**DECISION TREE LEARNING**

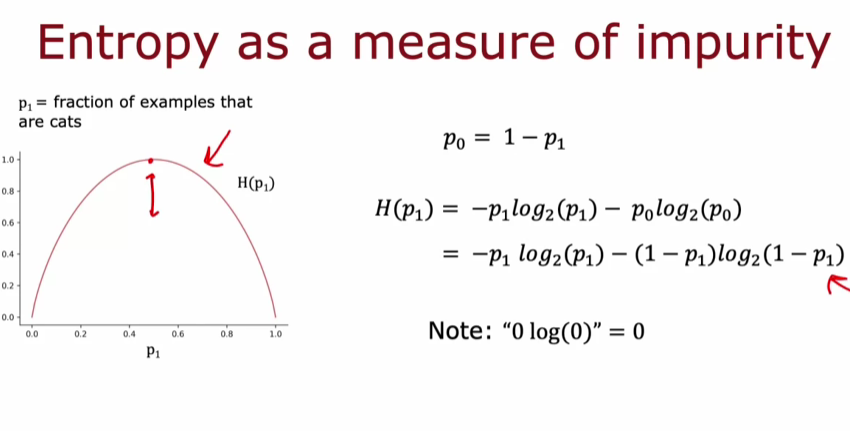
**MEASURING PURITY**

**Understanding Entropy**

* **Entropy quantifies the impurity of a dataset, with a value of zero indicating complete purity (all examples belong to one class) and a value of one indicating maximum impurity (equal representation of classes).**
* **The entropy function is defined as**

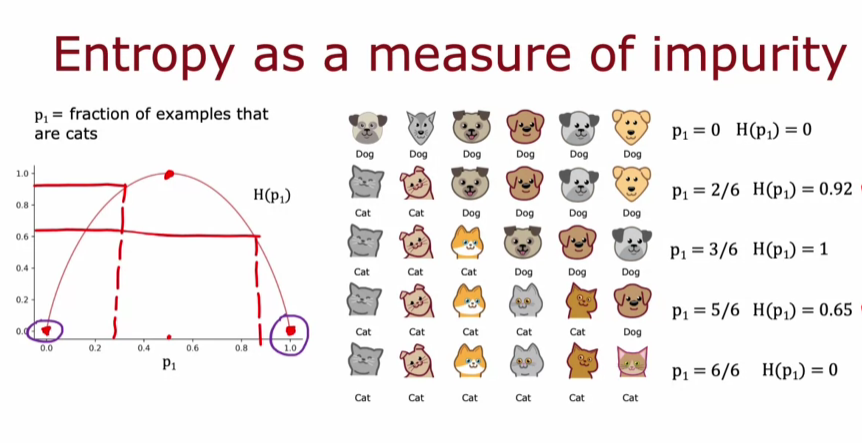
**,**

**where   is the fraction of positive examples (e.g., cats).**

****

**Examples of Entropy Calculation**

* **A dataset with three cats and three dogs (50-50 mix) has an entropy of 1, indicating maximum impurity.**
* **A dataset with five cats and one dog has a lower entropy of about 0.65, showing it is purer than the 50-50 mix.**

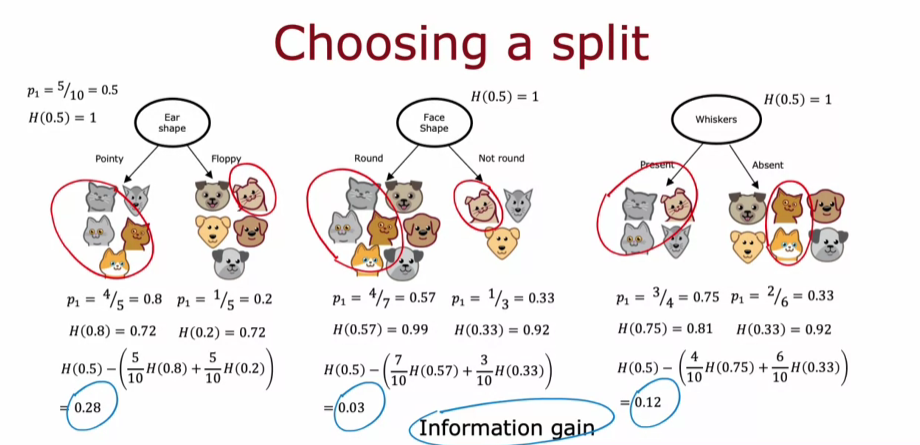
****

* **As the proportion of one class increases (e.g., all cats), the entropy decreases, indicating higher purity.**

**CHOOSING A SPLIT: INFORMATION GAIN**

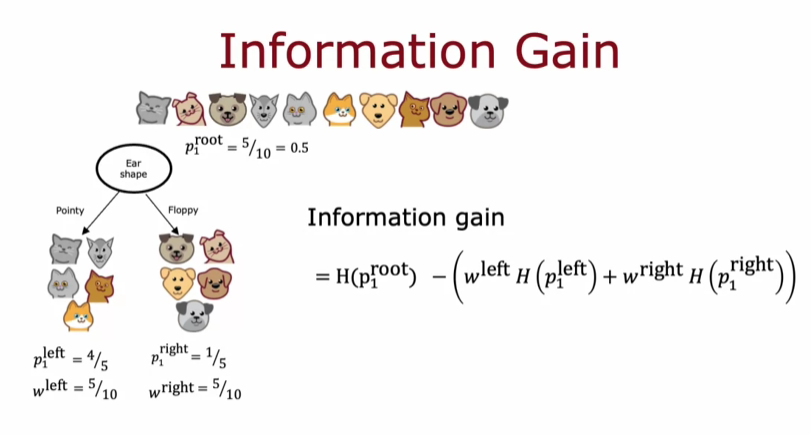
**Understanding Entropy and Information Gain**

* **Entropy is a measure of impurity in a dataset, and the goal is to reduce it when making splits in a decision tree. Information gain quantifies how much entropy is reduced by a particular split.**
* **For example, when splitting based on ear shape, the entropy values for the left and right branches were both 0.72, indicating a certain level of impurity.**
* **To determine the best feature to split on, a weighted average of the entropy values for the left and right branches is calculated, taking into account the number of examples in each branch.**
* **The split that results in the lowest average weighted entropy is preferred, as it indicates a more effective separation of the data.**

****

**Choosing the Best Split**

* **The final decision on which feature to split on is based on the information gain, which is calculated as the entropy of the root node minus the weighted entropy of the branches.**
* **A higher information gain indicates a more effective split, and the feature with the highest information gain is chosen to create the next node in the decision tree.**

****

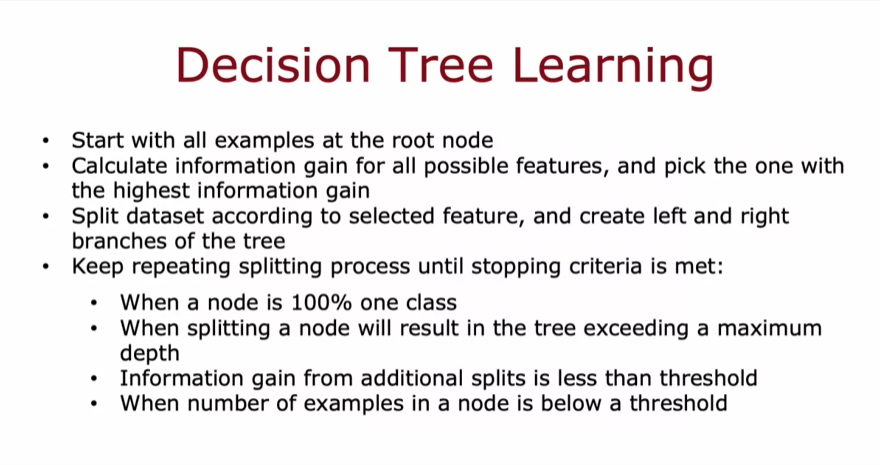
**PUTTING IT TOGETHER**

**Building the decision tree**

* **The process begins at the root node with all training examples, calculating information gain for all features to determine the best feature for splitting.**
* **After selecting a feature, the dataset is divided into subsets, creating left and right branches, and the splitting process is repeated recursively on each branch.**

**Stopping criteria for splitting**

* **Splitting continues until certain criteria are met, such as achieving a node with 100% of a single class, reaching entropy of zero, exceeding maximum depth, or having insufficient information gain.**
* **Other stopping conditions include having too few examples in a node or reaching a predefined threshold for information gain.**

****

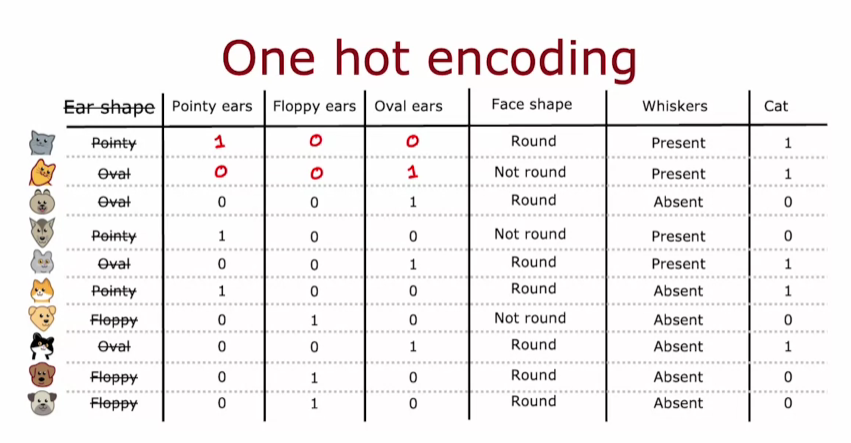
**Recursive nature of decision trees**

* **The construction of decision trees involves building smaller sub-decision trees for each branch, illustrating the recursive algorithm concept in computer science.**
* **Understanding recursion is beneficial, but not essential for using libraries or completing assignments related to decision trees.**

**USING ONE HOT ENCODING OF CATEGORICAL FEATURES**

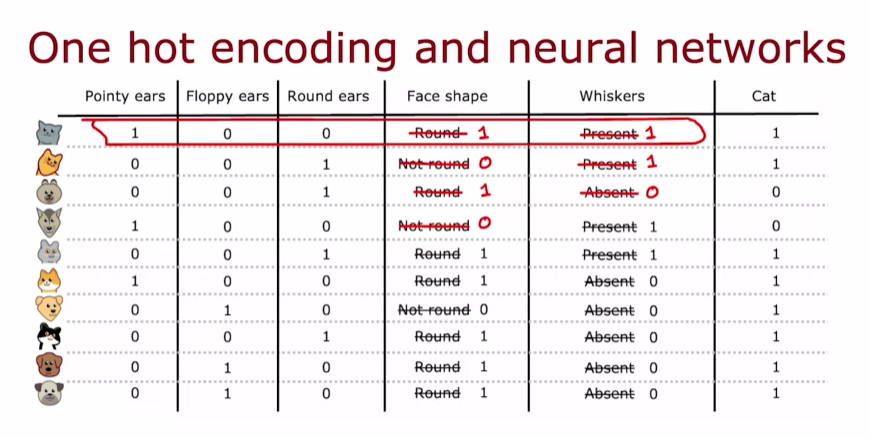
**Understanding one-hot encoding**

* **One-hot encoding transforms a categorical feature with multiple values into several binary features, each representing one possible value. For example, an ear shape feature with three values (pointy, floppy, oval) is converted into three binary features.**
* **Each binary feature can only take on values of 0 or 1, indicating the absence or presence of that specific feature.**

****

**Application in decision trees and neural networks**

* **With one-hot encoding, decision trees can effectively handle features that have more than two values, as each new feature now only has two possible values.**
* **This technique is also applicable to neural networks, allowing categorical features to be represented numerically, which is essential for model training.**

****

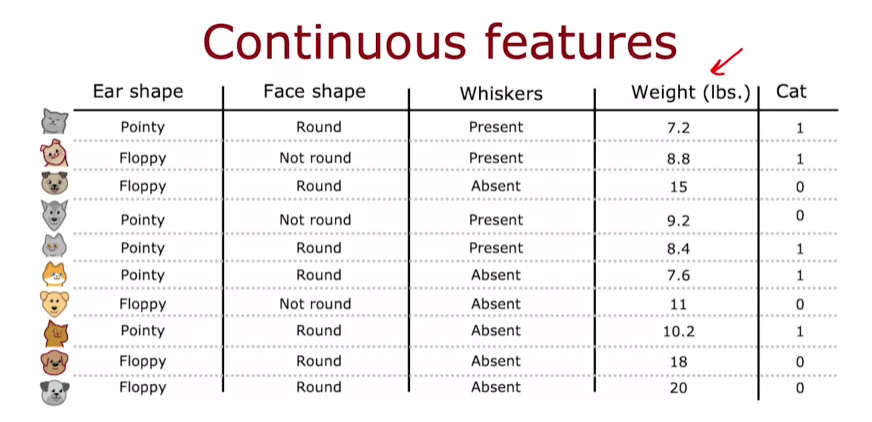
**Reinforcing learning**

* **One-hot encoding is a versatile method that enhances the ability of various algorithms, including decision trees and neural networks, to process categorical data.**

**CONTINUOUS VALUED FEATURES**

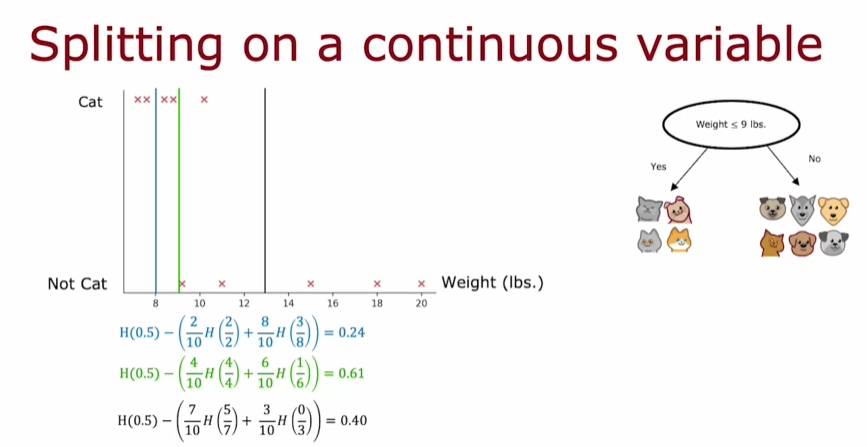
**Understanding Continuous Features in Decision Trees**

* **Decision trees can incorporate continuous features, such as weight, by considering various threshold values for splitting the data.**
* **The algorithm evaluates different thresholds to determine which one provides the highest information gain, thus improving the decision-making process.**

****

**Calculating Information Gain**

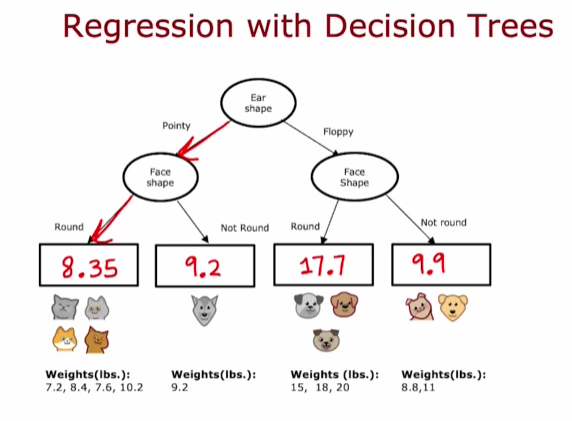
* **Information gain is calculated by comparing the entropy before and after the split based on a chosen threshold.**
* **For example, splitting at a weight of 9 pounds yields a higher information gain than splitting at 8 pounds, demonstrating the importance of selecting the optimal threshold.**

****

**Recursive Tree Building**

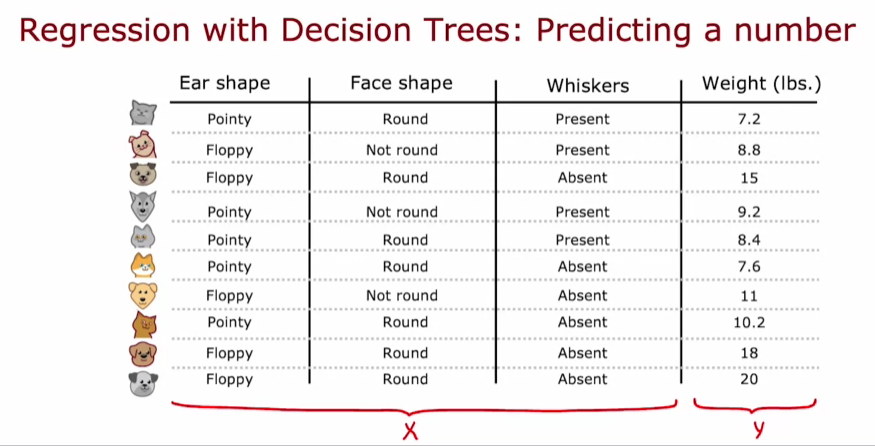
* **Once the best threshold is identified, the dataset is split into subsets, and the decision tree can be built recursively using these subsets.**
* **This process continues until the tree is fully developed, allowing for effective classification based on continuous features.**

**REGRESSION TREES**

****

**Understanding Regression Trees**

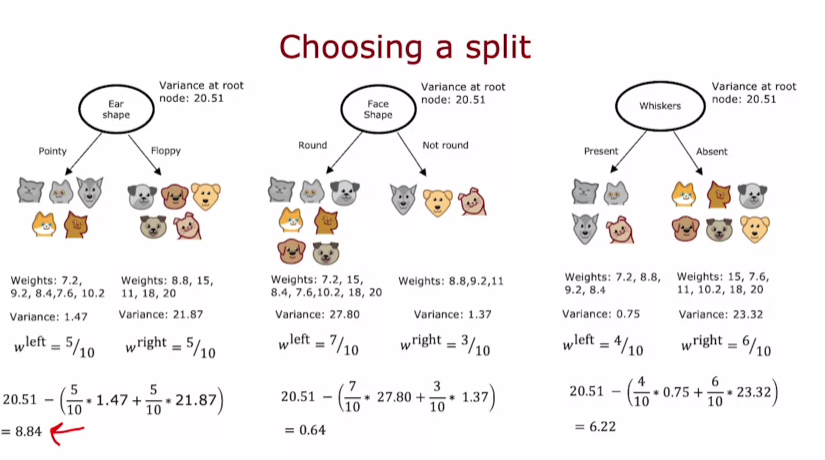
* **Regression trees predict a target output (Y), such as the weight of an animal, based on input features (X). Unlike classification, where the goal is to categorize, regression aims to estimate a continuous value.**

****

* **The decision tree splits based on features, such as ear shape and face shape, to create branches that lead to predictions based on the average of the target values in the leaf nodes.**

**Choosing Features for Splitting**

* **When building a regression tree, the goal is to reduce the variance of the target values in the subsets created by the splits, rather than reducing entropy as in classification.**
* **The quality of a split is evaluated by calculating the reduction in variance, with the feature that provides the largest reduction being chosen for the split.**

****

**Recursive Splitting and Ensemble Learning**

* **After selecting a feature to split on, the process is repeated recursively for the resulting subsets until a stopping criterion is met.**
* **Training multiple decision trees together, known as an ensemble of decision trees, can lead to improved predictive performance compared to a single tree.**