SILVER OAK UNIVERSITY

EDUCATION TO INNOVATION

SILVER OAK COLLEGE OF COMPUTER APPLICATION

SUBJECT :MACHINE LEARNING

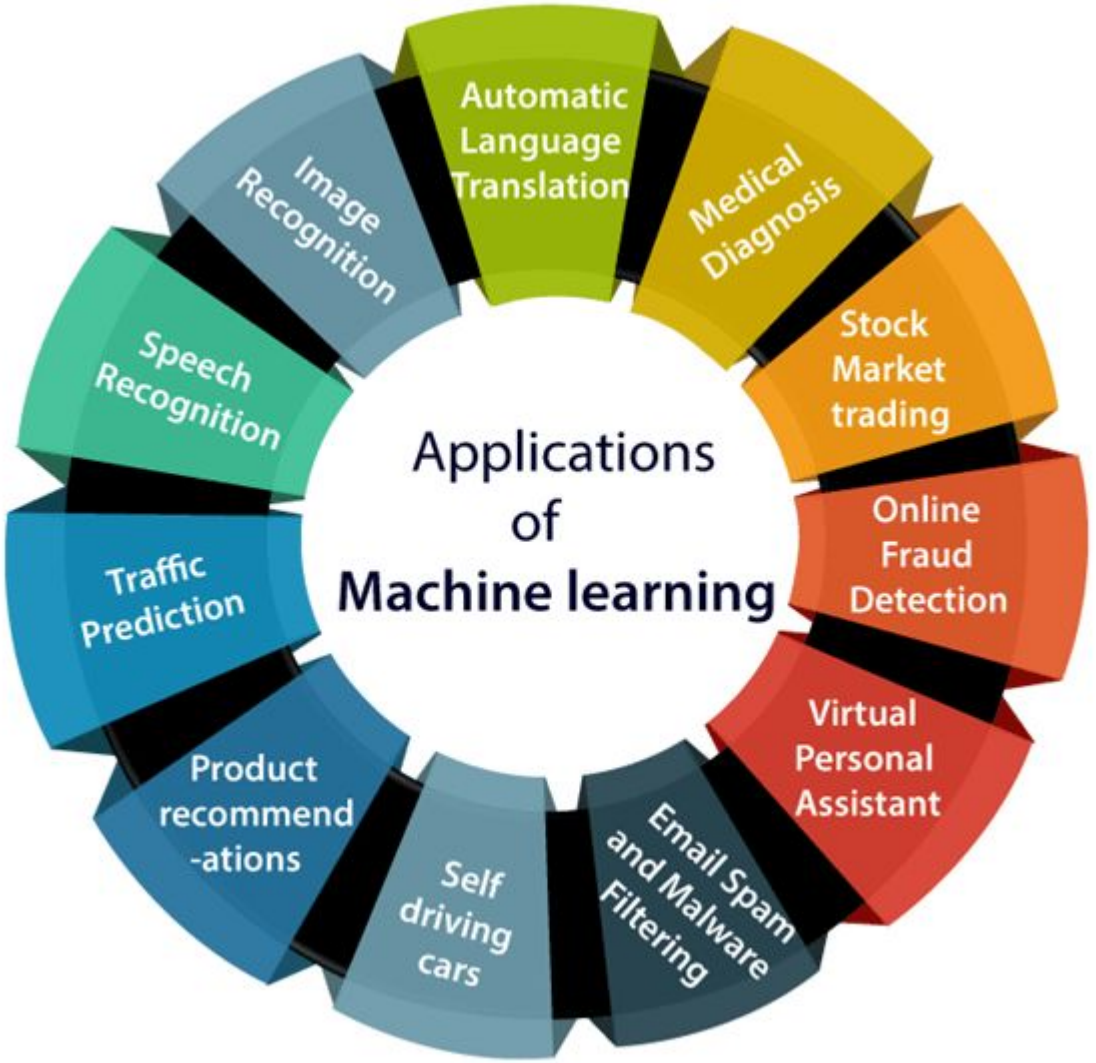
TOPIC : Unit:-3 Unsupervised Learning

Definition and its key characteristics and applications.

- **Definition**
- Machine learning (ML) is a branch of artificial intelligence (AI) that focuses on developing algorithms and statistical models that enable computers to perform specific tasks without using explicit instructions. Instead, they rely on patterns and inference derived from data.
- **Key Characteristics**

1. **Data-Driven**: ML systems learn from data rather than being programmed with explicit instructions. The quality and quantity of data play a crucial role in the performance of the model.
2. **Algorithms**: Various algorithms are used to analyze data, build models, and make predictions. These include supervised learning (e.g., regression, classification), unsupervised learning (e.g., clustering, dimensionality reduction), and reinforcement learning.
3. **Training and Testing**: ML models are trained on a dataset to learn patterns or relationships. After training, they are tested on a separate dataset to evaluate their performance and generalizability.

- Applications:



Describe the problem setup, which involves discovering patterns

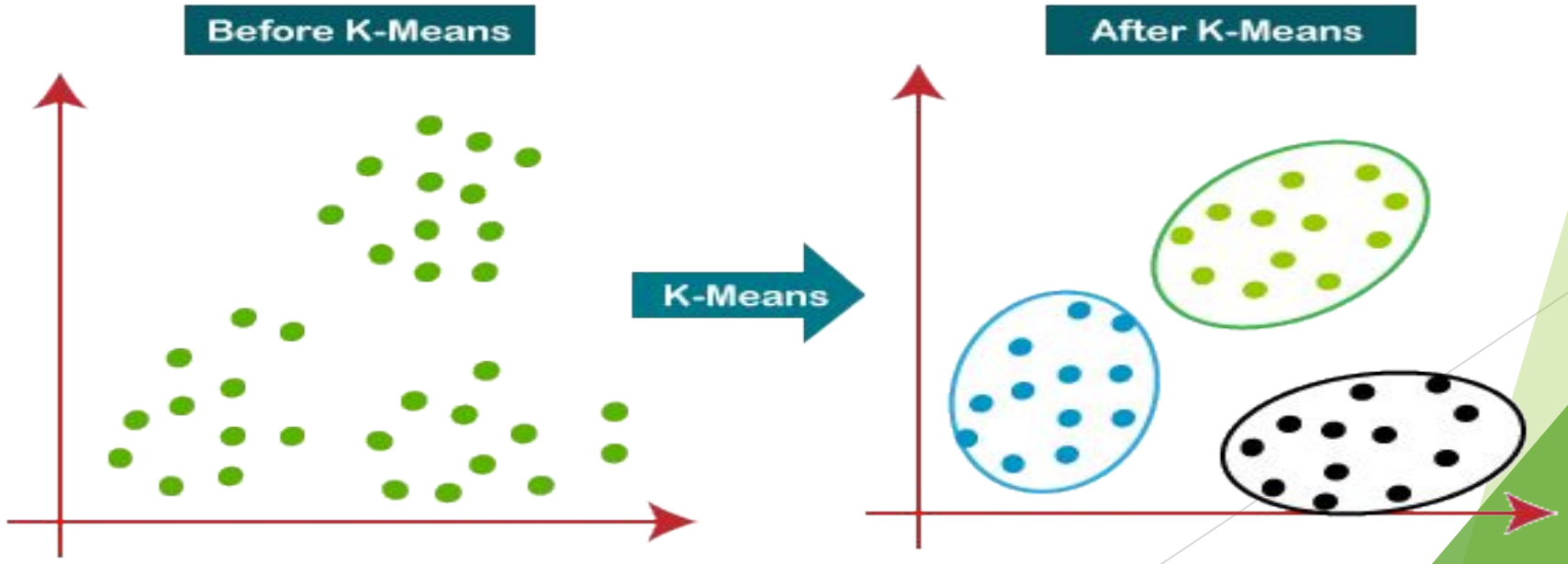
- Identifying patterns, trends, and correlations is an essential task that allows decision-makers to extract important insights from a sea of data. This blog digs into the complex art of identifying these critical characteristics in data, shining light on their importance in sectors such as banking, healthcare, marketing, and more.
- Patterns are recurring sequences or groupings seen in data and are frequently hidden under the surface. They give the predictive possibility of future events by providing an elementary understanding of the underpinning structure.
- Subsequently, recognising trends entails determining the trajectory of data points over time. This temporal viewpoint benefits forecasting and strategic decision-making.

Clustering algorithm: -K Means :

- K-Means is a popular clustering algorithm used in unsupervised machine learning to partition a dataset into KKK distinct, non-overlapping subsets or clusters. Here's a brief overview of how it works and some key points:
- **How K-Means Works**

1. **Initialization**: Choose KKK initial centroids (these could be selected randomly or using some heuristic).
2. **Assignment Step**: Assign each data point to the nearest centroid based on a distance metric (commonly Euclidean distance).
3. **Update Step**: Recalculate the centroids by taking the mean of all points assigned to each centroid.
4. **Repeat**: Repeat the assignment and update steps until the centroids no longer change significantly or a specified number of iterations is reached.
5. **Convergence**: The algorithm converges when the centroids stabilize, meaning there's no significant change in their positions or the cluster assignments.

- K-Means Clustering Algorithm
- K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. In this topic, we will learn what is K-means clustering algorithm, how the algorithm works, along with the Python implementation of k-means clustering.
- What is K-Means Algorithm?
- K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.
- The below diagram explains the working of the K-means Clustering Algorithm:
-

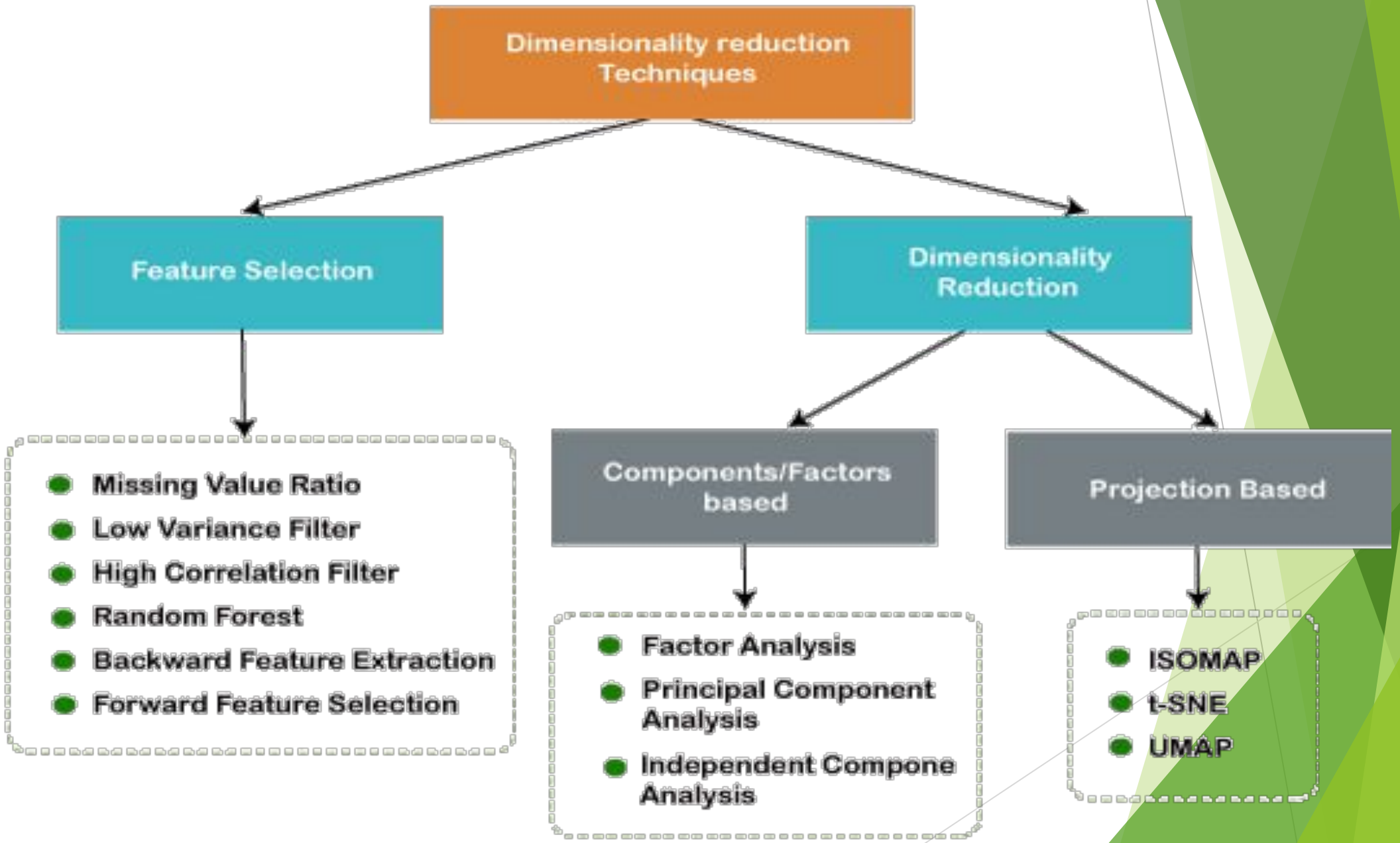


- How does the K-Means Algorithm Work?
- The working of the K-Means algorithm is explained in the below steps:
- **Step-1:** Select the number K to decide the number of clusters.
- **Step-2:** Select random K points or centroids. (It can be other from the input dataset).
- **Step-3:** Assign each data point to their closest centroid, which will form the predefined K clusters.
- **Step-4:** Calculate the variance and place a new centroid of each cluster.
- **Step-5:** Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.
- **Step-6:** If any reassignment occurs, then go to step-4 else go to FINISH.
- **Step-7**: The model is ready.

Dimensionality reduction: - Principal Component Analysis :

- Principal Component Analysis (PCA) is a linear technique used to reduce the dimensionality of data by projecting it onto a new set of orthogonal axes (principal components) that maximize variance. These principal components are linear combinations of the original features and are arranged in descending order of variance.
- **2. Key Characteristics**

1. **Variance Maximization**: PCA seeks to find the directions (principal components) that capture the maximum variance in the data. The first principal component accounts for the most variance, the second for the next highest variance orthogonal to the first, and so on.
2. **Orthogonality**: Principal components are orthogonal (uncorrelated) to each other. This ensures that each component represents a unique dimension of the data.
3. **Linear Transformation**: PCA is a linear method, meaning it transforms the data using linear combinations of the original features. It does not capture non-linear relationships in the data.
4. **Eigenvalues and Eigenvectors**: PCA involves calculating the eigenvalues and eigenvectors of the covariance matrix of the data. The eigenvectors represent the principal components, while the eigenvalues indicate the amount of variance captured by each component.
5. **Data Centering**: PCA requires the data to be centered (i.e., mean subtracted) so that the principal components are computed based on the variance from the mean.


Dimensionality reduction Techniques
Feature Selection
Dimensionality Reduction
Missing Value Ratio
Low Variance Filter
High Correlation Filter
Random Forest
Backward Feature Extraction
Forward Feature Selection
Components/Factors based
Projection Based
Factor Analysis
Principal Component Analysis
Independent Compone Analysis
ISOMAP
t-SNE
UMAP