over-fitting. We should terminate building branches, for example if there are less than five instances in the sub data set. Or standard deviation of the sub data set can be less than 5% of the entire data set. I prefer to apply the first one. I will terminate the branch if there are less than 5 instances in the current sub data set. If this termination condition is satisfied, then I will calculate the average of the sub data set. This operation is called pruning in decision trees.

**Overcast outlook**

The Overcast outlook branch has already 4 instances in the sub data set. We can terminate building branches for this leaf. Final decision will be average of the following table for an overcast outlook.

| Day | Outlook | Temp. | Humidity | Wind | Golf Players |
| --- | --- | --- | --- | --- | --- |
| 3 | Overcast | Hot | High | Weak | 46 |
| 7 | Overcast | Cool | Normal | Strong | 43 |
| 12 | Overcast | Mild | High | Strong | 52 |
| 13 | Overcast | Hot | Normal | Weak | 44 |

If the outlook is overcast, then there would be (46+43+52+44)/4 = 46.25 golf players.

**Rainy Outlook**

| Day | Outlook | Temp. | Humidity | Wind | Golf Players |
| --- | --- | --- | --- | --- | --- |
| 4 | Rain | Mild | High | Weak | 45 |
| 5 | Rain | Cool | Normal | Weak | 52 |
| 6 | Rain | Cool | Normal | Strong | 23 |
| 10 | Rain | Mild | Normal | Weak | 46 |
| 14 | Rain | Mild | High | Strong | 30 |

We need to find standard deviation reduction values for the rest of the features in the same way for the sub data set above.

Standard deviation for rainy outlook = 10.87

Notice that we will use this value as global standard deviation for this branch in the reduction step.

**Rainy outlook and temperature**

| Temperature | Standard deviation for golf players | instances |
| --- | --- | --- |
| Cool | 14.50 | 2 |
| Mild | 7.32 | 3 |

Weighted standard deviation for rainy outlook and temperature = (2/5)x14.50 + (3/5)x7.32 = 10.19

Standard deviation reduction for rainy outlook and temperature = 10.87 – 10.19 = 0.67

**Rainy outlook and humidity**

| Humidity | Standard deviation for golf players | instances |
| --- | --- | --- |
| High | 7.50 | 2 |
| Normal | 12.50 | 3 |

Weighted standard deviation for rainy outlook and humidity = (2/5)x7.50 + (3/5)x12.50 = 10.50

Standard deviation reduction for rainy outlook and humidity = 10.87 – 10.50 = 0.37

**Rainy outlook and  wind**

| Wind | Standard deviation for golf players | instances |
| --- | --- | --- |
| Weak | 3.09 | 3 |
| Strong | 3.5 | 2 |

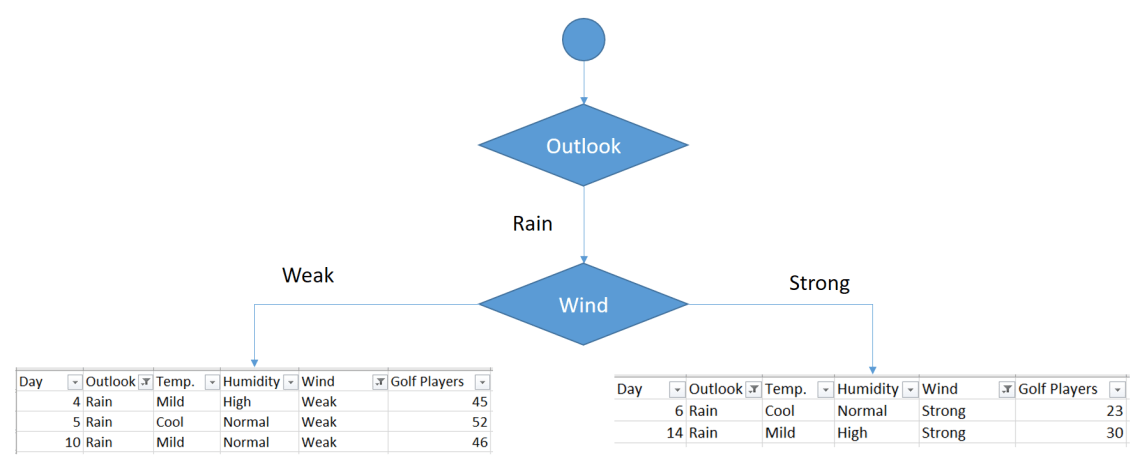
Weighted standard deviation for rainy outlook and wind = (3/5)x3.09 + (2/5)x3.5 = 3.25

Standard deviation reduction for rainy outlook and wind = 10.87 – 3.25 = 7.62

**Summarizing rainy outlook**

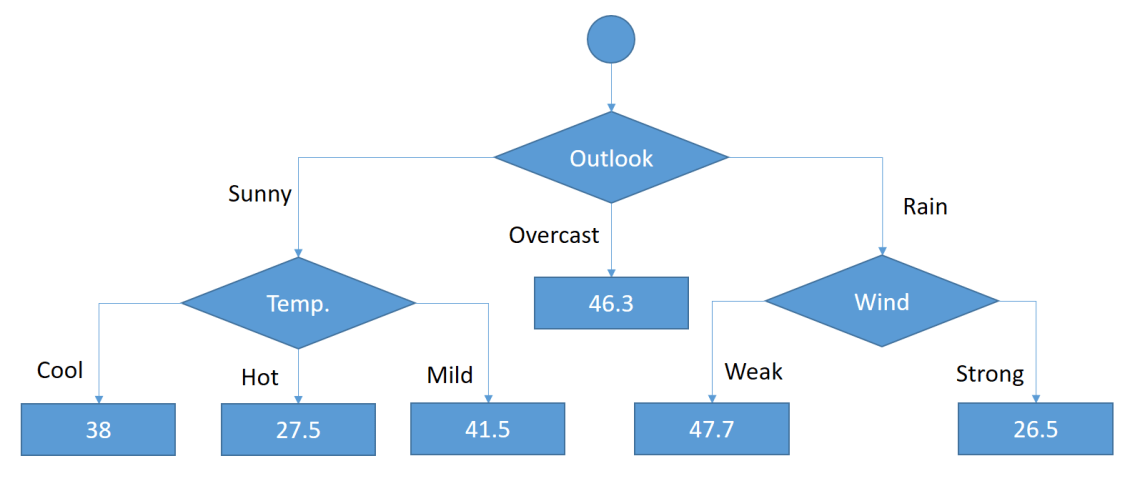
As illustrated below, the winner is the wind feature.

| Feature | Standard deviation reduction |
| --- | --- |
| Temperature | 0.67 |
| Humidity | 0.37 |
| Wind | 7.62 |

Sub data set for rainy outlook

As seen, both branches have items less than 5. Now, we can terminate these leaves based on the termination rule.

So, Final form of the decision tree is demonstrated below.

Final form of the regression tree

**Feature importance**

Decision trees are naturally explainable and interpretable algorithms. They also offer feature importance calculation to understand the model better.

**Conclusion**

So, we have mentioned how to build decision trees for regression problems. Even though decision trees are powerful ways to classify problems, they can be adapted into regression problems as mentioned. Regression trees tend to over-fit much more than classification trees. Termination rule should be tuned carefully to avoid over-fitting.

Ref. <https://www.youtube.com/watch?v=RCkekrLIP_4>