

Out[7]: KNeighborsClassifier()

Predicting the Test Result:

To predict the test set result, we will create a `y_pred` vector as we did in Logistic Regression.

Below is the code for it:

```
In [8]: #Predicting the test set result
y_pred= classifier.predict(x_test)
```

Creating the Confusion Matrix:

Now we will create the Confusion Matrix for our K-NN model to see the accuracy of the classifier. Below is the code for it:

```
In [10]: #Creating the Confusion matrix
from sklearn.metrics import confusion_matrix
cm= confusion_matrix(y_test, y_pred)
cm
```



```
Out[10]: array([[59,  9],
   [ 8, 24]], dtype=int64)
```

In the above image, we can see there are $59+24= 83$ correct predictions and $8+9= 17$ incorrect predictions,

Decision Tree

Decision Tree

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems.

Decision Tree Classification Algorithm / Decision Tree Classifier

It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

In a Decision tree, there are two nodes, which are the **Decision Node** and **Leaf Node**.

Decision nodes are used to make any decision and have multiple branches, whereas **Leaf nodes** are the output of those decisions and do not contain any further branches.

The decisions or the test are performed on the basis of features of the given dataset.

It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

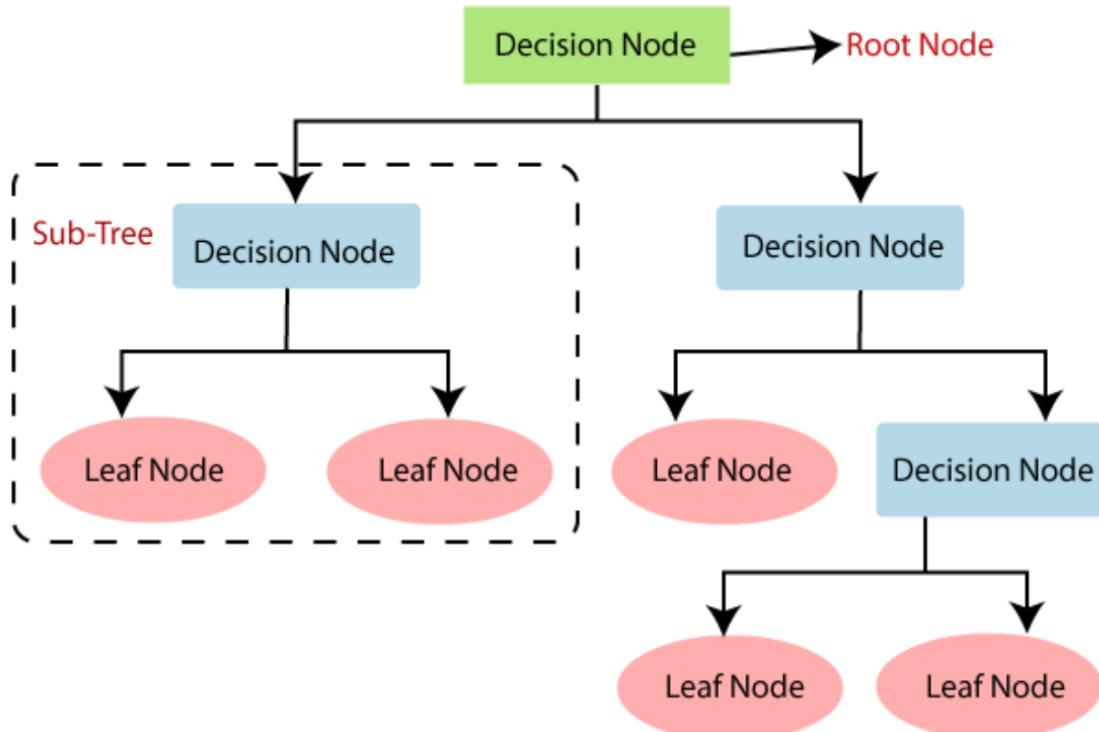
In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.

A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

Below diagram explains the general structure of a decision tree:

Note:

A decision tree can contain categorical data (YES/NO) as well as numeric data.



Why use Decision Trees?

There are various algorithms in Machine learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine learning model. Below are the two reasons for using the Decision tree:

- **Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.**
- **The logic behind the decision tree can be easily understood because it shows a tree-like structure.**

Decision Tree Terminologies

Root Node:

Root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.

Leaf Node:

Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.

Splitting:

Splitting is the process of dividing the decision node/root node into sub-nodes according to the given conditions.

Branch/Sub Tree:

A tree formed by splitting the tree.

Pruning:

Pruning is the process of removing the unwanted branches from the tree.

Parent/Child node:

The root node of the tree is called the parent node, and other nodes are called the child nodes.

How does the Decision Tree algorithm Work?

In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree.

This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node.

For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further.

It continues the process until it reaches the leaf node of the tree. The complete process can be better understood using the below algorithm:

Step-1:

Begin the tree with the root node, says S, which contains the complete dataset.

Step-2:

Find the best attribute in the dataset using Attribute Selection Measure (ASM).

Step-3:

Divide the S into subsets that contains possible values for the best attributes.

Step-4:

Generate the decision tree node, which contains the best attribute.

Step-5:

Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

Example:

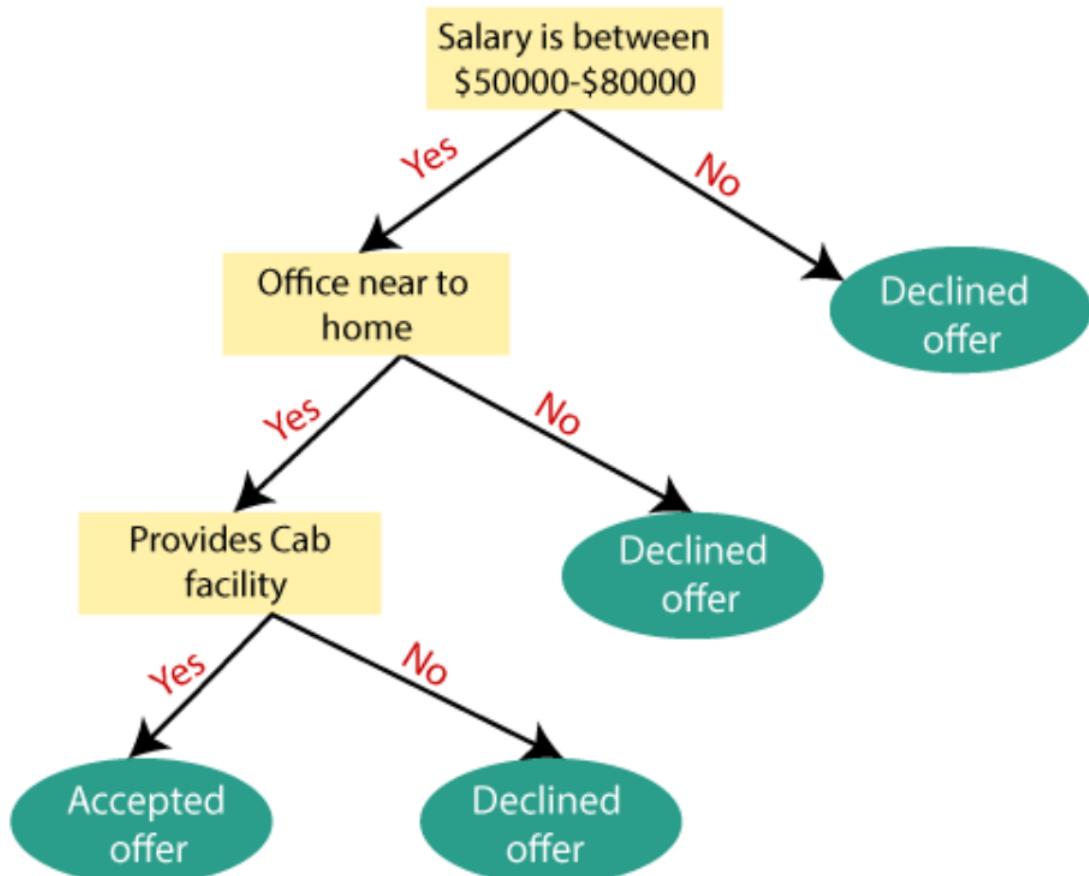
Suppose there is a candidate who has a job offer and wants to decide whether he should accept the offer or Not. So, to solve this problem, the decision tree starts with the root node (Salary attribute by ASM).

The root node splits further into the next decision node (distance from the office) and one leaf node based on the corresponding labels.

The next decision node further gets split into one decision node (Cab facility) and one leaf node.

Finally, the decision node splits into two leaf nodes (Accepted offers and Declined offer).

Consider the below diagram:



Decision Trees

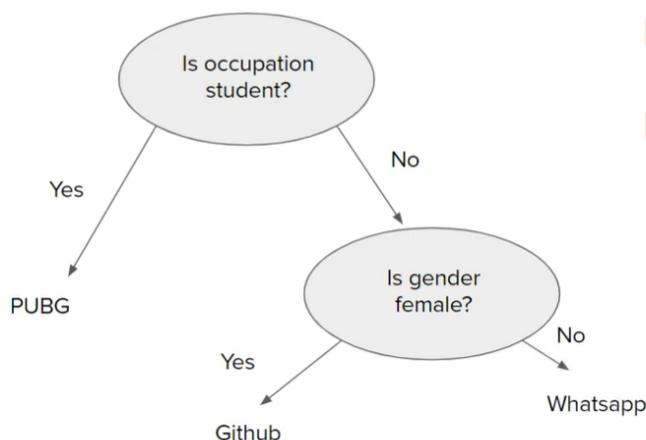
Introduction and Geometric Intuition

Example 1

Gender	Occupation	Suggestion
F	Student	PUBG
F	Programmer	Github
M	Programmer	Whatsapp
F	Programmer	Github
M	Student	PUBG
M	Student	PUBG

```
If occupation==student
    print(PUBG)
Else
    If gender==female
        print(Github)
    Else
        print(Whatsapp)
```

Where is the Tree?

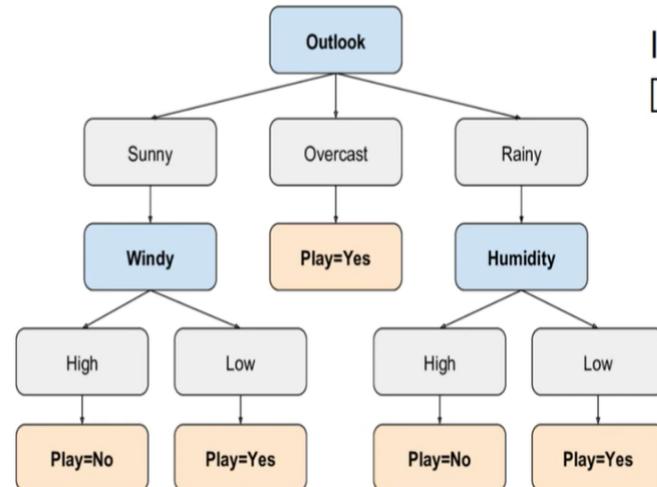
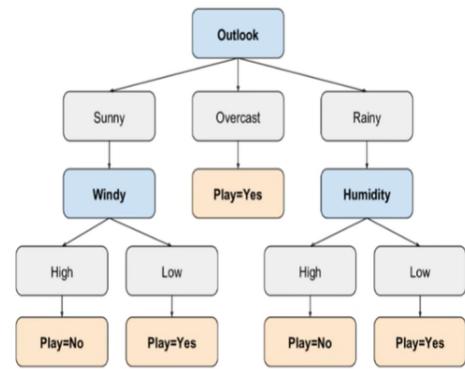


```
If occupation==student
    print(PUBG)
Else
    If gender==female
        print(Github)
    Else
        print(Whatsapp)
```

#

Example 2

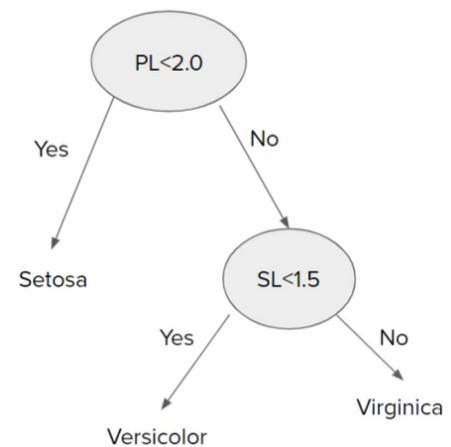
Day	Outlook	Temp	Humid	Wind	Play?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No



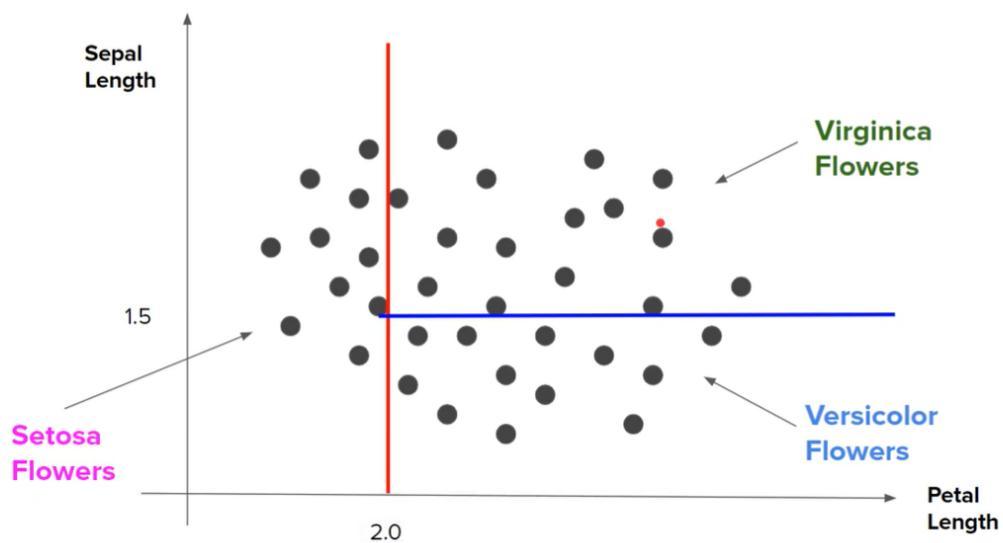
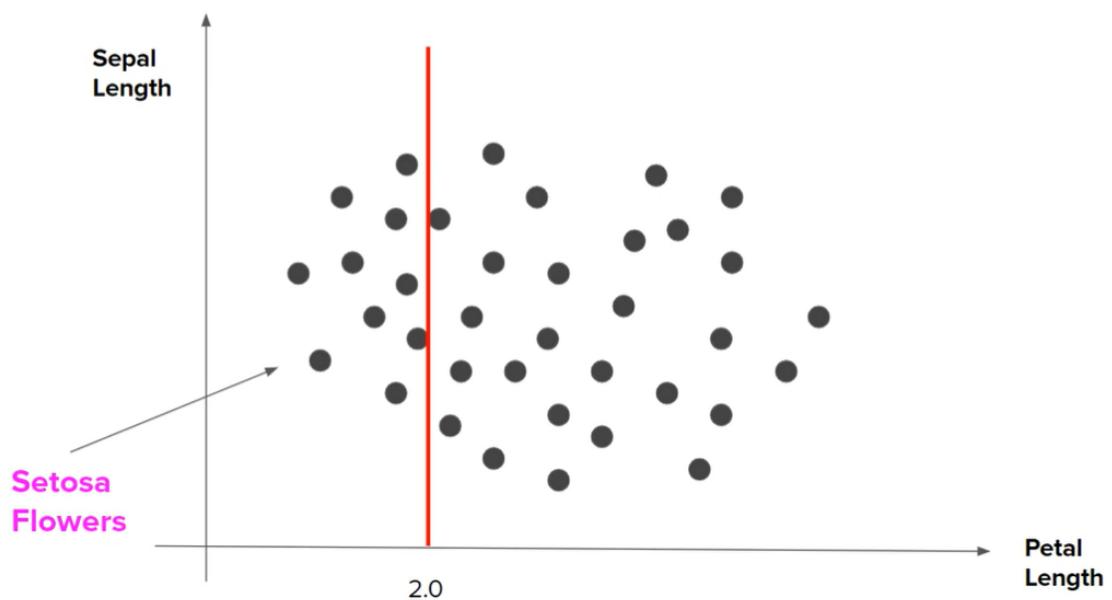
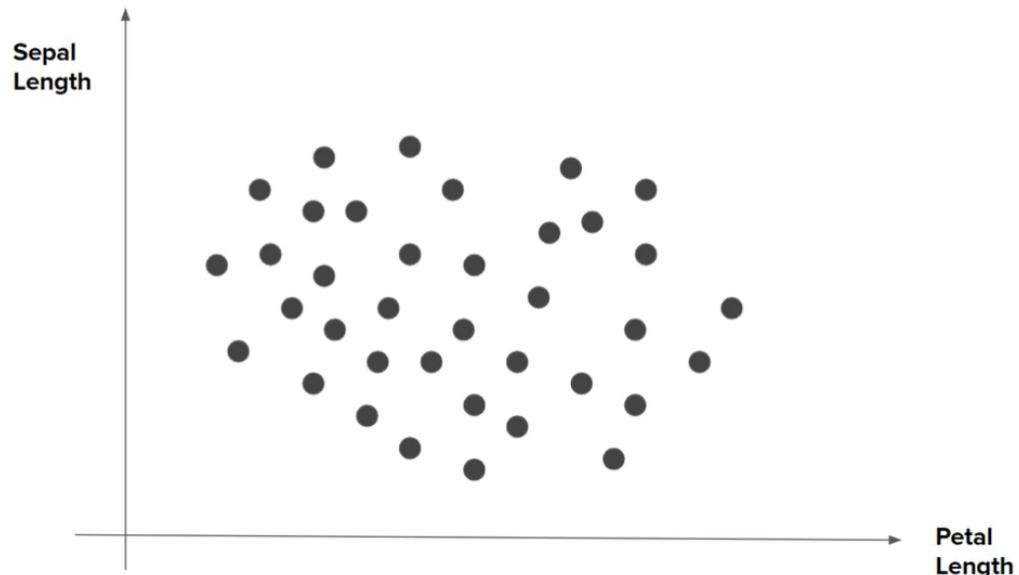
Input query point:
[Rainy, Mild, High, Strong]

What if we have numerical data?

Petal Length	Sepal Length	Type
1.34	0.34	Setosa
3.45	1.45	Versicolor
1.69	0.98	Setosa
2.56	1.79	Virginica
3.00	1.13	Versicolor
1.3	0.88	Setosa



Geometric Intuition



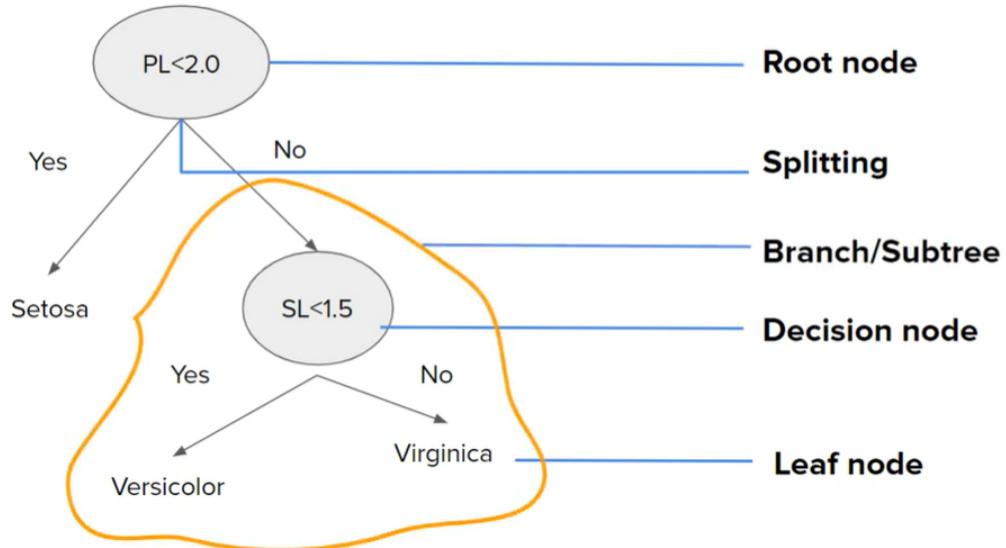
Pseudo code

- Begin with your training dataset, which should have some feature variables and classification or regression output.
- Determine the “best feature” in the dataset to split the data on; more on how we define “best feature” later
- Split the data into subsets that contain the correct values for this best feature. This splitting basically defines a node on the tree i.e each node is a splitting point based on a certain feature from our data.
- Recursively generate new tree nodes by using the subset of data created from step 3.

Programmatically speaking, Decision trees are nothing but a giant structure of nested if-else condition

Mathematically speaking, Decision trees use **hyperplanes** which run **parallel to any one of the axes** to cut your coordinate system into **hyper cuboids**

Terminology



Some unanswered questions

How to decide which column should be considered as root node?

How to select subsequent decision nodes?

How to decide splitting criteria in case of numerical columns?

Advantages

Intuitive and easy to understand

Minimal data preparation is required

The cost of using the tree for inference is **logarithmic** in the number of data points used to train the tree

Disadvantages

Overfitting

Prone to errors for imbalanced datasets

CART - Classification and Regression Trees

The logic of decision trees can also be applied to regression problems, hence the name CART

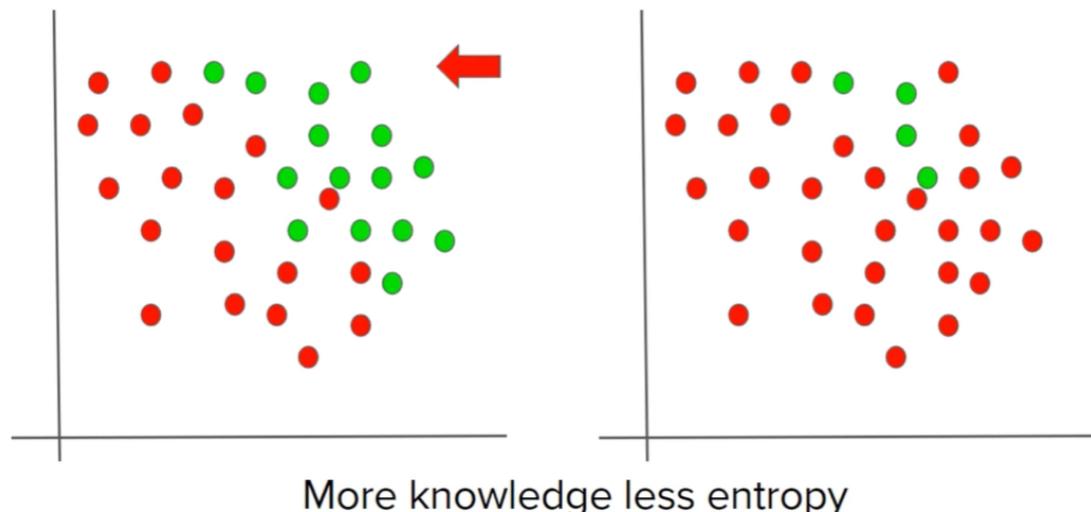
Entropy in Decision Trees

Decision Trees

Entropy

What is Entropy?

In the most layman terms, Entropy is nothing but the measure of disorder. Or you can also call it the measure of purity/impurity. Let's see an example...



How to calculate Entropy?

The mathematical formula for entropy is:

$$E(S) = \sum_{i=1}^c -p_i \log_2 p_i$$

For e.g if our data has only 2 class labels **Yes** and **No**.

Where 'Pi' is simply the frequentist probability of an element/class 'i' in our data.

$$E(D) = -p_{\text{yes}} \log_2(p_{\text{yes}}) - p_{\text{no}} \log_2(p_{\text{no}})$$

Example - Dataset

Salary	Age	Purchase
20000	21	Yes
10000	45	No
60000	27	Yes
15000	31	No
12000	18	No

Salary	Age	Purchase
34000	31	No
15000	25	No
69000	57	Yes
25000	21	No
32000	28	No

$$H(d) = -P_y \log_2(P_y) - P_n \log_2(P_n)$$

$$H(d) = -2/5 \log_2(2/5) - 3/5 \log_2(3/5)$$

$$H(d) = 0.97$$

$$H(d) = -P_y \log_2(P_y) - P_n \log_2(P_n)$$

$$H(d) = -1/5 \log_2(1/5) - 4/5 \log_2(4/5)$$

$$H(d) = 0.72$$

Calculating entropy for a 3 class problem

Salary	Age	Purchase
20000	21	Yes
10000	45	No
60000	27	Yes
15000	31	No
30000	30	Maybe
12000	18	No
40000	40	Maybe
20000	20	Maybe

$$H(d) = -P_y \log_2(P_y) - P_n \log_2(P_n) - P_m \log_2(P_m)$$

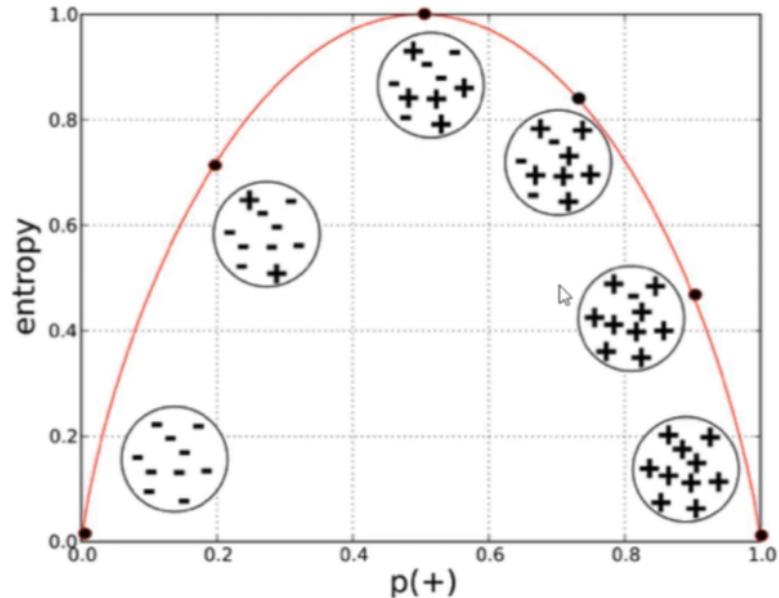
$$H(d) = -2/8 \log_2(2/8) - 3/8 \log_2(3/8) - 3/8 \log_2(3/8)$$

$$H(d) = 1.56$$

Observation

- More the uncertainty more is entropy
- For a 2 class problem the min entropy is 0 and the max is 1
- For more than 2 classes the min entropy is 0 but the max can be greater than 1
- Both \log_2 or \log_e can be used to calculate entropy

Entropy Vs Probability



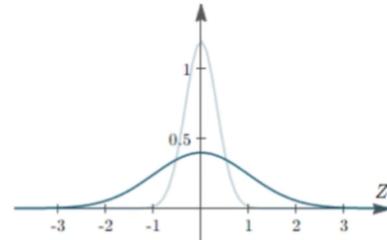
Entropy for continuous variables

Area	Built in	Price
1200	1999	3.5
1800	2011	5.6
1400	2000	7.3
...

Dataset 1

Area	Built in	Price
2200	1989	4.6
800	2018	6.5
1100	2005	12.8
...

Dataset 2



Quiz: Which of the above datasets have higher entropy?

Ans: Whichever is less peaked

Information Gain

Information Gain, is a metric used to train Decision Trees. Specifically, this metric measures the quality of a split.

The information gain is based on the decrease in entropy after a data-set is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain

$$\text{Information Gain} = E(\text{Parent}) - (\text{Weighted Average}) * E(\text{Children})$$

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

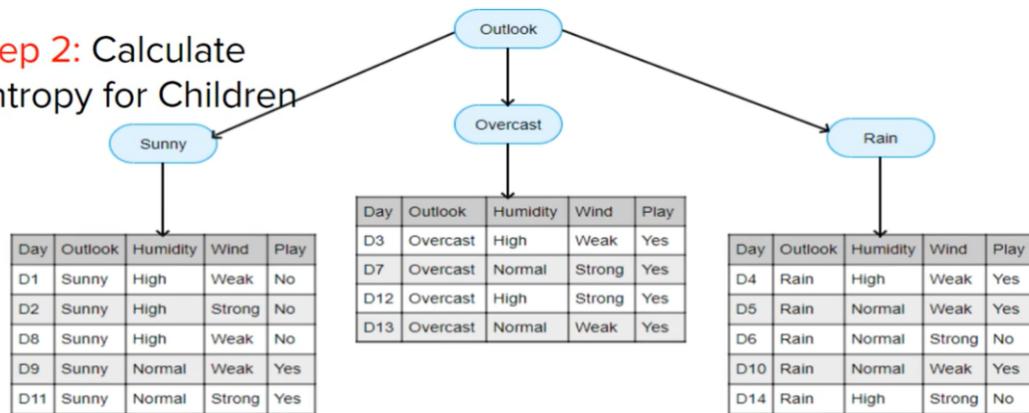
Step 1:Entropy of Parent

$$E(P) = -p_y \log_2(p_y) - p_n \log_2(p_n)$$

$$= 9/14 \log_2(9/14) - 5/14 \log_2(5/14)$$

$$E(P) = 0.94$$

Step 2: Calculate Entropy for Children



$$E(S) = -2/5 \log(2/5) - 3/5 \log(3/5)$$

$$E(S) = 0.97$$

$$E(O) = -5/5 \log(5/5) - 0/5 \log(0/5)$$

$$E(O) = 0$$

$$E(R) = -3/5 \log(3/5) - 2/5 \log(2/5)$$

$$E(S) = 0.97$$

Step 3 : Calculate weighted Entropy of Children

$$\text{Weighted Entropy} = 5/14 * 0.97 + 4/14 * 0 + 5/14 * 0.97$$

$$W.E(\text{Children}) = 0.69$$

P(Overcast) is a leaf node as it's entropy is 0

Step 4 : Calculate Information Gain

Information Gain = $E(\text{Parent}) - \{\text{Weighted Average}\} * E(\text{Children})$

$$\text{IG} = \mathbf{0.97 - 0.69 = 0.28}$$

So the information gain(or the decrease in entropy/impurity) when you split this data on the basis of **Outlook** condition/column is **0.28**

Step 5 : Calculate Information Gain for all the columns

Whichever column has the highest Information Gain(maximum decrease in entropy) the algorithm will select that column to split the data.

Step 6 : Find Information Gain recursively

Decision tree then applies a recursive greedy search algorithm in top bottom fashion to find Information Gain at every level of the tree.

Once a leaf node is reached (Entropy = 0), no more splitting is done.

Gini Impurity

Decision Trees

Gini Impurity

Gini index is a measure of impurity or purity used while creating a decision tree in the CART(Classification and Regression Tree) algorithm.

An attribute with the low Gini index should be preferred as compared to the high Gini index.

It only creates binary splits, and the CART algorithm uses the Gini index to create binary splits.

Gini index can be calculated using the below formula:

$$\text{Gini Index} = 1 - \sum_j P_j^2$$

Example - Dataset

Salary	Age	Purchase
20000	21	Yes
10000	45	No
60000	27	Yes
15000	31	No
12000	18	No

Salary	Age	Purchase
34000	31	No
15000	25	No
69000	57	Yes
25000	21	No
32000	28	No

$$G = 1 - (P_y^2 + P_n^2)$$

$$G = 1 - (4/25 + 9/25)$$

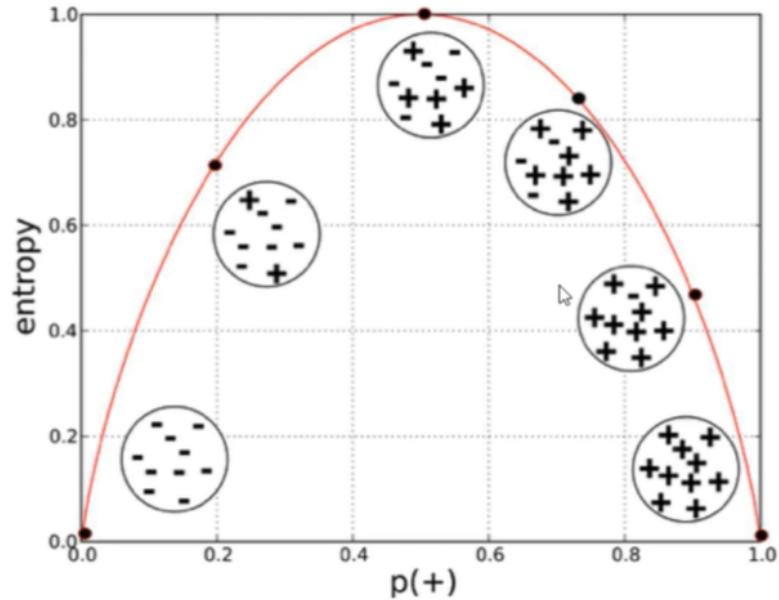
$$G = 0.48$$

$$G = 1 - (P_y^2 + P_n^2)$$

$$G = 1 - (1/25 + 16/25)$$

$$G = 0.32$$

Entropy Vs Probability



Handling Numerical Values in Decision Trees

Handling Numerical Data

S No	User Rating	Downloaded
1	3.5	Yes
2	4.6	Yes
3	2.2	No
4	1.6	Yes
5	4.1	No
6	3.9	No
7	3.2	No
9	2.9	Yes
10	4.8	Yes
11	3.3	No
12	2.5	Yes
13	1.9	Yes

Step 1:

Sort the data on the basis of numerical column

S No	User Rating	Downloaded
1	1.6	Yes
2	1.9	Yes
3	2.2	No
4	2.5	Yes
5	2.9	Yes
6	3.2	No
7	3.3	No
9	3.5	Yes
10	3.9	No
11	4.1	No
12	4.6	Yes
13	4.8	Yes

Step 2:

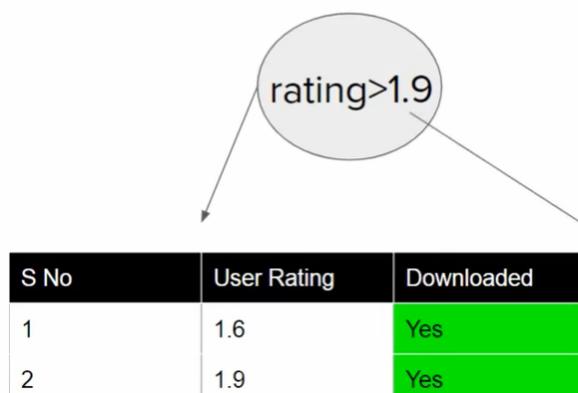
Split the entire data on the basis of every value of user_rating



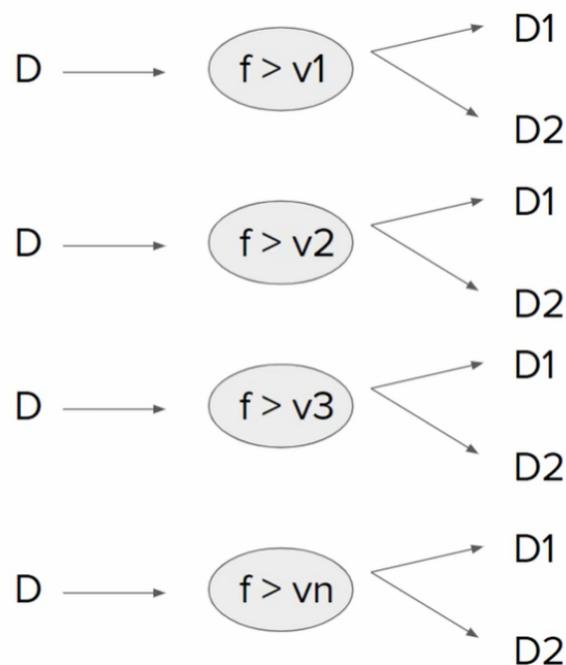
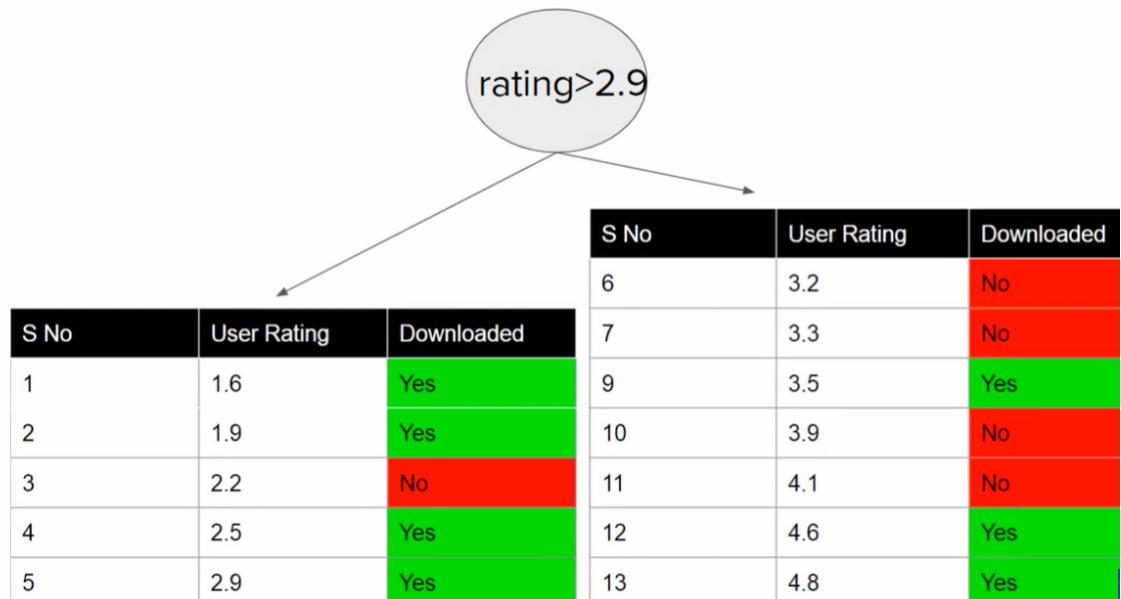
S No	User Rating	Downloaded
2	1.9	Yes
3	2.2	No
4	2.5	Yes
5	2.9	Yes
6	3.2	No
7	3.3	No
9	3.5	Yes
10	3.9	No
11	4.1	No
12	4.6	Yes
13	4.8	Yes

Step 2:

Split the entire data on the basis of every value of user_rating

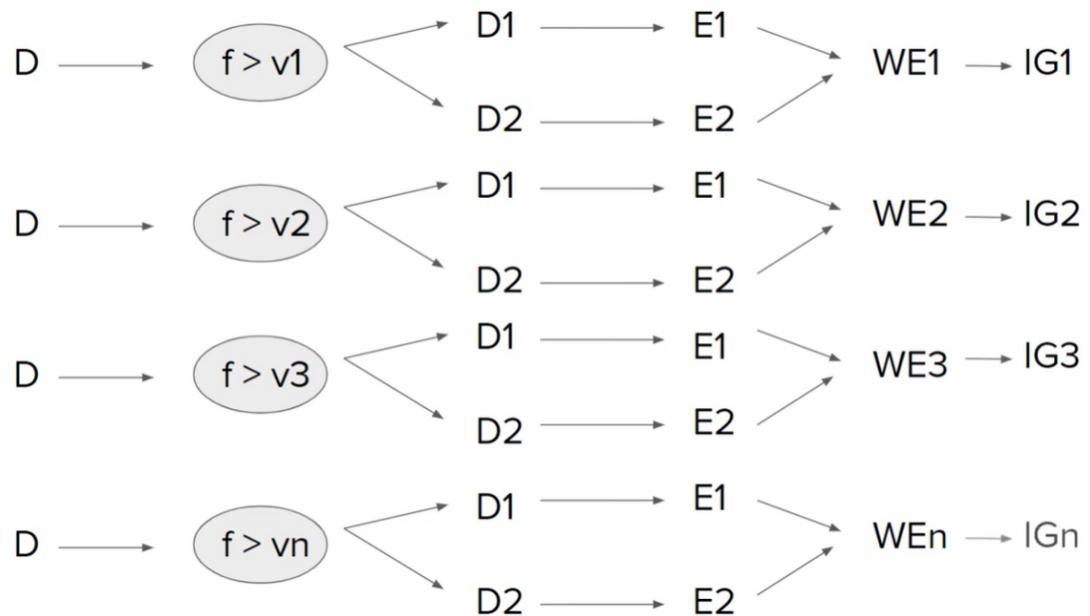


S No	User Rating	Downloaded
3	2.2	No
4	2.5	Yes
5	2.9	Yes
6	3.2	No
7	3.3	No
9	3.5	Yes
10	3.9	No
11	4.1	No
12	4.6	Yes
13	4.8	Yes



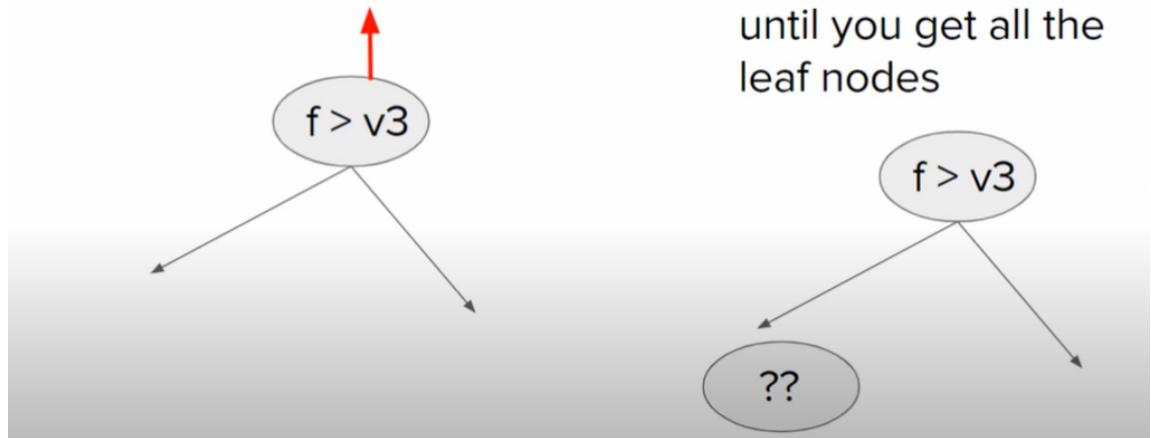
- Where D is the Dataset
- f is the column User Rating
- v_1, v_2, \dots, v_n are the values of various rows of User Rating

In []:



Step 5:

$\text{Max}\{ \text{IG}_1, \text{IG}_2, \text{IG}_3, \dots, \text{IG}_n \}$



Step 6:

Do this recursively until you get all the leaf nodes

Visualizing a Decision Tree- Decision Tree Classification Python Code

```
In [23]: from sklearn.datasets import load_iris
```

```
In [25]: iris=load_iris()
```

```
In [28]: X=iris.data
X
```

```
Out[28]: array([[5.1, 3.5, 1.4, 0.2],  
 [4.9, 3. , 1.4, 0.2],  
 [4.7, 3.2, 1.3, 0.2],  
 [4.6, 3.1, 1.5, 0.2],  
 [5. , 3.6, 1.4, 0.2],  
 [5.4, 3.9, 1.7, 0.4],  
 [4.6, 3.4, 1.4, 0.3],  
 [5. , 3.4, 1.5, 0.2],  
 [4.4, 2.9, 1.4, 0.2],  
 [4.9, 3.1, 1.5, 0.1],  
 [5.4, 3.7, 1.5, 0.2],  
 [4.8, 3.4, 1.6, 0.2],  
 [4.8, 3. , 1.4, 0.1],  
 [4.3, 3. , 1.1, 0.1],  
 [5.8, 4. , 1.2, 0.2],  
 [5.7, 4.4, 1.5, 0.4],  
 [5.4, 3.9, 1.3, 0.4],  
 [5.1, 3.5, 1.4, 0.3],  
 [5.7, 3.8, 1.7, 0.3],  
 [5.1, 3.8, 1.5, 0.3],  
 [5.4, 3.4, 1.7, 0.2],  
 [5.1, 3.7, 1.5, 0.4],  
 [4.6, 3.6, 1. , 0.2],  
 [5.1, 3.3, 1.7, 0.5],  
 [4.8, 3.4, 1.9, 0.2],  
 [5. , 3. , 1.6, 0.2],  
 [5. , 3.4, 1.6, 0.4],  
 [5.2, 3.5, 1.5, 0.2],  
 [5.2, 3.4, 1.4, 0.2],  
 [4.7, 3.2, 1.6, 0.2],  
 [4.8, 3.1, 1.6, 0.2],  
 [5.4, 3.4, 1.5, 0.4],  
 [5.2, 4.1, 1.5, 0.1],  
 [5.5, 4.2, 1.4, 0.2],  
 [4.9, 3.1, 1.5, 0.2],  
 [5. , 3.2, 1.2, 0.2],  
 [5.5, 3.5, 1.3, 0.2],  
 [4.9, 3.6, 1.4, 0.1],  
 [4.4, 3. , 1.3, 0.2],  
 [5.1, 3.4, 1.5, 0.2],  
 [5. , 3.5, 1.3, 0.3],  
 [4.5, 2.3, 1.3, 0.3],  
 [4.4, 3.2, 1.3, 0.2],  
 [5. , 3.5, 1.6, 0.6],  
 [5.1, 3.8, 1.9, 0.4],  
 [4.8, 3. , 1.4, 0.3],  
 [5.1, 3.8, 1.6, 0.2],  
 [4.6, 3.2, 1.4, 0.2],  
 [5.3, 3.7, 1.5, 0.2],  
 [5. , 3.3, 1.4, 0.2],  
 [7. , 3.2, 4.7, 1.4],  
 [6.4, 3.2, 4.5, 1.5],  
 [6.9, 3.1, 4.9, 1.5],  
 [5.5, 2.3, 4. , 1.3],  
 [6.5, 2.8, 4.6, 1.5],  
 [5.7, 2.8, 4.5, 1.3],  
 [6.3, 3.3, 4.7, 1.6],  
 [4.9, 2.4, 3.3, 1. ],  
 [6.6, 2.9, 4.6, 1.3],  
 [5.2, 2.7, 3.9, 1.4],  
 [5. , 2. , 3.5, 1. ],  
 [5.9, 3. , 4.2, 1.5],  
 [6. , 2.2, 4. , 1. ],  
 [6.1, 2.9, 4.7, 1.4],
```

```
[5.6, 2.9, 3.6, 1.3],  
[6.7, 3.1, 4.4, 1.4],  
[5.6, 3. , 4.5, 1.5],  
[5.8, 2.7, 4.1, 1. ],  
[6.2, 2.2, 4.5, 1.5],  
[5.6, 2.5, 3.9, 1.1],  
[5.9, 3.2, 4.8, 1.8],  
[6.1, 2.8, 4. , 1.3],  
[6.3, 2.5, 4.9, 1.5],  
[6.1, 2.8, 4.7, 1.2],  
[6.4, 2.9, 4.3, 1.3],  
[6.6, 3. , 4.4, 1.4],  
[6.8, 2.8, 4.8, 1.4],  
[6.7, 3. , 5. , 1.7],  
[6. , 2.9, 4.5, 1.5],  
[5.7, 2.6, 3.5, 1. ],  
[5.5, 2.4, 3.8, 1.1],  
[5.5, 2.4, 3.7, 1. ],  
[5.8, 2.7, 3.9, 1.2],  
[6. , 2.7, 5.1, 1.6],  
[5.4, 3. , 4.5, 1.5],  
[6. , 3.4, 4.5, 1.6],  
[6.7, 3.1, 4.7, 1.5],  
[6.3, 2.3, 4.4, 1.3],  
[5.6, 3. , 4.1, 1.3],  
[5.5, 2.5, 4. , 1.3],  
[5.5, 2.6, 4.4, 1.2],  
[6.1, 3. , 4.6, 1.4],  
[5.8, 2.6, 4. , 1.2],  
[5. , 2.3, 3.3, 1. ],  
[5.6, 2.7, 4.2, 1.3],  
[5.7, 3. , 4.2, 1.2],  
[5.7, 2.9, 4.2, 1.3],  
[6.2, 2.9, 4.3, 1.3],  
[5.1, 2.5, 3. , 1.1],  
[5.7, 2.8, 4.1, 1.3],  
[6.3, 3.3, 6. , 2.5],  
[5.8, 2.7, 5.1, 1.9],  
[7.1, 3. , 5.9, 2.1],  
[6.3, 2.9, 5.6, 1.8],  
[6.5, 3. , 5.8, 2.2],  
[7.6, 3. , 6.6, 2.1],  
[4.9, 2.5, 4.5, 1.7],  
[7.3, 2.9, 6.3, 1.8],  
[6.7, 2.5, 5.8, 1.8],  
[7.2, 3.6, 6.1, 2.5],  
[6.5, 3.2, 5.1, 2. ],  
[6.4, 2.7, 5.3, 1.9],  
[6.8, 3. , 5.5, 2.1],  
[5.7, 2.5, 5. , 2. ],  
[5.8, 2.8, 5.1, 2.4],  
[6.4, 3.2, 5.3, 2.3],  
[6.5, 3. , 5.5, 1.8],  
[7.7, 3.8, 6.7, 2.2],  
[7.7, 2.6, 6.9, 2.3],  
[6. , 2.2, 5. , 1.5],  
[6.9, 3.2, 5.7, 2.3],  
[5.6, 2.8, 4.9, 2. ],  
[7.7, 2.8, 6.7, 2. ],  
[6.3, 2.7, 4.9, 1.8],  
[6.7, 3.3, 5.7, 2.1],  
[7.2, 3.2, 6. , 1.8],  
[6.2, 2.8, 4.8, 1.8],  
[6.1, 3. , 4.9, 1.8],
```

```
[6.4, 2.8, 5.6, 2.1],  
[7.2, 3. , 5.8, 1.6],  
[7.4, 2.8, 6.1, 1.9],  
[7.9, 3.8, 6.4, 2. ],  
[6.4, 2.8, 5.6, 2.2],  
[6.3, 2.8, 5.1, 1.5],  
[6.1, 2.6, 5.6, 1.4],  
[7.7, 3. , 6.1, 2.3],  
[6.3, 3.4, 5.6, 2.4],  
[6.4, 3.1, 5.5, 1.8],  
[6. , 3. , 4.8, 1.8],  
[6.9, 3.1, 5.4, 2.1],  
[6.7, 3.1, 5.6, 2.4],  
[6.9, 3.1, 5.1, 2.3],  
[5.8, 2.7, 5.1, 1.9],  
[6.8, 3.2, 5.9, 2.3],  
[6.7, 3.3, 5.7, 2.5],  
[6.7, 3. , 5.2, 2.3],  
[6.3, 2.5, 5. , 1.9],  
[6.5, 3. , 5.2, 2. ],  
[6.2, 3.4, 5.4, 2.3],  
[5.9, 3. , 5.1, 1.8]])
```

```
In [27]: y=iris.target  
y
```

```
Out[27]: array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,  
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,  
0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,  
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,  
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,  
2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,  
2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2])
```

```
In [29]: from sklearn.model_selection import train_test_split  
X_train, X_test, y_train, y_test=train_test_split(X,y, test_size=0.2,random_state=42)
```

```
In [30]: X_train.shape
```

```
Out[30]: (120, 4)
```

```
In [31]: X_test.shape
```

```
Out[31]: (30, 4)
```

```
In [32]: from sklearn.tree import DecisionTreeClassifier
```

```
In [33]: clf=DecisionTreeClassifier()
```

```
In [34]: clf.fit(X_train,y_train)
```

```
Out[34]: DecisionTreeClassifier()
```

```
In [35]: y_pred=clf.predict(X_test)
```

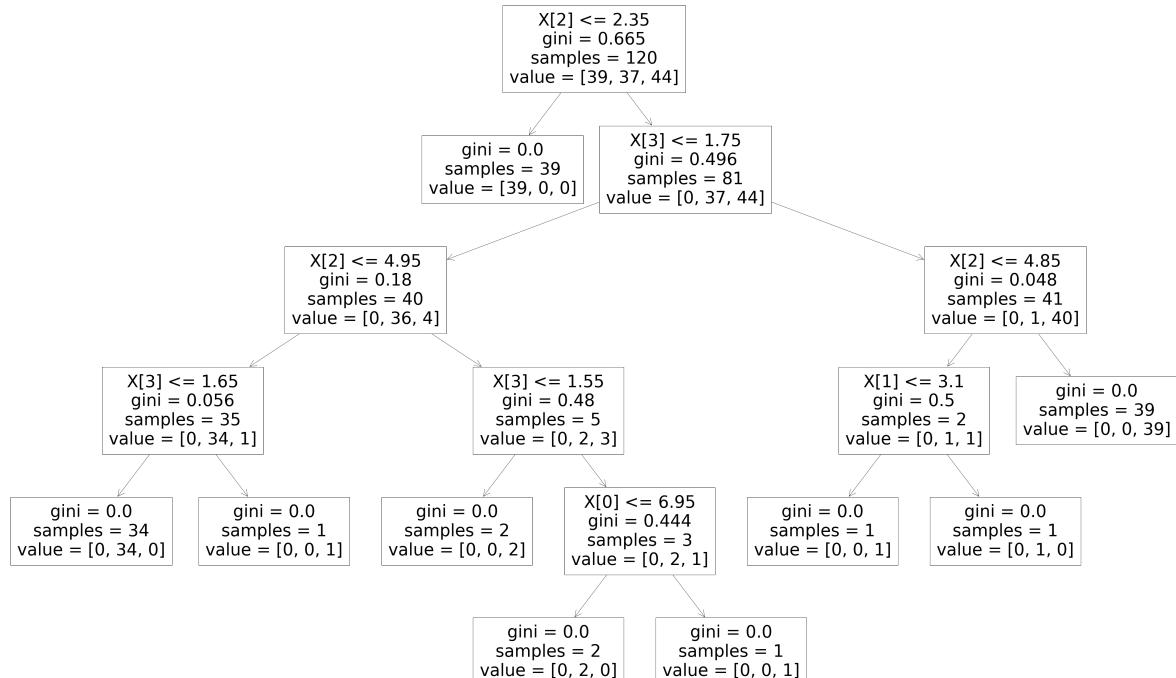
```
In [37]: from sklearn.metrics import accuracy_score  
accuracy_score(y_test,y_pred)
```

```
Out[37]: 1.0
```

```
In [39]: from sklearn.tree import plot_tree
```

```
In [41]: from matplotlib.pyplot import rcParams
rcParams['figure.figsize'] = 80,50
plot_tree(clf)
```

```
Out[41]: [Text(0.5, 0.9166666666666666, 'X[2] <= 2.35\ngini = 0.665\nsamples = 120\nvalue = [39, 37, 44]'),
Text(0.4230769230769231, 0.75, 'gini = 0.0\nsamples = 39\nvalue = [39, 0, 0]'),
Text(0.5769230769230769, 0.75, 'X[3] <= 1.75\ngini = 0.496\nsamples = 81\nvalue = [0, 37, 44]'),
Text(0.3076923076923077, 0.5833333333333334, 'X[2] <= 4.95\ngini = 0.18\nsamples = 40\nvalue = [0, 36, 4]'),
Text(0.15384615384615385, 0.4166666666666667, 'X[3] <= 1.65\ngini = 0.056\nsamples = 35\nvalue = [0, 34, 1]'),
Text(0.07692307692307693, 0.25, 'gini = 0.0\nsamples = 34\nvalue = [0, 34, 0]'),
Text(0.23076923076923078, 0.25, 'gini = 0.0\nsamples = 1\nvalue = [0, 0, 1]'),
Text(0.46153846153846156, 0.4166666666666667, 'X[3] <= 1.55\ngini = 0.48\nsamples = 5\nvalue = [0, 2, 3]'),
Text(0.38461538461538464, 0.25, 'gini = 0.0\nsamples = 2\nvalue = [0, 0, 2]'),
Text(0.5384615384615384, 0.25, 'X[0] <= 6.95\ngini = 0.444\nsamples = 3\nvalue = [0, 2, 1]'),
Text(0.46153846153846156, 0.0833333333333333, 'gini = 0.0\nsamples = 2\nvalue = [0, 2, 0]'),
Text(0.6153846153846154, 0.0833333333333333, 'gini = 0.0\nsamples = 1\nvalue = [0, 0, 1]'),
Text(0.8461538461538461, 0.5833333333333334, 'X[2] <= 4.85\ngini = 0.048\nsamples = 41\nvalue = [0, 1, 40]'),
Text(0.7692307692307693, 0.4166666666666667, 'X[1] <= 3.1\ngini = 0.5\nsamples = 2\nvalue = [0, 1, 1]'),
Text(0.6923076923076923, 0.25, 'gini = 0.0\nsamples = 1\nvalue = [0, 0, 1]'),
Text(0.8461538461538461, 0.25, 'gini = 0.0\nsamples = 1\nvalue = [0, 1, 0]'),
Text(0.9230769230769231, 0.4166666666666667, 'gini = 0.0\nsamples = 39\nvalue = [0, 0, 39]')]
```



```
In [20]: #Program 2
```

```
In [42]: import numpy as np
import pandas as pd
```

```
In [43]: data=pd.read_csv('Social_Network_Ads.csv')
data.head()
```

	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	Male	19	19000	0
1	15810944	Male	35	20000	0
2	15668575	Female	26	43000	0
3	15603246	Female	27	57000	0
4	15804002	Male	19	76000	0

```
In [46]: data['Gender'].replace({'Male':0,'Female':1},inplace=True)
data.head()
```

	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	0	19	19000	0
1	15810944	0	35	20000	0
2	15668575	1	26	43000	0
3	15603246	1	27	57000	0
4	15804002	0	19	76000	0

```
In [47]: X=data.iloc[:, 1:4].values
y=data.iloc[:, -1].values
```

```
In [48]: X.shape
```

```
Out[48]: (400, 3)
```

```
In [49]: y.shape
```

```
Out[49]: (400,)
```

```
In [50]: #clf1=DecisionTreeClassifier(max_depth=3)
clf1=DecisionTreeClassifier()
```

```
In [52]: clf1.fit(X,y)
```

```
Out[52]: DecisionTreeClassifier()
```

```
In [53]: rcParams['figure.figsize'] = 80,50
plot_tree(clf1)
```

```

Out[53]: [Text(0.44502314814814814, 0.96666666666666667, 'X[1] <= 42.5\ngini = 0.459\nsamples = 400\nvalue = [257, 143']),
Text(0.24189814814814814, 0.9, 'X[2] <= 90500.0\ngini = 0.271\nsamples = 285\nvalue = [239, 46']),
Text(0.11342592592592593, 0.83333333333333334, 'X[1] <= 36.5\ngini = 0.072\nsamples = 241\nvalue = [232, 9']),
Text(0.09490740740740741, 0.76666666666666667, 'gini = 0.0\nsamples = 162\nvalue = [162, 0']),
Text(0.13194444444444445, 0.76666666666666667, 'X[2] <= 83500.0\ngini = 0.202\nsamples = 79\nvalue = [70, 9']),
Text(0.11342592592592593, 0.7, 'X[2] <= 67500.0\ngini = 0.165\nsamples = 77\nvalue = [70, 7']),
Text(0.09490740740740741, 0.6333333333333333, 'gini = 0.0\nsamples = 40\nvalue = [40, 0']),
Text(0.13194444444444445, 0.6333333333333333, 'X[2] <= 70500.0\ngini = 0.307\nsamples = 37\nvalue = [30, 7']),
Text(0.0833333333333333, 0.56666666666666667, 'X[0] <= 0.5\ngini = 0.5\nsamples = 2\nvalue = [1, 1']),
Text(0.06481481481481481, 0.5, 'gini = 0.0\nsamples = 1\nvalue = [0, 1']),
Text(0.10185185185185185, 0.5, 'gini = 0.0\nsamples = 1\nvalue = [1, 0']),
Text(0.18055555555555555, 0.56666666666666667, 'X[1] <= 41.5\ngini = 0.284\nsamples = 35\nvalue = [29, 6']),
Text(0.1388888888888889, 0.5, 'X[1] <= 40.5\ngini = 0.231\nsamples = 30\nvalue = [26, 4']),
Text(0.12037037037037036, 0.4333333333333335, 'X[2] <= 77500.0\ngini = 0.287\nsamples = 23\nvalue = [19, 4']),
Text(0.05555555555555555, 0.36666666666666664, 'X[1] <= 38.5\ngini = 0.219\nsamples = 16\nvalue = [14, 2']),
Text(0.037037037037037035, 0.3, 'gini = 0.0\nsamples = 7\nvalue = [7, 0']),
Text(0.07407407407407407, 0.3, 'X[2] <= 71500.0\ngini = 0.346\nsamples = 9\nvalue = [7, 2']),
Text(0.037037037037037035, 0.2333333333333334, 'X[1] <= 39.5\ngini = 0.444\nsamples = 3\nvalue = [2, 1']),
Text(0.018518518518518517, 0.16666666666666666, 'gini = 0.0\nsamples = 2\nvalue = [2, 0']),
Text(0.05555555555555555, 0.16666666666666666, 'gini = 0.0\nsamples = 1\nvalue = [0, 1']),
Text(0.1111111111111111, 0.2333333333333334, 'X[1] <= 39.5\ngini = 0.278\nsamples = 6\nvalue = [5, 1']),
Text(0.09259259259259259, 0.16666666666666666, 'X[0] <= 0.5\ngini = 0.444\nsamples = 3\nvalue = [2, 1']),
Text(0.07407407407407407, 0.1, 'gini = 0.0\nsamples = 1\nvalue = [1, 0']),
Text(0.1111111111111111, 0.1, 'X[2] <= 74000.0\ngini = 0.5\nsamples = 2\nvalue = [1, 1']),
Text(0.09259259259259259, 0.03333333333333333, 'gini = 0.0\nsamples = 1\nvalue = [1, 0']),
Text(0.12962962962962962, 0.03333333333333333, 'gini = 0.0\nsamples = 1\nvalue = [0, 1']),
Text(0.12962962962962962, 0.16666666666666666, 'gini = 0.0\nsamples = 3\nvalue = [3, 0']),
Text(0.18518518518518517, 0.36666666666666664, 'X[1] <= 37.5\ngini = 0.408\nsamples = 7\nvalue = [5, 2']),
Text(0.16666666666666666, 0.3, 'X[2] <= 79500.0\ngini = 0.5\nsamples = 4\nvalue = [2, 2']),
Text(0.14814814814814814, 0.2333333333333334, 'gini = 0.0\nsamples = 2\nvalue = [0, 2']),
Text(0.18518518518518517, 0.2333333333333334, 'gini = 0.0\nsamples = 2\nvalue = [2, 0]),
Text(0.2037037037037037, 0.3, 'gini = 0.0\nsamples = 3\nvalue = [3, 0']),
Text(0.1574074074074074, 0.4333333333333335, 'gini = 0.0\nsamples = 7\nvalue = [7, 0']),
Text(0.2222222222222222, 0.5, 'X[2] <= 74000.0\ngini = 0.48\nsamples = 5\nvalue = [3, 2']),
Text(0.2037037037037037, 0.4333333333333335, 'gini = 0.0\nsamples = 1\nvalue =

```

```
[0, 1']),
Text(0.24074074074074073, 0.4333333333333335, 'X[2] <= 79500.0\ngini = 0.375\nsa
mple = 4\nvalue = [3, 1']),
Text(0.2222222222222222, 0.3666666666666666, 'gini = 0.0\nsamples = 2\nvalue =
[2, 0']),
Text(0.25925925925925924, 0.3666666666666666, 'X[0] <= 0.5\ngini = 0.5\nsamples
= 2\nvalue = [1, 1']),
Text(0.24074074074074073, 0.3, 'gini = 0.0\nsamples = 1\nvalue = [1, 0']),
Text(0.2777777777777778, 0.3, 'gini = 0.0\nsamples = 1\nvalue = [0, 1']),
Text(0.15046296296296297, 0.7, 'gini = 0.0\nsamples = 2\nvalue = [0, 2']),
Text(0.37037037037037035, 0.8333333333333334, 'X[2] <= 119000.0\ngini = 0.268\nsa
mple = 44\nvalue = [7, 37']),
Text(0.35185185185185186, 0.7666666666666667, 'X[2] <= 107500.0\ngini = 0.413\nsa
mple = 24\nvalue = [7, 17']),
Text(0.3333333333333333, 0.7, 'gini = 0.0\nsamples = 11\nvalue = [0, 11']),
Text(0.37037037037037035, 0.7, 'X[1] <= 30.5\ngini = 0.497\nsamples = 13\nvalue =
[7, 6']),
Text(0.35185185185185186, 0.6333333333333333, 'gini = 0.0\nsamples = 2\nvalue =
[2, 0]),
Text(0.3888888888888889, 0.6333333333333333, 'X[2] <= 117500.0\ngini = 0.496\nsa
mple = 11\nvalue = [5, 6']),
Text(0.37037037037037035, 0.5666666666666667, 'X[1] <= 40.0\ngini = 0.494\nsample
s = 9\nvalue = [5, 4']),
Text(0.35185185185185186, 0.5, 'X[2] <= 110000.0\ngini = 0.469\nsamples = 8\nvalu
e = [5, 3']),
Text(0.3333333333333333, 0.4333333333333335, 'gini = 0.0\nsamples = 1\nvalue =
[1, 0]),
Text(0.37037037037037035, 0.4333333333333335, 'X[2] <= 114000.0\ngini = 0.49\nsa
mple = 7\nvalue = [4, 3']),
Text(0.3333333333333333, 0.3666666666666664, 'X[1] <= 33.5\ngini = 0.5\nsamples
= 4\nvalue = [2, 2']),
Text(0.3148148148148148, 0.3, 'gini = 0.0\nsamples = 1\nvalue = [1, 0]),
Text(0.35185185185185186, 0.3, 'X[1] <= 36.0\ngini = 0.444\nsamples = 3\nvalue =
[1, 2']),
Text(0.3333333333333333, 0.2333333333333334, 'gini = 0.0\nsamples = 1\nvalue =
[0, 1]),
Text(0.37037037037037035, 0.2333333333333334, 'X[2] <= 112500.0\ngini = 0.5\nsa
mple = 2\nvalue = [1, 1']),
Text(0.35185185185185186, 0.1666666666666666, 'gini = 0.0\nsamples = 1\nvalue =
[1, 0]),
Text(0.3888888888888889, 0.1666666666666666, 'gini = 0.0\nsamples = 1\nvalue =
[0, 1]),
Text(0.4074074074074074, 0.3666666666666664, 'X[1] <= 33.0\ngini = 0.444\nsample
s = 3\nvalue = [2, 1']),
Text(0.3888888888888889, 0.3, 'gini = 0.5\nsamples = 2\nvalue = [1, 1']),
Text(0.42592592592592593, 0.3, 'gini = 0.0\nsamples = 1\nvalue = [1, 0]),
Text(0.3888888888888889, 0.5, 'gini = 0.0\nsamples = 1\nvalue = [0, 1]),
Text(0.4074074074074074, 0.5666666666666667, 'gini = 0.0\nsamples = 2\nvalue =
[0, 2]),
Text(0.3888888888888889, 0.7666666666666667, 'gini = 0.0\nsamples = 20\nvalue =
[0, 20]),
Text(0.6481481481481481, 0.9, 'X[1] <= 46.5\ngini = 0.264\nsamples = 115\nvalue =
[18, 97]),
Text(0.5231481481481481, 0.8333333333333334, 'X[2] <= 35500.0\ngini = 0.444\nsa
mple = 24\nvalue = [8, 16]),
Text(0.46296296296296297, 0.7666666666666667, 'X[2] <= 22500.0\ngini = 0.219\nsa
mple = 8\nvalue = [1, 7]),
Text(0.4444444444444444, 0.7, 'X[1] <= 45.5\ngini = 0.444\nsamples = 3\nvalue =
[1, 2]),
Text(0.42592592592592593, 0.6333333333333333, 'gini = 0.0\nsamples = 2\nvalue =
[0, 2]),
Text(0.46296296296296297, 0.6333333333333333, 'gini = 0.0\nsamples = 1\nvalue =
[1, 0]),
Text(0.48148148148148145, 0.7, 'gini = 0.0\nsamples = 5\nvalue = [0, 5]),
```

```

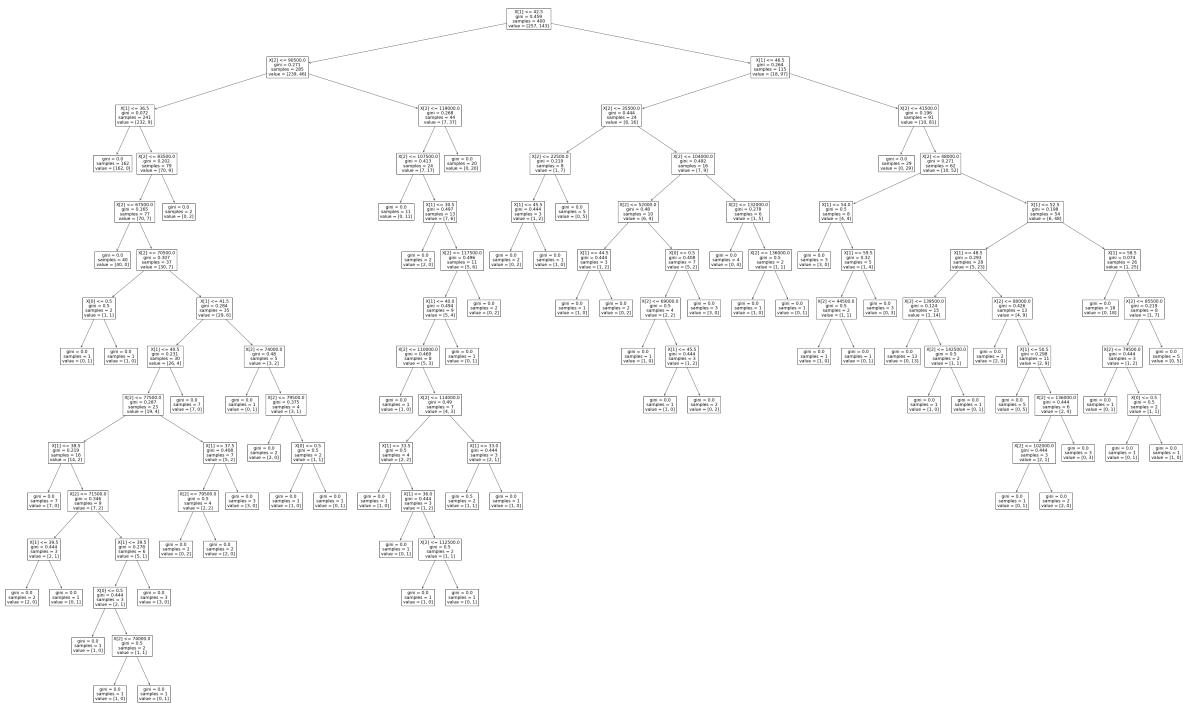
Text(0.5833333333333334, 0.766666666666666667, 'X[2] <= 104000.0\ngini = 0.492\nsamples = 16\nvalue = [7, 9]'),
Text(0.5370370370370371, 0.7, 'X[2] <= 52000.0\ngini = 0.48\nsamples = 10\nvalue = [6, 4]'),
Text(0.5, 0.6333333333333333, 'X[1] <= 44.5\ngini = 0.444\nsamples = 3\nvalue = [1, 2]'),
Text(0.48148148148148145, 0.566666666666666667, 'gini = 0.0\nsamples = 1\nvalue = [1, 0]'),
Text(0.5185185185185185, 0.566666666666666667, 'gini = 0.0\nsamples = 2\nvalue = [0, 2]'),
Text(0.5740740740740741, 0.6333333333333333, 'X[0] <= 0.5\ngini = 0.408\nsamples = 7\nvalue = [5, 2]'),
Text(0.5555555555555556, 0.566666666666666667, 'X[2] <= 69000.0\ngini = 0.5\nsamples = 4\nvalue = [2, 2]'),
Text(0.5370370370370371, 0.5, 'gini = 0.0\nsamples = 1\nvalue = [1, 0]'),
Text(0.5740740740740741, 0.5, 'X[1] <= 45.5\ngini = 0.444\nsamples = 3\nvalue = [1, 2]'),
Text(0.5555555555555556, 0.4333333333333335, 'gini = 0.0\nsamples = 1\nvalue = [1, 0]'),
Text(0.5925925925925926, 0.4333333333333335, 'gini = 0.0\nsamples = 2\nvalue = [0, 2]'),
Text(0.5925925925925926, 0.566666666666666667, 'gini = 0.0\nsamples = 3\nvalue = [3, 0]'),
Text(0.6296296296296297, 0.7, 'X[2] <= 132000.0\ngini = 0.278\nsamples = 6\nvalue = [1, 5]'),
Text(0.6111111111111112, 0.6333333333333333, 'gini = 0.0\nsamples = 4\nvalue = [0, 4]'),
Text(0.6481481481481481, 0.6333333333333333, 'X[2] <= 136000.0\ngini = 0.5\nsamples = 2\nvalue = [1, 1]'),
Text(0.6296296296296297, 0.566666666666666667, 'gini = 0.0\nsamples = 1\nvalue = [1, 0]'),
Text(0.6666666666666666, 0.566666666666666667, 'gini = 0.0\nsamples = 1\nvalue = [0, 1]'),
Text(0.7731481481481481, 0.8333333333333334, 'X[2] <= 41500.0\ngini = 0.196\nsamples = 91\nvalue = [10, 81]'),
Text(0.7546296296296297, 0.766666666666666667, 'gini = 0.0\nsamples = 29\nvalue = [0, 29]'),
Text(0.7916666666666666, 0.766666666666666667, 'X[2] <= 48000.0\ngini = 0.271\nsamples = 62\nvalue = [10, 52]'),
Text(0.7037037037037037, 0.7, 'X[1] <= 54.0\ngini = 0.5\nsamples = 8\nvalue = [4, 4]'),
Text(0.6851851851851852, 0.6333333333333333, 'gini = 0.0\nsamples = 3\nvalue = [3, 0]'),
Text(0.7222222222222222, 0.6333333333333333, 'X[1] <= 59.5\ngini = 0.32\nsamples = 5\nvalue = [1, 4]'),
Text(0.7037037037037037, 0.566666666666666667, 'X[2] <= 44500.0\ngini = 0.5\nsamples = 2\nvalue = [1, 1]'),
Text(0.6851851851851852, 0.5, 'gini = 0.0\nsamples = 1\nvalue = [1, 0]'),
Text(0.7222222222222222, 0.5, 'gini = 0.0\nsamples = 1\nvalue = [0, 1]'),
Text(0.7407407407407407, 0.566666666666666667, 'gini = 0.0\nsamples = 3\nvalue = [0, 3]'),
Text(0.8796296296296297, 0.7, 'X[1] <= 52.5\ngini = 0.198\nsamples = 54\nvalue = [6, 48]'),
Text(0.8148148148148148, 0.6333333333333333, 'X[1] <= 48.5\ngini = 0.293\nsamples = 28\nvalue = [5, 23]'),
Text(0.7777777777777778, 0.566666666666666667, 'X[2] <= 139500.0\ngini = 0.124\nsamples = 15\nvalue = [1, 14]'),
Text(0.7592592592592593, 0.5, 'gini = 0.0\nsamples = 13\nvalue = [0, 13]'),
Text(0.7962962962962963, 0.5, 'X[2] <= 142500.0\ngini = 0.5\nsamples = 2\nvalue = [1, 1]'),
Text(0.7777777777777778, 0.4333333333333335, 'gini = 0.0\nsamples = 1\nvalue = [1, 0]'),
Text(0.8148148148148148, 0.4333333333333335, 'gini = 0.0\nsamples = 1\nvalue = [0, 1]),

```

```

Text(0.8518518518518519, 0.566666666666666667, 'X[2] <= 80000.0\ngini = 0.426\nsamples = 13\nvalue = [4, 9']),
Text(0.8333333333333334, 0.5, 'gini = 0.0\nsamples = 2\nvalue = [2, 0']),
Text(0.8703703703703703, 0.5, 'X[1] <= 50.5\ngini = 0.298\nsamples = 11\nvalue = [2, 9']),
Text(0.8518518518518519, 0.4333333333333335, 'gini = 0.0\nsamples = 5\nvalue = [0, 5']),
Text(0.8888888888888888, 0.4333333333333335, 'X[2] <= 136000.0\ngini = 0.444\nsamples = 6\nvalue = [2, 4']),
Text(0.8703703703703703, 0.36666666666666664, 'X[2] <= 102000.0\ngini = 0.444\nsamples = 3\nvalue = [2, 1']),
Text(0.8518518518518519, 0.3, 'gini = 0.0\nsamples = 1\nvalue = [0, 1']),
Text(0.8888888888888888, 0.3, 'gini = 0.0\nsamples = 2\nvalue = [2, 0']),
Text(0.9074074074074074, 0.36666666666666664, 'gini = 0.0\nsamples = 3\nvalue = [0, 3']),
Text(0.9444444444444444, 0.6333333333333333, 'X[1] <= 58.5\ngini = 0.074\nsamples = 26\nvalue = [1, 25']),
Text(0.9259259259259259, 0.5666666666666667, 'gini = 0.0\nsamples = 18\nvalue = [0, 18']),
Text(0.9629629629629629, 0.5666666666666667, 'X[2] <= 85500.0\ngini = 0.219\nsamples = 8\nvalue = [1, 7']),
Text(0.9444444444444444, 0.5, 'X[2] <= 79500.0\ngini = 0.444\nsamples = 3\nvalue = [1, 2']),
Text(0.9259259259259259, 0.4333333333333335, 'gini = 0.0\nsamples = 1\nvalue = [0, 1']),
Text(0.9629629629629629, 0.4333333333333335, 'X[0] <= 0.5\ngini = 0.5\nsamples = 2\nvalue = [1, 1']),
Text(0.9444444444444444, 0.36666666666666664, 'gini = 0.0\nsamples = 1\nvalue = [0, 1']),
Text(0.9814814814814815, 0.36666666666666664, 'gini = 0.0\nsamples = 1\nvalue = [1, 0']),
Text(0.9814814814814815, 0.5, 'gini = 0.0\nsamples = 5\nvalue = [0, 5]')

```



```
In [54]: clf1=DecisionTreeClassifier(max_depth=3)
```

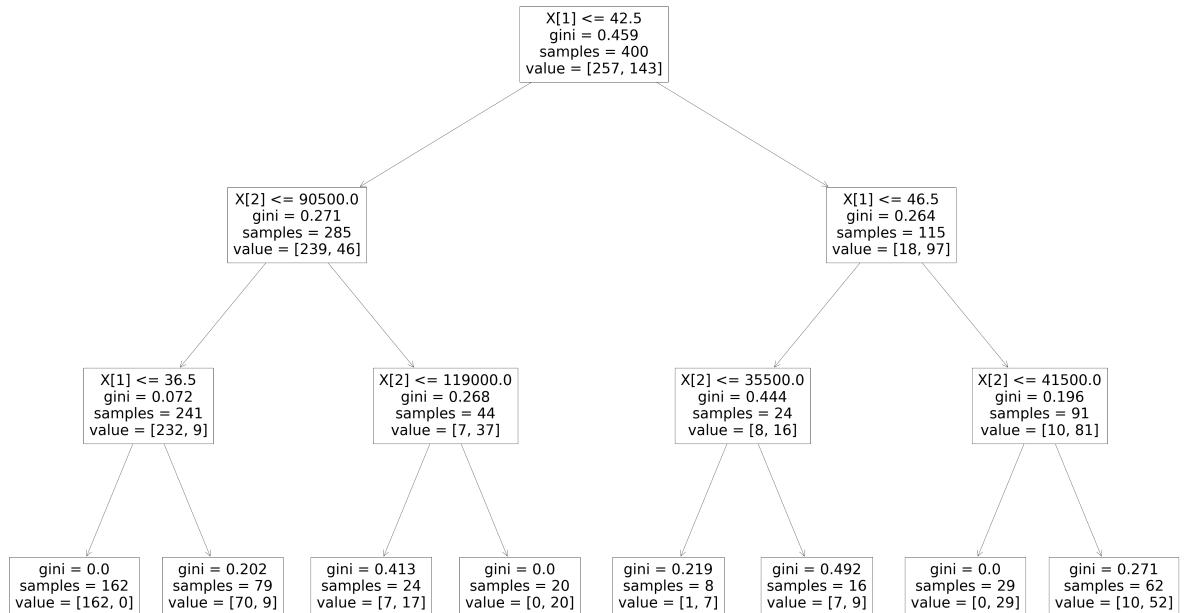
```
In [55]: clf1.fit(X,y)
```

```
Out[55]: DecisionTreeClassifier(max_depth=3)
```

```
In [56]: rcParams['figure.figsize'] = 80,50
```

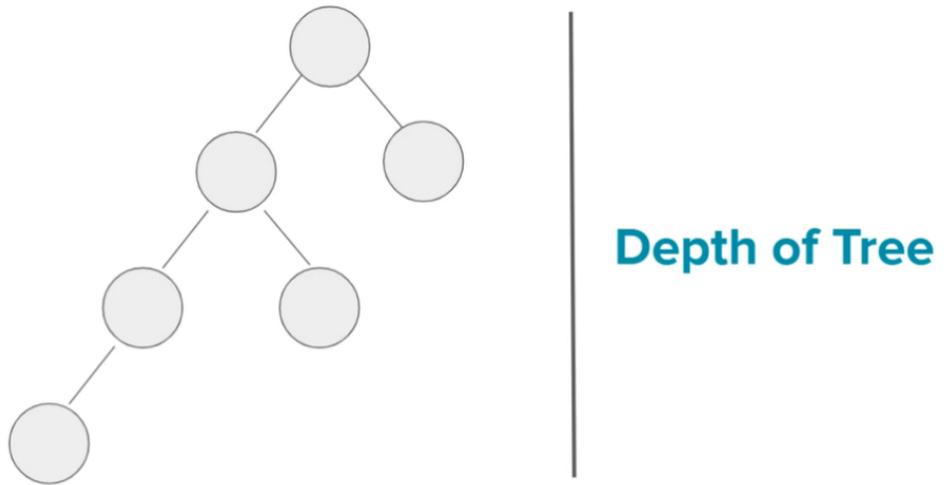
```
plot_tree(clf1)

Out[56]: [Text(0.5, 0.875, 'X[1] <= 42.5\ngini = 0.459\nsamples = 400\nvalue = [257, 143]'),
Text(0.25, 0.625, 'X[2] <= 90500.0\ngini = 0.271\nsamples = 285\nvalue = [239, 46]'),
Text(0.125, 0.375, 'X[1] <= 36.5\ngini = 0.072\nsamples = 241\nvalue = [232, 9]'),
Text(0.0625, 0.125, 'gini = 0.0\nsamples = 162\nvalue = [162, 0]'),
Text(0.1875, 0.125, 'gini = 0.202\nsamples = 79\nvalue = [70, 9]'),
Text(0.375, 0.375, 'X[2] <= 119000.0\ngini = 0.268\nsamples = 44\nvalue = [7, 37]'),
Text(0.3125, 0.125, 'gini = 0.413\nsamples = 24\nvalue = [7, 17]'),
Text(0.4375, 0.125, 'gini = 0.0\nsamples = 20\nvalue = [0, 20]'),
Text(0.75, 0.625, 'X[1] <= 46.5\ngini = 0.264\nsamples = 115\nvalue = [18, 97]'),
Text(0.625, 0.375, 'X[2] <= 35500.0\ngini = 0.444\nsamples = 24\nvalue = [8, 16]'),
Text(0.5625, 0.125, 'gini = 0.219\nsamples = 8\nvalue = [1, 7]'),
Text(0.6875, 0.125, 'gini = 0.492\nsamples = 16\nvalue = [7, 9]'),
Text(0.875, 0.375, 'X[2] <= 41500.0\ngini = 0.196\nsamples = 91\nvalue = [10, 81]'),
Text(0.8125, 0.125, 'gini = 0.0\nsamples = 29\nvalue = [0, 29]'),
Text(0.9375, 0.125, 'gini = 0.271\nsamples = 62\nvalue = [10, 52]')]
```

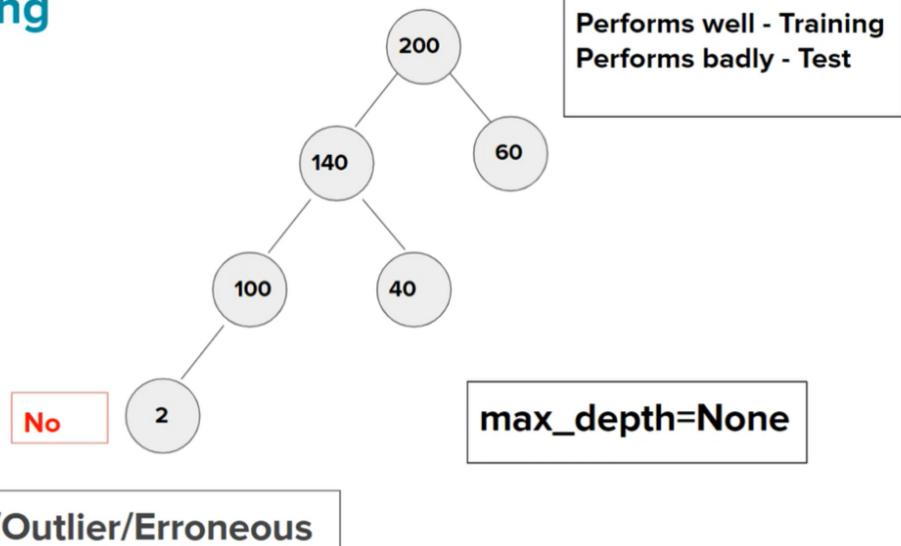


Overfitting in Decision Tree

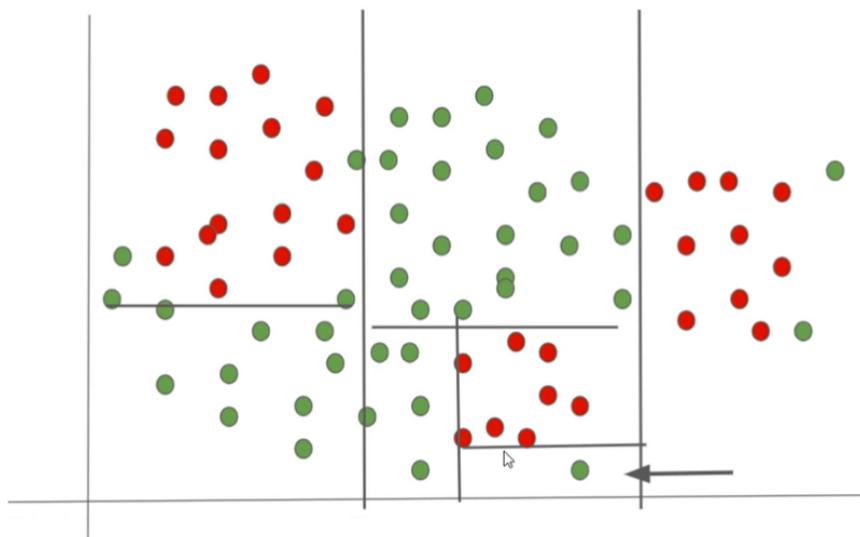
Overfitting/Underfitting



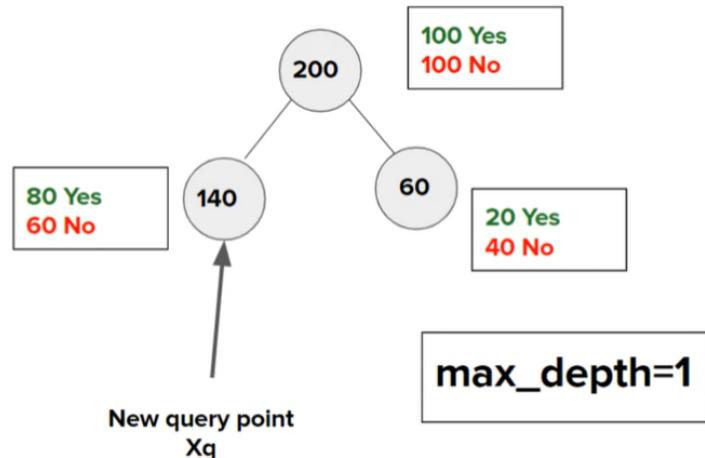
Overfitting



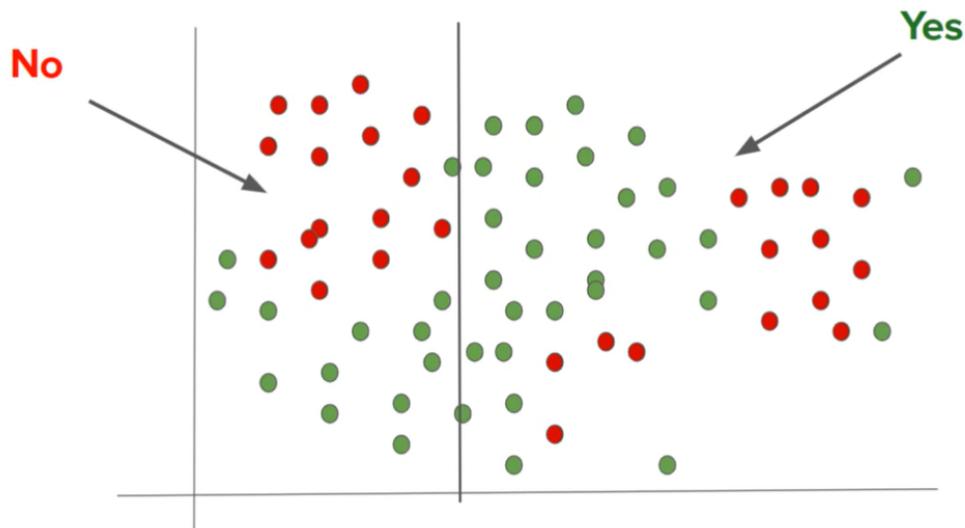
Geometric Intuition of Overfitting



Underfitting



Geometric Intuition of Underfitting



Decision Tree Hyperparameters In-depth Intuition

<https://dt-visualise.herokuapp.com/>

In []: <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>

Hyper-parameter Tuning using GridSearchCV

In [1]: `import numpy as np`

```
import pandas as pd
```

In [2]: `data=pd.read_csv('Social_Network_Ads.csv')
data.head()`

Out[2]:

	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	Male	19	19000	0
1	15810944	Male	35	20000	0
2	15668575	Female	26	43000	0
3	15603246	Female	27	57000	0
4	15804002	Male	19	76000	0

In [3]: `data['Gender'].replace({'Male':0,'Female':1},inplace=True)
data.head()`

Out[3]:

	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	0	19	19000	0
1	15810944	0	35	20000	0
2	15668575	1	26	43000	0
3	15603246	1	27	57000	0
4	15804002	0	19	76000	0

In [4]: `X=data.iloc[:, 1:4].values
y=data.iloc[:, -1].values`

In [5]: `from sklearn.preprocessing import StandardScaler
scaler=StandardScaler()`

In [6]: `X=scaler.fit_transform(X)
X`

Out[6]: `array([[-1.02020406, -1.78179743, -1.49004624],
 [-1.02020406, -0.25358736, -1.46068138],
 [0.98019606, -1.11320552, -0.78528968],
 ...,
 [0.98019606, 1.17910958, -1.46068138],
 [-1.02020406, -0.15807423, -1.07893824],
 [0.98019606, 1.08359645, -0.99084367]])`

In [13]: `from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.2,random_state=1)`

In [14]: `from sklearn.tree import DecisionTreeClassifier
clf=DecisionTreeClassifier()`

In [15]: `clf.fit(X_train,y_train)`

Out[15]: `DecisionTreeClassifier()`

In [16]: `y_pred=clf.predict(X_test)`

In [17]: `from sklearn.metrics import accuracy_score`

```
accuracy_score(y_test,y_pred)
```

Out[17]: 0.775

```
In [20]: param_dist={  
    "criterion": ["gini", "entropy"],  
    "max_depth": [1, 2, 3, 4, 5, 6, 7, None]  
}
```

```
In [21]: from sklearn.model_selection import GridSearchCV  
grid=GridSearchCV(clf,param_grid=param_dist, cv=10, n_jobs=-1)
```

```
In [22]: grid.fit(X_train,y_train)
```

```
Out[22]: GridSearchCV(cv=10, estimator=DecisionTreeClassifier(), n_jobs=-1,  
param_grid={'criterion': ['gini', 'entropy'],  
'max_depth': [1, 2, 3, 4, 5, 6, 7, None]})
```

```
In [37]: grid.best_estimator_
```

Out[37]: DecisionTreeClassifier(max_depth=2)

```
In [38]: grid.best_score_
```

Out[38]: 0.91875

```
In [39]: grid.best_params_
```

Out[39]: {'criterion': 'gini', 'max_depth': 2}

Decision Tree Regressor

Python Code-

```
In [44]: import pandas as pd  
#from pandas_datareader import data  
import numpy as np  
from sklearn.tree import DecisionTreeRegressor  
from sklearn.model_selection import train_test_split  
from sklearn import metrics  
from sklearn.metrics import r2_score  
from sklearn.datasets import load_boston  
from sklearn.model_selection import GridSearchCV
```

```
In [47]: boston = load_boston()  
df = pd.DataFrame(boston.data)  
df
```

Out[47]:	0	1	2	3	4	5	6	7	8	9	10	11	12	
	0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98
	1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14
	2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03
	3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94
	4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33
	
	501	0.06263	0.0	11.93	0.0	0.573	6.593	69.1	2.4786	1.0	273.0	21.0	391.99	9.67
	502	0.04527	0.0	11.93	0.0	0.573	6.120	76.7	2.2875	1.0	273.0	21.0	396.90	9.08
	503	0.06076	0.0	11.93	0.0	0.573	6.976	91.0	2.1675	1.0	273.0	21.0	396.90	5.64
	504	0.10959	0.0	11.93	0.0	0.573	6.794	89.3	2.3889	1.0	273.0	21.0	393.45	6.48
	505	0.04741	0.0	11.93	0.0	0.573	6.030	80.8	2.5050	1.0	273.0	21.0	396.90	7.88

506 rows × 13 columns

In [48]: `df.columns = boston.feature_names
df['MEDV'] = boston.target`

In [49]: `df.head()`

Out[49]:	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LST	
	0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.
	1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.
	2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.
	3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.
	4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.

In [50]: `X = df.iloc[:,0:13]
y = df.iloc[:,13]`

In [51]: `X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)`

In [37]: `rt = DecisionTreeRegressor(criterion = 'mse', max_depth=5)`

In [52]: `rt.fit(X_train,y_train)`

```
C:\Users\Shreeji\anaconda3\lib\site-packages\sklearn\tree\_classes.py:359: FutureWarning: Criterion 'mse' was deprecated in v1.0 and will be removed in version 1.2. Use `criterion='squared_error'` which is equivalent.
```

```
    warnings.warn(
```

Out[52]: `DecisionTreeRegressor(criterion='mse', max_depth=5)`

In [53]: `y_pred = rt.predict(X_test)`

In [54]: `r2_score(y_test,y_pred)`

Out[54]: 0.8833565347917997

Hyperparameter Tuning

```
In [56]: param_grid = {  
    'max_depth':[2,4,8,10,None],  
    'criterion':['mse','mae'],  
    'max_features':[0.25,0.5,1.0],  
    'min_samples_split':[0.25,0.5,1.0]  
}
```

```
In [42]: reg = GridSearchCV(DecisionTreeRegressor(),param_grid=param_grid)
```

```
In [57]: reg.fit(X_train,y_train)
```



```
C:\Users\Shreeji\anaconda3\lib\site-packages\sklearn\tree\_classes.py:366: FutureWarning: Criterion 'mae' was deprecated in v1.0 and will be removed in version 1.2.
Use `criterion='absolute_error'` which is equivalent.
    warnings.warn(
C:\Users\Shreeji\anaconda3\lib\site-packages\sklearn\tree\_classes.py:366: FutureWarning: Criterion 'mae' was deprecated in v1.0 and will be removed in version 1.2.
Use `criterion='absolute_error'` which is equivalent.
    warnings.warn(
C:\Users\Shreeji\anaconda3\lib\site-packages\sklearn\tree\_classes.py:359: FutureWarning: Criterion 'mse' was deprecated in v1.0 and will be removed in version 1.2.
Use `criterion='squared_error'` which is equivalent.
    warnings.warn(
Out[57]: GridSearchCV(estimator=DecisionTreeRegressor(),
                     param_grid={'criterion': ['mse', 'mae'],
                                 'max_depth': [2, 4, 8, 10, None],
                                 'max_features': [0.25, 0.5, 1.0],
                                 'min_samples_split': [0.25, 0.5, 1.0]})
```

In [58]: `reg.best_score_`

Out[58]: 0.6591758393991596

In [59]: `reg.best_params_`

Out[59]: {'criterion': 'mse',
 'max_depth': 8,
 'max_features': 0.5,
 'min_samples_split': 0.25}

Feature Importance

```
In [23]: for importance, name in sorted(zip(rt.feature_importances_, X_train.columns), reverse=True):
    print (name, importance)

RM 0.6344993240692439
LSTAT 0.20562720153418443
DIS 0.06744514557703217
CRIM 0.03550388785641197
NOX 0.02531597355560238
PTRATIO 0.01669708628286102
INDUS 0.007018566233811912
AGE 0.006176126174365166
CHAS 0.00117395935157392
B 0.0005427293649131195
ZN 0.0
TAX 0.0
RAD 0.0
```

In []:

In []:

In []:

2. Decision Trees for Interview call

For example,

- a human resources application contains a process for assessing a job candidate. The candidate receives a set of ratings during the interviews. These ratings are evaluated to determine whether to extend a job offer to the candidate.
- A decision tree is configured to automatically use the ratings as test conditions to decide whether the candidate is qualified.
- The decision starts at the top of the tree and proceeds downward. Each yes advances the evaluation.
- The result is either Not qualified or Eligible for job offer.



Attribute Selection Measures

While implementing a Decision tree, the main issue arises that how to select the best attribute for the root node and for sub-nodes.

So, to solve such problems there is a technique which is called as Attribute selection measure or ASM.

By this measurement, we can easily select the best attribute for the nodes of the tree.

There are two popular techniques for ASM, which are:

- Information Gain

- Gini Index

1. Information Gain:

Information gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute.

It calculates how much information a feature provides us about a class.

According to the value of information gain, we split the node and build the decision tree.

A decision tree algorithm always tries to maximize the value of information gain, and a node/attribute having the highest information gain is split first. It can be calculated using the below formula:

Information Gain= Entropy(S)- [(Weighted Avg) *Entropy(each feature)]

1. Entropy:

Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data. Entropy can be calculated as:

$$\text{Entropy}(S) = -P(\text{yes})\log_2 P(\text{yes}) - P(\text{no}) \log_2 P(\text{no})$$

Where,

- **S= Total number of samples**
- **P(yes)= probability of yes**
- **P(no)= probability of no**

2. Gini Index:

Gini index is a measure of impurity or purity used while creating a decision tree in the CART(Classification and Regression Tree) algorithm.

An attribute with the low Gini index should be preferred as compared to the high Gini index.

It only creates binary splits, and the CART algorithm uses the Gini index to create binary splits.

Gini index can be calculated using the below formula:

$$\text{Gini Index} = 1 - \sum_j P_j^2$$

Python Implementation of Decision Tree

Now we will implement the Decision tree using Python.

For this, we will use the dataset "user_data.csv," which we have used in previous classification models.

By using the same dataset, we can compare the Decision tree classifier with other classification models such as

- KNN
- SVM,
- LogisticRegression, etc.

Steps will also remain the same, which are given below:

- Data Pre-processing step
- Fitting a Decision-Tree algorithm to the Training set
- Predicting the test result
- Test accuracy of the result(Creation of Confusion matrix)
- Visualizing the test set result.

1. Data Pre-Processing Step:

Below is the code for the pre-processing step:

```
In [15]: # importing Libraries
import numpy as nm
import matplotlib.pyplot as mtp
import pandas as pd
```

```
In [16]: #importing datasets
data_set= pd.read_csv('user_data.csv')
data_set
```

	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	Male	19	19000	0
1	15810944	Male	35	20000	0
2	15668575	Female	26	43000	0
3	15603246	Female	27	57000	0
4	15804002	Male	19	76000	0
...
395	15691863	Female	46	41000	1
396	15706071	Male	51	23000	1
397	15654296	Female	50	20000	1
398	15755018	Male	36	33000	0
399	15594041	Female	49	36000	1

400 rows × 5 columns

```
In [17]: #Extracting Independent and dependent Variable
x= data_set.iloc[:, [2,3]].values
```

```
y= data_set.iloc[:, 4].values
```

In [18]: # Splitting the dataset into training and test set.
`from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test= train_test_split(x, y, test_size= 0.25, random_s`

In [19]: #feature Scaling
`from sklearn.preprocessing import StandardScaler
st_x= StandardScaler()
x_train= st_x.fit_transform(x_train)
x_test= st_x.transform(x_test)`

2. Fitting a Decision-Tree algorithm to the Training set

Now we will fit the model to the training set.

For this, we will import the `DecisionTreeClassifier` class from `sklearn.tree` library.

Below is the code for it:

In [20]: #Fitting Decision Tree classifier to the training set
`from sklearn.tree import DecisionTreeClassifier
classifier= DecisionTreeClassifier(criterion='entropy', random_state=0)
classifier.fit(x_train, y_train)`

Out[20]: `DecisionTreeClassifier(criterion='entropy', random_state=0)`

In the above code, we have created a classifier object, in which we have passed two main parameters;

"criterion='entropy':

Criterion is used to measure the quality of split, which is calculated by information gain given by entropy.

`random_state=0":`

For generating the random states.

3. Predicting the test result

Now we will predict the test set result.

We will create a new prediction vector `y_pred`. Below is the code for it:

In [21]: #Predicting the test set result
`y_pred= classifier.predict(x_test)`

4. Test accuracy of the result (Creation of Confusion matrix)

In the above output, we have seen that there were some incorrect predictions, so if we want to know the number of correct and incorrect predictions, we need to use the confusion matrix.

Below is the code for it:

```
In [22]: #Creating the Confusion matrix
from sklearn.metrics import confusion_matrix
cm= confusion_matrix(y_test, y_pred)
cm

Out[22]: array([[62,  6],
               [ 3, 29]], dtype=int64)
```

In the above output image, we can see the confusion matrix, which has $6+3=9$ incorrect predictions and $62+29=91$ correct predictions.

Therefore, we can say that compared to other classification models, the Decision Tree classifier made a good prediction.

Decision Tree Regression in Python

We will now go through a step-wise Python implementation of the Decision Tree Regression algorithm that we just discussed.

1. Importing necessary libraries

First, let us import some essential Python libraries.

```
In [23]: # Importing the Libraries
import numpy as np # for array operations
import pandas as pd # for working with DataFrames

# scikit-learn modules
from sklearn.model_selection import train_test_split # for splitting the data
from sklearn.metrics import mean_squared_error # for calculating the cost function
from sklearn.tree import DecisionTreeRegressor # for building the model
```

2. Importing the data set

The dataset consists of data related to petrol consumptions (in millions of gallons) for 48 US states.

This value is based upon several features such as the petrol tax (in cents), Average income (dollars), paved highways (in miles), and the proportion of the population with a driver's license.

We will be loading the data set using the `read_csv()` function from the `pandas` module and store it as a `pandas DataFrame` object.

```
In [25]: # Reading the data
dataset = pd.read_csv('petrol_consumption.csv')
dataset.head()
```

Out[25]:	Petrol_tax	Average_income	Paved_Highways	Population_Driver_licence(%)	Petrol_Consumption
0	9.0	3571	1976	0.525	541
1	9.0	4092	1250	0.572	524
2	9.0	3865	1586	0.580	561
3	7.5	4870	2351	0.529	414
4	8.0	4399	431	0.544	410

< >

3. Separating the features and the target variable

After loading the dataset, the independent variable (x) and the dependent variable (y) need to be separated.

Our concern is to model the relationships between the features (Petrol_tax, Average_income, etc.) and the target variable (Petrol_consumption) in the dataset.

```
In [26]: x = dataset.drop('Petrol_Consumption', axis = 1) # Features
y = dataset['Petrol_Consumption'] # Target
```

4. Splitting the data into a train set and a test set

We use the `train_test_split()` module of scikit-learn for splitting the data into a train set and a test set.

We will be using 20% of the available data as the testing set and the remaining data as the training set.

```
In [27]: # Splitting the dataset into training and testing set (80/20)
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.2, random_
```

5. Fitting the model to the training dataset

After splitting the data, let us initialize a Decision Tree Regressor model and fit it to the training data.

This is done with the help of `DecisionTreeRegressor()` module of scikit-learn.

```
In [28]: # Initializing the Decision Tree Regression model
model = DecisionTreeRegressor(random_state = 0)
# Fitting the Decision Tree Regression model to the data
model.fit(x_train, y_train)
```

```
Out[28]: DecisionTreeRegressor(random_state=0)
```

6. Calculating the loss after training

Let us now calculate the loss between the actual target values in the testing set and the values predicted by the model with the use of a cost function called the Root Mean Square Error (RMSE).

$$RMSE = \sqrt{\left(\frac{1}{n}\right) \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

where,

y_i is the actual target value,

\hat{y}_i is the predicted target value, and

n is the total number of data points.

The RMSE of a model determines the absolute fit of the model to the data.

In other words, it indicates how close the actual data points are to the model's predicted values.

A low value of RMSE indicates a better fit and is a good measure for determining the accuracy of the model's predictions.

```
In [29]: # Predicting the target values of the test set
y_pred = model.predict(x_test)
# RMSE (Root Mean Square Error)
rmse = np.sqrt(mean_squared_error(y_test, y_pred))
print("\nRMSE: ", rmse)
```

RMSE: 133.35066554014645

Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique.

It can be used for both Classification and Regression problems in ML.

It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

Random Forest Classifier

As the name suggests,