Calculate IG of Humiditv

- **Step1: Entropy of entire dataset**

$$S\{+9,-5\} = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94$$

- **Step2: Entropy of all attributes:**
- **Entropy of High** $\{+3,-4\} = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = 0.98$
- **Entropy of Normal** $\{+6,-1\} = -\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7} = 0.59$

- **Information Gain= Entropy(whole data)** $-\frac{7}{14}$**Ent(H)** $-\frac{7}{14}$**Ent(N)**

$= 0.15$

Calculate IG of Wind

Step1: Entropy of entire dataset

$$S\{+9,-5\} = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94$$

Step2: Entropy of all attributes:

$$\text{Entropy of Strong } \{+3,-3\} = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1.0$$

$$\text{Entropy of Normal } \{+6,-2\} = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8} = 0.81$$

$$\text{Information Gain} = \text{Entropy(whole data)} - \frac{6}{14}\text{Ent(S)} - \frac{8}{14}\text{Ent(W)}$$

$$= 0.0478$$

- Gain (S, Weather) = 0.246
- Gain (S, Temp) = 0.029
- Gain (S, Humidity) = 0.15
- Gain (S, Wind) = 0.0478

data set for sunny

Day	Weather	Temperature	Humidity	Wind	Play Football?
Day 1	Sunny	Hot	High	Weak	No
Day 2	Sunny	Hot	High	Strong	No
Day 8	Sunny	Mild	High	Weak	No
Day 9	Sunny	Cool	Normal	Weak	Yes
Day 11	Sunny	Mild	Normal	Strong	Yes

Calculate IG of Temperature

Step1: Entropy of Sunny {+2,-3} $= -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.97$

- **Step2: Entropy of all attributes:**
- **Entropy of Hot {+0,-2}** $= -\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2} = 0$
- **Entropy of Mild {+1,-1}** $= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1.0$
- **Entropy of Cool {+1,-0}** $= -\frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1} = 0$
- **Information Gain= Entropy(Sunny)** $-\frac{2}{5}$**Ent(H)** $-\frac{2}{5}$**Ent(M)** $-\frac{1}{5}$**Ent(C)**

$= 0.57$

Calculate IG of Humidity.

Step 1:- Entropy of sunny {+2, -3} =

$$-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.97$$

Step 2: Entropy of all attributes.

Entropy of High {+0, -3} =

$$-\frac{0}{3}\log_2\frac{0}{3} - \frac{3}{3}\log_2\frac{3}{3} = 0$$

Entropy of Normal {+2, -0}

$$-\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2} = 0$$

Information gain = Entropy (Sunny

$$-\frac{3}{5}\text{Ent}(H) - \frac{2}{5}\text{Ent}(\text{N})$$

$$= 0.97.$$

Calculate IG of wind.

step 1:

Entropy of sunny $\{+2, -3\}$ =

$$-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.97$$

step 2: Entropy of all attributes.

- Entropy of strong $\{+1, -1\} = \frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$

Entropy of weak $\{+1, -2\}$ =

$$-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.918.$$

Information gain = Entropy (sunny) $-\frac{3}{5}$ Ent(S) $-\frac{3}{5}$ Ent(W).

$= 0.019.$

- Gain (S_{sunny}, Temp) = 0.57
- Gain (S_{sunny}, Humidity) = 0.97
- Gain (S_{sunny}, Wind) = 0.019

Data set for rain

Day	Weather	Temperature	Humidity	Wind	Play Football?
Day 4	Rain	Mild	High	Weak	Yes
Day 5	Rain	Cool	Normal	Weak	Yes
Day 6	Rain	Cool	Normal	Strong	No
Day 10	Rain	Mild	Normal	Weak	Yes
Day 14	Rain	Mild	High	Strong	No

Calculate IG of Tempreture

Step1: Entropy of Rain {+3,-2} $= -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.97$

Step2: Entropy of all attributes:

Entropy of Hot {+0,-0} $= -\frac{0}{2}\log_2\frac{0}{2} - \frac{0}{2}\log_2\frac{0}{2} = 0$

Entropy of Mild {+2,-1} $= -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = 0.918$

Entropy of Cool {+1,-1} $= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1.0$

Information Gain= Entropy(Rain) $-\frac{0}{5}$**Ent(H)** $-\frac{3}{5}$**Ent(M)** $-\frac{2}{5}$**Ent(C)**

$= 0.019$

Calculate IG of Humidity.

Step 1 :- Entropy of Rain $\{+3, 2\} = \frac{-3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$

Step 2 :- Entropy of all attributes :- $= 0.97$

Entropy of High $\{+1, -1\} = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$

" " Normal $\{+2, -1\} = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = 0.918$

Information Gain = Entropy (Rain) $- \frac{2}{5}$ Ent(S) $- \frac{3}{5}$ En

$= 0.019$

Calculate IG of wind

Step 1: Entropy of Rain

Step 2: Entropy of all attributes:-

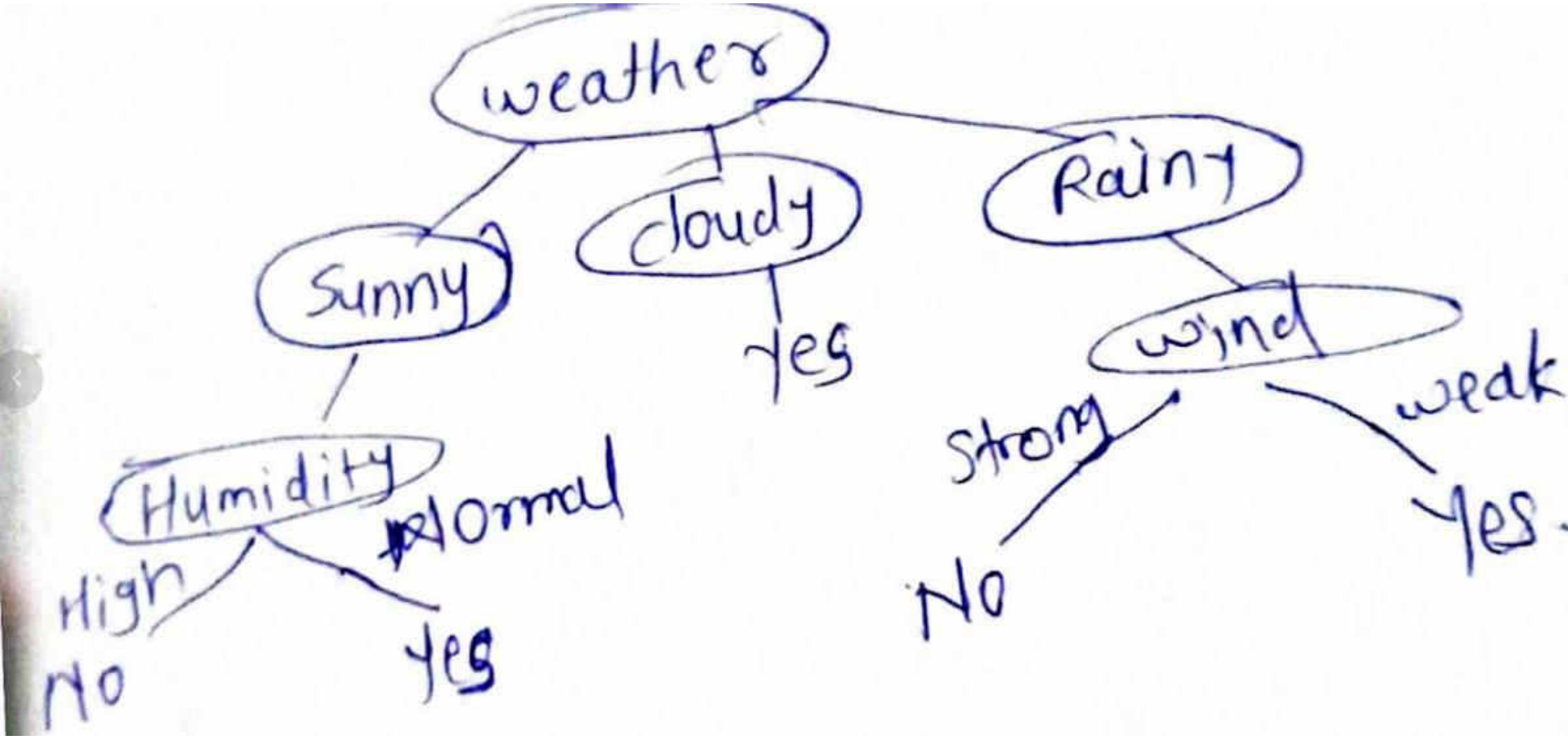
Entropy of strong $\{+0, -2\} = -\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2} = 0$

weak $\{+3, -0\} = \frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3} = 0$

Information gain = Entropy(Rain) $- \frac{2}{5}$ EH(S) $+ \frac{3}{5}$ EH(W

$= 0.97$

- Gain (S_{Rain}, Temp) = 0.019
- Gain (S_{Rain}, Humidity) = 0.019
- Gain (S_{Rain}, Wind) = 0.97


weather
Sunny
Cloudy
Rainy
Humidity
Yes
Wind
High
Normal
Strong
Weak
No
Yes
No
Yes

Decision tree Algorithms

2. CART(Classification And Regression Tree)

It is a variation of the decision tree algorithm. It can handle both classification and

regression tasks.

The CART algorithm works via the following process:

- The best split point of each input is obtained.
- Based on the best split points of each input in Step 1, the new “best” split point

is identified.

- Split the chosen input according to the “best” split point.
- Continue splitting until a stopping rule is satisfied or no further desirable

splitting is available

CART algorithm uses Gini Impurity to split the dataset into a decision tree .

Gini index/Gini impurity

- In Decision Tree, the major challenge is to identify the attribute of the root node in each level. This process is known as attribute selection.
- The Gini impurity measure is one of the methods used in decision tree algorithms to decide the optimal split from a root node and subsequent splits.
- Gini index calculates the amount of probability of a specific feature that is classified incorrectly when selected randomly.
- If all the elements are linked with a single class then it is called pure.

The degree of the Gini index varies from 0 to 1,

- Where 0 depicts that all the elements are allied to a certain class, or only one class exists there.
- The Gini index of value 1 signifies that all the elements are randomly distributed across various classes, and
- A value of 0.5 denotes the elements are uniformly distributed into some classes.

Mathematically, we can write Gini Impurity as follows:

$$Gini = 1 - \sum_{i=1}^{n} (pi)^2$$

where pi is the probability of an object being classified to a particular class.

Random Forest

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems.

- It can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.
- Before understanding the working of the random forest we must look into the ensemble technique. Ensemble simply means combining multiple models. Thus a collection of models is used to make predictions rather than an individual model.

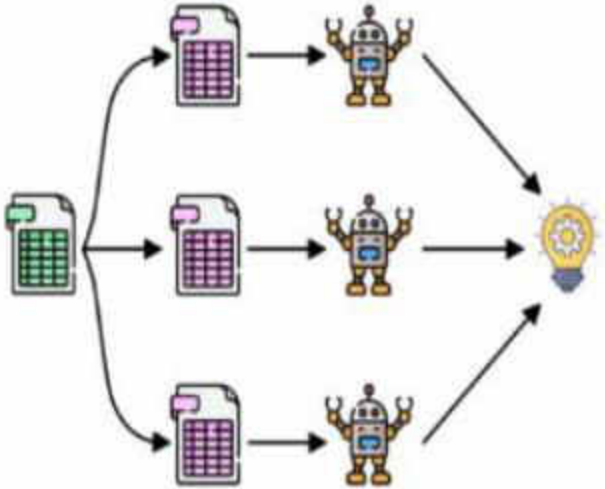
Ensemble uses two types of methods:

1. Bagging– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest.

2. Boosting– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST

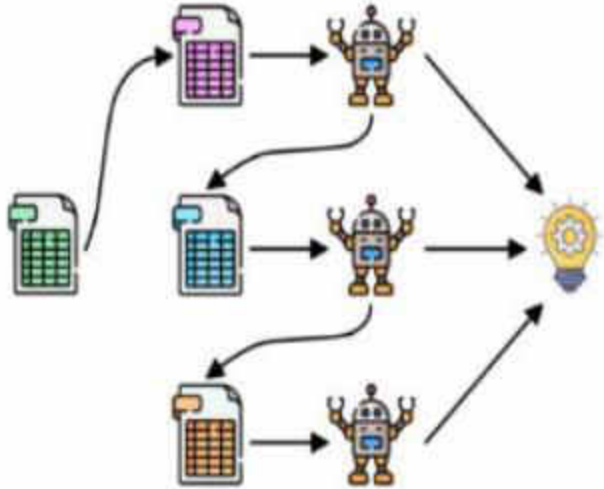
NOTE: Random forest works on the Bagging principle.

Bagging



Parallel

Boosting



Sequential

Bagging

- Bagging, also known as **Bootstrap Aggregation** is the ensemble technique used by random forest.
- Bagging chooses a random sample from the data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as row sampling. This step of row sampling with replacement is called bootstrap.
- Now each model is trained independently which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting is known as aggregation.
- E.g. In figure, the bootstrap sample is taken from actual data (Bootstrap sample 01, Bootstrap sample 02, and Bootstrap sample 03) with a replacement which means there is a high possibility that each sample won't contain unique data. Now the model (Model 01, Model 02, and Model 03) obtained from this bootstrap sample is trained independently. Each model generates results as shown. Now Happy emoji is having a majority when compared to sad emoji. Thus based on majority voting final output is obtained as Happy emoji.

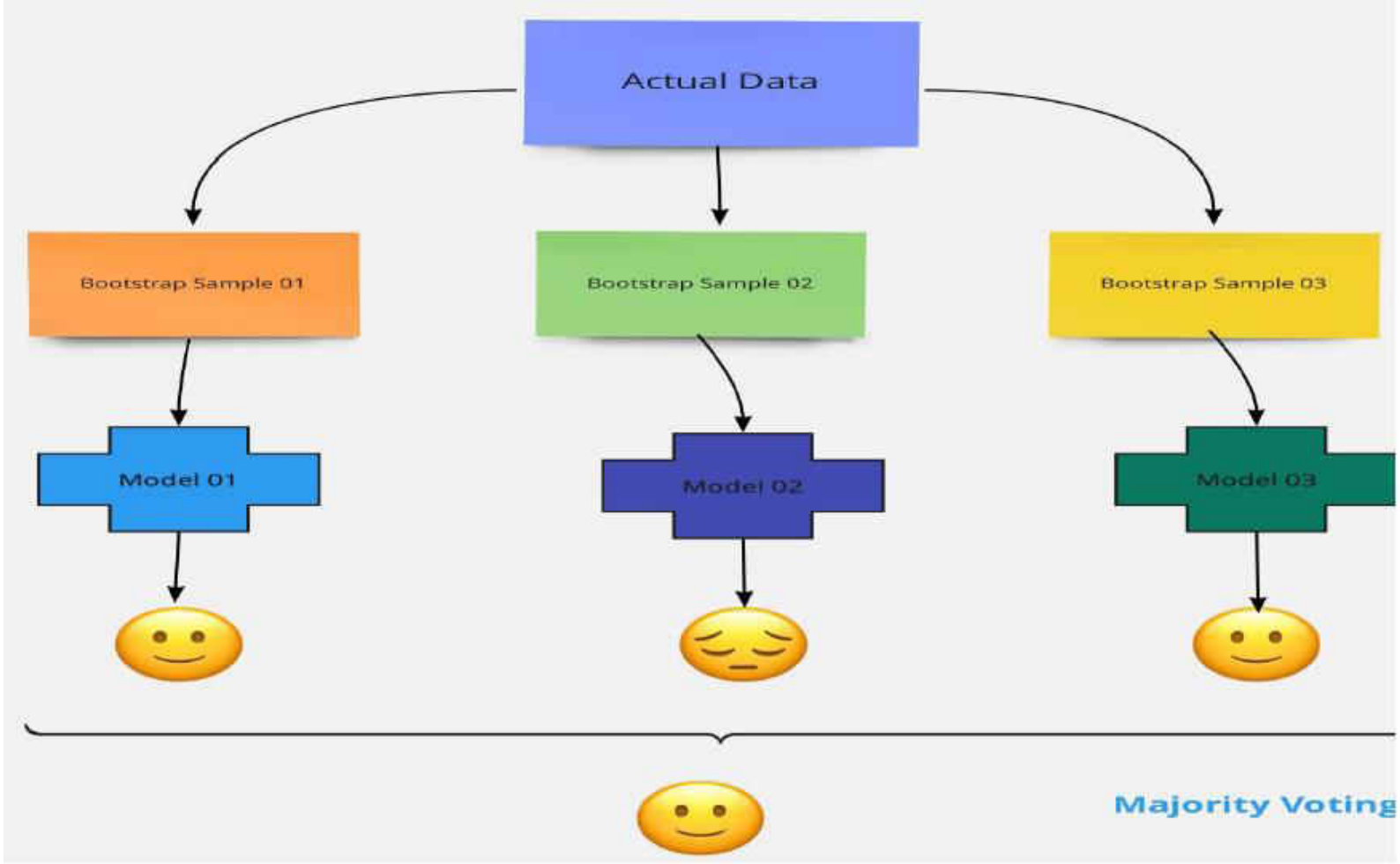


Fig.Bagging Ensemble Method

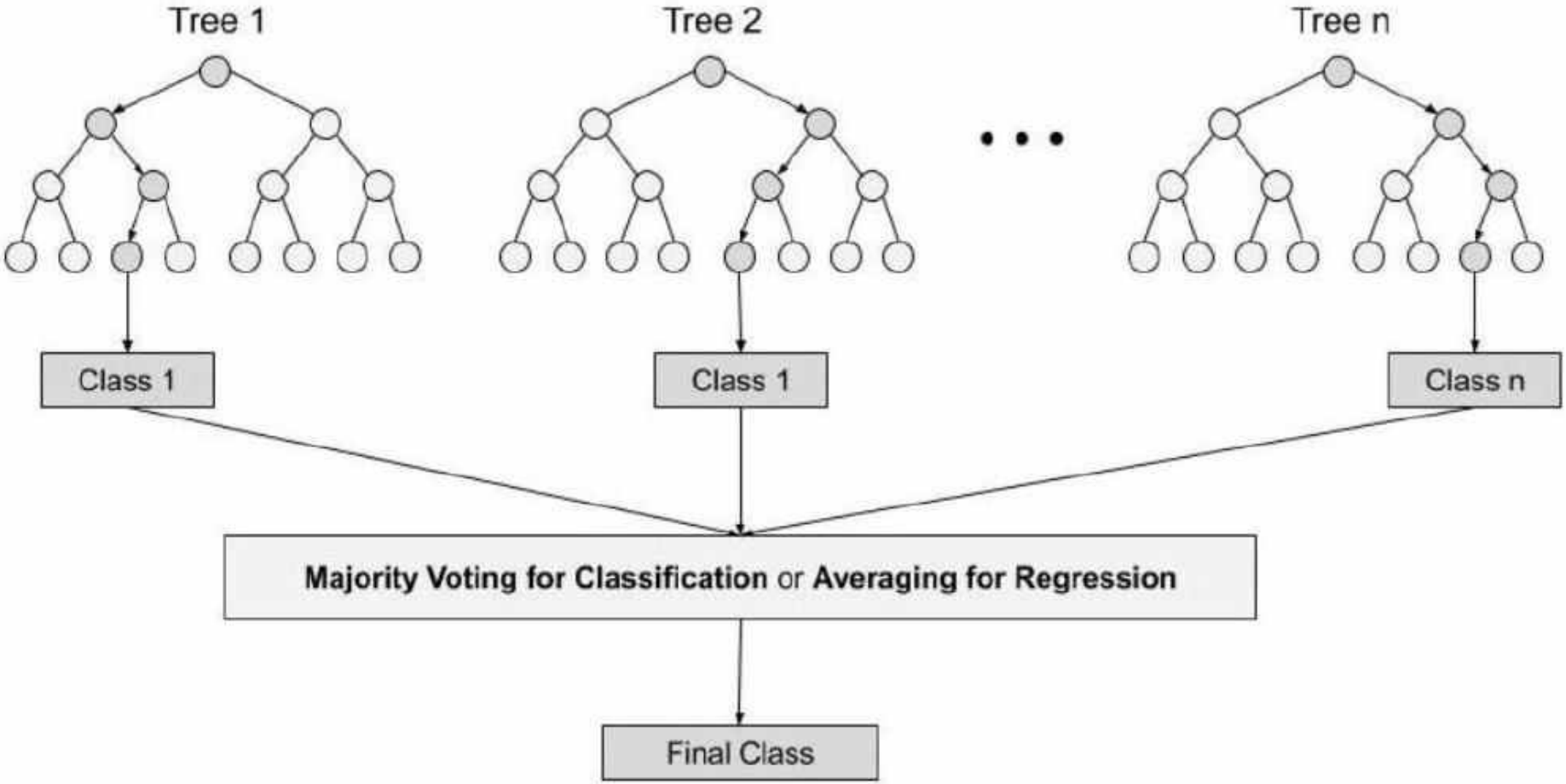
Steps involved in random forest algorithm:

Step 1: In Random forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.


Tree 1
Tree 2
Tree n
Class 1
Class 1
Class n
Majority Voting for Classification or Averaging for Regression
Final Class

For example: consider the fruit basket as the data as shown in the figure below. Now n number of samples are taken from the fruit basket and an individual decision tree is constructed for each sample. Each decision tree will generate an output as shown in the figure. The final output is considered based on majority voting. In the below figure you can see that the majority decision tree gives output as an apple when compared to a banana, so the final output is taken as an apple.