SNS COLLEGE OF ENGINEERING

Kurumbapalayam (Po), Coimbatore – 641 107

An Autonomous Institution

Accredited by NAAC – UGC with 'A' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF COMPUTER SCIENCE AND TECHNOLOGY

COURSE NAME : 19CS407-DATA ANALYTICS WITH R

II YEAR /IV SEMESTER

Unit II – Statistics and Prescriptive Analytics
Topic : Random Forest

Brain Storming

What is classification and clustering?

Forms of Learning

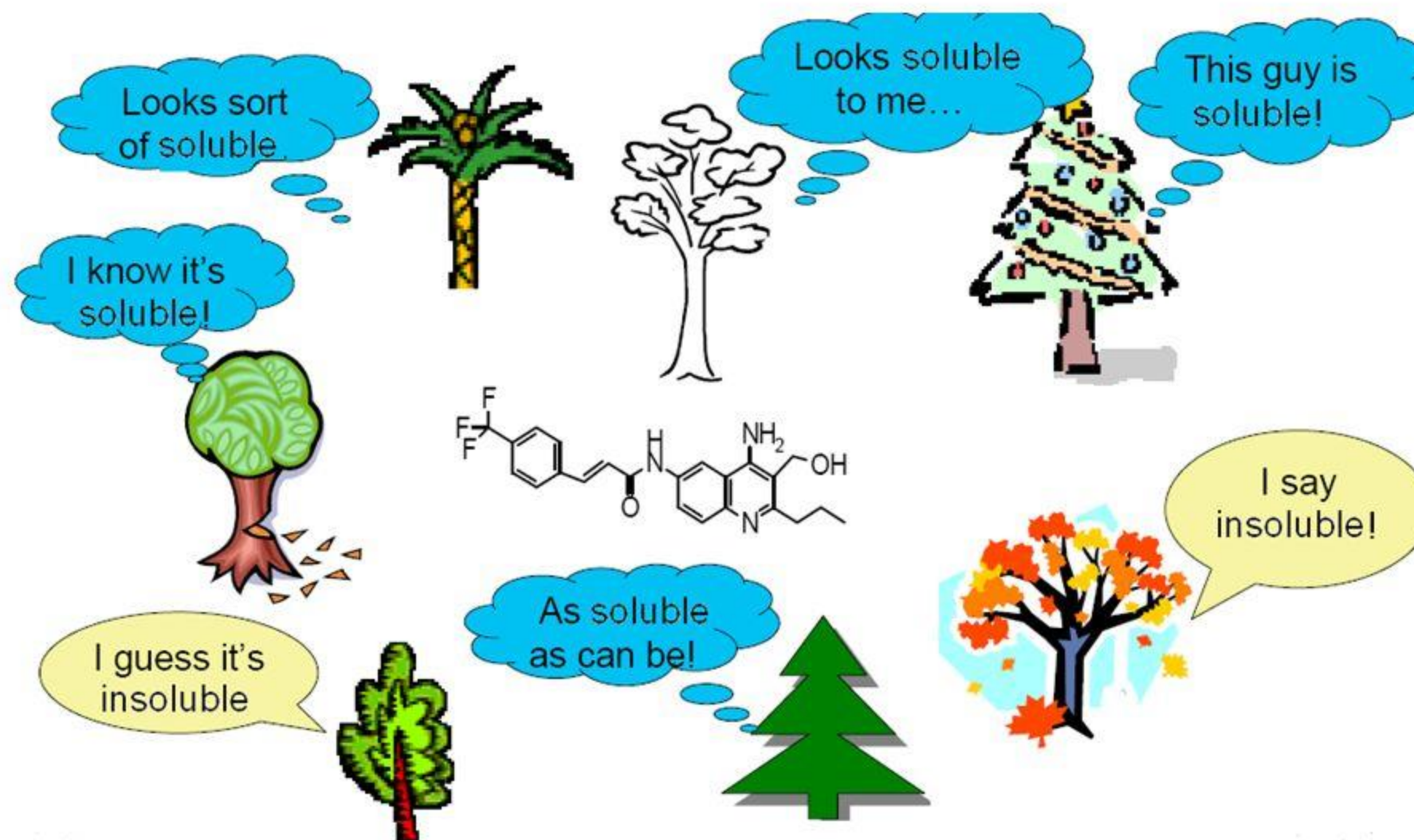At a broad level, machine learning can be classified into three types:

- Supervised learning

- Unsupervised learning

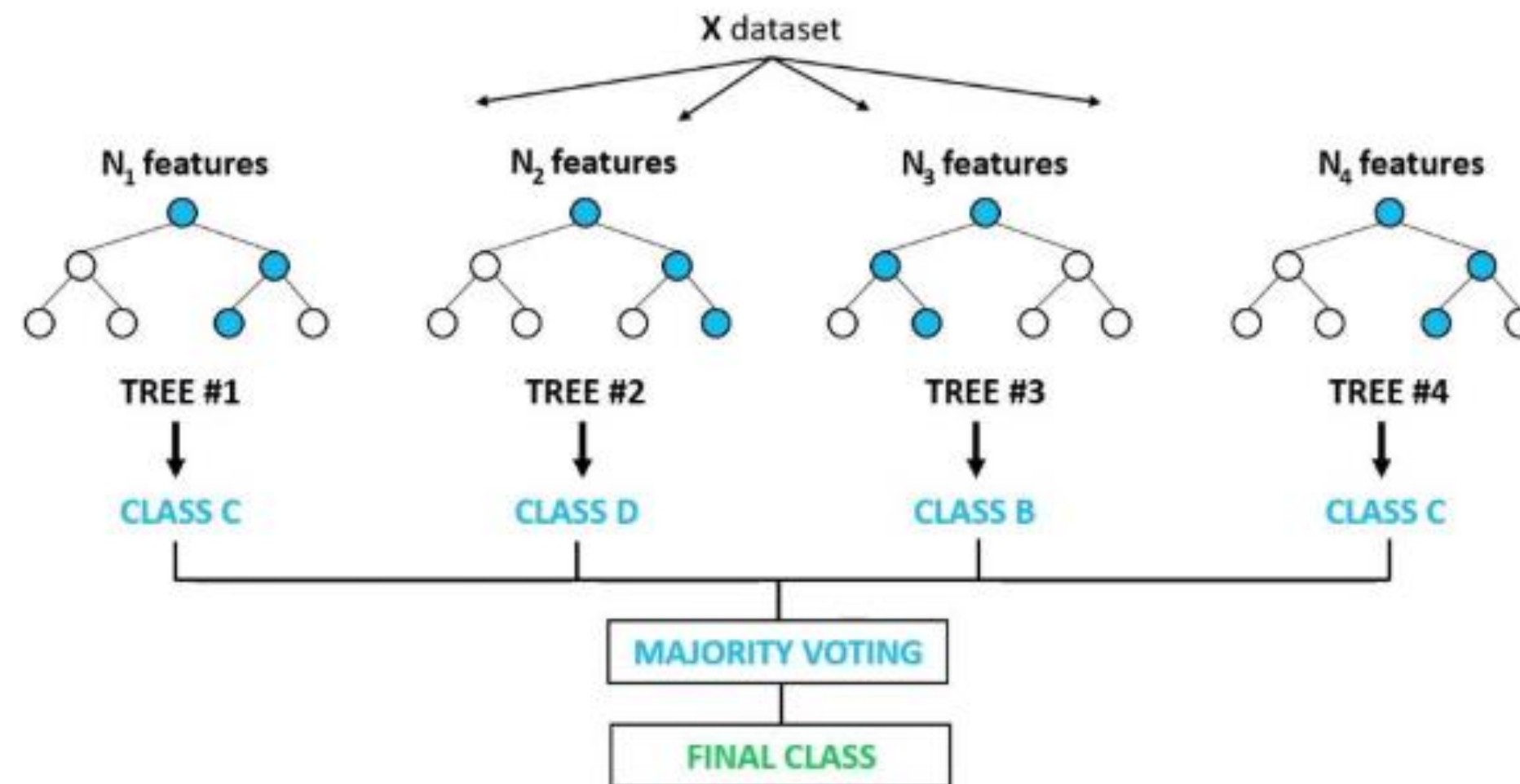- Reinforcement learning

Random Forest Algorithm

19CS407-Data Analytics/M.Kanchana/CST/SNSCE

Random Forest Algorithm


Random Forest Classifier

Random Forest Algorithm

- "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."

- The random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

- The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Random Forest Algorithm

The below diagram explains the working of the Random Forest algorithm:
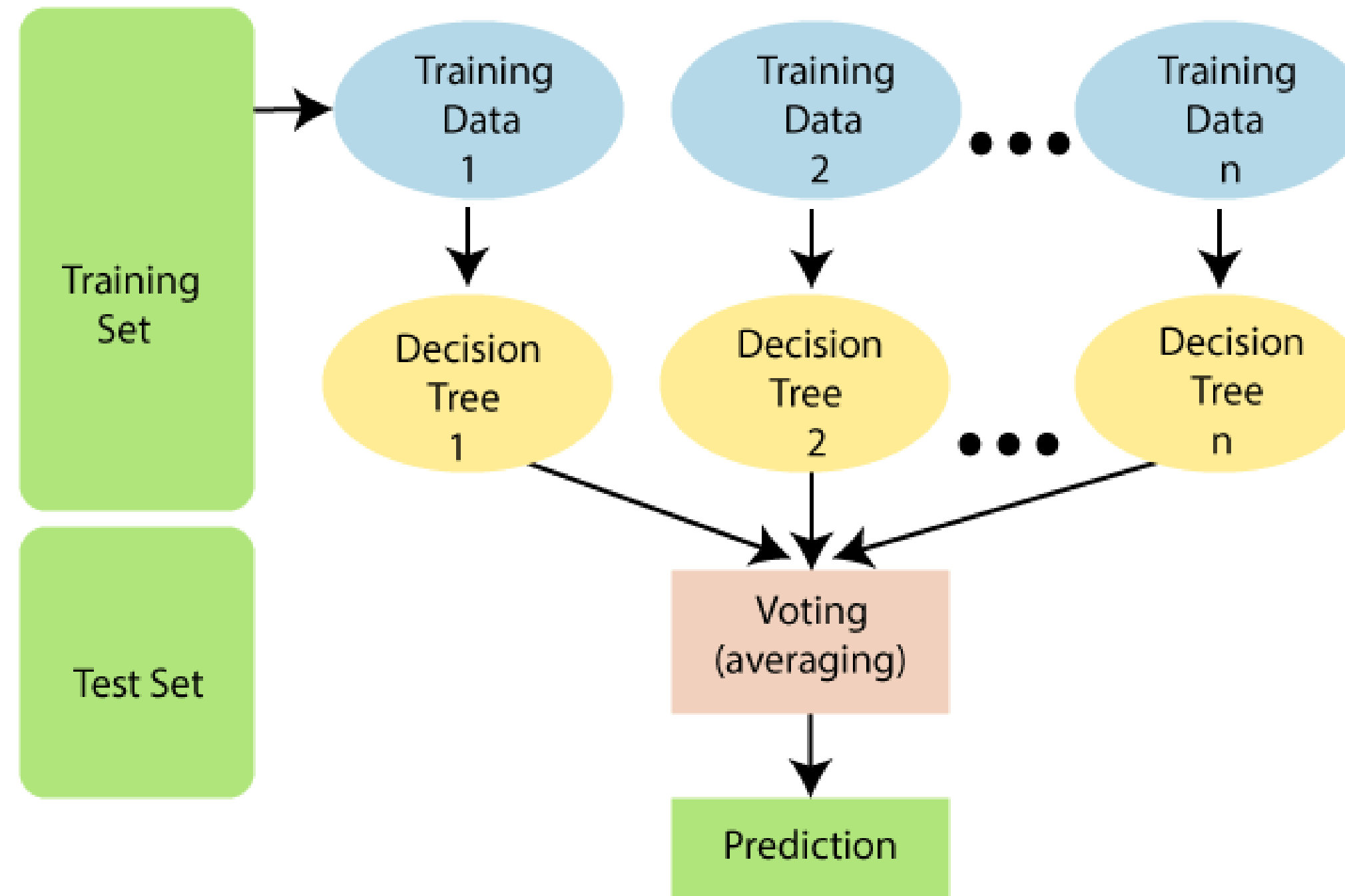
Random Forest Algorithm

How does Random Forest algorithm work?

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Random Forest Algorithm

How does Random Forest algorithm work?

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.
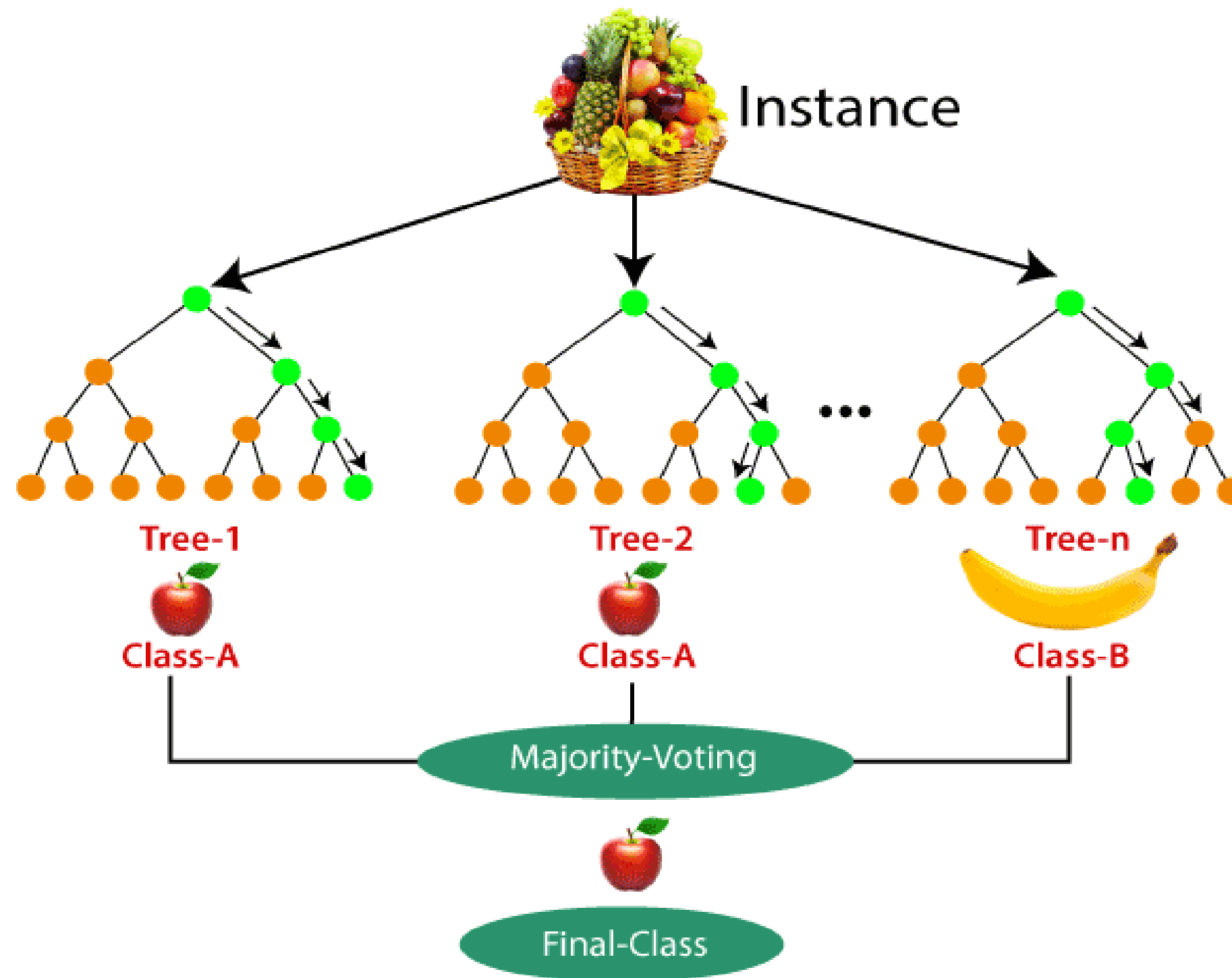
Random Forest Algorithm

Example:

Example: Suppose there is a dataset that contains multiple fruit images. So, this dataset is given to the Random forest classifier. The dataset is divided into subsets and given to each decision tree.

During the training phase, each decision tree produces a prediction result, and when a new data point occurs, then based on the majority of results, the Random Forest classifier predicts the final decision. Consider the below image:

Random Forest Algorithm

Example:

Random Forest Algorithm

Applications of Random Forest

There are mainly four sectors where Random forest mostly used:

Banking: Banking sector mostly uses this algorithm for the identification of loan risk.

Medicine: With the help of this algorithm, disease trends and risks of the disease can be identified.

Land Use: We can identify the areas of similar land use by this algorithm.

Marketing: Marketing trends can be identified using this algorithm.

Random Forest Algorithm

Advantages of Random Forest

- Random Forest is capable of performing both Classification and Regression tasks.

- It is capable of handling large datasets with high dimensionality.

- It enhances the accuracy of the model and prevents the overfitting issue.

- Disadvantages of Random Forest

- Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks.

- https://blockgeni.com/random-forest-algorithm-in-ml/

Assessment 1

1. What is called univariate statistics?.

Ans : _____

2. What is dispersion univariate statistics?

Ans : _____

References

TEXT BOOKS

1.João Moreira, Andre Carvalho, Tomás Horvath – "A General Introduction to Data Analytics" – Wiley - 2018

2.An Introduction to R, Notes on R: A Programming Environment for Data Analysis and Graphics. W. N. Venables, D.M. Smith and the R Development Core Team. Version 3.0.1 (2013-05-16). URL: https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf

References:

1**. Dean J, —**_Big Data, Data Mining and Machine learning_**,** Wiley publications, 2014.

2. Provost F and Fawcett T, —_Data Science for Business_, O'Reilly Media Inc, 2013.

3. Janert PK, —_Data Analysis with Open Source Tools_, O'Reilly Media Inc, 2011.

4. Weiss SM, Indurkhya N and Zhang T, —_Fundamentals of Predictive Text Mining_, Springer-Verlag London Limited, 2010.

5.Marz N and Warren J,- Big Data, Manning Publications,2015

Thank You