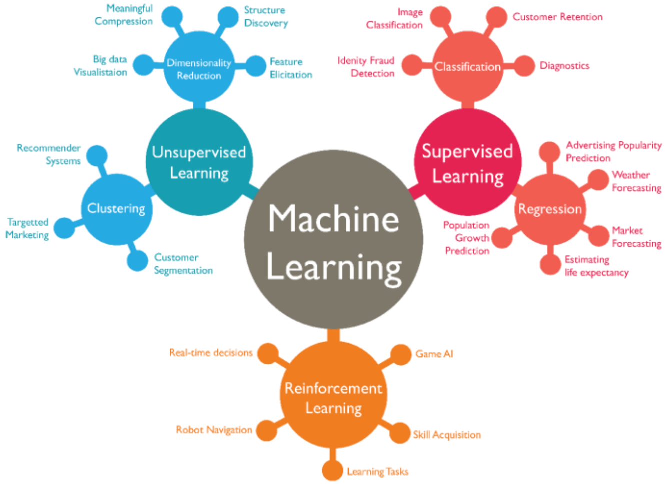
Pyspark
MLlib

Agenda

- What is Machine Learning
- Introduction of MLlib
- Difference between python machine learning and Spark Mllib
- When we need to use Spark Mllib
- Classification Algorithms
- UnSupervised Learning
- Data Pre-processing
- Implementation Of Algorithms

What is Machine Learning

- Machine learning (ML) is the study of computer algorithms that can improve automatically through experience and by the use of data
- It is seen as a part of artificial intelligence
- Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed
- Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks
- Machine Learning subsets:-
  - Deep Learning
  - Data Mining
  - Artificial Intelligence (A.I)



Introduction of MLlib

- MLlib is Apache Spark's scalable machine learning (ML) library
- Spark ML standardizes APIs for machine learning algorithms to make it easier to combine multiple algorithms into a single pipeline
- Spark ML uses the SchemaRDD from Spark SQL as a dataset that can hold a variety of data types. E.g., a dataset could have different columns storing text, feature vectors, true labels, and predictions
- A Transformer is an algorithm that can transform one SchemaRDD into another SchemaRDD. E.g., an ML model is a Transformer that transforms an RDD with features into an RDD with predictions
- An Estimator is an algorithm that can be fit on a SchemaRDD to produce a Transformer E.g., a learning algorithm is an Estimator which trains on a dataset and produces a model

Machine Learning Vs Spark MLlib

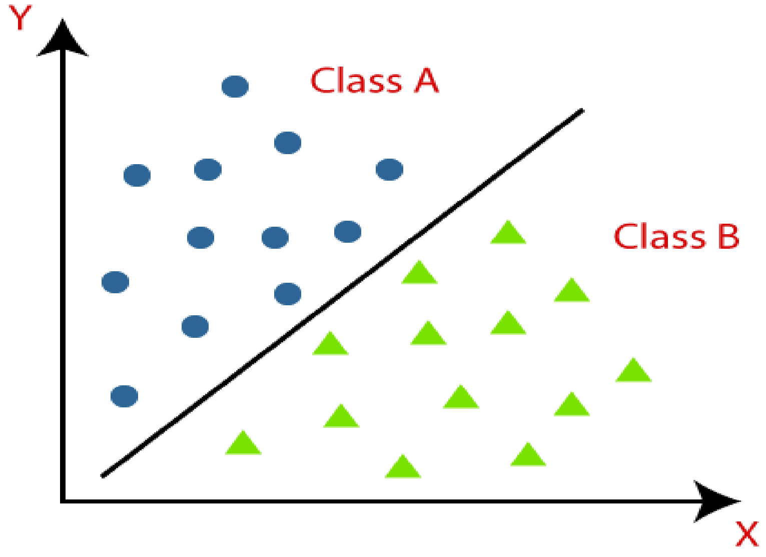
- Machine learning is getting popular in solving real-world problems in almost every business domain
- It helps solve the problems using the data which is often unstructured, noisy, and in huge size
- With the increase in data sizes and various sources of data, solving machine learning problems using standard techniques pose a big challenge
- Spark is a distributed processing engine using the MapReduce framework to solve problems related to big data and processing of it

Need Of Spark MLlib

- When data is huge we need to use spark MLlib
- Train model from different data sources
- Due to parallel computation it is so faster
- MLlib in Spark is a scalable Machine learning library that discusses both high-quality algorithms and high speed
- Easily to train the model on the particular data source and In-Memory

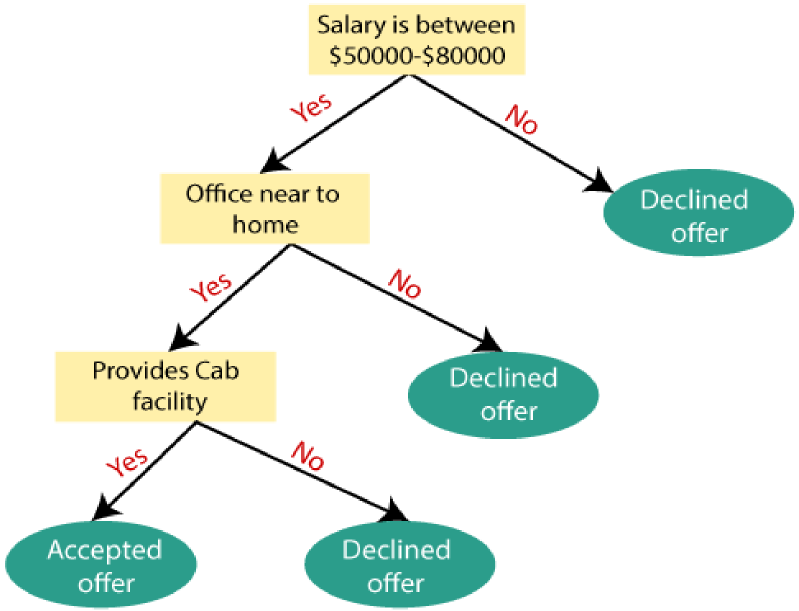
What is Classification

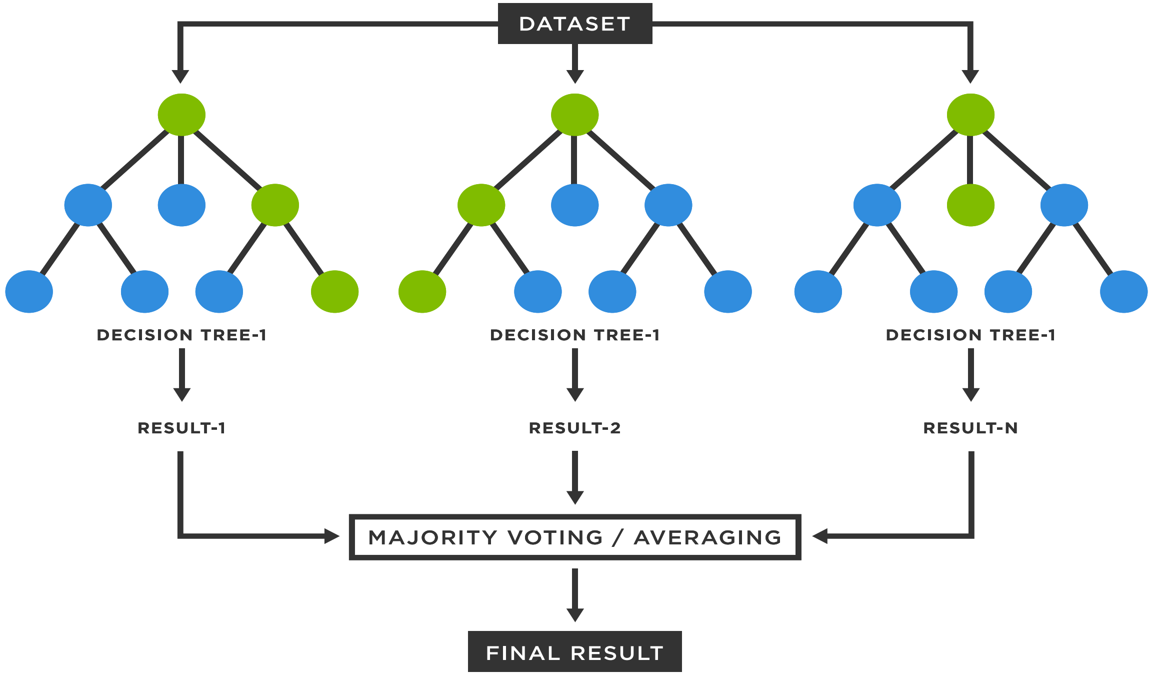
- Classification is a process of categorizing a given set of data into classes, It can be performed on both structured and unstructured data
- The process starts with predicting the class of given data points
- The classes are often referred to as target, label, or categories
- The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data



Supervised Classification Algorithms

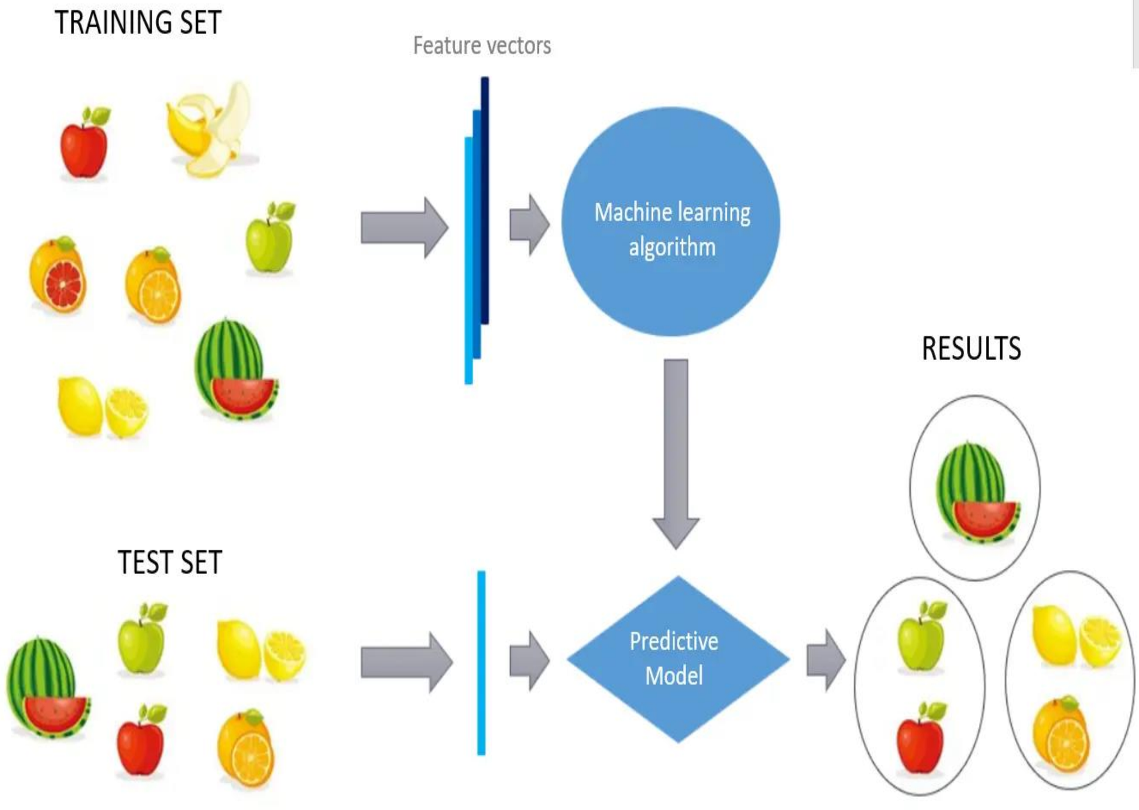
- Classification Algorithms
  - Decision Tree Classification
  - Random Forest Classification





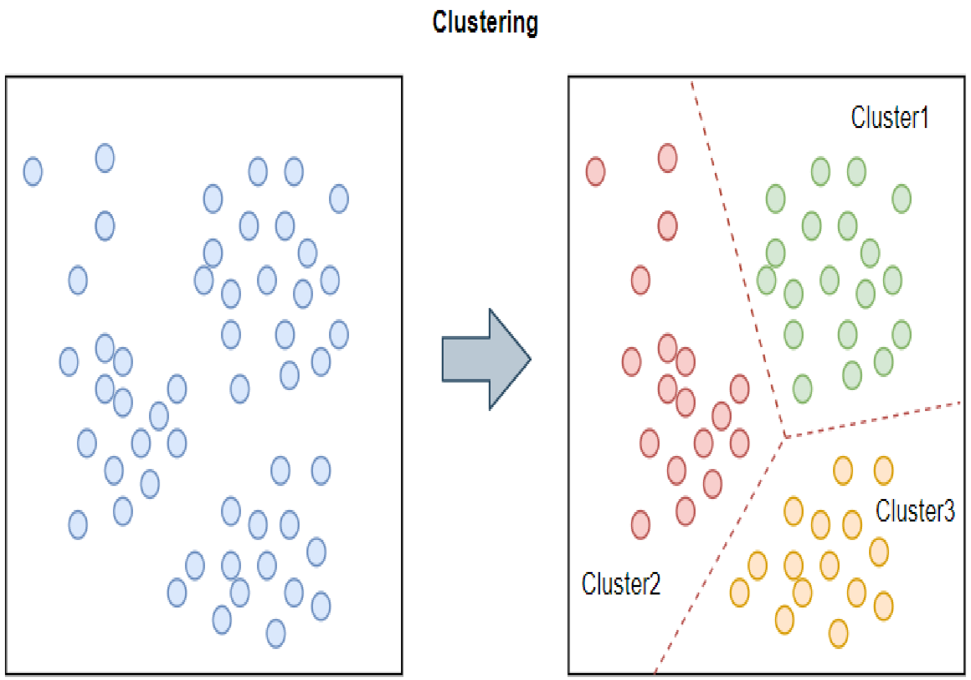
UnSupervised Learning

- Unsupervised learning is where you only have input data (X) and no corresponding output variables
- All data is unlabeled and the algorithms learn to inherent structure from the input data
- Make supervised data from unsupervised data using clustering algorithms
- Unsupervised learning problems can be further grouped into clustering and then apply supervised algorithms to clustered data



UnSupervised Algorithms

- K-means clustering
- k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid)
- k-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances



Data Pre-processing

- Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model
- Data preprocessing Phases:-
    - Data Sampling:-
        - Sampling is a method that allows us to get information about the population based on the statistics from a subset of the population (sample), without having to investigate every individual
    - Data Cleaning:-
        - Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset
    - Training Data:-
        - Training data is an extremely large dataset that is used to teach a machine learning model
    - Testing Data:-
        - Once your machine learning model is built (with your training data), you need unseen data to test your model. This data is called testing data, and you can use it to evaluate the performance and progress of your algorithms' training and adjust or optimize it for improved results

Implementation

- Implement Decision Tree Classification
- Implement Random Forest Classification
- Implement K-means clustering