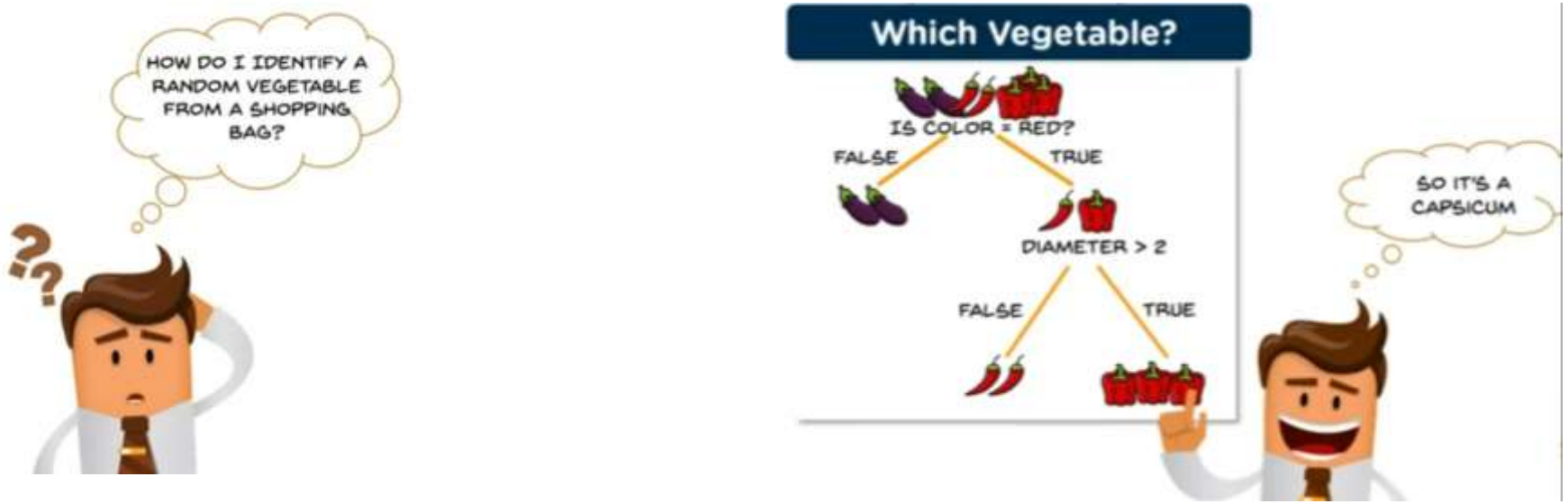
Decision Tree

Why Decision Tree !

Decision Tree

Decision Tree is a tree shaped diagram used to determine a course of action . Each branch of the tree represent a possible decision, occurrence or reaction.



Where we use Decision Tree

Classification

A classification tree will determine a set of logical if-then conditions to classify problems.
For example, discriminating between three types of flowers based on certain features

Regression

Regression tree is used when the target variable is numerical or continuous in nature. We fit a regression model to the target variable using each of the independent variables. Each split is made based on the sum of squared error.

Important Terms in Decision Tree

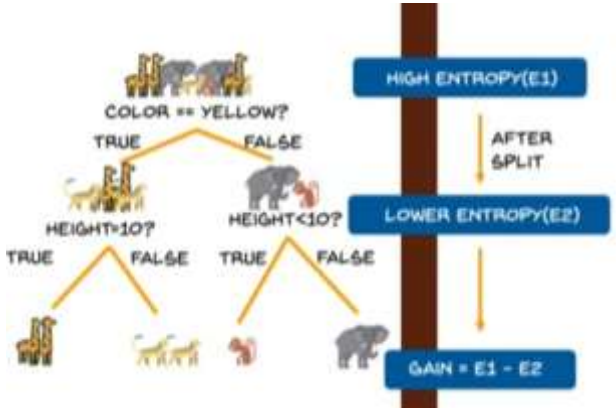
ENTROPY

ENTROPY IS THE MEASURE OF RANDOMNESS OR UNPREDICTABILITY IN THE DATASET

THIS DATASET HAS A VERY HIGH ENTROPY

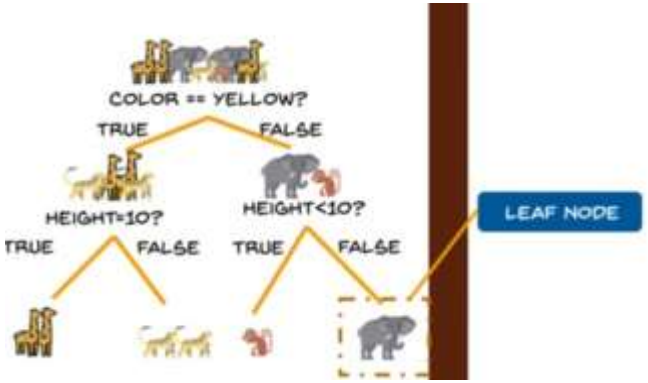
INFORMATION GAIN

IT IS THE MEASURE OF DECREASE IN ENTROPY AFTER THE DATASET IS SPLIT



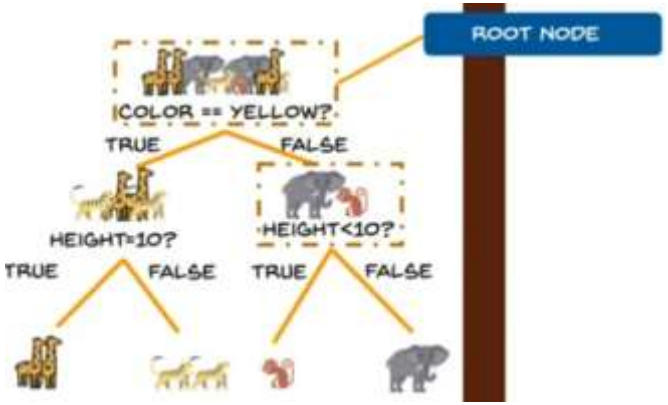
LEAF NODE

LEAF NODE CARRIES THE CLASSIFICATION OR THE DECISION

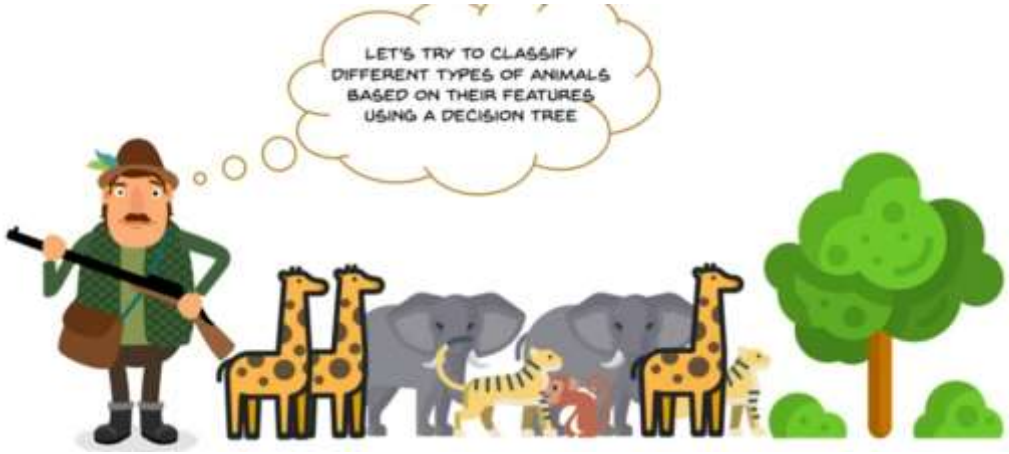


ROOT NODE

THE TOP MOST DECISION NODE IS KNOWN AS THE ROOT NODE



How Does Decision Tree Works ?



PROBLEM STATEMENT

TO CLASSIFY THE DIFFERENT TYPES OF ANIMALS BASED ON THEIR FEATURES USING DECISION TREE

THE DATASET IS LOOKING QUITE MESSY AND THE ENTROPY IS HIGH IN THIS CASE

TRAINING DATASET

COLOR	HEIGHT	LABEL
GREY	10	ELEPHANT
YELLOW	10	GIRAFFE
BROWN	5	MONKEY
GREY	10	ELEPHANT
YELLOW	4	TIGER

HOW TO SPLIT THE DATA

WE HAVE TO FRAME THE CONDITIONS THAT SPLIT THE DATA IN SUCH A WAY THAT THE INFORMATION GAIN IS THE HIGHEST

NOTE

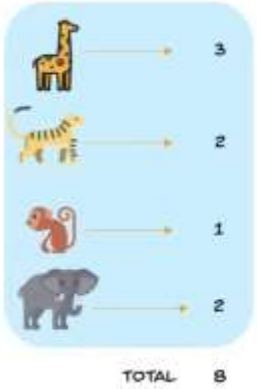
GAIN IS THE MEASURE OF DECREASE IN ENTROPY AFTER SPLITTING

LET'S TRY TO CALCULATE THE ENTROPY FOR THE CURRENT DATASET

FORMULA FOR ENTROPY

$$\sum_{i=1}^{k} P(valuei).log_2(P(value_i)$$

How Does Decision Tree Works ?



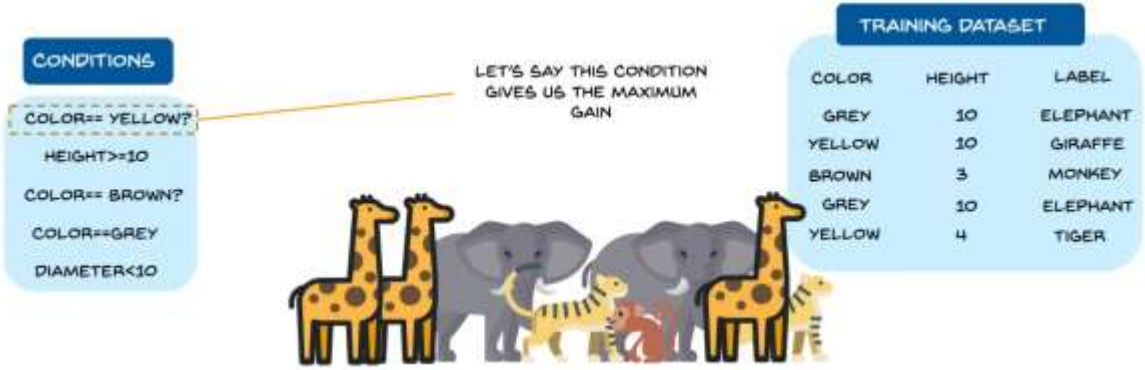
LET'S USE THE FORMULA

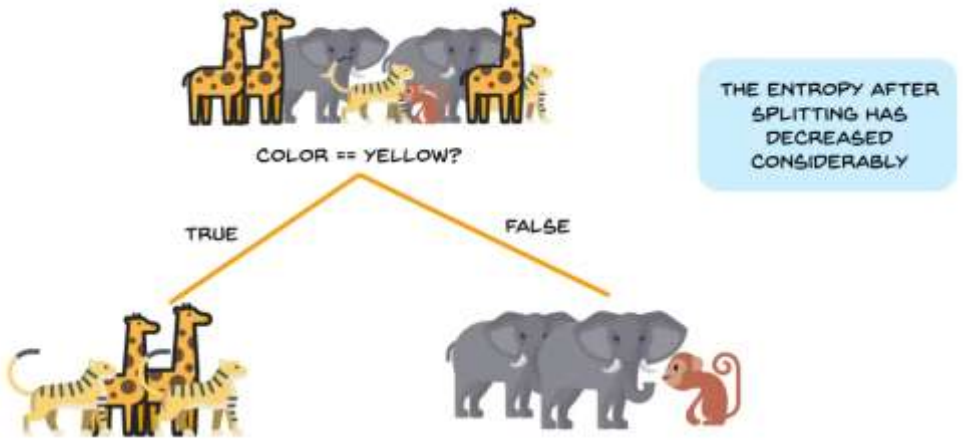
$$\sum_{i=1}^{k} P(valuei).log_2(P(value_i)$$

$$\text{ENTROPY} = \left(\frac{3}{8}\right) log_2\left(\frac{3}{8}\right) + \left(\frac{2}{8}\right) log_2\left(\frac{2}{8}\right) + \left(\frac{1}{8}\right) log_2\left(\frac{1}{8}\right) + \left(\frac{2}{8}\right) log_2\left(\frac{2}{8}\right)$$

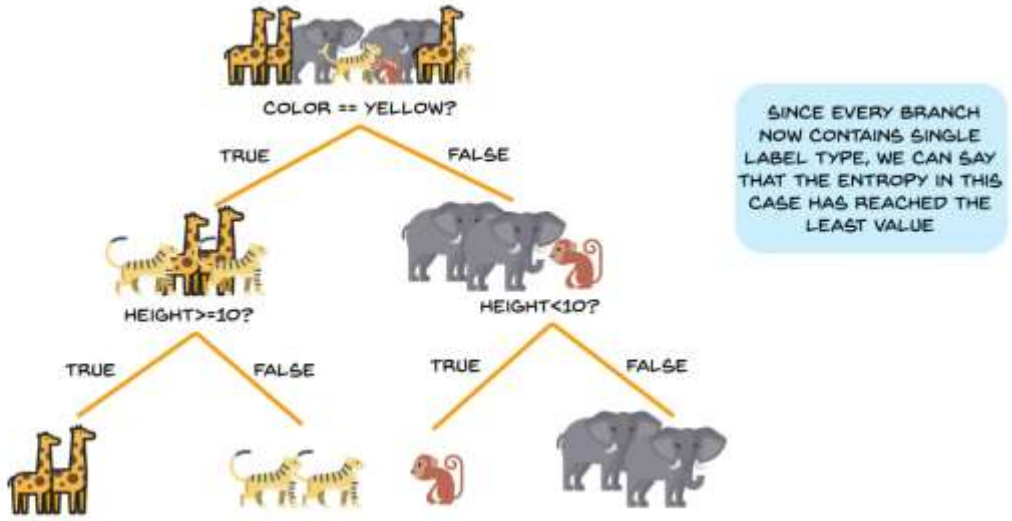
ENTROPY=0.571

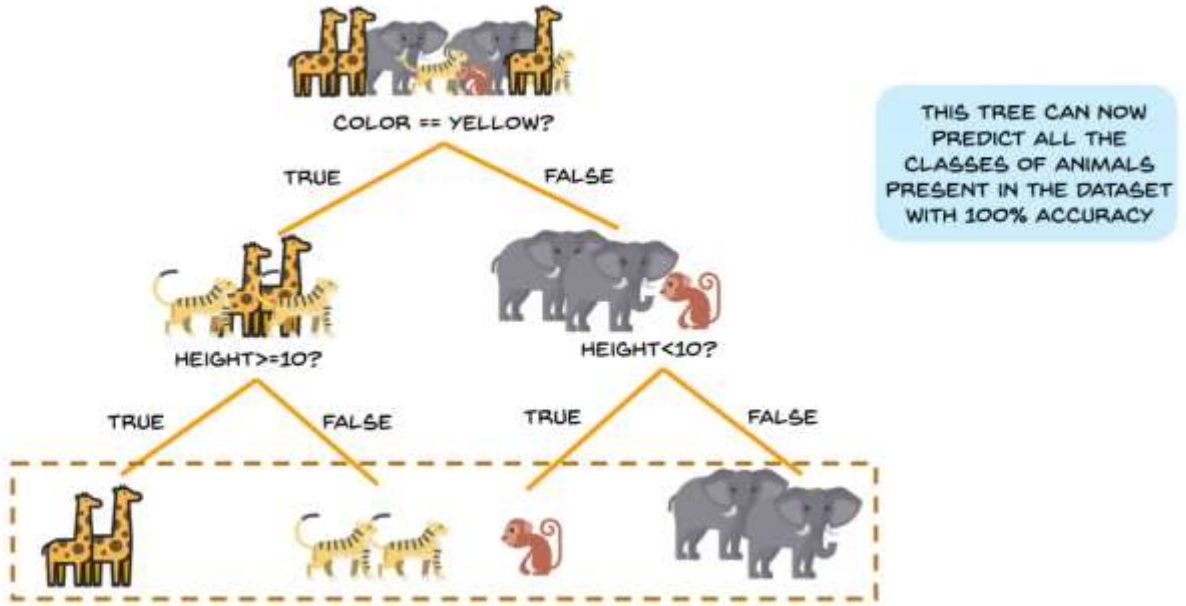
How Does Decision Tree Works ?





How Does Decision Tree Works ?





Confusion Metrics

A confusion matrix is a table used to evaluate the performance of a classification model by comparing the predicted labels with the actual labels. It provides a detailed breakdown of the model's prediction results and helps identify the types of errors the model is making.

N = 15	Predicted: No	Predicted: Yes
Actual: No	TN = 3	FP = 2
Actual: Yes	FN = 1	TP = 9

The Four main metrics

Accuracy : The accuracy is used to find the portion of correctly classified values.
It tells us how often our classifier is right

It is the sum of all true values divided by total values

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

The Four main metrics

Precision : Precision is used to calculate the model's ability to classify positive values correctly. It answers the question, "When the model predicts a positive value, how often is it right?"

It is the true positives divided by total number of predicted positive values

$$\text{Precision} = \frac{TP}{TP + FP}$$

The Four main metrics

Recall : It is used to calculate the model's ability to predict positive values. "How often does the model actually predict the correct positive values?"

It is the true positives divided by total number of actual positive values

$$\text{Recall} = \frac{TP}{TP + FN}$$

The Four main metrics

F1-Score : It is the harmonic mean of Recall and Precision. It is useful when you need to take both Precision and Recall into account

$$\text{F1-Score} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision} + \text{Recall}}$$

Thank You