Machine Learning for Beginners

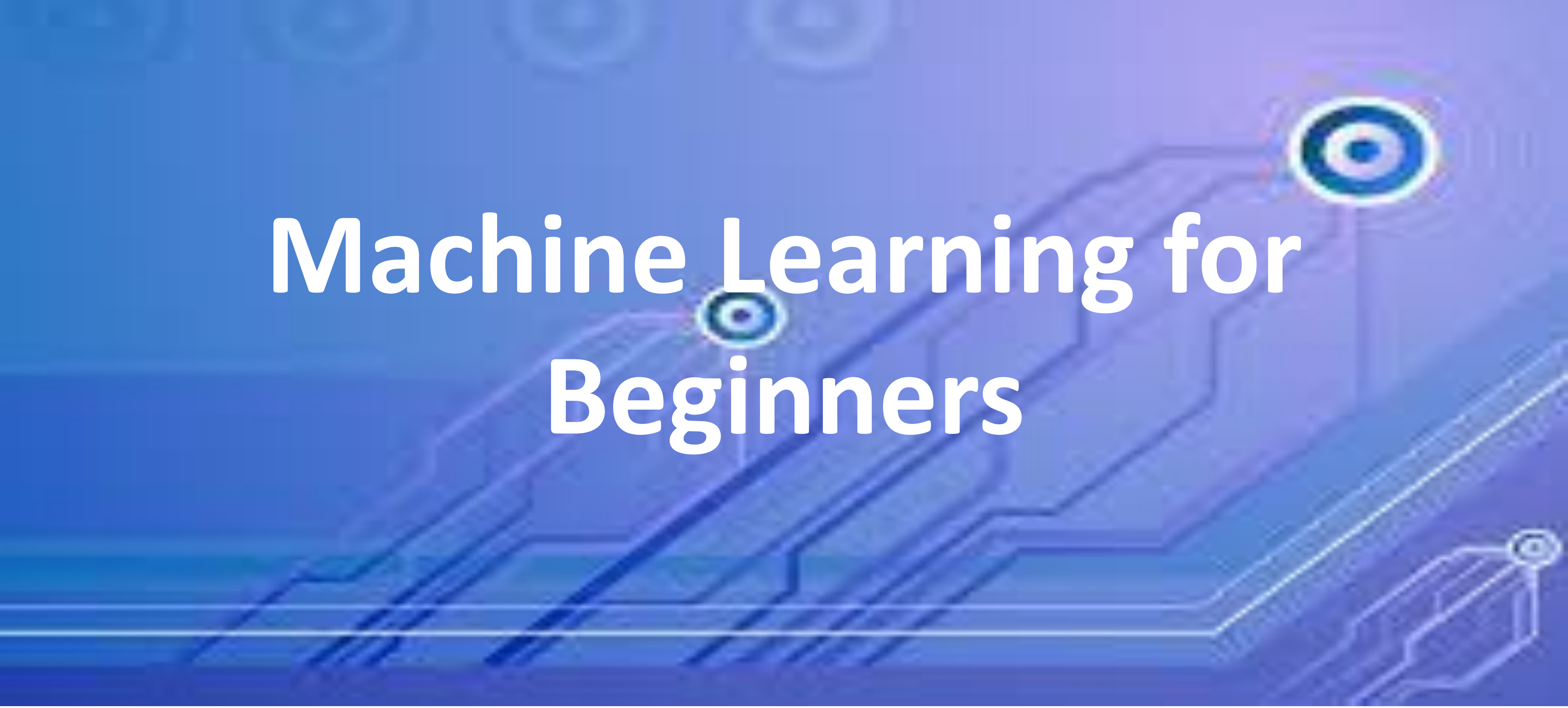
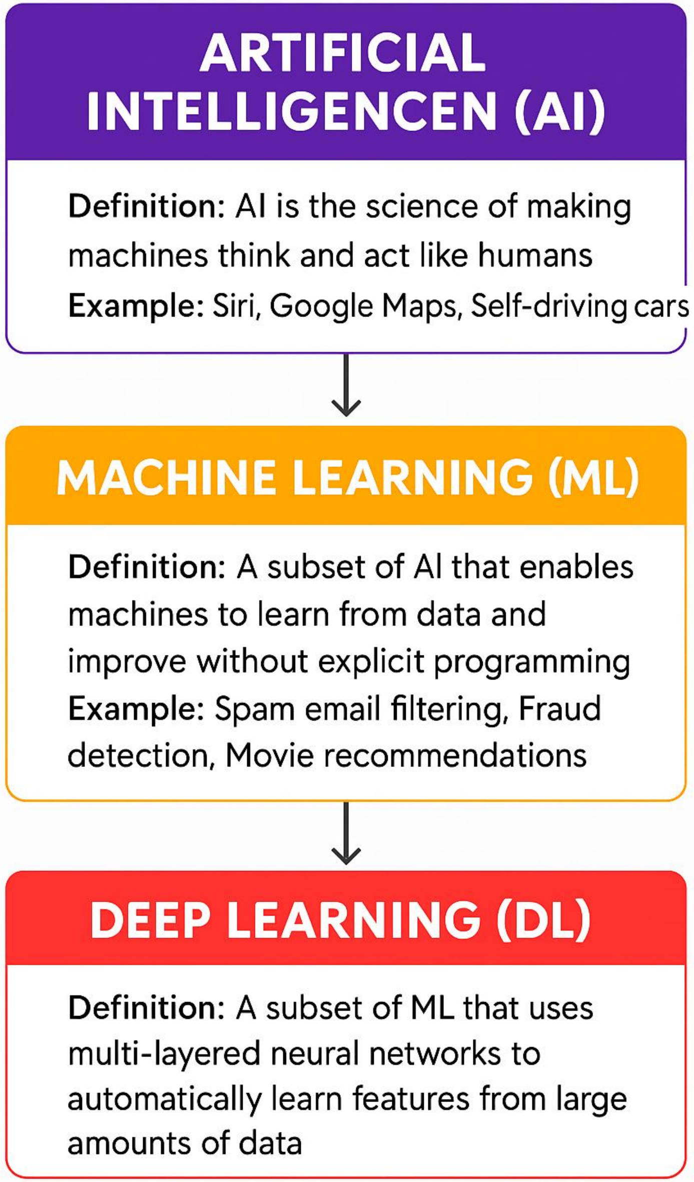


Table of contents

- Overview
- Stages of ML
- Types of ML Algorithm
- Classification
- Regression
- Clustering
- Dimensionality reduction

Overview



ARTIFICIAL INTELLIGENCEN (AI)

Definition: AI is the science of making machines think and act like humans

Example: Siri, Google Maps, Self-driving cars

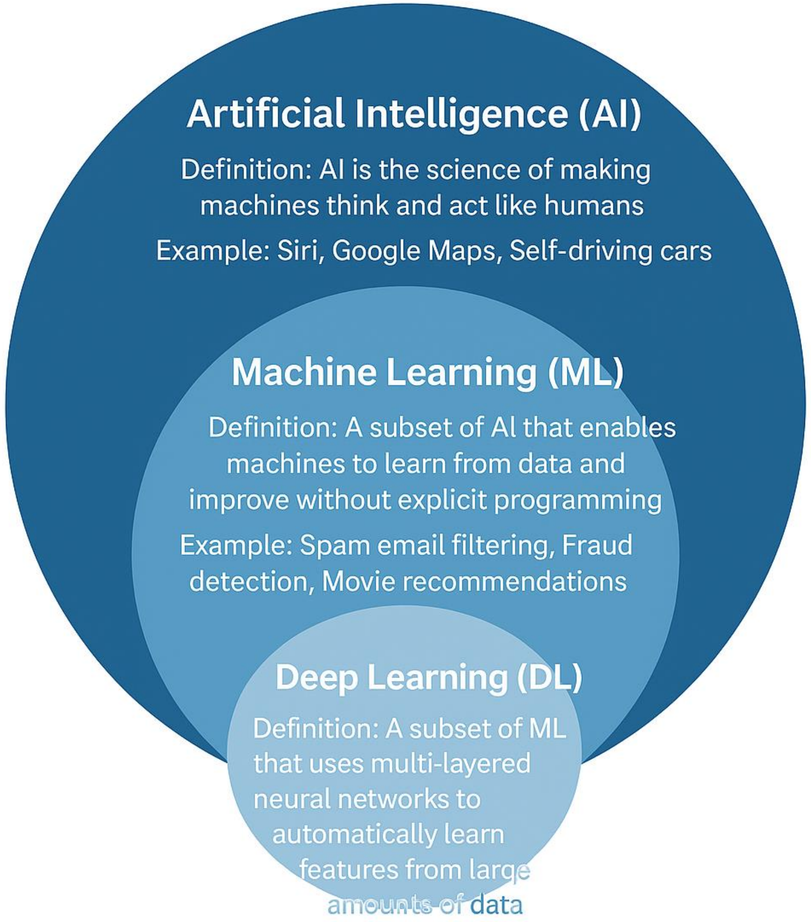
MACHINE LEARNING (ML)

Definition: A subset of AI that enables machines to learn from data and improve without explicit programming

Example: Spam email filtering, Fraud detection, Movie recommendations

DEEP LEARNING (DL)

Definition: A subset of ML that uses multi-layered neural networks to automatically learn features from large amounts of data



Artificial Intelligence (AI)

Definition: AI is the science of making machines think and act like humans

Example: Siri, Google Maps, Self-driving cars

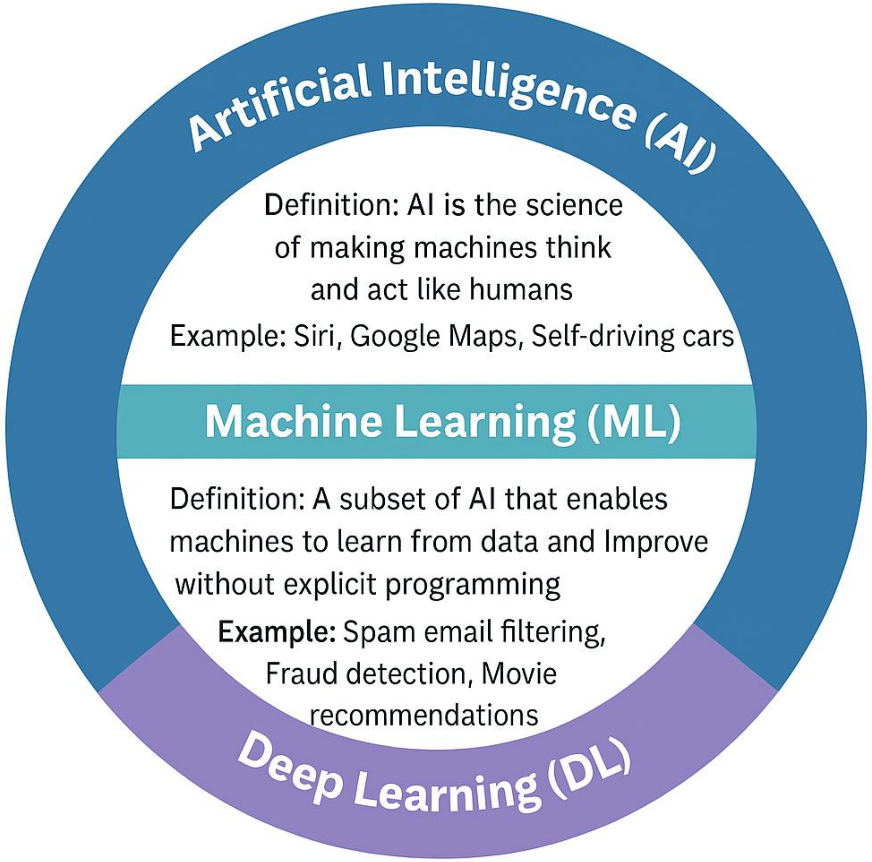
Machine Learning (ML)

Definition: A subset of AI that enables machines to learn from data and improve without explicit programming

Example: Spam email filtering, Fraud detection, Movie recommendations

Deep Learning (DL)

Definition: A subset of ML that uses multi-layered neural networks to automatically learn features from large amounts of data



Artificial Intelligence (AI)

Definition: AI is the science of making machines think and act like humans

Example: Siri, Google Maps, Self-driving cars

Machine Learning (ML)

Definition: A subset of AI that enables machines to learn from data and Improve without explicit programming

Example: Spam email filtering, Fraud detection, Movie recommendations

Deep Learning (DL)

Definition: A subset of ML that uses multi-layered neural networks to automatically learn features from large amounts of data

Example: Image recognition (face ID), Speech-to-text Self-driving perception

Overview

Machine learning (ML) is a field of artificial intelligence (AI) focused on enabling systems to learn from data and improve their performance on specific tasks without explicit programming. Essentially, it empowers computers to identify patterns, make predictions, and improve their accuracy as they are exposed to more data.

Key Aspects:

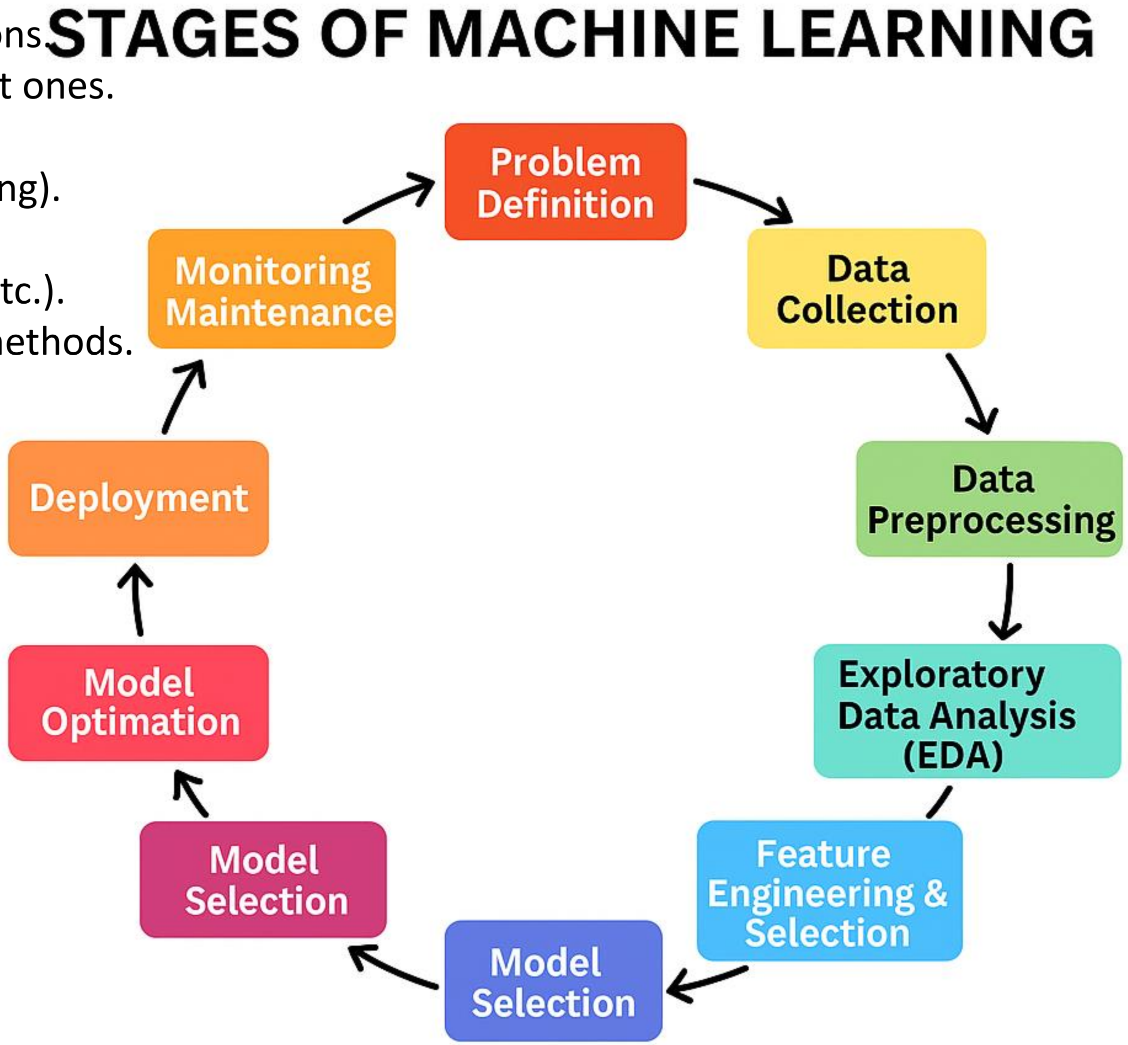
Data-driven:	ML relies heavily on data to train models. The more data available, the better the model's performance is likely to be.
Pattern Recognition:	Algorithms are designed to identify recurring patterns and correlations within the data, which are then used to make predictions or classifications.
Experience-based:	As ML models are exposed to more data, they refine their understanding of patterns and improve their accuracy over time, without requiring constant human intervention.
Subset of AI:	Machine learning is a specific area within the broader field of artificial intelligence, focusing on learning from data to achieve intelligent behavior.

7 stages of ML model development

- Data collection and preparation. ...
- Feature engineering and selection. ...
- Model selection and architecture. ...
- Training and validation. ...
- Model evaluation and testing. ...
- Deployment and integration. ...
- Monitoring and maintenance.

Here are the **12 stages of ML workflow**:

1. **Problem Definition** – Define business objective or research question.
2. **Data Collection** – Gather raw data from databases, APIs, sensors, etc.
3. **Data Pre-processing** – Handle missing values, duplicates, normalization, etc.
4. **Data Exploration (EDA)** – Visualize data, find patterns, distributions, correlations.
5. **Feature Engineering & Selection** – Create new features, select most important ones.
6. **Splitting Data** – Train / Validation / Test sets.
7. **Model Selection** – Choose algorithms (e.g., Regression, Classification, Clustering).
8. **Model Training** – Fit model on training data.
9. **Model Evaluation** – Use metrics (Accuracy, RMSE, Precision, Recall, F1, AUC, etc.).
10. **Model Optimization** – Hyperparameter tuning, cross-validation, ensemble methods.
11. **Deployment** – Put model into production (API, App, Dashboard).
12. **Monitoring & Maintenance** – Track performance drift, retrain with new data



Stage	Description	Key Activities	Tools/Techniques
1. Problem Definition	Identify the business or research problem. Define objectives.	Define goals, success metrics, ML type (classification, regression, clustering, etc.)	Brainstorming, Requirement Analysis
2. Data Collection	Gather raw data from different sources.	Collect structured/unstructured data, APIs, sensors, databases	SQL, APIs, Web Scraping, IoT
3. Data Preprocessing (Data Cleaning)	Prepare raw data for modeling.	Handling missing values, removing duplicates, dealing with outliers, feature engineering	Pandas, NumPy, Excel, ETL tools
4. Data Exploration & Visualization (EDA)	Understand data patterns and relationships.	Statistical analysis, visualization, correlation analysis	Matplotlib, Seaborn, Power BI, Tableau
5. Feature Engineering & Selection	Select/create best input features.	Encoding categorical variables, scaling, feature extraction, dimensionality reduction	PCA, Lasso, Feature Importance
6. Splitting Data	Divide data for training and testing.	Train/Test split, Cross-validation	scikit-learn, K-fold CV
7. Model Selection	Choose suitable ML model.	Select algorithm based on task (classification, regression, clustering, etc.)	Logistic Regression, Decision Trees, Random Forest, Neural Networks
8. Model Training	Train model on training dataset.	Hyperparameter tuning, optimization	scikit-learn, TensorFlow, PyTorch
9. Model Evaluation	Test model on unseen data.	Accuracy, Precision, Recall, F1-score, RMSE, AUC-ROC	scikit-learn, confusion matrix, metrics
10. Model Optimization & Tuning	Improve model performance.	Hyperparameter tuning, Regularization, Ensemble methods	GridSearchCV, RandomizedSearch, Bayesian Optimization
11. Deployment	Integrate model into production system.	Model serving, APIs, cloud deployment	Flask/Django (APIs), FastAPI, AWS, Azure, GCP
12. Monitoring & Maintenance	Track performance post-deployment.	Monitor drift, retrain with new data	MLflow, Prometheus, CI/CD pipelines

Data types

In machine learning, data is typically divided into distinct subsets for specific purposes in model development and evaluation. The two primary subsets are training data and testing data.

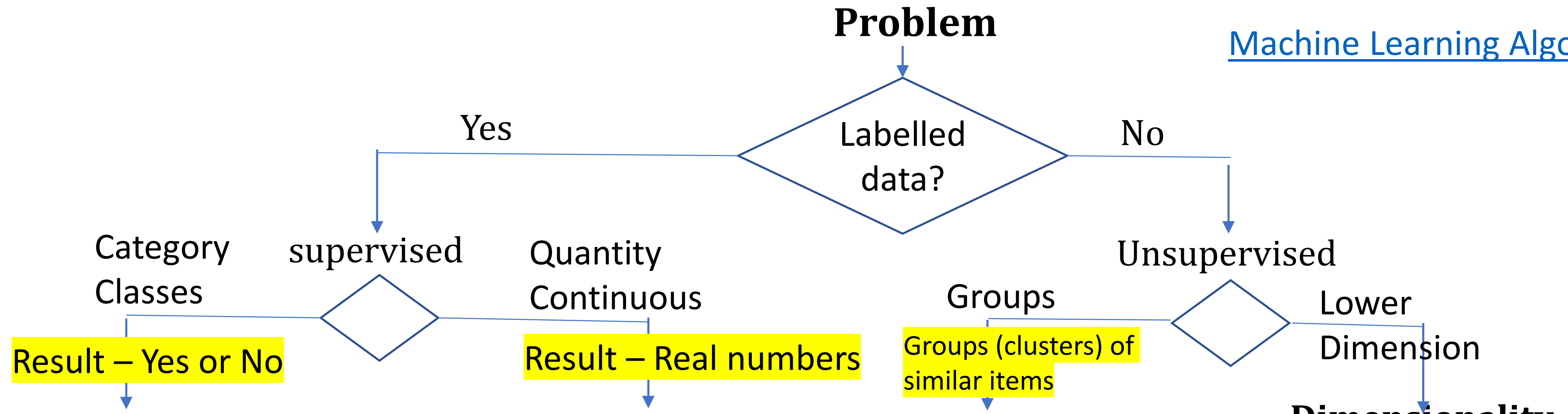
	Train Data	Test data
Purpose:	The training data is used to "teach" the machine learning model. It consists of input features and their corresponding known output labels or values.	The testing data is used to evaluate the performance of the trained model on unseen data. It assesses how well the model generalizes to new examples it has not encountered during the training phase.
Usage:	The model analyzes patterns, relationships, and trends within this data to learn how to map inputs to outputs. During training, the model adjusts its internal parameters and weights to minimize errors and improve its predictive accuracy on the training examples.	After the model has been trained, it is presented with the testing data (which also contains input features and known output labels). The model makes predictions on this data, and these predictions are then compared to the actual labels to measure the model's accuracy, precision, recall, and other performance metrics.
Characteristics:	It typically comprises the larger portion of the overall dataset to provide a comprehensive learning experience for the model.	It is a separate and distinct subset from the training data, ensuring an unbiased evaluation of the model's ability to perform on real-world, novel data. It is typically a smaller portion of the overall dataset.

Key Differences Summarized:

- **Role:** Training data is for learning, while testing data is for evaluation.
- **Exposure:** The model sees and learns from training data; it only predicts on testing data.
- **Goal:** Training aims to optimize model parameters; testing aims to assess generalization and performance.
- **Separation:** It is crucial to keep training and testing data separate to ensure an accurate and unbiased assessment of the model's performance on new data.

Types Algorithm

Machine Learning Algorithms - Python Geeks



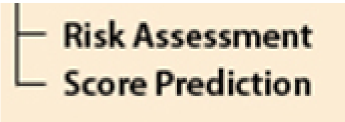
Classification

Logistic Regression:
Decision Trees:
Random Forest:
Support Vector Machines (SVMs)
K-Nearest Neighbors (KNN)
Naive Bayes:

- Fraud detection
- Email Spam Detection
- Diagnostics
- Image Classification

Regression

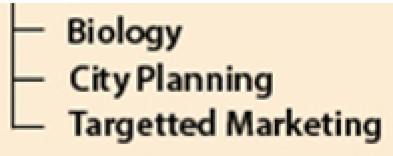
Linear Regression:
Polynomial Regression:
Ridge Regression:
Lasso Regression:
Ensemble methods
Elastic Net Regression:
Decision Tree Regression:
Random Forest Regression:
Support Vector Regression (SVR):
K-Nearest Neighbors (KNN) Regression:



- Risk Assessment
- Score Prediction

Clustering

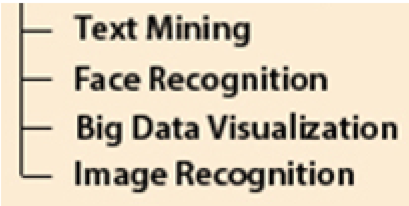
K-means
DBSCAN
Special clustering
Hierarchical clustering
Gaussian mixture
Hidden Markav model



- Biology
- City Planning
- Targetted Marketing

Dimensionality reduction

PCA
LDA
Isomap
Autoencoder
Policy grading
Trust region policy
Hindsight

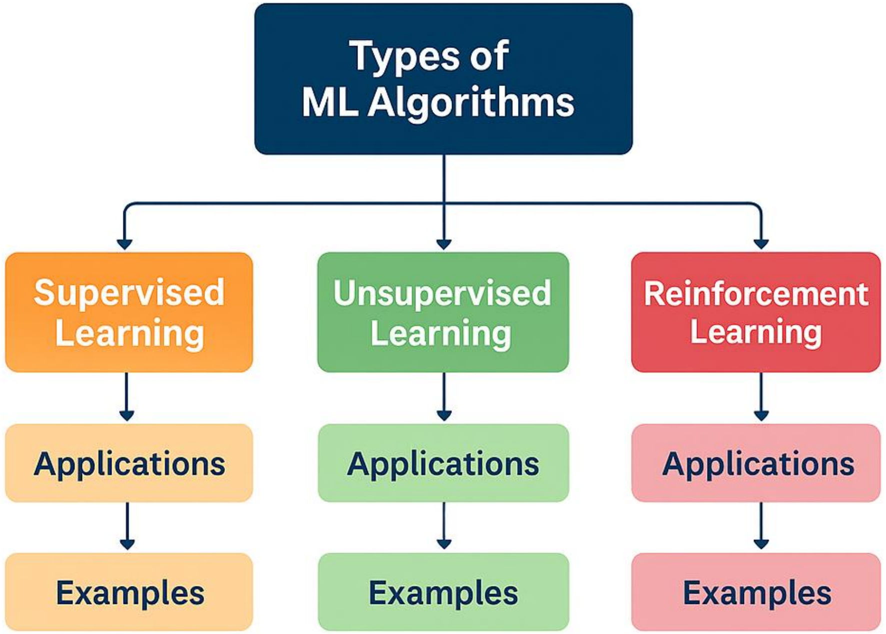


- Text Mining
- Face Recognition
- Big Data Visualization
- Image Recognition

Reinforcement Learning

- Gaming
- Finance Sector
- Manufacturing
- Inventory Management
- Robot Navigation

Types of ML Algorithms



Types of ML Algorithms

Supervised Learning

Applications

Examples

Unsupervised Learning

Applications

Examples

Reinforcement Learning

Applications

Examples

- Image classification
- Spam detection
- Sales forecasting

- Linear regression
- Decision trees
- Support vector machines

- Anomaly detection
- Customer segmentation
- Market basket analysis

- K-means clustering
- Hierarchical clustering
- Principal component analysis

- Game playing
- Robotics
- Autonomous vehicles

- Q-Learning
- Deep Q-Networks
- Policy gradients

Classification

Classification is a **supervised machine learning task** where the goal is to **predict a categorical label (class)** for given input data.

- Input: Features (numerical, text, image, etc.)
- Output: Discrete class/category (e.g., "spam" or "not spam").

👉 Example:

- Email filtering → classify emails into **Spam** or **Not Spam**.
- Medical diagnosis → classify a tumor as **Benign** or **Malignant**

Algorithms Used for Classification

Some commonly used ML algorithms for classification are:

- **Logistic Regression** → For binary/multi-class classification.
- **Decision Trees** → Easy to interpret.
- **Random Forest** → Ensemble method for higher accuracy.
- **Support Vector Machine (SVM)** → Works well with high-dimensional data.
- **Naïve Bayes** → Good for text classification (spam, sentiment).
- **k-Nearest Neighbors (kNN)** → Instance-based classification.
- **Neural Networks (Deep Learning)** → For complex problems (image, speech, NLP).

◈ **Applications of Classification in Real Life**

1. **Healthcare** → Disease prediction, cancer detection.
2. **Finance** → Credit scoring, fraud detection.
3. **E-commerce** → Product recommendation, customer segmentation.
4. **Email Filtering** → Spam vs. Non-Spam.
5. **Image Recognition** → Face recognition, object detection.
6. **Natural Language Processing (NLP)** → Sentiment analysis, chatbot intent classification.
7. **Cybersecurity** → Malware detection, intrusion detection.

In short:

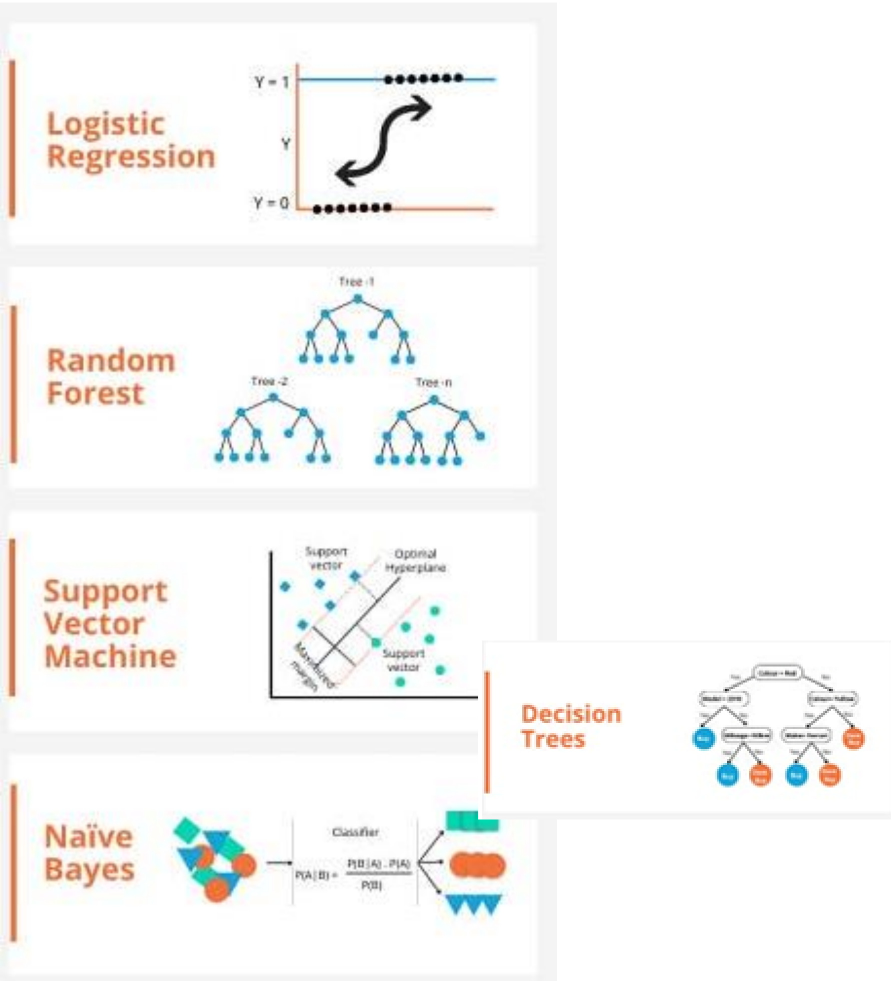
Classification = Predicting labels from data.

Types = Binary, Multi-class, Multi-label, Imbalanced, Ordinal.

Applications = From spam filters to medical diagnosis & fraud detection.

Classification

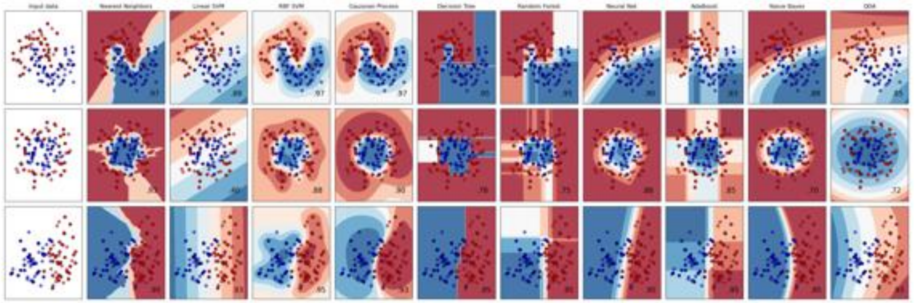
Sl no	Types	Description	Syntex
1	**Logistic Regression:**	A linear algorithm used primarily for binary classification, though it can be extended for multi-class problems. It models the probability of a binary outcome.	
2	**Decision Trees:**	Non-linear algorithms that recursively partition the data based on features, creating a tree-like structure where each leaf node represents a class label.	
3	**Random Forest:**	An ensemble method that builds multiple decision trees and combines their predictions to improve accuracy and reduce overfitting.	
4	**Support Vector Machines (SVM):**	Algorithms that find an optimal hyperplane to separate data points into different classes, effective in high-dimensional spaces.	
5	**Naive Bayes:**	A family of probabilistic algorithms based on Bayes' theorem, assuming independence between features.	
6	**K-Nearest Neighbors (KNN):**	A non-parametric, instance-based algorithm that classifies a data point based on the majority class among its 'k' nearest neighbors in the feature space.	
7	**Artificial Neural Networks (ANNs):**	Inspired by the human brain, these algorithms consist of interconnected nodes (neurons) organized in layers, capable of learning complex patterns.	
8	**Gradient Boosting Algorithms (e.g., AdaBoost, XGBoost, LightGBM):**	Ensemble methods that build a strong learner by sequentially adding weak learners, with each new learner correcting the errors of the previous ones.	



Classification ; Identifying which category an object belongs to.

Applications: Spam detection, image recognition.

Algorithms: Gradient boosting, nearest neighbors, random forest, logistic regression, and more...



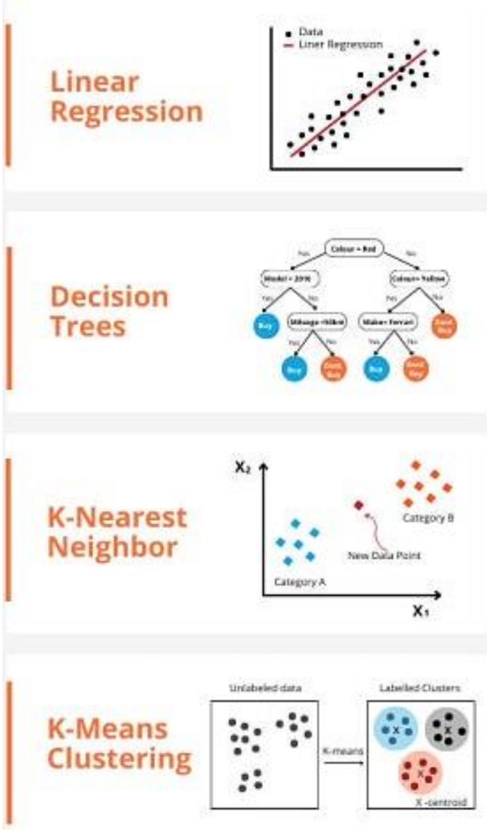
Classification Algorithm in Machine Learning - Types & Examples

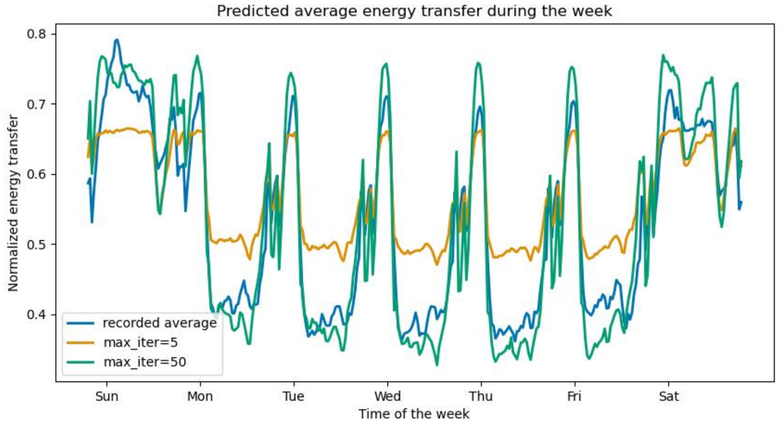
Classification

Type	Common Algorithms	Examples	Applications
Binary Classification	Logistic Regression, Support Vector Machine (SVM), Decision Trees, Random Forest, Naïve Bayes	Spam vs. Not Spam, Fraud vs. Genuine	Email filtering, Fraud detection, Medical diagnosis
Multi-class Classification	k-Nearest Neighbors (kNN), Random Forest, Neural Networks, Decision Trees, Logistic Regression (One-vs-Rest)	Digit recognition (0–9), Fruit classification	Handwriting recognition, Image classification
Multi-label Classification	Binary Relevance, Classifier Chains, Neural Networks, SVM	Movie genres (Comedy + Drama), News topics	Tagging movies/music, Document categorization
Imbalanced Classification	Random Forest with class weights, XGBoost, SMOTE + Logistic Regression, Ensemble Methods	Fraud detection, Rare disease detection	Anomaly detection, Healthcare, Cybersecurity
Ordinal Classification	Ordinal Logistic Regression, Decision Trees, Gradient Boosting	Movie ratings (Poor < Good < Excellent), Education level	Customer satisfaction surveys, Credit scoring

Regression:

Sl no	Types	Description	Syntex
1	**Linear Regression:**	Models the linear relationship between a dependent variable and one or more independent variables.	
2	**Polynomial Regression:**	Models the relationship between variables as an nth-degree polynomial, allowing for non-linear relationships.	
3	**Ridge Regression:**	A regularization technique that adds an L2 penalty to linear regression, helping to prevent overfitting and handle multicollinearity.	
4	**Lasso Regression:**	Another regularization technique that adds an L1 penalty, which can lead to sparse models by shrinking some coefficients to zero, effectively performing feature selection.	
5	**Elastic Net Regression:**	Combines both L1 (Lasso) and L2 (Ridge) regularization, offering the benefits of both.	
6	**Decision Tree Regression:**	Uses a tree-like structure to make predictions by recursively partitioning the data based on feature values.	
7	**Random Forest Regression:**	An ensemble method that builds multiple decision trees and averages their predictions to improve accuracy and reduce variance.	
8	**Support Vector Regression (SVR):**	An extension of Support Vector Machines (SVMs) for regression tasks, aiming to find a hyperplane that best fits the data within a specified margin.	
9	**Gradient Boosting Regressors (e.g., XGBoost, LightGBM, CatBoost):**	Ensemble methods that build a sequence of weak prediction models, typically decision trees, where each new model corrects the errors of the previous ones.	
10	**Bayesian Linear Regression:**	Incorporates prior knowledge about the model parameters into the regression analysis, providing a probabilistic approach to linear regression.	
11	**Quantile Regression:**	Focuses on modeling the conditional median or other quantiles of the dependent variable, rather than just the conditional mean.	
12	**Principal Components Regression (PCR):**	Uses principal component analysis (PCA) to reduce the dimensionality of the independent variables before performing linear regression.	





Regression : Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, stock prices.

Algorithms: Gradient boosting, nearest neighbors, random forest, ridge, and more...

Regression

Regression is a **supervised machine learning technique** used to **predict continuous numerical values** based on input features.

•Input: Features (numbers, categorical variables, etc.)

•Output: Continuous value (real number).

👉 Example: Predicting **house price**, **stock price**, or **temperature**.

Common Algorithms

•**Linear Models**: Linear, Multiple, Polynomial

•**Regularized Models**: Ridge, Lasso, ElasticNet

•**Tree-based Models**: Decision Trees, Random Forest, Gradient Boosting (XGBoost, LightGBM, CatBoost)

•**Kernel-based Models**: Support Vector Regression

•**Deep Learning Models**: Neural Networks

Applications of Regression

1. **Finance** → Stock price prediction, credit scoring.
2. **Real Estate** → House price prediction.
3. **Healthcare** → Predicting disease progression (blood sugar levels, tumor growth).
4. **Economics** → Predicting GDP, inflation rates.
5. **Business** → Sales forecasting, demand prediction.
6. **Engineering** → Predicting equipment failure, energy consumption.
7. **Weather** → Temperature, rainfall prediction.

Type	**Description**	**Example**
Linear Regression	Models relationship between independent variable(s) and dependent variable as a straight line.	Predicting house prices from area & location.
Multiple Linear Regression	Uses more than one independent variable to predict output.	Predicting salary based on education, age, experience.
Polynomial Regression	Captures non-linear relationships using polynomial terms.	Predicting growth curves, population growth.
Ridge Regression (L2 Regularization)	Linear regression with penalty to reduce overfitting.	Predicting stock prices with many features.
Lasso Regression (L1 Regularization)	Similar to Ridge but also does feature selection.	Selecting important medical factors for disease prediction.
ElasticNet Regression	Combination of Lasso + Ridge.	Predicting product demand with many correlated variables.
Logistic Regression (technically classification, but often grouped)	Predicts probability of categorical outcome (0/1).	Spam / Not Spam, Fraud detection.
Quantile Regression	Estimates conditional quantiles (like median).	Predicting 90th percentile of house prices.
Support Vector Regression (SVR)	Uses Support Vector Machine for regression.	Predicting electricity consumption.
Decision Tree / Random Forest Regression	Tree-based regression for non-linear relationships.	Predicting sales, crop yield.
Neural Network Regression	Deep learning-based regression for complex patterns.	Predicting stock prices, weather forecasting.

Types:

Type of Regression	Common Algorithms	Examples	Applications
Linear Regression	Linear Regression (OLS)	Predicting house price based on size	Real estate price prediction
Multiple Linear Regression	Multiple Linear Regression	Salary prediction (education + experience + age)	HR analytics, business forecasting
Polynomial Regression	Polynomial Regression	Population growth curve	Trend analysis, growth prediction
Ridge Regression (L2)	Ridge Regression	Stock price with many features	Finance, stock market
Lasso Regression (L1)	Lasso Regression	Selecting key medical predictors	Healthcare, genetics
ElasticNet Regression	ElasticNet	Product demand prediction	Retail, supply chain
Quantile Regression	Quantile Regression	Median/percentile of house price	Risk analysis, economics
Support Vector Regression (SVR)	SVR (linear, polynomial, RBF kernels)	Electricity consumption prediction	Energy forecasting
Decision Tree Regression	Decision Tree, CART	Sales prediction	Business analytics
Random Forest Regression	Random Forest	Crop yield prediction	Agriculture, food industry
Gradient Boosting Regression	XGBoost, LightGBM, CatBoost	Insurance claim cost prediction	Insurance, healthcare
Neural Network Regression	Deep Learning (ANN, RNN, LSTM)	Stock price prediction, weather forecasting	Finance, meteorology

Clustering

Clustering is an **unsupervised machine learning technique** used to **group similar data points together** without predefined labels.

•Output: Groups (clusters) of similar items.

•Example: Grouping customers by purchasing behavior, grouping documents by topic.

Common Algorithms

Partitioning → K-Means, K-Medoids (PAM)
Hierarchical → Agglomerative, Divisive
Density-based → DBSCAN, OPTICS
Grid-based → STING, CLIQUE
Fuzzy → Fuzzy C-Means
Model-based → Gaussian Mixture Models (GMM), Expectation-Maximization

Applications of Clustering

Customer Segmentation → Grouping customers by behavior.
Image Segmentation → Dividing images into meaningful regions.
Market Research → Identifying buyer personas.
Healthcare → Grouping patients by symptoms or genetic similarity.
Anomaly Detection → Identifying outliers in fraud or network intrusion.
Search Engines → Document/topic clustering.
Social Network Analysis → Community detection.

Type of Clustering	Description	Example
Partitioning Clustering	Divides data into non-overlapping subsets (clusters).	K-Means clustering for customer segmentation
Hierarchical Clustering	Builds a tree (dendrogram) of clusters.	Gene sequence analysis
Density-based Clustering	Groups dense regions and marks sparse points as noise.	DBSCAN for anomaly detection
Grid-based Clustering	Divides data into grid cells and clusters cells.	STING for spatial data
Fuzzy Clustering	A point can belong to multiple clusters with probabilities.	Fuzzy C-Means for market segmentation
Model-based Clustering	Assumes data is generated from a mixture of distributions.	Gaussian Mixture Models (GMM)

Clustering

Type of Clustering	**Common Algorithms**	**Examples**	**Applications**
Partitioning	K-Means, K-Medoids	Customer groups by spending	Customer segmentation
Hierarchical	Agglomerative, Divisive	Gene sequence tree	Bioinformatics, taxonomy
Density-based	DBSCAN, OPTICS	Outlier detection	Fraud detection, anomaly detection
Grid-based	STING, CLIQUE	Spatial data grouping	Geographic information systems
Fuzzy Clustering	Fuzzy C-Means	Overlapping customer profiles	Market segmentation
Model-based	Gaussian Mixture Models (GMM), EM Algorithm	Mixed distribution modeling	Pattern recognition, speech processing

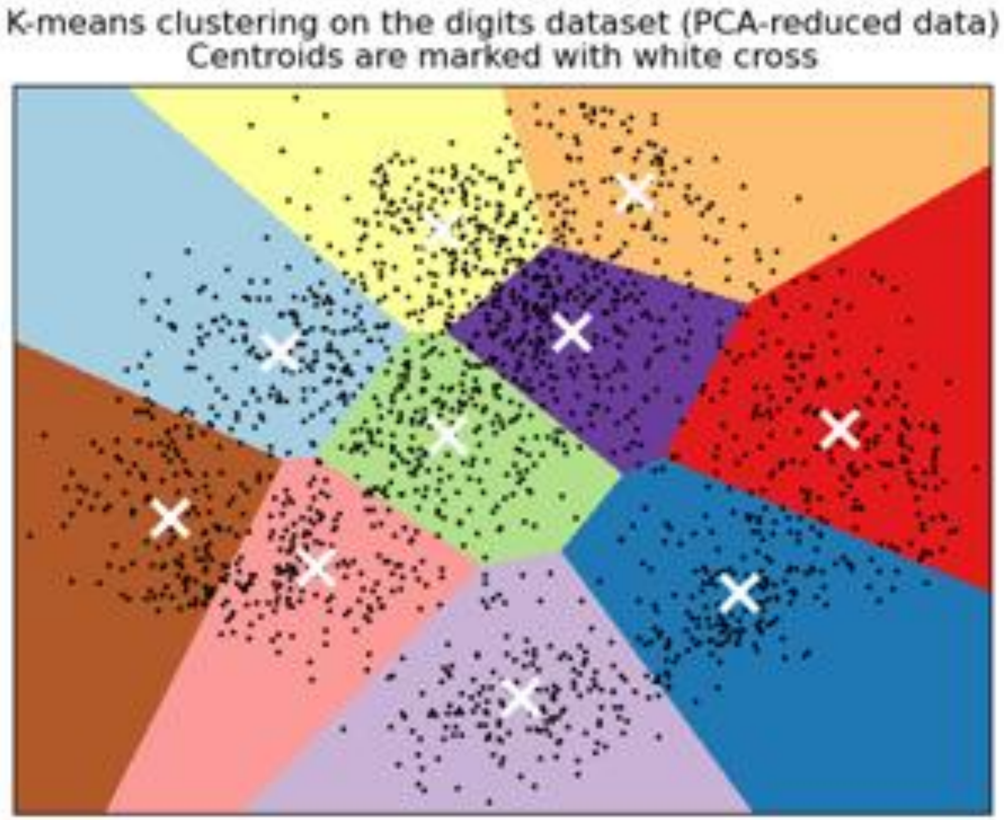
Clustering:

Sl no	Types	Description	Syntex
1	**K-Means Clustering:**	A centroid-based algorithm that partitions data into a predefined number of k clusters, where each data point belongs to the cluster with the nearest mean (centroid).	
2	**K-Medoids Clustering:**	Similar to K-Means, but uses actual data points (medoids) as cluster centers instead of means, making it more robust to outliers.	
3	**Hierarchical Clustering:**	Builds a hierarchy of clusters, either by starting with individual data points and merging them (agglomerative) or by starting with a single cluster and recursively splitting it (divisive).	
4	**DBSCAN (Density-Based Spatial Clustering of Applications with Noise):**	A density-based algorithm that identifies clusters based on the density of data points and can discover clusters of arbitrary shapes while identifying outliers.	
5	**OPTICS (Ordering Points to Identify the Clustering Structure):**	An extension of DBSCAN that builds a reachability plot, allowing for the identification of clusters with varying densities.	
6	**HDBSCAN (Hierarchical DBSCAN):**	A hierarchical version of DBSCAN that can find clusters of varying densities and is less sensitive to parameter choices.	
7	**Mean-Shift Clustering:**	A non-parametric, density-based algorithm that seeks modes (peaks) in the data density to identify clusters.	
8	**Affinity Propagation:**	A graph-based algorithm that does not require a pre-defined number of clusters and identifies "exemplars" as cluster representatives.	
9	**Gaussian Mixture Models (GMM):**	A probabilistic model that assumes data points are generated from a mixture of Gaussian distributions, allowing for soft assignments of data points to clusters.	
10	**Spectral Clustering:**	Uses the eigenvalues of a similarity matrix to perform dimensionality reduction before clustering, often with K-Means on the lower-dimensional representation.	
11	**BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies):**	Designed for large datasets, it builds a CF-tree (Clustering Feature Tree) to summarize data and then applies a clustering algorithm.	
12	**Mini-Batch K-Means:**	A variant of K-Means that uses mini-batches of data to update cluster centroids, making it more efficient for large datasets.	

Clustering: Automatic grouping of similar objects into sets.

Applications: Customer segmentation, grouping experiment outcomes.

Algorithms: k-Means, HDBSCAN, hierarchical clustering, and more...



Dimensionality Reduction

What is Dimensionality Reduction?

Dimensionality Reduction is an **unsupervised ML technique** used to **reduce the number of input features** while retaining important information.

- Input: High-dimensional data
- Output: Lower-dimensional representation (fewer features)

👉 Why?

- Reduce computation cost
- Remove noise & redundancy
- Avoid **curse of dimensionality**
- Improve visualization
- (e.g., compress data to 2D/3D)

- **Linear Methods**:
- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)
- Singular Value Decomposition (SVD)
- **Non-linear (Manifold Learning)**:
- t-SNE (t-distributed Stochastic Neighbor Embedding)
- UMAP (Uniform Manifold Approximation and Projection)
- Isomap
- **Feature Selection Methods**:
- Filter methods (Chi-square test, ANOVA, Information Gain)
- Wrapper methods (RFE – Recursive Feature Elimination)
- Embedded methods (Lasso, Decision Trees feature importance)
- **Deep Learning-based**:
- Autoencoders

Type	Description	Example
Feature Selection	Selects the most important features and discards the rest.	Removing irrelevant survey questions
Feature Extraction	Creates new features by transforming original ones.	PCA converts correlated features into uncorrelated principal components
Linear Methods	Assume linear relationships in data.	PCA, LDA
Non-linear Methods (Manifold Learning)	Preserve non-linear structures in data.	t-SNE, UMAP
Supervised Methods	Use class labels to reduce dimensions.	Linear Discriminant Analysis (LDA)
Unsupervised Methods	Do not use class labels.	PCA, Autoencoders

Dimensionality reduction

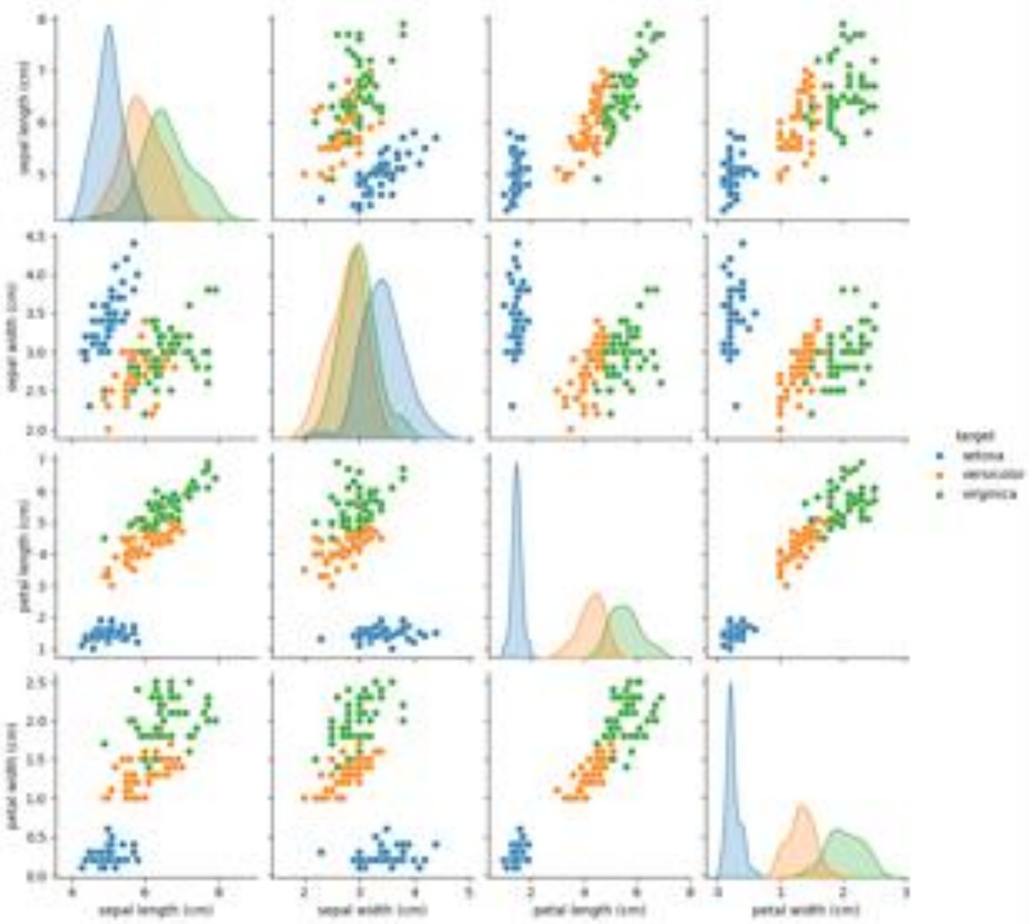
Sl no	Types	Description	Syntex
1	**K-Means Clustering:**	A centroid-based algorithm that partitions data into a predefined number of k clusters, where each data point belongs to the cluster with the nearest mean (centroid).	
2	**K-Medoids Clustering:**	Similar to K-Means, but uses actual data points (medoids) as cluster centers instead of means, making it more robust to outliers.	
3	**Hierarchical Clustering:**	Builds a hierarchy of clusters, either by starting with individual data points and merging them (agglomerative) or by starting with a single cluster and recursively splitting it (divisive).	
4	**DBSCAN (Density-Based Spatial Clustering of Applications with Noise):**	A density-based algorithm that identifies clusters based on the density of data points and can discover clusters of arbitrary shapes while identifying outliers.	
5	**OPTICS (Ordering Points to Identify the Clustering Structure):**	An extension of DBSCAN that builds a reachability plot, allowing for the identification of clusters with varying densities.	
6	**HDBSCAN (Hierarchical DBSCAN):**	A hierarchical version of DBSCAN that can find clusters of varying densities and is less sensitive to parameter choices.	
7	**Mean-Shift Clustering:**	A non-parametric, density-based algorithm that seeks modes (peaks) in the data density to identify clusters.	
8	**Affinity Propagation:**	A graph-based algorithm that does not require a pre-defined number of clusters and identifies "exemplars" as cluster representatives.	
9	**Gaussian Mixture Models (GMM):**	A probabilistic model that assumes data points are generated from a mixture of Gaussian distributions, allowing for soft assignments of data points to clusters.	
10	**Spectral Clustering:**	Uses the eigenvalues of a similarity matrix to perform dimensionality reduction before clustering, often with K-Means on the lower-dimensional representation.	
11	**BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies):**	Designed for large datasets, it builds a CF-tree (Clustering Feature Tree) to summarize data and then applies a clustering algorithm.	
12	**Mini-Batch K-Means:**	A variant of K-Means that uses mini-batches of data to update cluster centroids, making it more efficient for large datasets	

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, increased efficiency.

Algorithms: PCA, feature selection, non-negative matrix factorization, and more...



Commonly used machine learning algorithms

1. Linear regression

Consider x variables and y variables. The independent variable is on the x-axis, and the dependent variable, y, is on the y-axis. We try to form a relation between these two variables and draw a straight line.

As the independent variable changes on this line, the dependent variable either goes up or down accordingly.

Suppose the independent variable increases with an increase in the dependent variable. In that case, there is said to be a positive relationship. On the other hand, if the dependent variable decreases with an increase, the variables have a negative relationship.

2. Logistic regression

Logistic regression is a special case of regression analysis. It is calculated when the dependent variable is nominal or ordinally scaled.

Dichotomous variables (0 or 1) can be predicted using logistic regression.

The probability of occurrence of a characteristic (=1 character is present) is estimated.

For example, a common goal in medicine is determining which variables impact the disease.

In this case, 0 could stand for “not disease” and 1 for “disease”, and the influence of age, gender and smoking status on this particular disease is estimated.

The logistic model is based on the logistic function. The important thing about the logistic function is that only values between 0 and 1 are entertained.

3. Decision trees

Decision trees are a type of supervised machine-learning algorithm we use for classification problems. It can operate on both continuous and categorical variables.

The population in the decision trees is divided into two or more homogeneous sets.

In the above picture, we decide whether the child should play based on multiple attributes. First, we have the outlook attributes: sunny, overcast and rainy.

4. SVM- support vector machine

SVM is a classification method. Each object you want to classify is represented as a point in an n-dimensional space. The coordinates of this point are usually called features.

SVMs perform the classification test by drawing a hyperplane line in a 2D plane or a 3D plane so that all points of one category are on one of the sides of the hyperplane. All points of the other category are on one of the sides of the hyperplane, and all points of the other category are on the other side, while there could be multiple such hyperplanes.

The name support vector classifier comes from the fact that the observation on edge and within the soft margin are called support vectors.

5. Naive Bayes

Naive Bayes is another classification technique based on the Bayes theorem. It assumes independence among features. We can make a simplifying assumption that the elements of the feature vector are conditionally independent of each other, given the classification.

This is a great simplification over evaluating the full probability, so it might be surprising that the naive Bayes classifier has shown comparable results to other classification methods in certain domains.

Types:

6. KNN- K- nearest neighbours

The idea behind K- nearest Neighbours (KNN) is very simple. For each record to be classified or predicted:

•Find K records that have similar features.

•For classification, find out the majority of issues among similar records and assign that class to the new record.

•For prediction, find the average among those similar records, and predict that average for the new record.

7. K-means

Clustering is a technique to divide data into different groups where the records in each group are similar. The goal of clustering is to identify meaningful groups of data. The groups can be used directly, analysed in more depth, or passed as a feature or an outcome of a predictive regression and classification model.

K- means was the first clustering method to be developed. It is still widely used owing to its popularity, the relative simplicity of the algorithm, and its ability to scale to large datasets.

8. Dimensionality Reduction Algorithms

PCA is a technique to find how numeric variables covary. Covary means when they vary together. Some variations in one variable are caused by variations in another—for example, restaurant checks and tips.

It helps you reduce the number of dimensions into other lower number dimensions. Then we can apply machine learning algorithms.

As for the number of dimensions, it is considered a curse since it directly impacts the accuracy.

9. Random forest

A random forest refers to a collection of multiple decision trees and is much less sensitive to the training data.

We use multiple trees, and hence it has the name forest.

Process of creating a random forest:

The first step is to build new datasets from our original data. Then, we randomly select rows from the original data to build the ne dataset.

Here we perform random sampling with replacement. After selecting a row, we are putting it back into the data.

The process that we just followed to create new data is called Bootstrapping.

First, we train a decision tree on each of the datasets separately.

We randomly select a subset of the features for each tree and use only them for training.

Make predictions

We pass this new data point through each tree and note down the prediction. Finally, we combine all the predictions, and the majority voting is taken.

This process of combining multiple results is called bagging.

10. Gradient Boosting Algorithms

It is another boosting technique. The learning happens with the help of optimising the loss function. These use two types of base estimators first is the average type model, and second is the decision tree in full depth.

It is used for classification and regression.

Data set available link

[Find Open Datasets and Machine Learning Projects | Kaggle]

[Home | Open Government Data (OGD) Platform India]

[Untitled0.ipynb - Colab]

Thankyou