CLUSTERING

3rd Sem, MCA

CONTENT

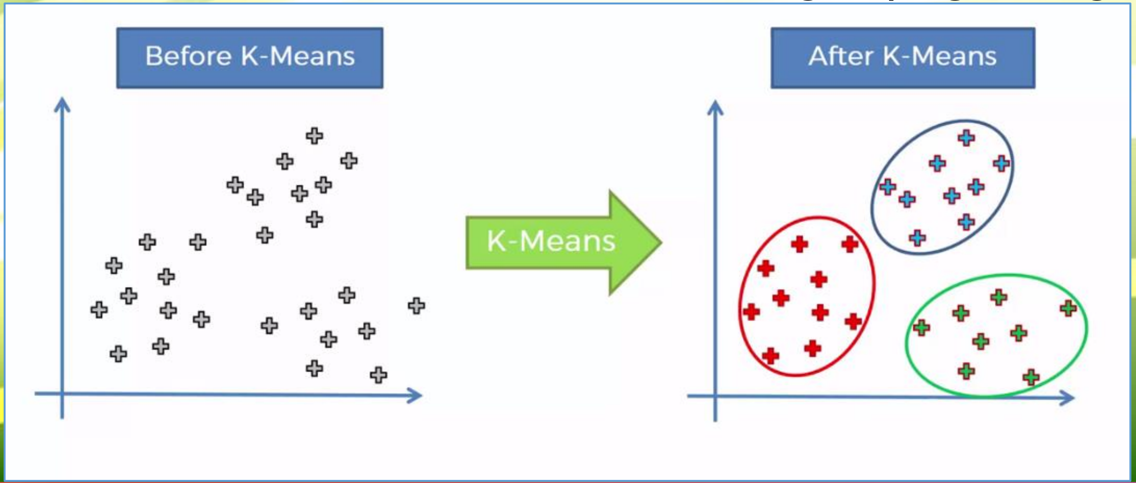
- Clustering
  - Types of Clustering
  - Hierarchical Clustering
  - K-Means Clustering
- Decision Tree
- Random Forest

GROUPING

- Grouping and classification techniques are very important methods in predictive data analysis.
- **Grouping Analysis** methods helps to determine natural groupings in data.
- Useful to decompose data set into simpler subsets → helps to make sense of entire collection of observations.
- For each group summary statistics, variety of graphs may help in better analysis
- Different ways to visualize and group observations,
  - ***Clustering***: based on similarities of overall set of variables of interest.
  - ***Association rule***: identify groups based on interesting combinations of predefined categories
  - ***Decision tree / Random forest***: groups observation based on combination of ranges of continuous variables or of specific categories.

CLUSTERING

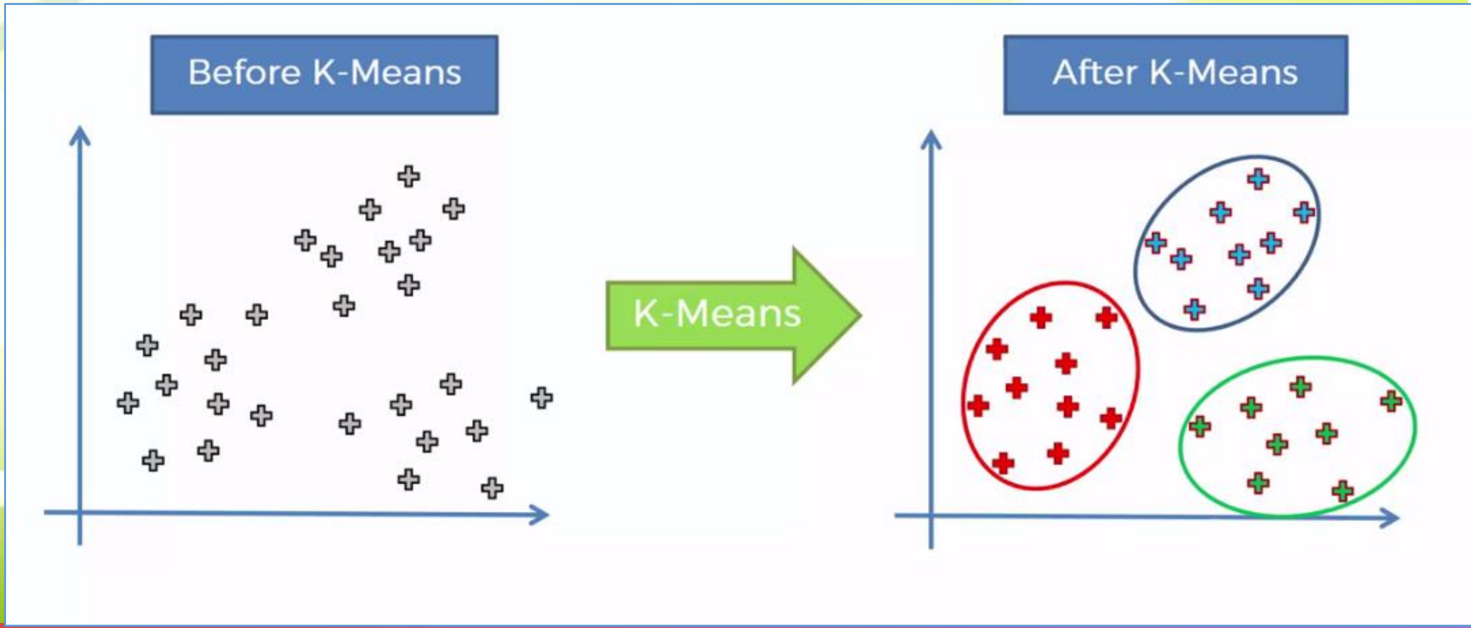
- **Cluster**: group of (similar) objects that belongs to same class.
- **Clustering**: process of making a group of abstract objects into classes of similar objects.
- Given a data set of items, with certain features, and values for these features; the task is to categorize those items into groups.
    - Used to find similarity as well as relationship patterns among data samples and then cluster those samples into groups having similarity based on features.
    - Clustering is important because it determines the intrinsic grouping among the present unlabeled data.



CLUSTERING

Clustering methods –

- Partitioning Method
- Hierarchical Method; Agglomerative Approach, Divisive Approach
- Constraint-based Method
- Density-based Method
- Grid-Based Method
- Distribution Model-Based Method
- Fuzzy Method



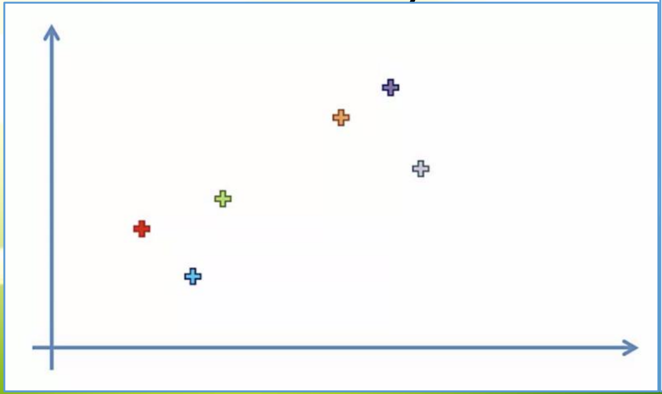
Clustering

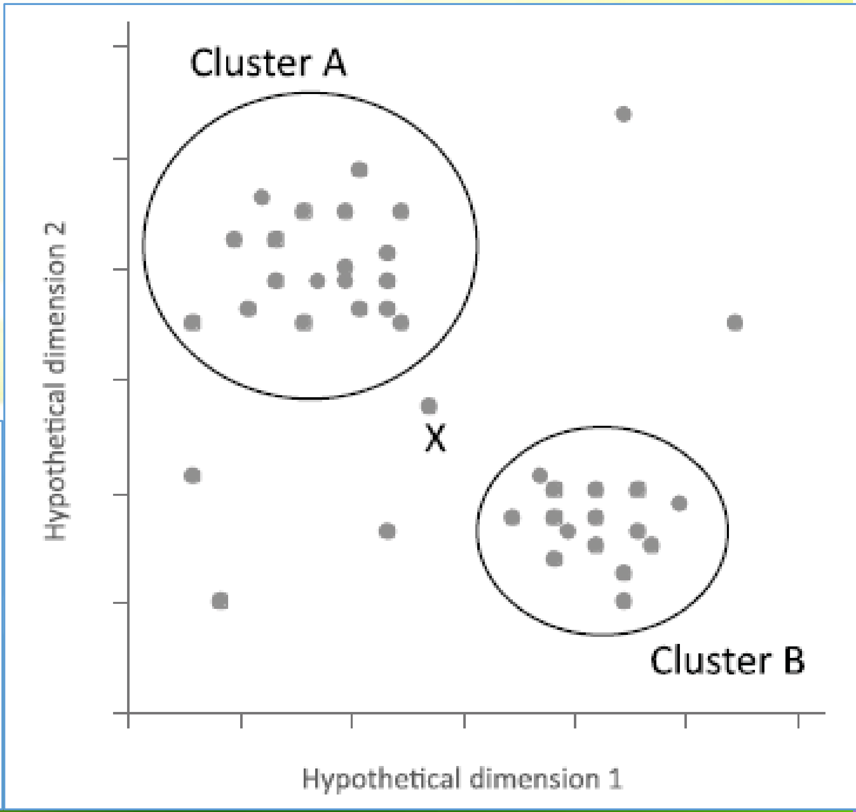
Applications of Cluster Analysis

- Market Segmentation: help marketers discover distinct groups in their customer base → characterize customer groups based on purchasing patterns.
- Anomaly detection, Outlier detection applications; example detection of credit card fraud.
- *Biological data analysis:* used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations.
- Social network analysis
- Image segmentation

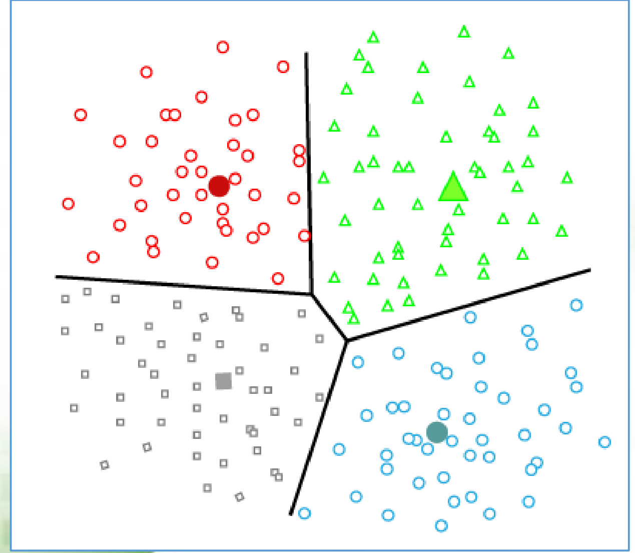
CLUSTERING

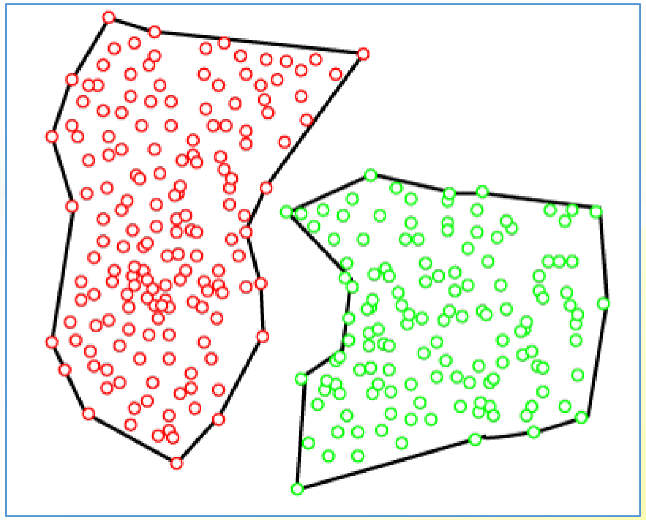
- Clustering is an *unsupervised* method for grouping.
- **Unsupervised**: groups are not known in advance.
- Clustering method chosen to subdivide data into groups applies automated procedure to discover groups based on some criteria.
- Many clustering methods.
- Each method will group data differently based on criteria it uses.
- For clustering, there is no way to measure accuracy (usefulness matters).
- **Distance** between two observations defines how similar they are to be in same cluster or not.





CLUSTERING





Partitioning Clustering

- Also known as centroid-based method.
- Divides the data into non-hierarchical groups.
- Example: K-Means Clustering algorithm.
- The dataset is divided into a set of k groups, where K is used to define the number of pre-defined groups.
- The cluster center is created in such a way that the distance between the data points of one cluster is minimum as compared to another cluster centroid.

Density-Based Clustering

- Connects the highly-dense areas into clusters, and the arbitrarily shaped distributions are formed as long as the dense region can be connected.
- Identify different clusters in the dataset and connects the areas of high densities into clusters.
- The dense areas in data space are divided from each other by sparser areas.
- Face difficulty if the dataset has varying densities and high dimensions.

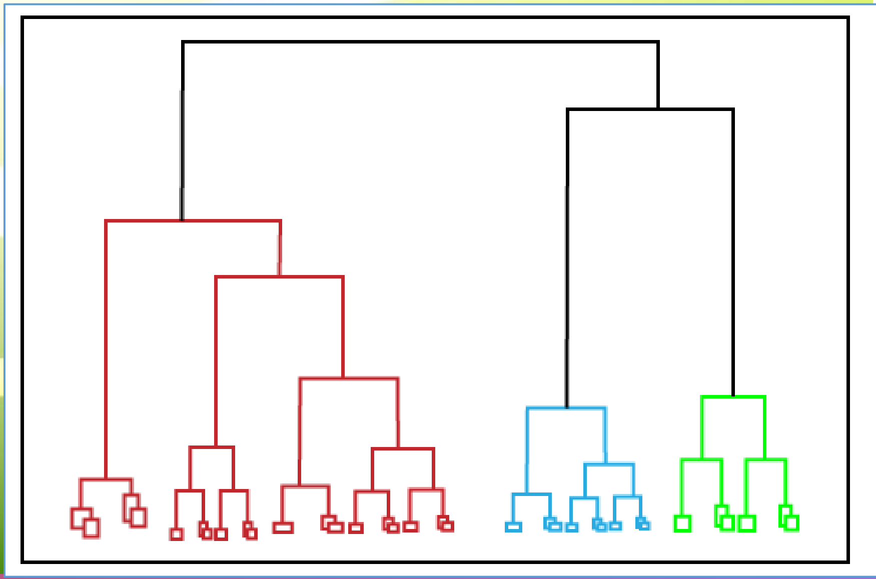
CLUSTERING

Hierarchical Clustering

- Used as an alternative for partitioned clustering, as there is no requirement of pre-specifying the number of clusters (k) to be created.
- Dataset is divided into clusters to create a tree-like structure (dendrogram).
- The observations or any number of clusters can be selected by cutting the tree at the correct level.
- Example: Agglomerative Hierarchical algorithm.

Fuzzy Clustering

- Data object may belong to more than one group or cluster.
- Each dataset has a set of membership coefficients, which depend on the degree of membership to be in a cluster.
- Example: Fuzzy k-means algorithm.



CLUSTERING

Clustering Algorithms

- **K-Means algorithm:** It classifies dataset by dividing samples into different clusters of equal variances. Number of clusters must be specified. It is fast with lower computation time required**.**

- **Agglomerative Hierarchical algorithm:** Performs the bottom-up hierarchical clustering. In this, each data point is treated as a single cluster at the outset and then successively merged. The cluster hierarchy can be represented as a tree-structure.

- **Affinity Propagation:** It is different from other clustering algorithms as it does not require to specify the number of clusters. In this, each data point sends a message between the pair of data points until convergence. It has higher time complexity, which is the main drawback of this algorithm.

CLUSTERING

Clustering Algorithms

- **Mean-shift algorithm:** Tries to find dense areas in smooth density of data points. It is example of centroid-based model, that works on updating candidates for centroid to be center of points within given region.

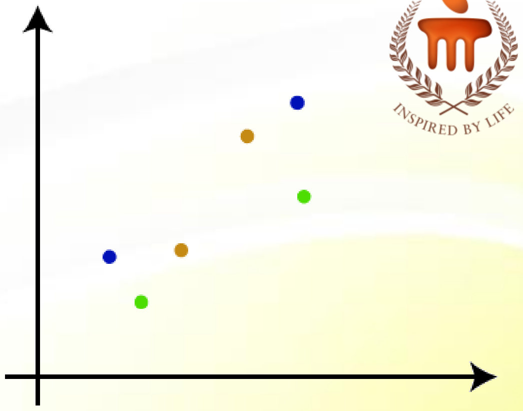
- **DBSCAN Algorithm: Density-Based Spatial Clustering of Applications with Noise**. It is an example of a density-based model similar to the mean-shift, but with some remarkable advantages. In this algorithm, the areas of high density are separated by the areas of low density. Because of this, the clusters can be found in any arbitrary shape.

- **Expectation-Maximization Clustering using GMM:** This algorithm can be used as an alternative for the k-means algorithm or for those cases where K-means can be failed. In GMM, it is assumed that the data points are Gaussian distributed.

HIERARCHICAL CLUSTERING

- **Hierarchical Clustering**: creates hierarchical decomposition of given set of data objects.
- Two approaches;
- **Agglomerative Approach**: (bottom-up approach) "*AGNES*" (*Agglomerative Nesting*)
  - Start with each object forming a separate/singleton group/cluster.
  - Keeps on merging objects or groups that are **close/similar to one another**. *(Euclidian distance)*
  - Keep on doing so until all of groups are merged into one or until termination condition holds.
  - normally limited to data sets with fewer (< 10,000 observations) → computational cost to generate hierarchical tree can be high for larger numbers of observations
  - result is a tree-based representation of the objects, named *dendrogram*.
- **Divisive Approach**: (top-down approach) "DIANA" (*Divise Analysis*)
  - Start with all of objects in same cluster.
  - In continuous iteration, a cluster is split up into smaller clusters.
  - Keep doing until each object in one cluster or termination condition holds.

Hierarchical clustering



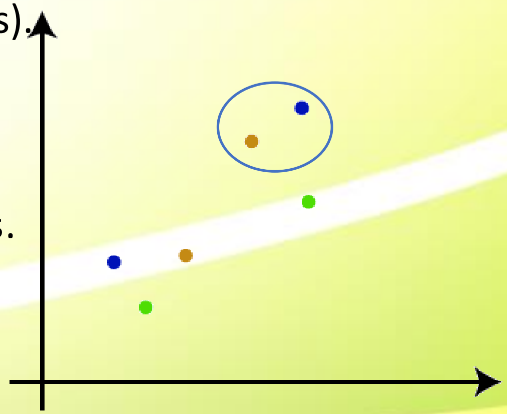
Step-1: Create each data point as a single cluster (for N data points, number of clusters will also be N).

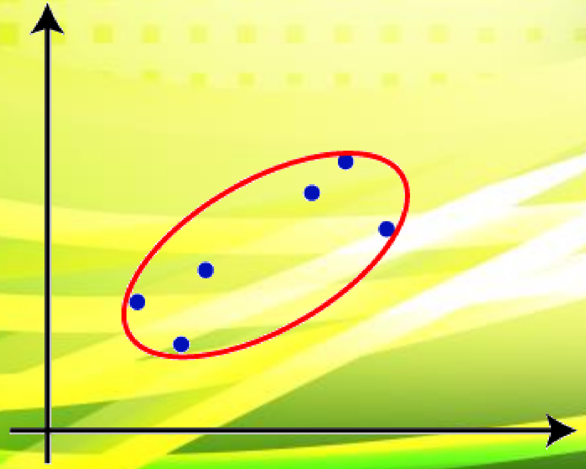
Step-2: Take two closest data points/clusters and merge them to form one cluster (N-1 clusters remains).

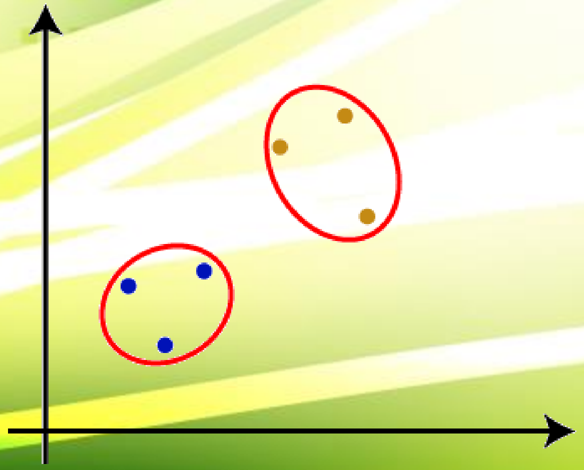
Step-3: Again, take two closest clusters and merge them together to form one cluster (remaining N-2 clusters).

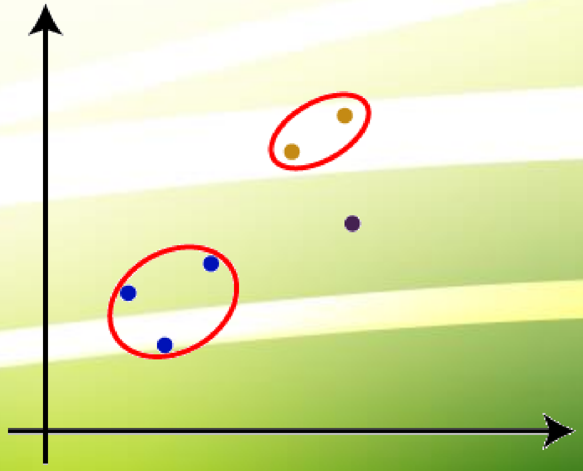
Step-4: **Repeat Step 3** until only one cluster left *(or termination condition meets).*

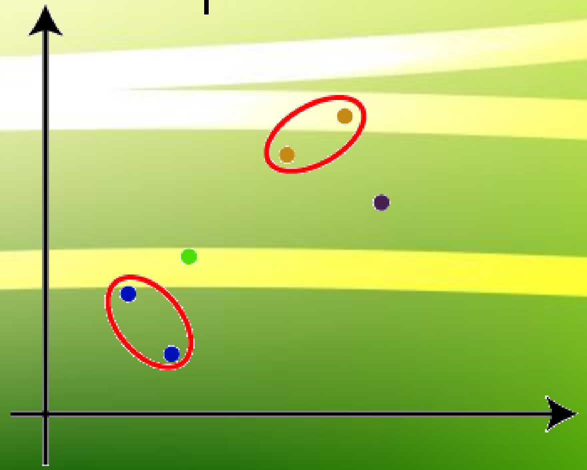
Step-5: Once all clusters are combined into one big cluster, develop the dendrogram to find required clusters.



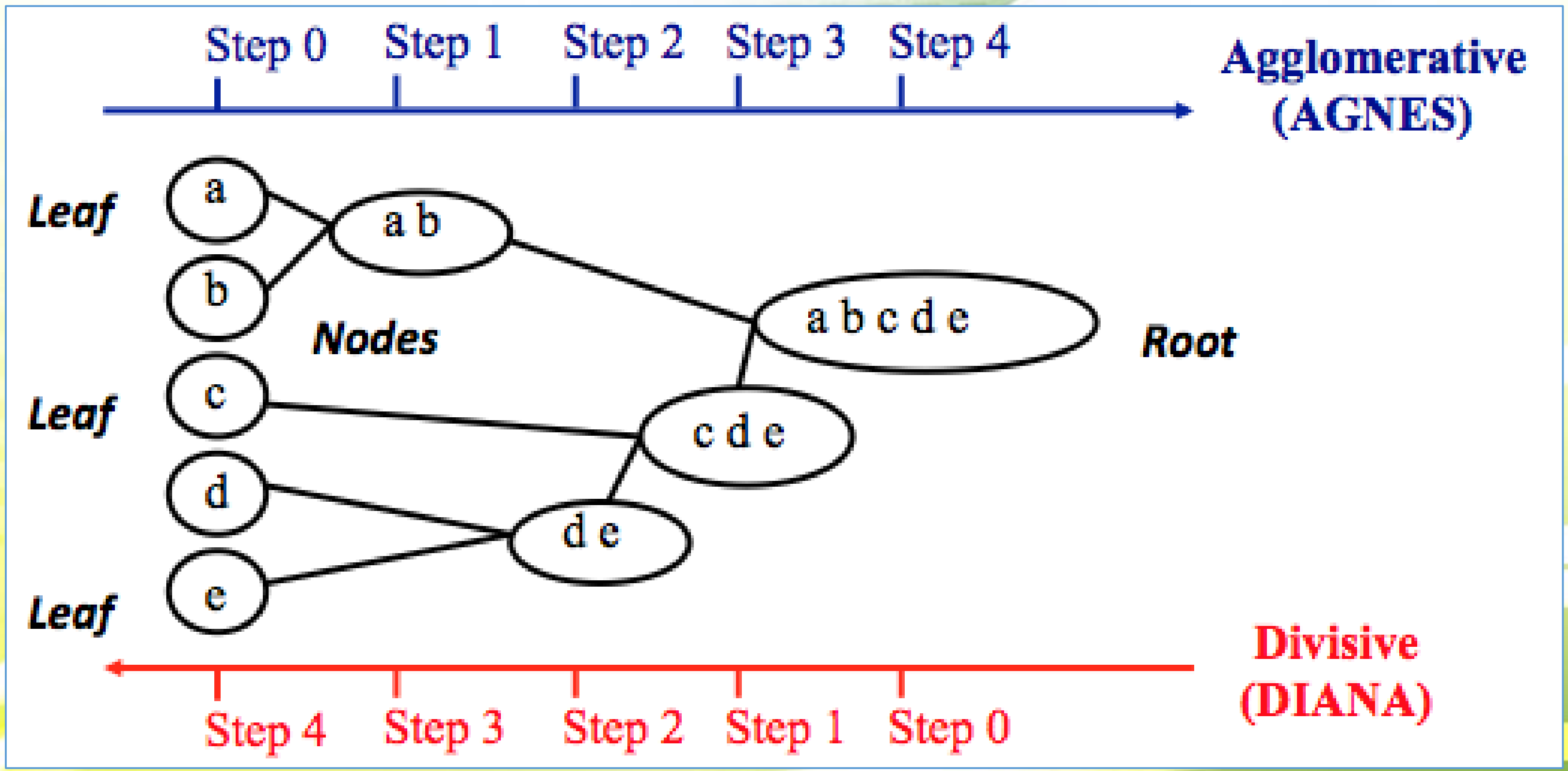




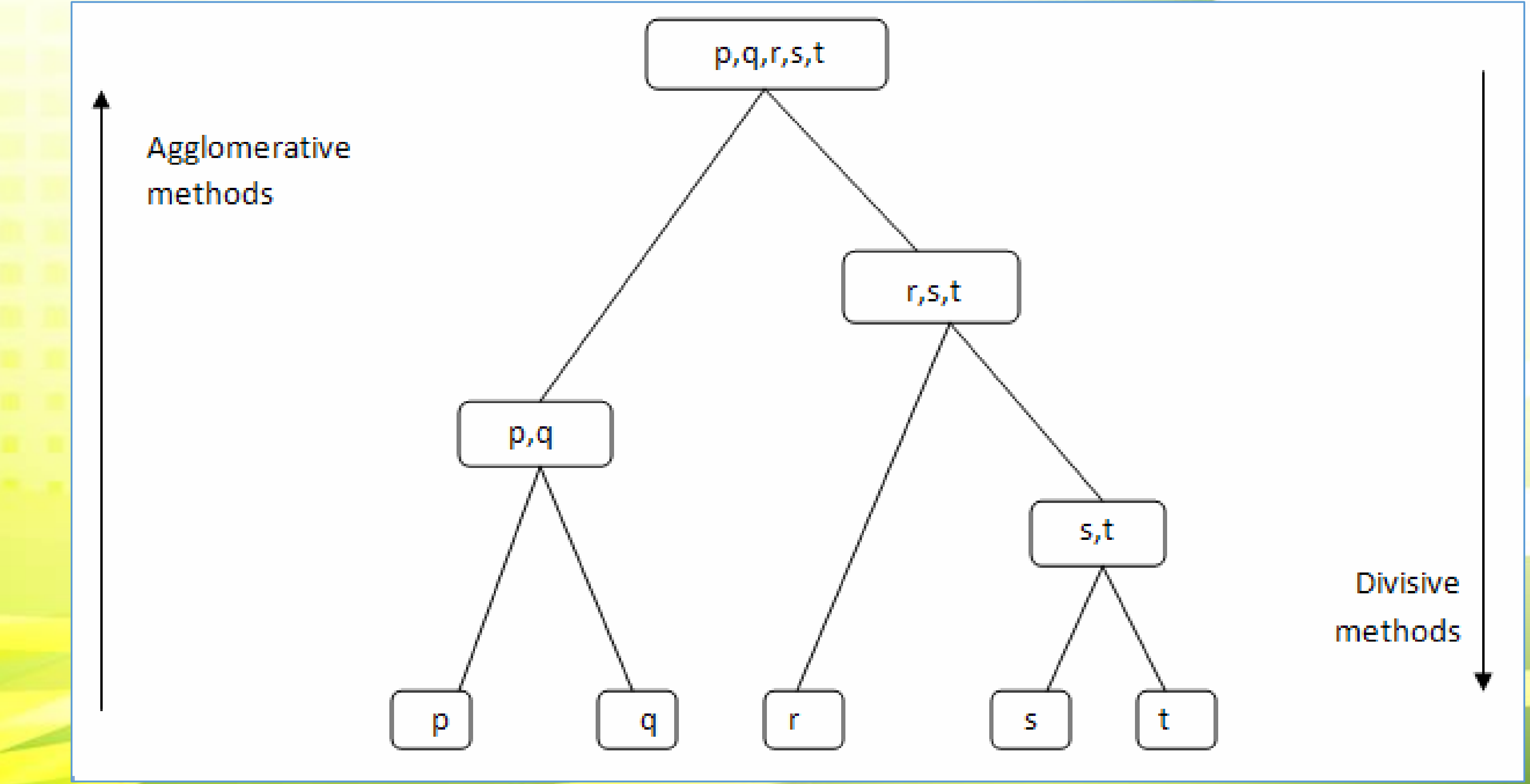


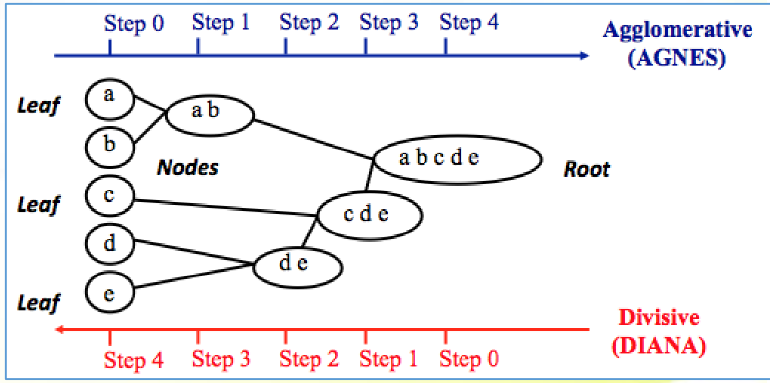


DATA ANALYSIS – HIERARCHICAL CLUSTERING

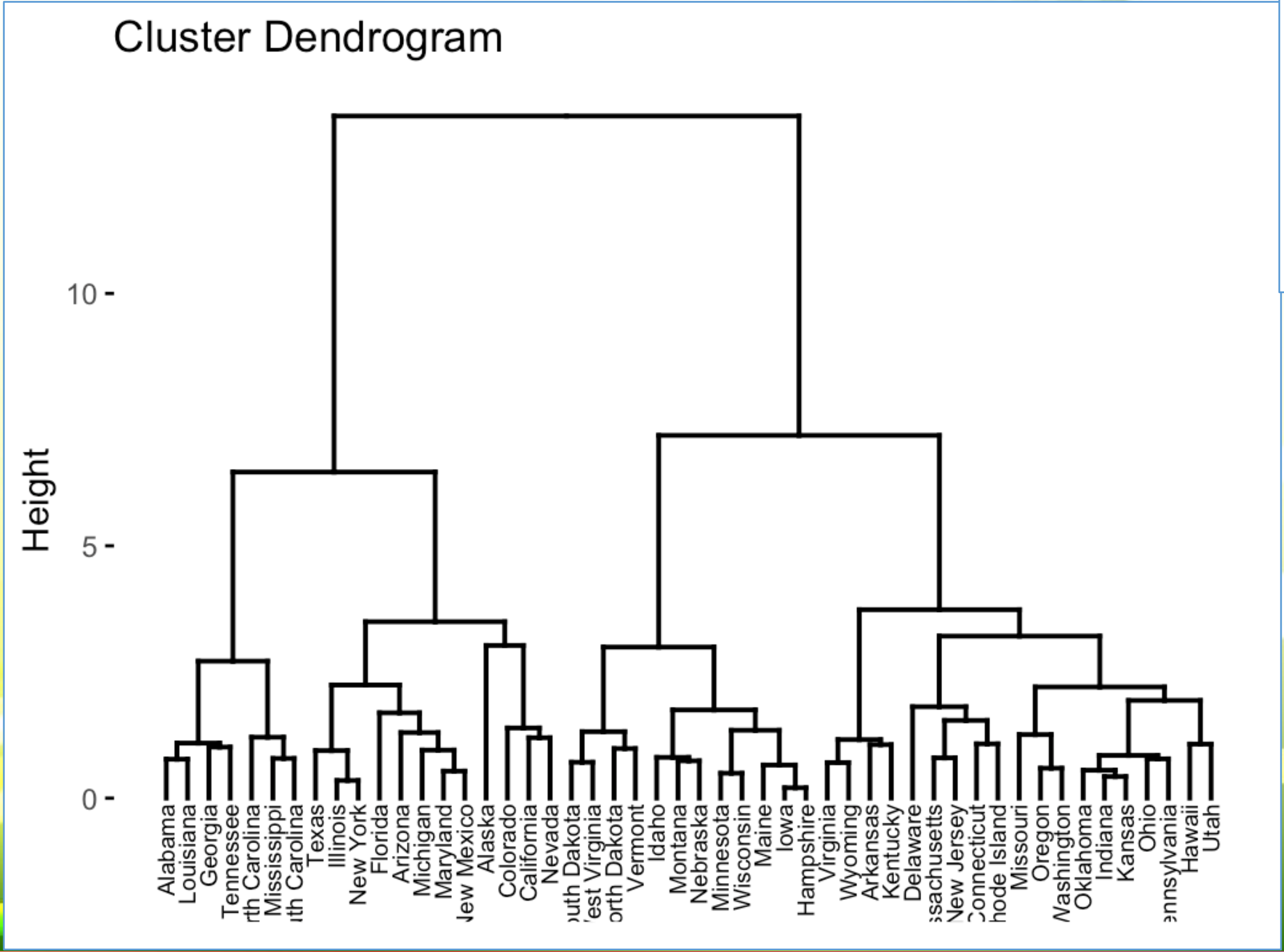


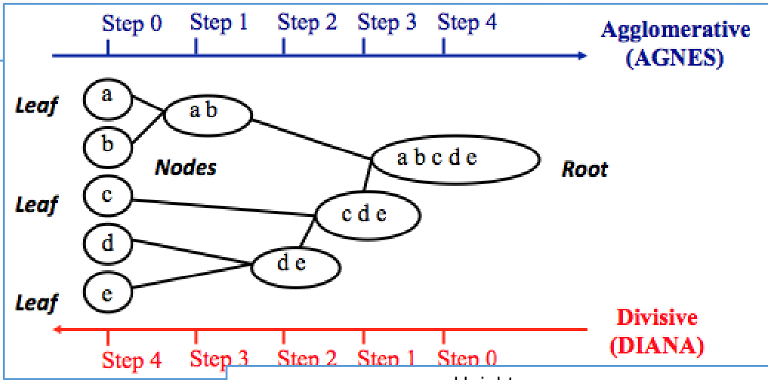
DATA ANALYSIS – HIERARCHICAL CLUSTERING

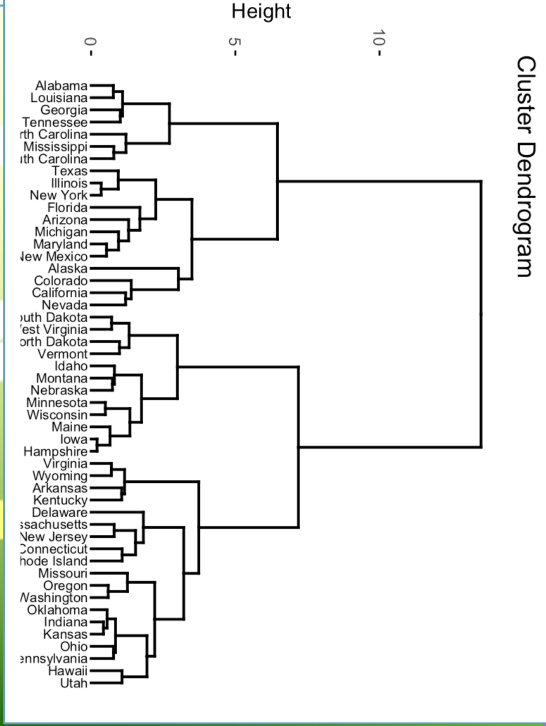




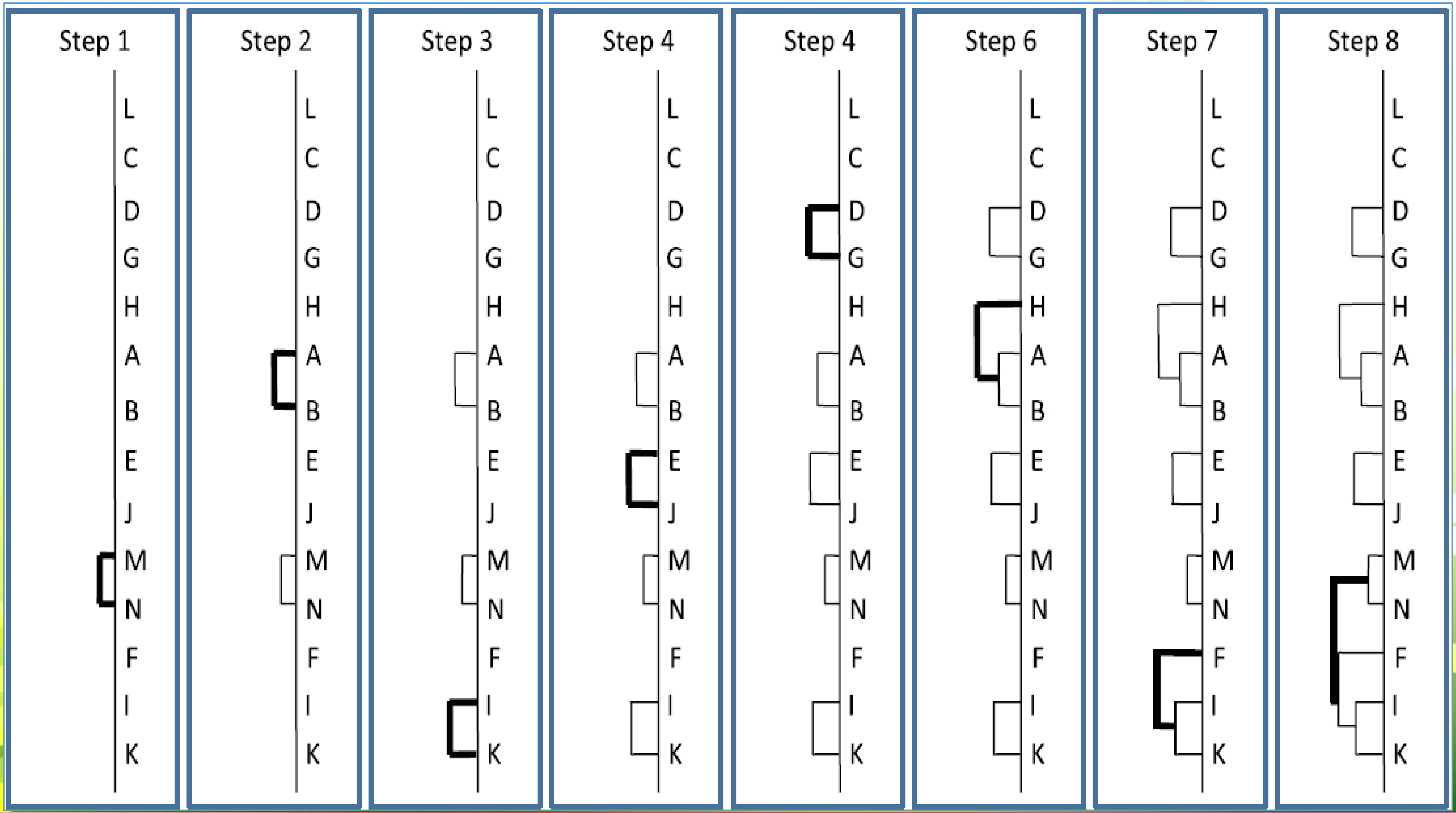
DATA ANALYSIS – HIERARCHICAL CLUSTERING

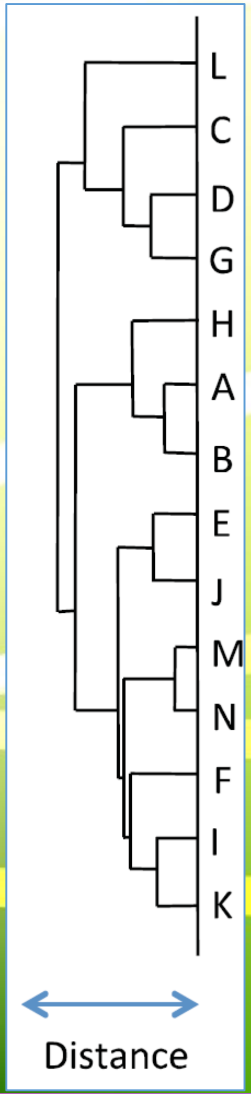




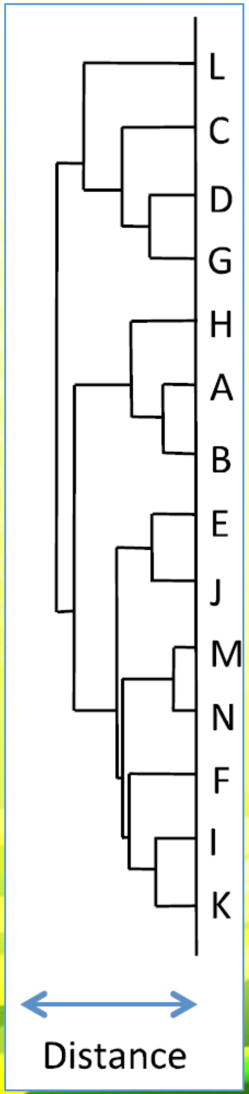


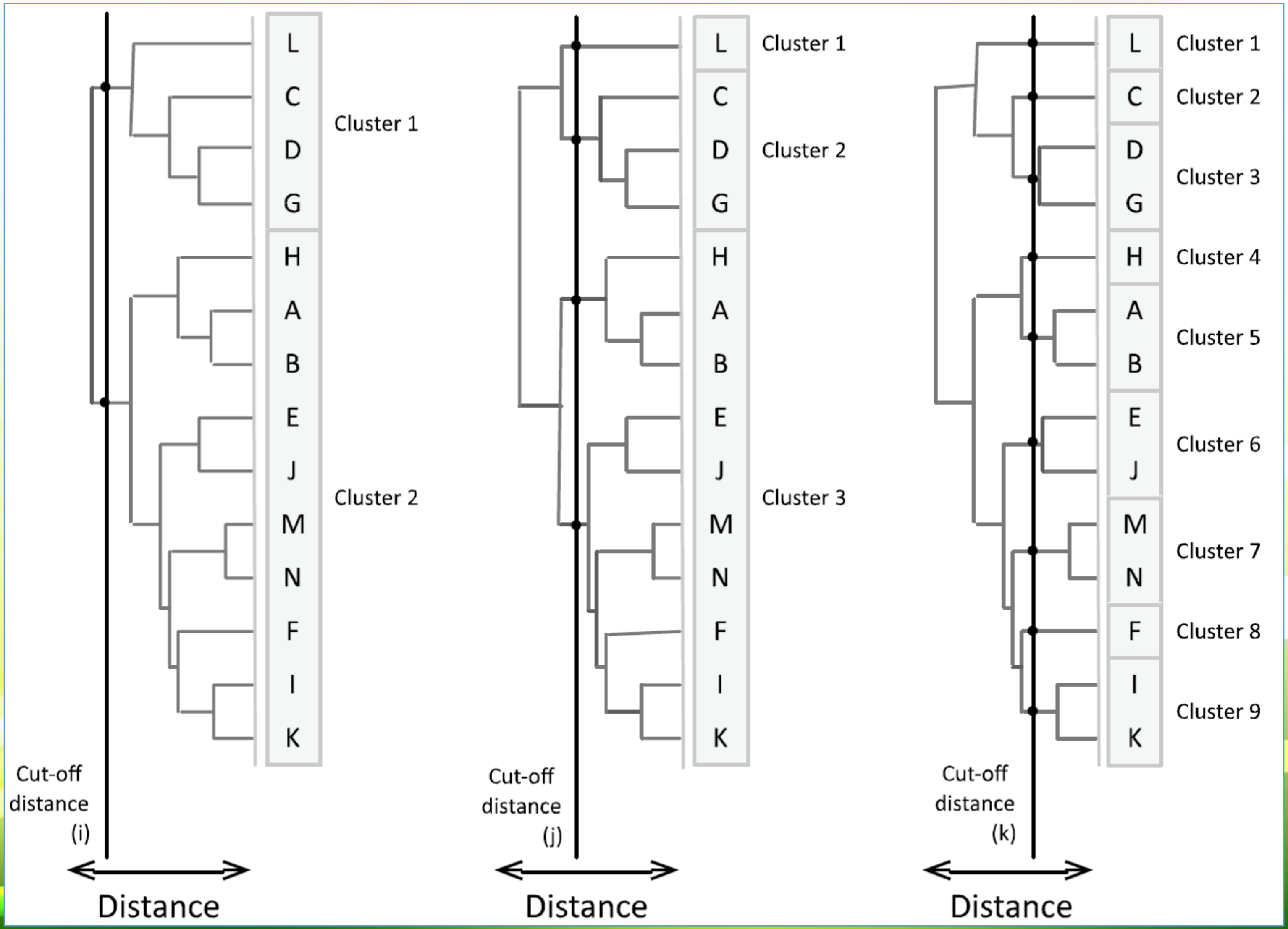
DATA ANALYSIS – HIERARCHICAL CLUSTERING





DATA ANALYSIS – HIERARCHICAL CLUSTERING





Distance cut-offs toward left result in fewer clusters with more diverse observations within each cluster.

Cut-offs toward right result in greater number of clusters with more similar observations within each cluster.

DATA ANALYSIS – K-MEANS CLUSTERING

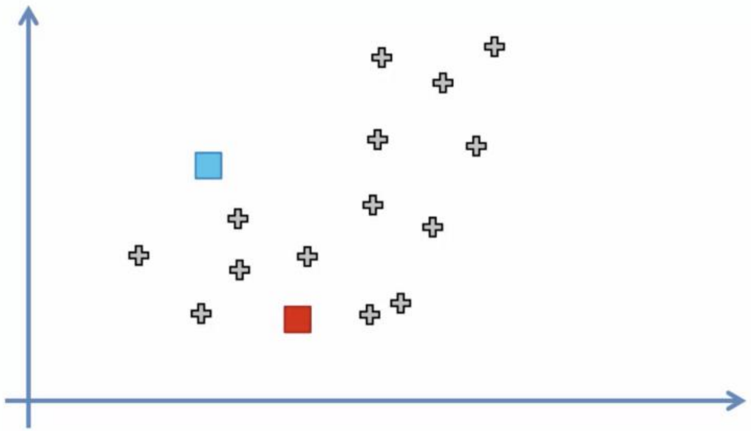
- K-Means Clustering is an unsupervised machine learning algorithm.
- Non-hierarchical method for grouping a data set.
- K-Means attempts to classify data without having first been trained with labeled data.
  - Once algorithm has been run and groups are defined, any new data can be easily assigned to most relevant group.
- *Top-down* approach: starts with predefined number of clusters and assigns all observations to each of them.
- Computationally faster and can handle greater numbers of observations than agglomerative hierarchical clustering.

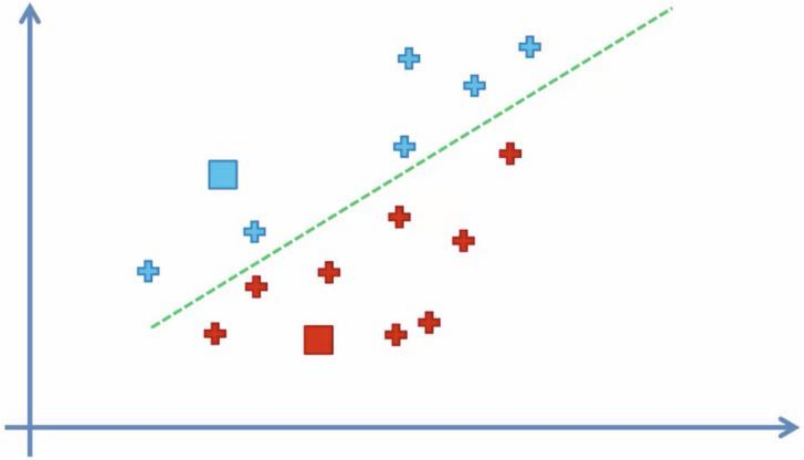
<u>Disadvantages:</u>

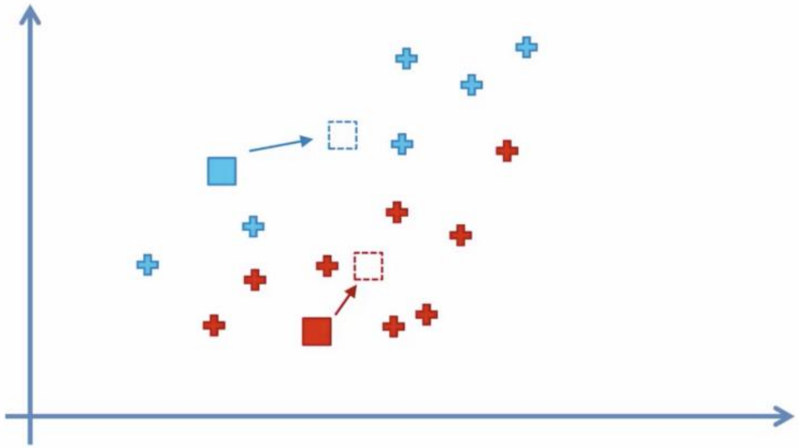
- o Number of groups (K) must be specified before creating clusters and this number is not guaranteed to provide best partitioning of observations.
- o when dataset contains many outliers, k-means may not create an optimal grouping; because reassignment of observations is based on closeness to cluster center and outliers pull cluster center in their direction *(WRONG)*.
- o No hierarchical organization is generated using k-means clustering and hence there is no ordering of individual observations.

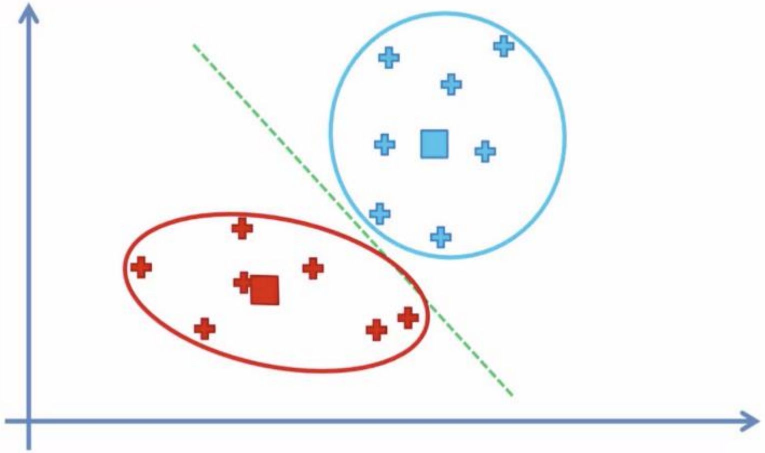
DATA ANALYSIS – K-MEANS CLUSTERING

1. Select **K** (= 2) random points as cluster centers called centroids.
2. Assign each data point to the closest cluster by calculating its distance with respect to each centroid.
3. Determine the new cluster center by computing the average of the assigned points.
4. Repeat steps 2 and 3 until none of the cluster assignments change.



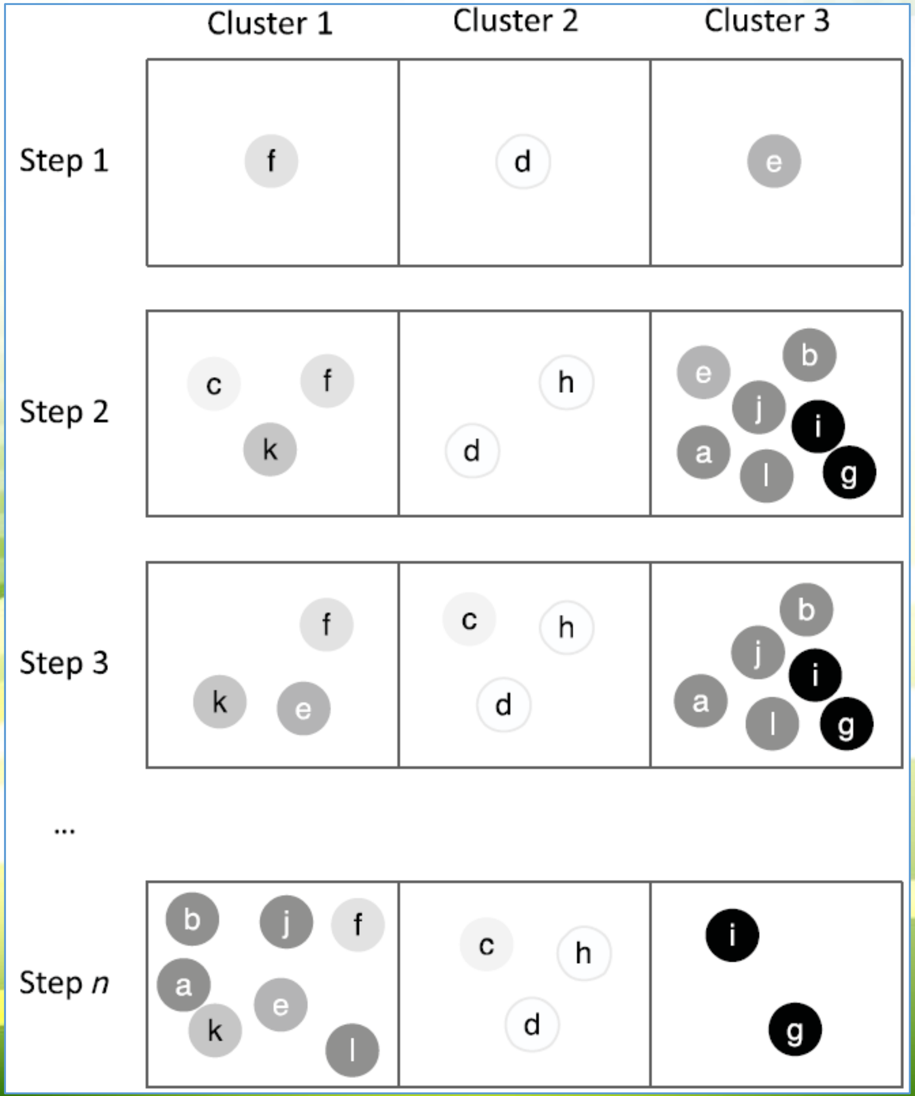






DATA ANALYSIS – K-MEANS CLUSTERING

Example



	Cluster 1	Cluster 2	Cluster 3
Step 1	f	d	e
Step 2	c f k	h d	e b j i a l g
Step 3	f k e	c h d	b j i a l g
...			
Step *n*	b j f a e k l	c h d	i g

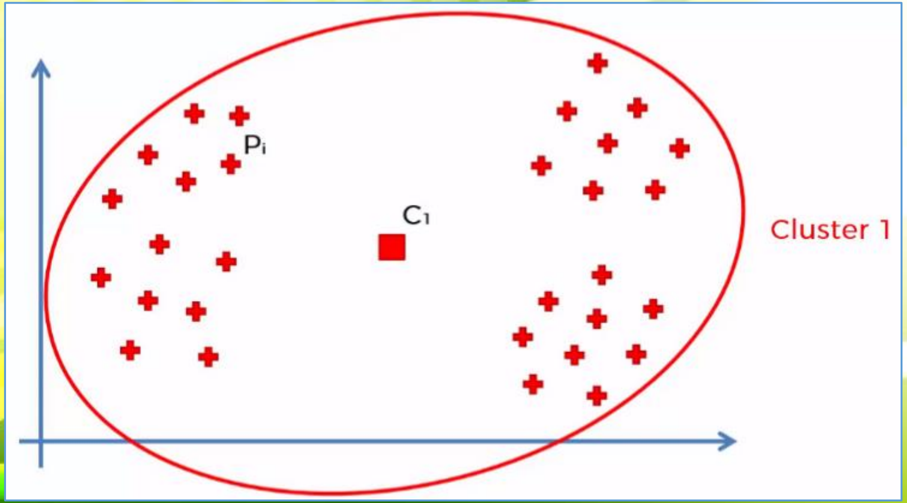
DATA ANALYSIS – K-MEANS CLUSTERING

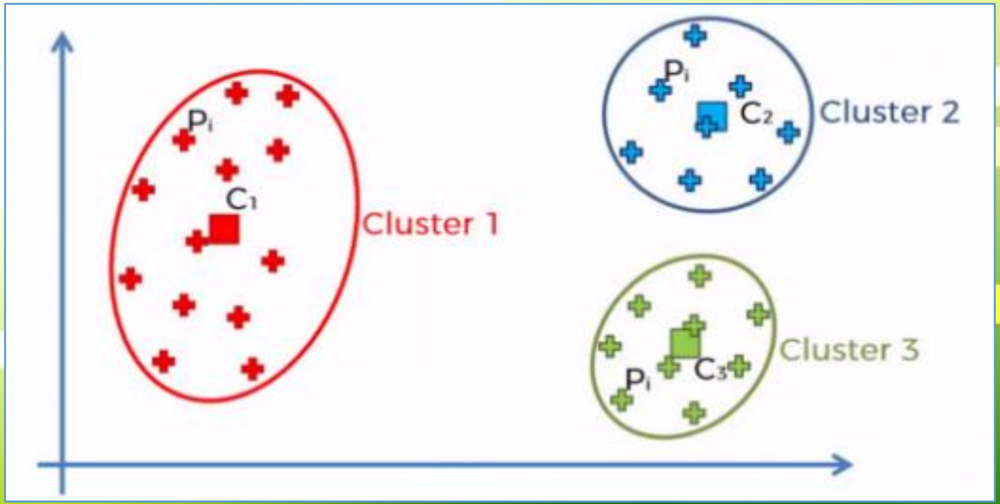
Choosing the right number of clusters

- Correct idea of number of cluster is very important for model performance.
- Choosing right number of cluster is very difficult, at first.
- Within Cluster Sum of Squares (WCSS) helps finding ‘K’ value.
  - WCSS: sum of squares of distances of data points in each and every cluster from its centroid.
  - Main idea is to minimize distance between data points and centroid of clusters.
- **Example**, computed WCSS for K=1 is greater than WCSS calculated for *K=3*.

$$WSS = \sum_{i=1}^{m} (x_i - c_i)^2$$

Where x_i = data point and c_i = closest point to centroid

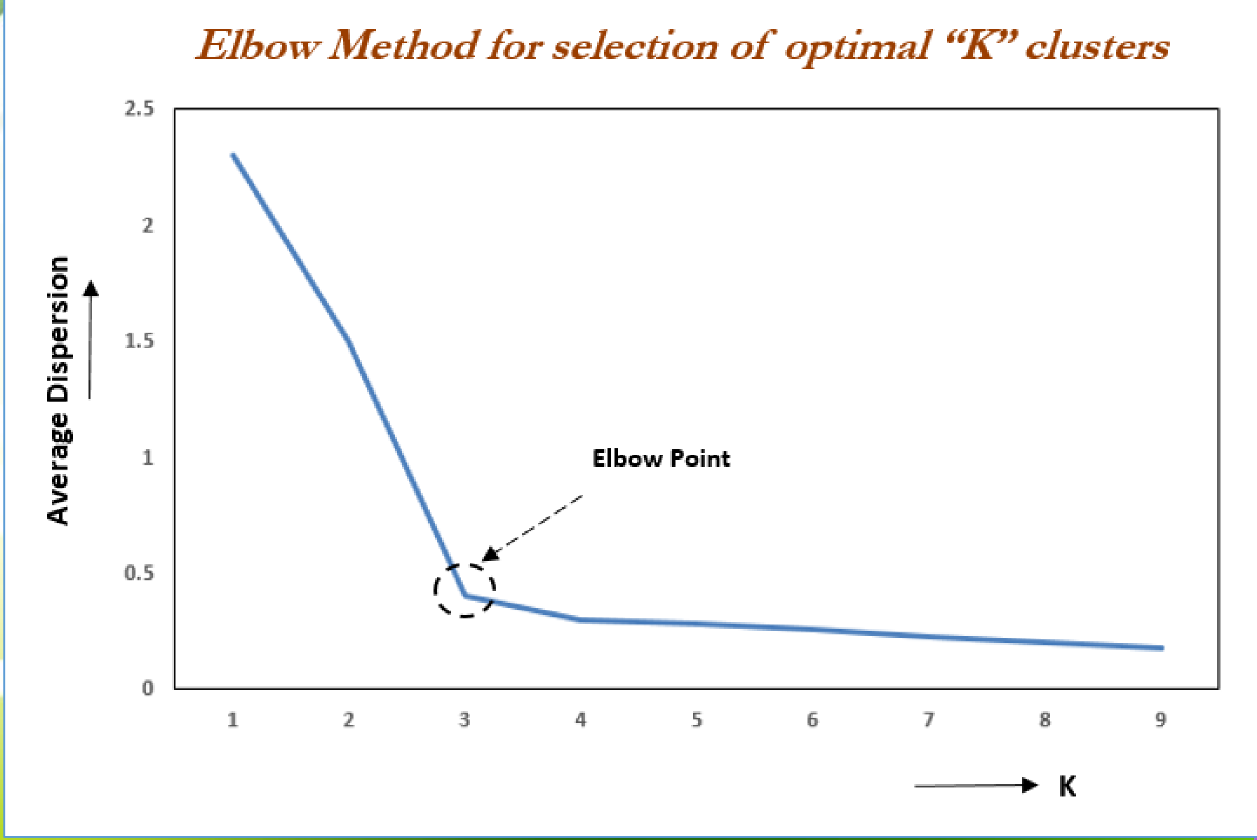




DATA ANALYSIS – K-MEANS CLUSTERING

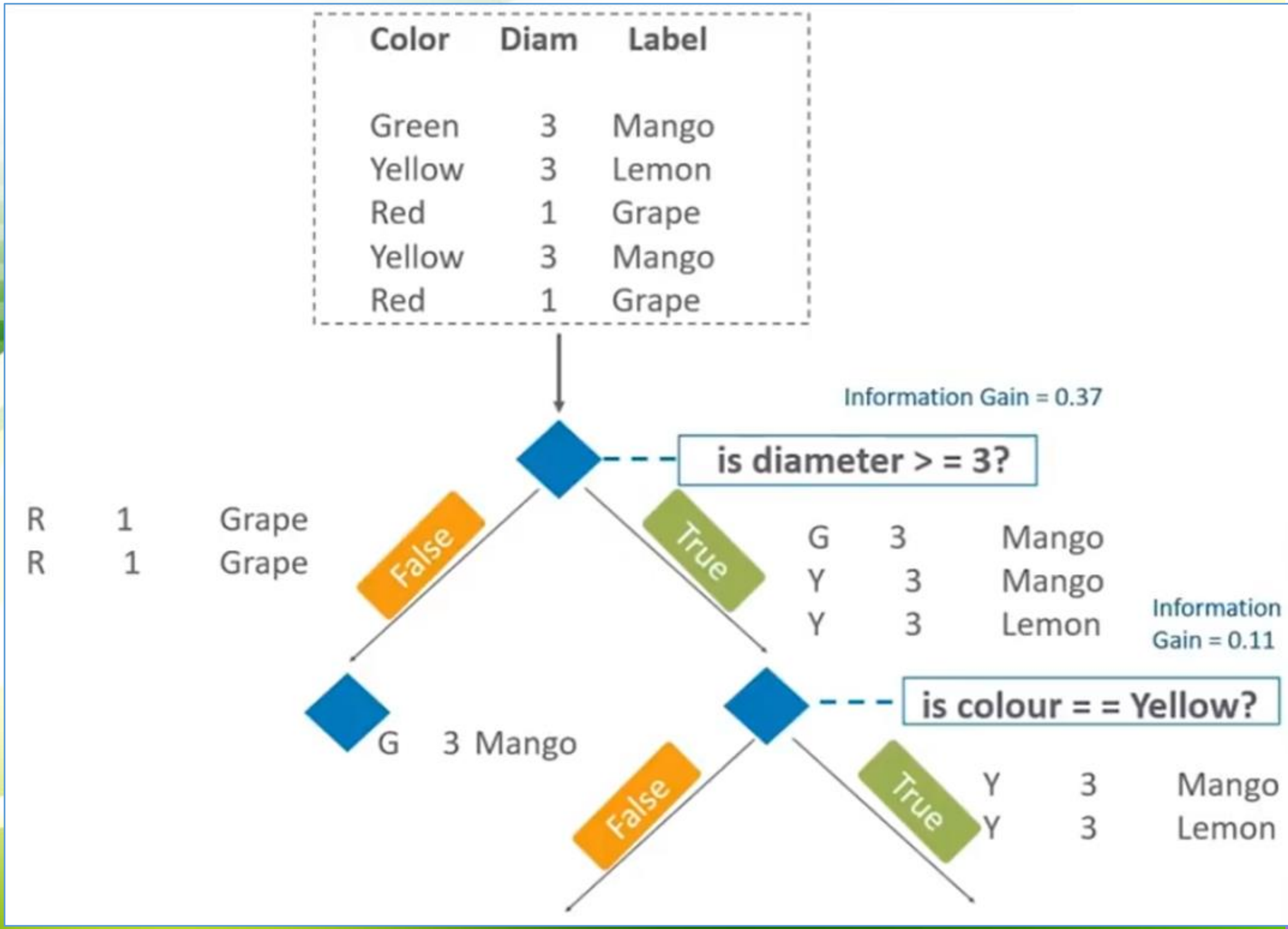
Choosing right number of clusters

- Most common technique is **elbow method**.
- Elbow method is used to determine optimal number of clusters in K-means clustering.
- Elbow method plots the value of cost function produced by different values of K.
- Value of K at which improvement in distortion declines the most is called the elbow.
- At this point, STOP division of data into further clusters.



DECISION TREE

- Supervised learning algorithm.
- Supervised methods attempt to place (classify) each observation into interesting groups (based on selected variable).
- These methods iterate over training set of observations and adjust parameters as the classifier correctly or incorrectly classifies each observation.

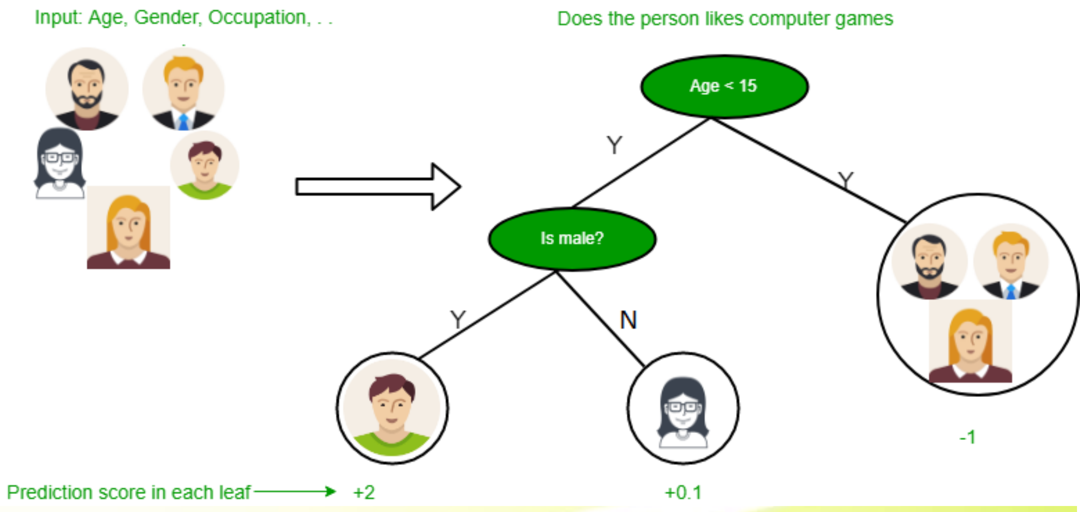


Color	Diam	Label
Green	3	Mango
Yellow	3	Lemon
Red	1	Grape
Yellow	3	Mango
Red	1	Grape

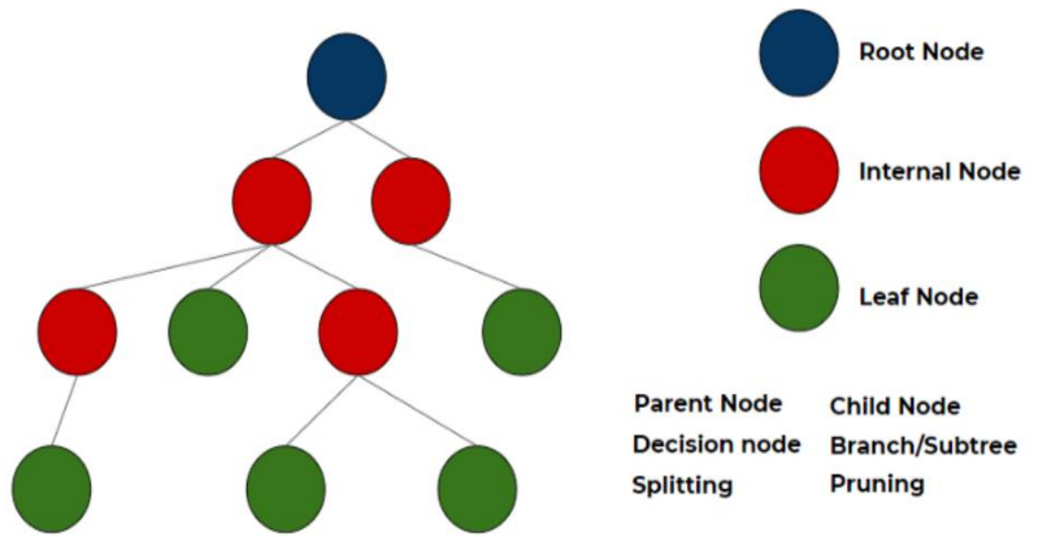
Decision Tree

- Building decision trees is computationally expensive for large data set with many continuous variables
- Decision tree uses tree representation to solve problem.
- Leaf node corresponds to a class label and attributes are represented on internal node of the tree.

- *parent–child relationship*: relationship between two connected nodes
- *splitting variable:* used as potential decision points
- *response* variable: used to guide construction of tree.
- Response variable used to guide which splitting variables are selected and at what value the split is made.
- DT splits data set into increasingly smaller, nonoverlapping subsets.
- *Root node* contains all observations.
- Splitting Condition; Termination/stop condition.


Input: Age, Gender, Occupation, . .
Does the person likes computer games
Age < 15
Y
Y
Is male?
Y
N
-1
Prediction score in each leaf
+2
+0.1

Terminologies associated with decision tree



Decision Tree

Advantages of a decision tree

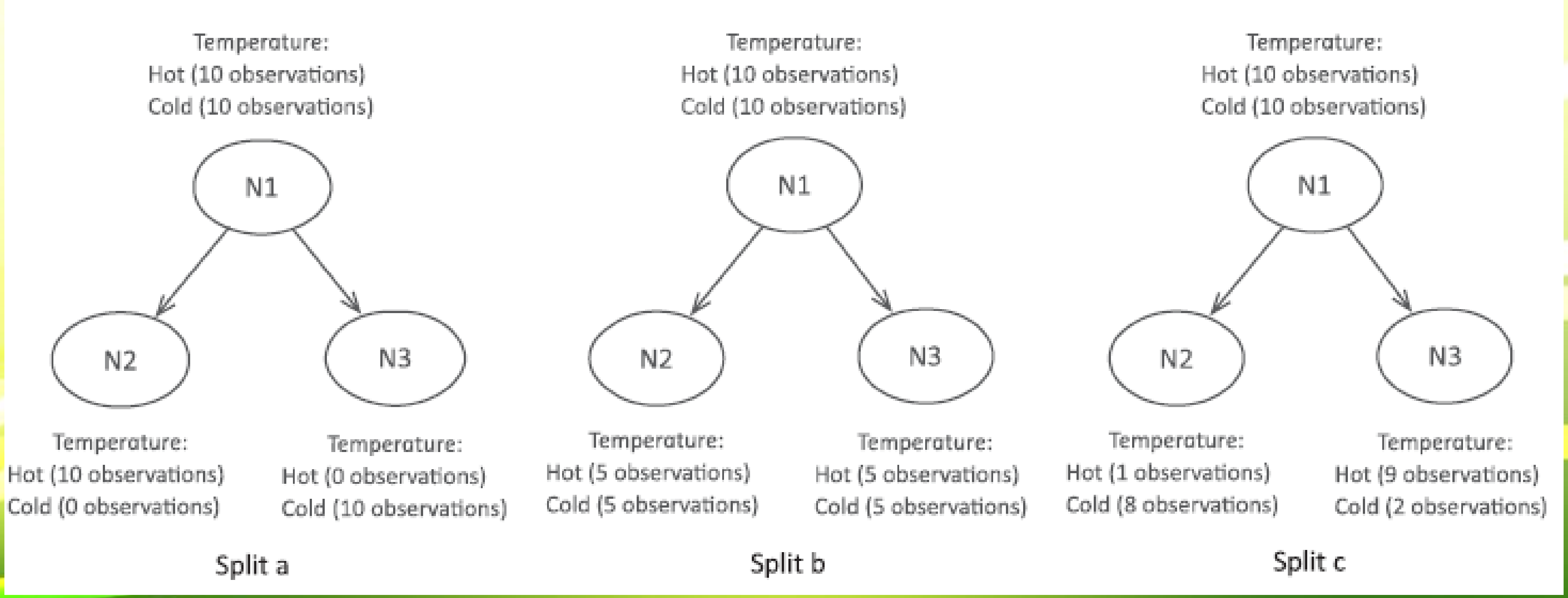
- **Easy to visualize and interpret**: Its graphical representation is very intuitive to understand and it does not require any knowledge of statistics to interpret it.
- **Useful in data exploration**: We can easily identify the most significant variable and the relation between variables with a decision tree. It can help us create new variables or put some features in one bucket.
- **Less data cleaning required**: It is fairly immune to outliers and missing data, hence less data cleaning is needed.
- **The data type is not a constraint**: It can handle both categorical and numerical data.

Disadvantages of decision tree

- **Overfitting**: single decision tree tends to overfit the data which is solved by setting constraints on model parameters and pruning.
- **Not exact fit for continuous data**: It losses some of the information associated with numerical variables when it classifies them into different categories.

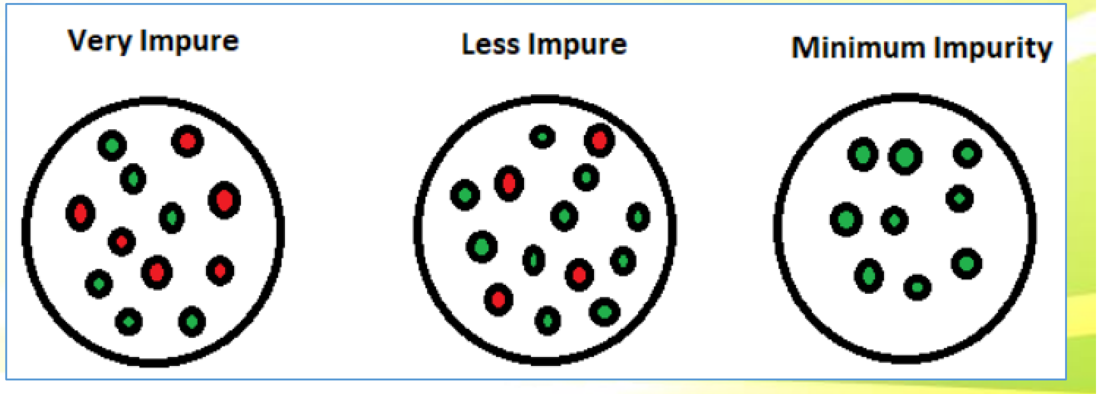
Decision Tree

- Data can be split in n-number of ways.
- To determine best split (condition), a ranking is made of all possible splits using scores calculated for each split.
  - Split a best split; each node is all of same category.
  - Split b (50% "hot," 50% "cold") not a good split.
  - Split c is good split (though not as clean as A); Good proportion of similarity with minimal impurity



Decision Tree

- “Goodness” of splitting criteria is determined by how clean each split is.
- **Impurity**: proportion of different categories of response variable.
- Cleaner splits result in lower scores.
- As the tree is being generated, it is desirable to decrease level of impurity until ideally there is only one category at a terminal node (a node with no children).
- Three primary methods for calculating impurity:
  - Misclassification,
  - Gini,
  - Entropy.
    - S: set of observations.
    - Pi : fraction of observations that belong to a particular value
    - C : number of different possible values of response variable.



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

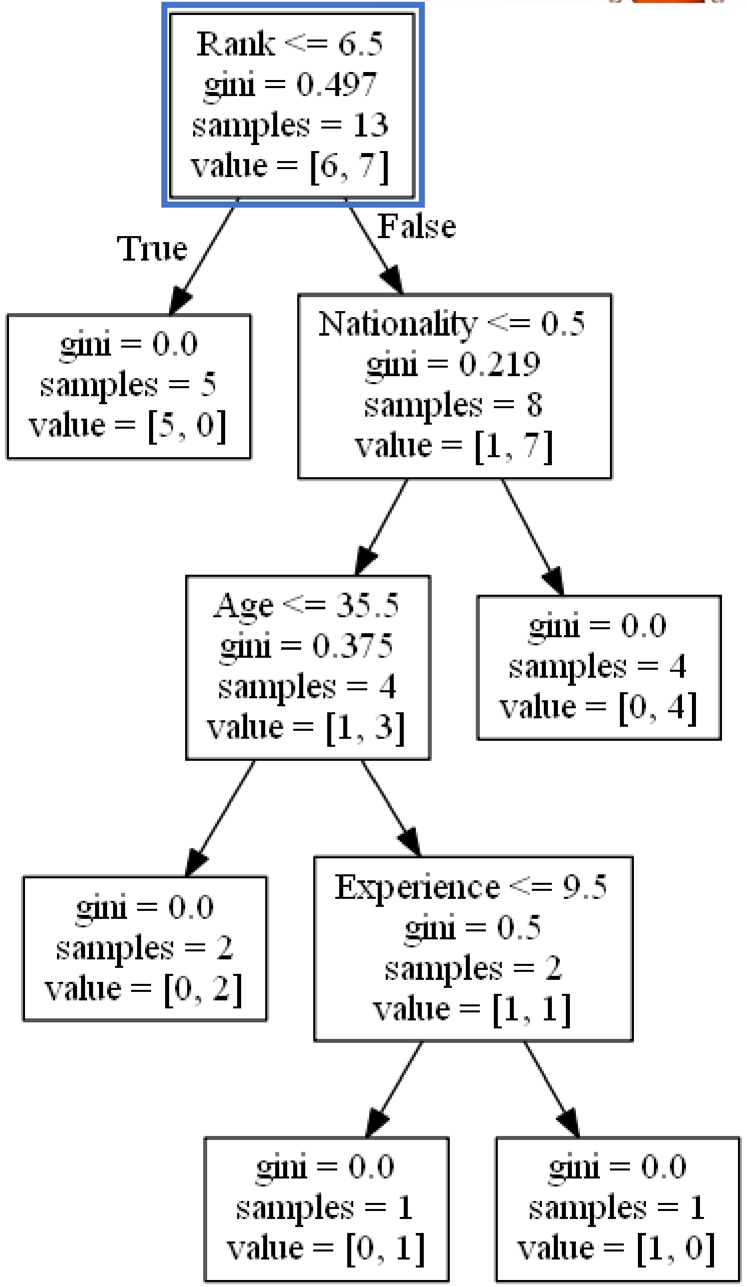
Decision Tree

- **Rank** <= 6.5 means that every candidate with a rank of 6.5 or lower will follow the True arrow (to the left), and the rest will follow the False arrow (to the right).

- **gini** = 0.497 refers to the quality of the split,
  - always a number between 0.0 and 0.5,
  - where 0.0 would mean all of the samples got the same result,
  - 0.5 would mean that the split is done exactly in the middle.

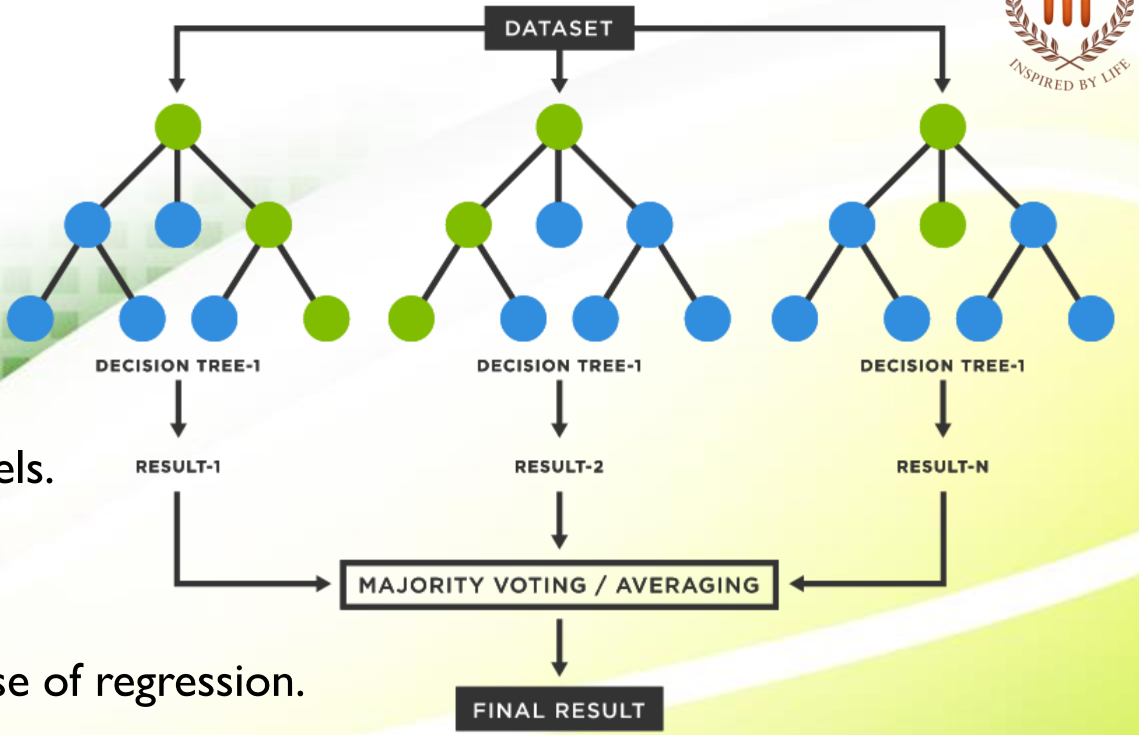
- **samples** = 13 means that there are 13 candidate left at this point in the decision, which is all of them since this is the first step.

- **value** = [6, 7] means that of these 13 candidate, 6 will get "NO", and 7 will get "GO".



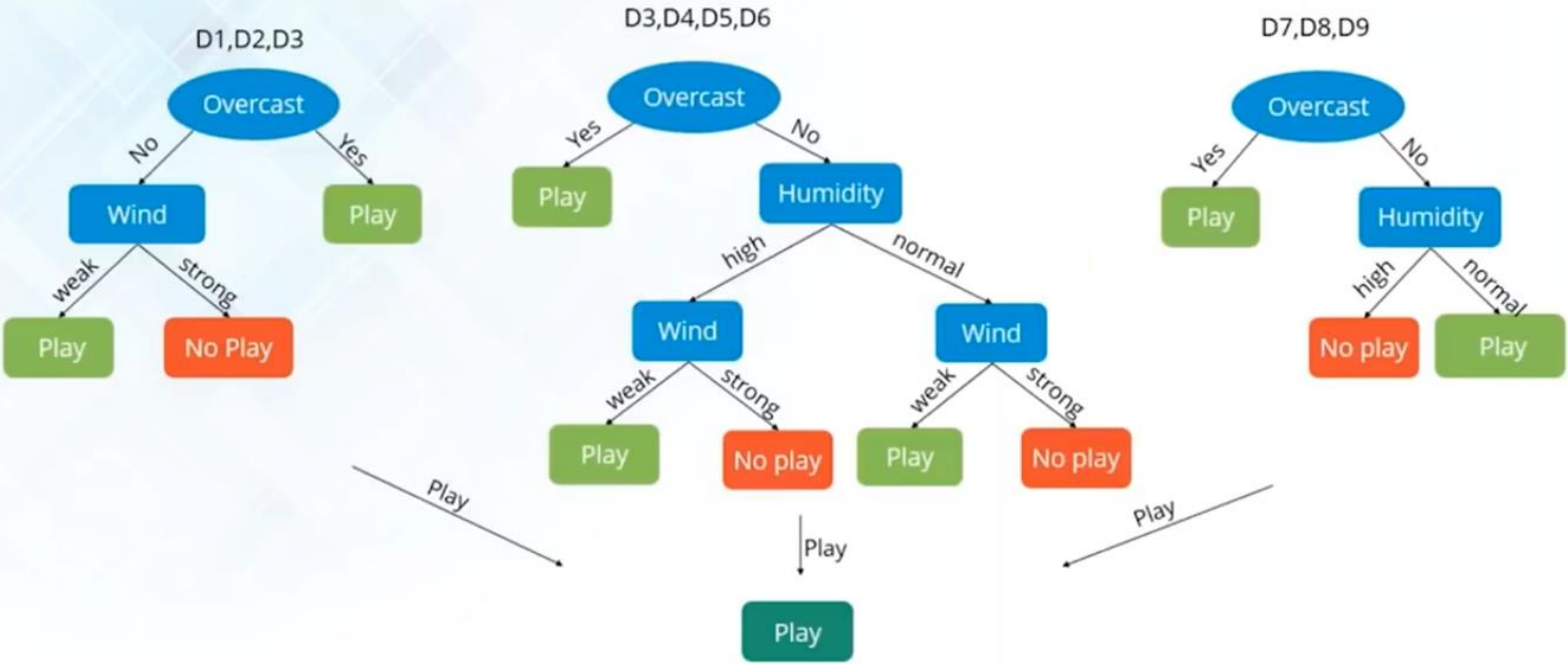
Random Forest

- *Supervised Learning Algorithm.*
- *used for Classification and Regression problems.*
- Random Forest has multiple decision trees as base learning models.
- builds decision trees on different samples
    - takes their majority vote for classification and average in case of regression.



- One important features of Random Forest Algorithm is that it can handle data set containing ***continuous variables*** as in case of regression and ***categorical variables*** as in case of classification.
    - performs better results for classification problems.
- basic idea behind this is to combine multiple decision trees in determining final output rather than relying on individual decision trees.

Random Forest


D1,D2,D3
Overcast
No
Yes
Wind
Play
weak
strong
Play
No Play
D3,D4,D5,D6
Overcast
Yes
No
Play
Humidity
high
normal
Wind
Wind
weak
strong
weak
strong
Play
No play
Play
No play
D7,D8,D9
Overcast
Yes
No
Play
Humidity
high
normal
No play
Play
Play
Play
Play
Play