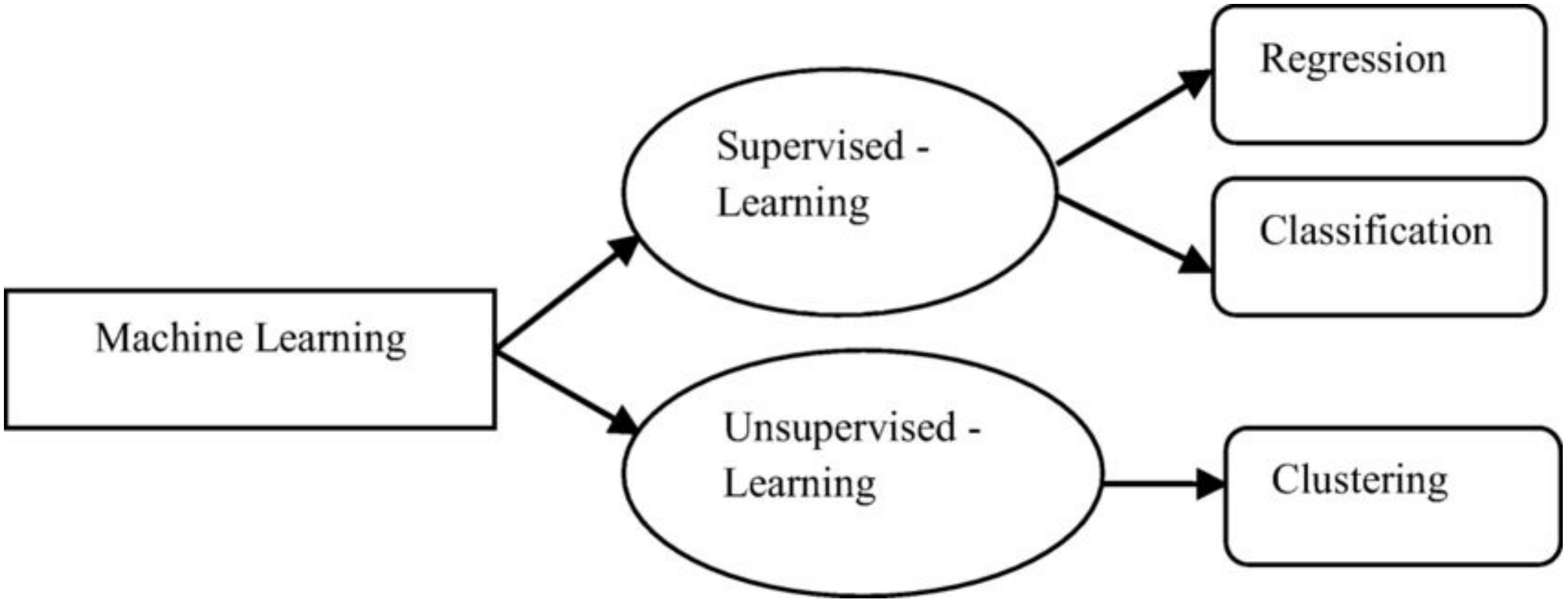
Programming for AI

Abdul Haseeb

BS(AI)-IV

What is Machine Learning?



Machine Learning Library

- **Scikit-learn** (often abbreviated as **sklearn**) is a popular open-source machine learning library for Python.

Sklearn

Scikit-learn offers a wide range of machine learning algorithms and tools for tasks such as:

Classification: Identifying which category an object belongs to (e.g., spam vs. non-spam).

Regression: Predicting a continuous value (e.g., predicting house prices).

Clustering: Grouping similar data points (e.g., customer segmentation).

Dimensionality reduction: Reducing the number of features while retaining essential information (e.g., Principal Component Analysis - PCA).

Model selection: Tuning model parameters and cross-validation.

Preprocessing: Scaling, transforming, and normalizing data.

Scaling

Scaling is a technique to bring all the features in the dataset on a same scale

ML Algorithms then perform better and don’t act biased, because algorithm gives equal importance to all features.

ML Process


Step -1
Collection of Data from Various source
Step -2
Data cleaning and Feature Engineering
Step -3
Model building for selecting correct ML Algorithm
Step -4
Evaluate Model
Step -5
Model Deployment

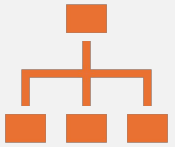
Logistic Regression

Used for: Classification problems (e.g., yes/no, spam/not spam).

Idea: Predicts the probability of a class (like 0 or 1). Despite the name, it's used for **classification**, not regression.

Why use it: Simple, fast, and works well when data is linearly separable.

K-Nearest Neighbors (KNN)



Used for: Classification and regression.

Idea: Looks at the "K" closest data points (neighbors) and decides the label based on majority vote (in classification).

Why use it: Very intuitive and easy to implement. No training is needed; it just stores the data.

Support Vector Machine (SVM)

Used for: Classification and regression.

Idea: Finds the best boundary (hyperplane) that separates different classes. Why use it:

Very powerful in high-dimensional spaces and works well with clear margin separation.

Naive Bayes

Used for: Classification.

Idea: Based on Bayes' Theorem. Assumes that features are independent (hence "naive").

Why use it: Fast and effective, especially good for text data like spam filtering or sentiment analysis.

Decision Tree

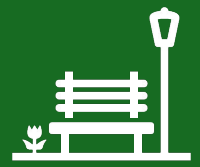
Used for: Classification and regression.

Idea: Builds a tree where each decision splits the data based on a feature to reach a final prediction.

Why use it: Easy to understand and interpret; works well with both numerical and categorical data.

Random Forest

Used for: Classification and regression.



Idea: An ensemble of many decision trees. It combines their predictions to improve accuracy and reduce overfitting.

Why use it: More accurate and stable than a single decision tree.

Gradient Boosting

Used for: Classification.

Idea: Builds trees one after another, each trying to fix the mistakes of the previous one.

Why use it: High performance in complex datasets. Often used in competitions (e.g., Kaggle).

MLP Classifier (Multi-Layer Perceptron)

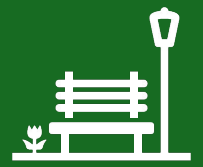
Used for: Classification.

Idea: A type of neural network with layers of nodes (neurons) that can learn complex patterns.

Why use it: Good for learning non-linear patterns, especially when simpler models don't perform well.

AdaBoost Classifier

Used for: Classification.



Idea: Combines several weak models (like shallow trees) to create a strong one by focusing more on difficult examples.

Why use it: Simple, yet powerful. Works well with clean, structured data.

Linear Regression

Linear Regression is a simple and commonly used machine learning algorithm for **predicting a continuous value** (like predicting price, temperature, height, etc.).

It finds the best straight line that fits through the data points to predict a value based on the input.

Clustering

Clustering is an **unsupervised learning** technique used to **group similar data points** together based on their features — **without using any labels**.

Think of it like sorting a basket of mixed fruits into groups: apples, bananas, oranges — based on their shape, color, or size — **without being told which is which**.

K-Means is one of the popular clustering algorithm