greatlearning
Learning for Life

DATA MINING (Week3)

greatlearning
Learning for Life

DSBA CURRICULUM DESIGN

FOUNDATIONS

Data Science Using Python

Statistical Methods for Decision Making

CORE COURSES

Advanced Statistics

Data Mining (Week-3/5)

Predictive Modelling

Machine Learning

Time Series Forecasting

Data Visualization

DOMAIN APPLICATIONS

Financial Risk Analytics

Web & Social Media Analytics

Marketing Retail Analytics

greatlearning
Learning for Life

LEARNING OBJECTIVE OF THIS MODULE

- Clustering
- CART & Model Performance Measures
- Random Forest
- Neural Network

greatlearning
Learning for Life

LEARNING OBJECTIVES OF THIS SESSION -

- Ensemble Modeling
- Bagging
- Random Forest Algorithm
- Finding Optimal Number of Trees

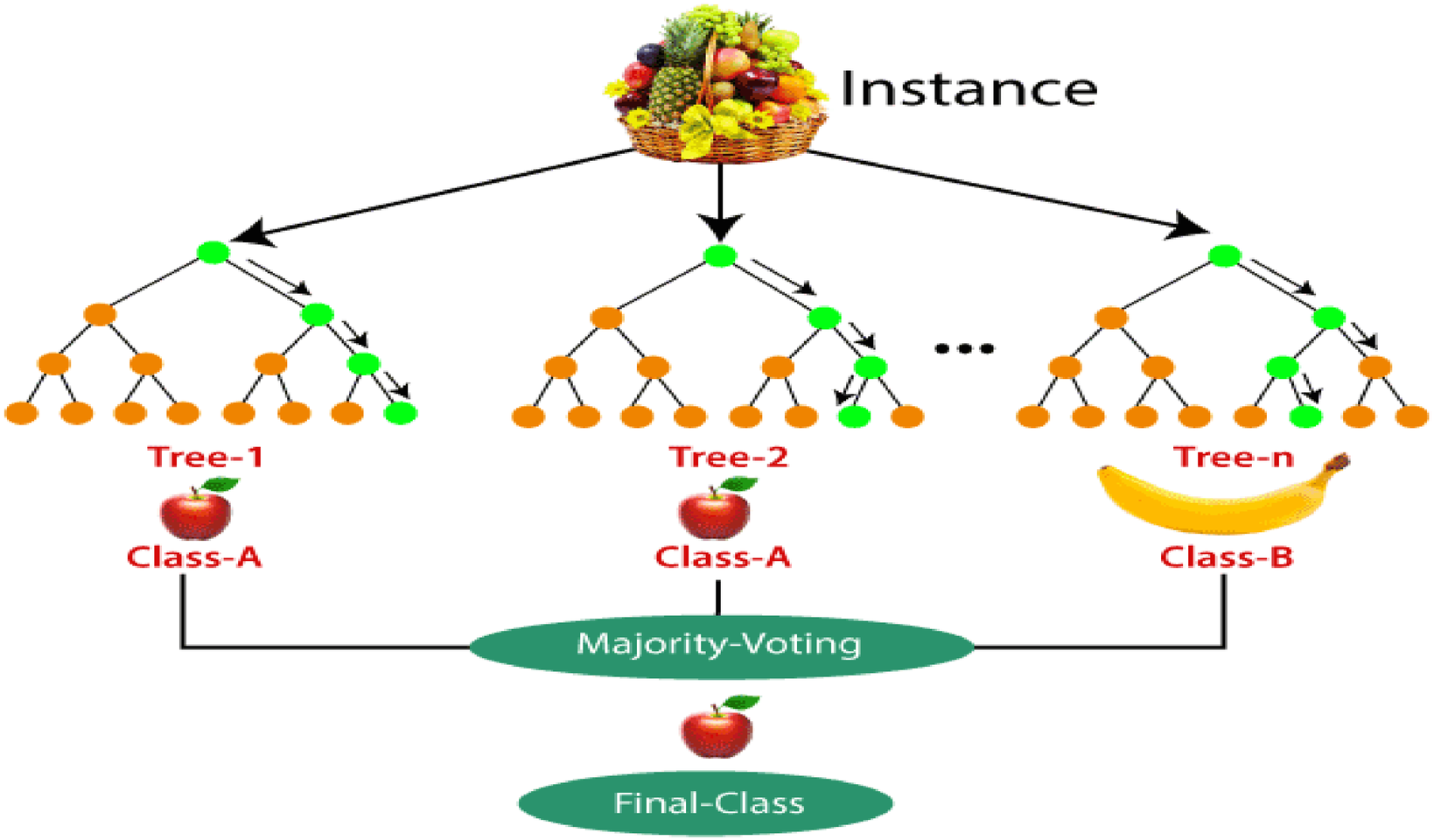
greatlearning
Learning for Life

TRY ANSWERING THE FOLLOWING

- Do we perform Pruning on Random Forest?
- Name a few Ensemble Techniques.
- In Random Forest, which class becomes the model prediction? Is it class with maximum votes or class with lowest votes?

greatlearning
Learning for Life

BROAD OVERVIEW- Random Forest



greatlearning
Learning for Life

Industry Application - Random Forest for urban planning

Over the years the importance as well as the complexity of urban planning have grown exponentially. One of the main problems affecting urban planning is the appropriate choice of location to host a particular activity (either commercial activity or common welfare service).

In terms of feature extraction, a variety of factors can be explored such as: Attributes of the actual building or location, distance from landmarks, distance from specific similar spots, density of occurrence per specific area, distance from means of public transport along with their plurality in a certain area, distance and density of occurrence with respect to points of touristic interest, economic, and monetary points of interest. Given these inputs, the dependent variable we are trying to predict is Parking (1) and No parking (0).

To solve this particular problem a number of algorithms were tried but the best accuracy of 94% was achieved using Random Forest.

Reference: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6567884/

greatlearning
Learning for Life

CASE STUDY - US Heart Patient

Five million Americans are currently living with heart diseases, and the numbers are expected to rise. It is very important to understand the factors which causes Heart-attacks so that certain precaution can be taken by individuals. In-order to understand the reasons of the Heart-attack, a data was collected from various hospitals across US which is given in US_Heart_Patients.csv. In the data set there are Heart-Att indicates whether the person suffered from Heart attack or not.

Perform EDA on the data and build a model which will predict whether the person will suffer from Heart-attack or not.

Data Science @ Work

greatlearning
Power Ahead

Apply **Data Science at your workplace** to gain some instant benefits:

- Get noticed by your management with your outstanding analysis backed by data science.
- Create an impact in your organization by taking up small projects/initiatives to solve critical issues using data science.
- Network with members from the data science vertical of your organization and seek opportunities to contribute in small projects.
- Share your success stories with us and the world to position yourself as a subject matter expert in data science.

greatlearning
Learning for Life

ANY QUESTIONS

greatlearning
Learning for Life

HAPPY LEARNING