Entropy in Decision Tree
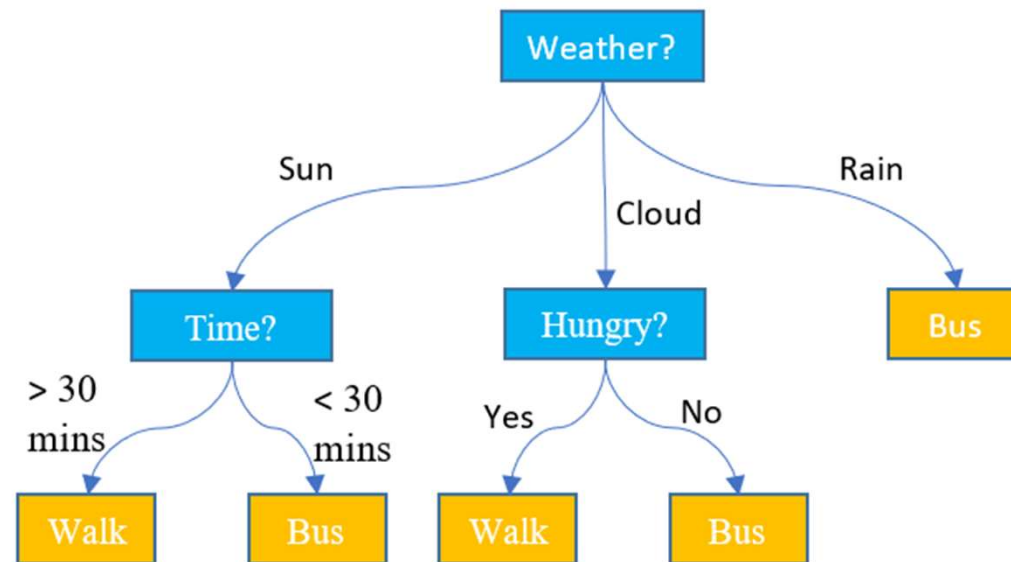
Getting Started with ML

Decisions Tree

```
model.fit(X_train, y_train)
```

```
DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='gini',
                       max_depth=None, max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort='deprecated',
                       random_state=None, splitter='best')
```

Entropy

- Entropy is a measure of purity.
- For Decision Tree it will tell us where to split.
- It is a prerequisite for Information Gain.

Formula for Entropy

$$Entropy\ (Situation) = \sum_{i=1}^{c} -p_i log_2 p_i$$

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Sunny	Mild	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Sunny	Mild	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No

$$Entropy\ (Situation) = \sum_{i=1}^{c} -p_i log_2 p_i$$

$$Entropy\ (Play\ Golf) = \sum_{i=1}^{c} -p_i log_2 p_i$$

$Entropy(9,5)$ where we have 5 NO and 9 YES

$$Entropy\ (9,5) = \{\left(-\frac{9}{14}log_2\frac{9}{14}\right) + \left(-\frac{5}{14}log_2\frac{5}{14}\right)\}$$

$$Entropy\ (9,5) = \{0.53 + 0.41\}$$

$$Entropy\ (9,5) = 0.94$$

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Sunny	Mild	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No

$$Entropy\ (Situation) = \sum_{i=1}^{c} -p_i log_2 p_i$$

$$Entropy\ (Play\ Golf, Outlook) = \sum_{i=1}^{c} -p_i log_2 p_i$$

$$Entropy\ (Play\ Golf, Outlook) = P(Sunny) * E(3,2) + P(Overcast) * E(4,0) + P(Rainy) * E(2,3)$$

$$Entropy\ (Play\ Golf, Outlook) = \{\left(\frac{5}{14}\right) * 0.971 + \left(\frac{4}{14}\right) * 0 + \left(\frac{5}{14}\right) * 0.971\}$$

$$Entropy\ (Play\ Golf, Outlook) = 0.693$$

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Sunny	Mild	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No

$$Entropy\ (Situation) = \sum_{i=1}^{c} -p_i log_2 p_i$$

$$Entropy\ (Play\ Golf, Temp) = \sum_{i=1}^{c} -p_i log_2 p_i$$

$$Entropy\ (Play\ Golf, Temp) = P(Hot) * E(2,2) + P(Mild) * E(4,2) + P(Cool) * E(3,1)$$

$$Entropy\ (Play\ Golf, Temp) = \{\left(\frac{4}{14}\right) * 1 + \left(\frac{6}{14}\right) * 0.92 + \left(\frac{4}{14}\right) * 0.81\}$$

$$Entropy\ (Play\ Golf, Temp) = 0.9114$$

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Sunny	Mild	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No

$$Entropy\ (Situation) = \sum_{i=1}^{c} -p_i log_2 p_i$$

$$Entropy\ (Play\ Golf, Humidity) = \sum_{i=1}^{c} -p_i log_2 p_i$$

$$Entropy\ (Play\ Golf, Humidity) = P(High) * E(3,4) + P(Normal) * E(6,1)$$

$$Entropy\ (Play\ Golf, Humidity) = \{\left(\frac{7}{14}\right) * 0.985 + \left(\frac{7}{14}\right) * 0.592\}$$

$$Entropy\ (Play\ Golf, Humidity) = 0.7885$$

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Sunny	Mild	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No

$$Entropy\ (Situation) = \sum_{i=1}^{c} -p_i log_2 p_i$$

$$Entropy\ (Play\ Golf, Windy) = \sum_{i=1}^{c} -p_i log_2 p_i$$

$$Entropy\ (Play\ Golf, Windy) = P(TRUE) * E(3,3) + P(FALSE) * E(6,2)$$

$$Entropy\ (Play\ Golf, Windy) = \{\left(\frac{6}{14}\right) * 1 + \left(\frac{8}{14}\right) * 0.811\}$$

$$Entropy\ (Play\ Golf, Windy) = 0.892$$

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Sunny	Mild	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No

$$Entropy\ (9,5) = 0.94$$

$$Entropy\ (Play\ Golf, Outlook) = 0.693$$

$$Entropy\ (Play\ Golf, Temp) = 0.9114$$

$$Entropy\ (Play\ Golf, Humidity) = 0.7885$$

$$Entropy\ (Play\ Golf, Windy) = 0.892$$