Machine Learning

Methodology & Basics

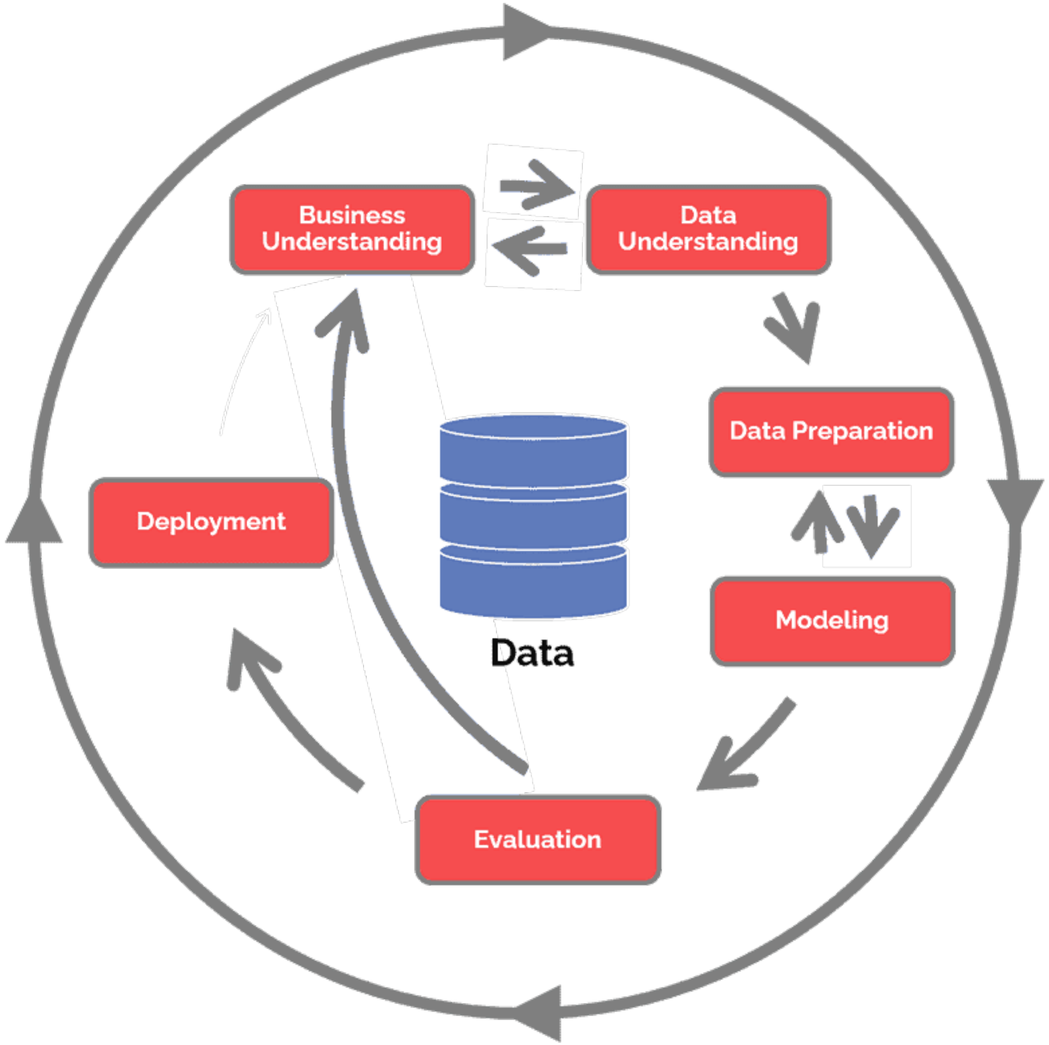
Rohan

5 Steps

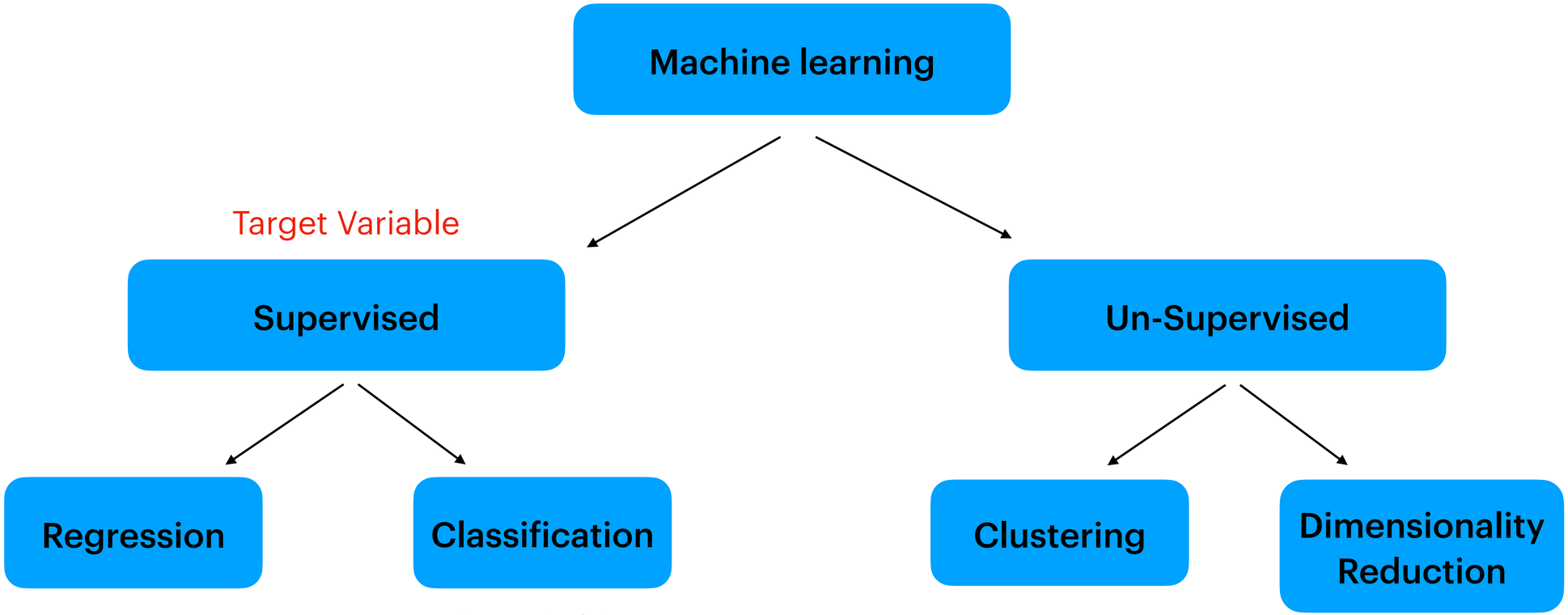
Data Extraction	Web Data, CSV, Excel Sheet etc...
Data Analysis	Pattern & Trend Analysis
Feature Engineering	Handling Missing Values, Unknowns etc ...
Model Building	Model Building based on use case
Model Evaluation	Model Performance Metrics like RMSE,MSE etc...

CRISP - DM

Cross Industry Standard Process for Data Mining



5 Steps


Machine learning
Target Variable
Supervised
Un-Supervised
Regression
Classification
Clustering
Dimensionality Reduction
- numerical Data
- Categorical Data
- Yes/No, -> Binary
- Multi- Class

Features and Target Variable

Examples

Supervised Learning

X_1	X_2	X_3	X_p	Y

Target

Un-Supervised Learning

X_1	X_2	X_3	X_p	Y

No Target

Training data				
var1	var2	var3	var4	result
5	2	2	1	dead
5	1	2	2	allve
3	3	1	2	dead
5	3	2	1	alive
4	3	2	2	alive

Numerical feature

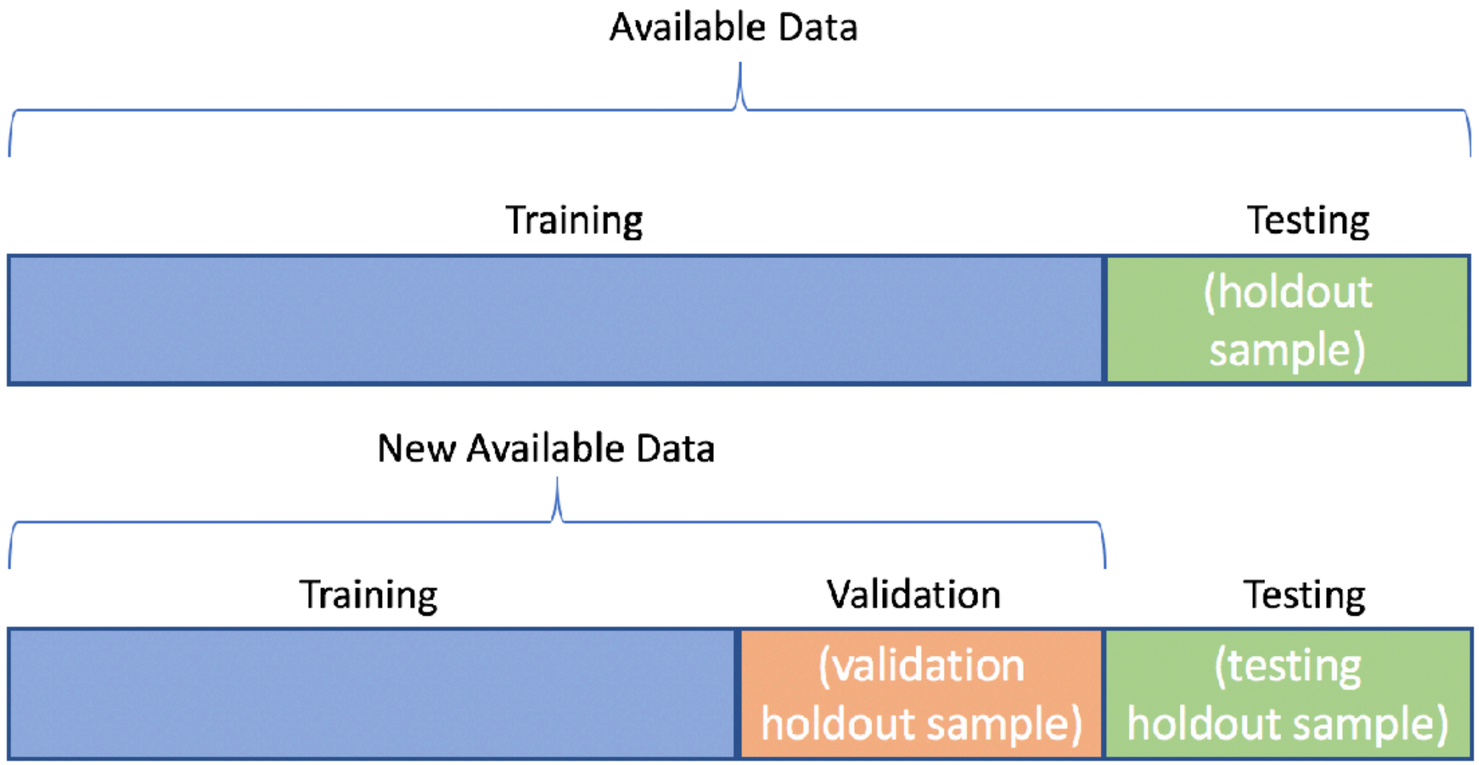
Categorical features

Target variable

	hour_min	day_of_week	ph_eve	parking_zone
0	20.6	Mon	nil	Zone 1
1	9.2	Tue	nil	Zone 1
2	20.6	Tue	nil	Zone 1
3	8.9	Wed	nil	Zone 1
4	18.7	Wed	nil	Zone 1

Data For Model Building

Train - Test Split



Python

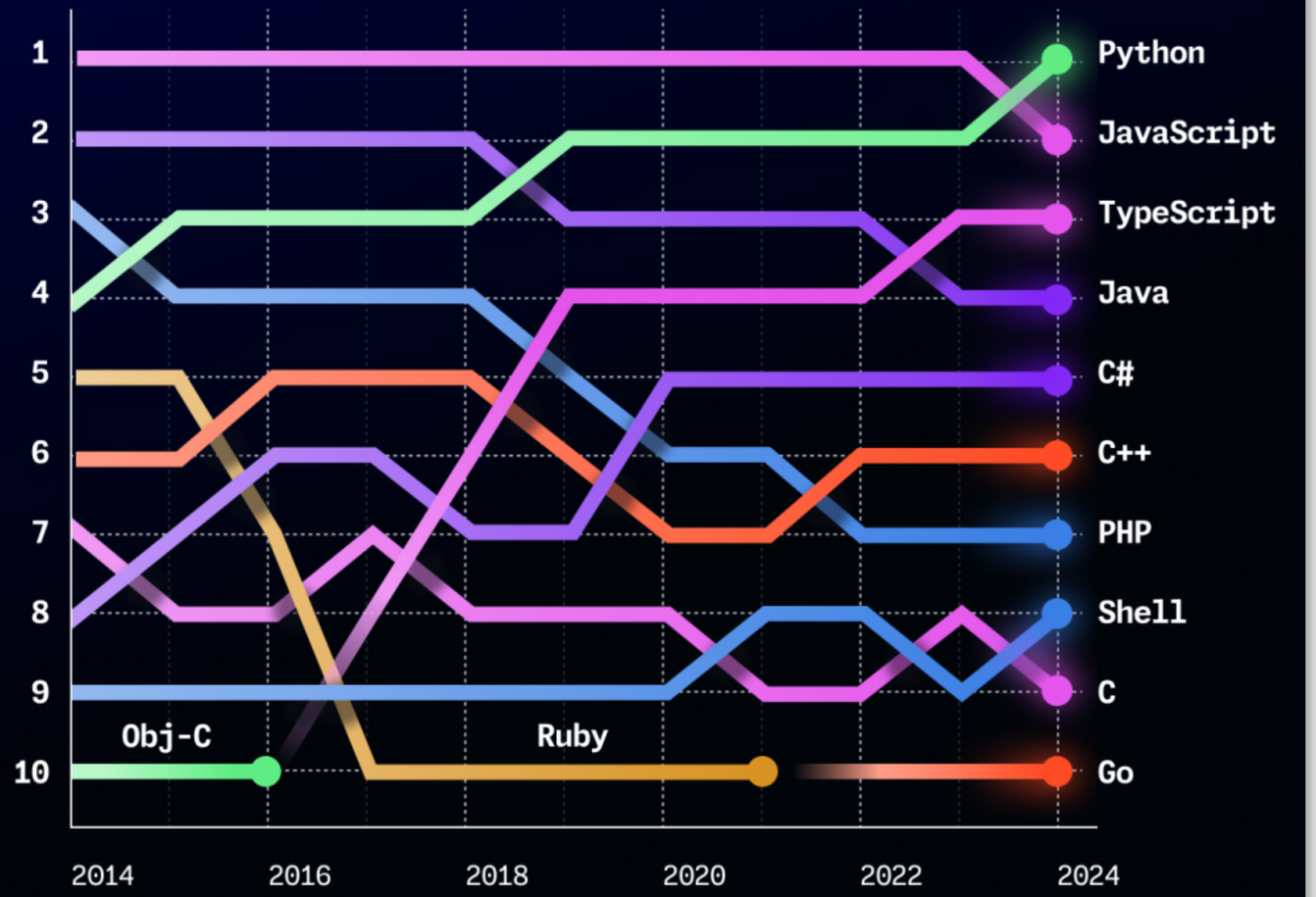
Python Language

History

1. Python is a high-level, general-purpose programming language known for its emphasis on code readability, achieved through the use of significant indentation.
2. Developed by “Guido van Rossum” and first released in 1991, Python's design philosophy prioritizes clarity and simplicity, making it accessible for beginners while remaining powerful enough for complex applications.

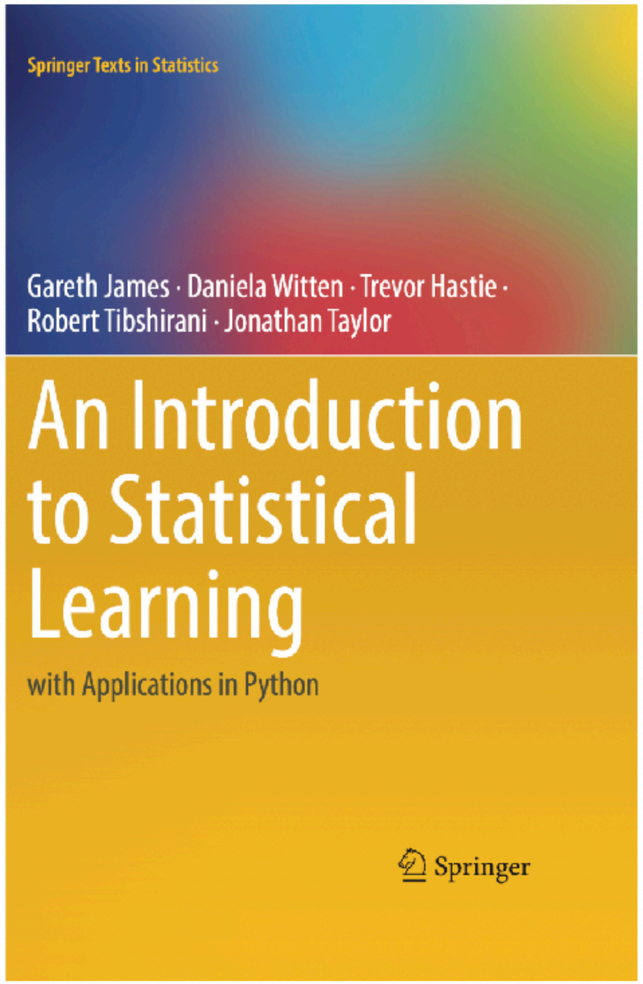
Top programming languages on GitHub

RANKED BY COUNT OF DISTINCT USERS CONTRIBUTING TO PROJECTS OF EACH LANGUAGE.



Python Use Cases

Best Resources for ML



Authors

Gareth James

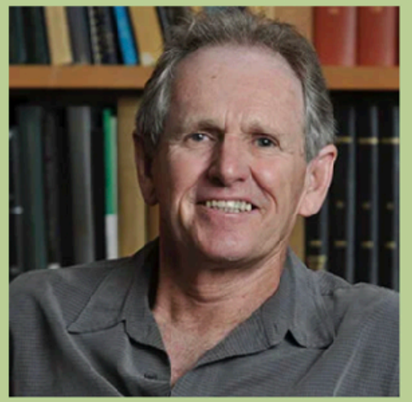
John H. Harland Dean
Goizueta Business School

Emory University

Daniela Witten

Dorothy Gilford Endowed Chair
Professor of Statistics
Professor of Biostatistics

University of Washington



Trevor Hastie

The John A. Overdeck Professor
Professor of Statistics
Professor of Biomedical Data Science

Stanford University

Rob Tibshirani

Professor of Biomedical Data Science
Professor of Statistics

Stanford University

GodFathers of AI

Andrew Ng

Geoffrey Hinton

Demis Hassabis

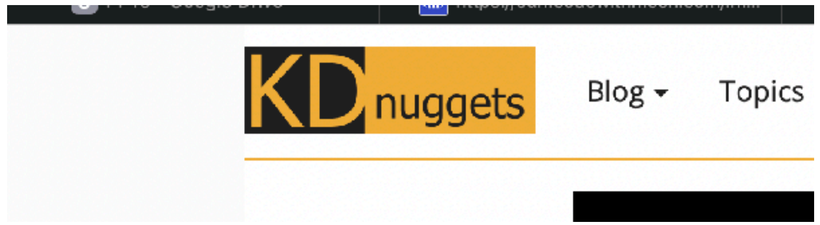
Jim Simson

Renaissance

"Since 1988, his flagship Medallion fund has generated average annual returns of 66% before charging hefty investor fees—39% after fees—racking up trading gains of more than $100 billion. No one in the investment world comes close. Warren Buffett, George Soros, Peter Lynch, Steve Cohen, and Ray Dalio all fall short."

Best Resources for ML

Newsletters



Data Repos

NumPy

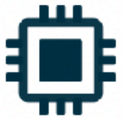
A fundamental package to manipulate arrays

NumPy

The fundamental package for scientific computing with Python

LATEST RELEASE: NUMPY 2.3. VIEW ALL RELEASES

Quantum Computing



QuTiP
PyQuil
Qiskit
PennyLane

Statistical Computing

Pandas
statsmodels
Xarray
Seaborn

Signal Processing

SciPy
PyWavelets
python-control
HyperSpy

Image Processing

Scikit-image
OpenCV
Mahotas

Graphs and Networks

NetworkX
graph-tool
igraph
PyGSP

Astronomy

AstroPy
SunPy
SpacePy

Cognitive Psychology

PsychoPy

Bioinformatics

BioPython
Scikit-Bio
PyEnsembl
ETE

Bayesian Inference

PyStan
PyMC
ArviZ
emcee

Mathematical Analysis

SciPy
SymPy
cvxpy
FEniCS

Chemistry

Cantera
MDAnalysis
RDKit
PyBaMM

Geoscience

Pangeo
Simpeg
ObsPy
Fatiando a Terra

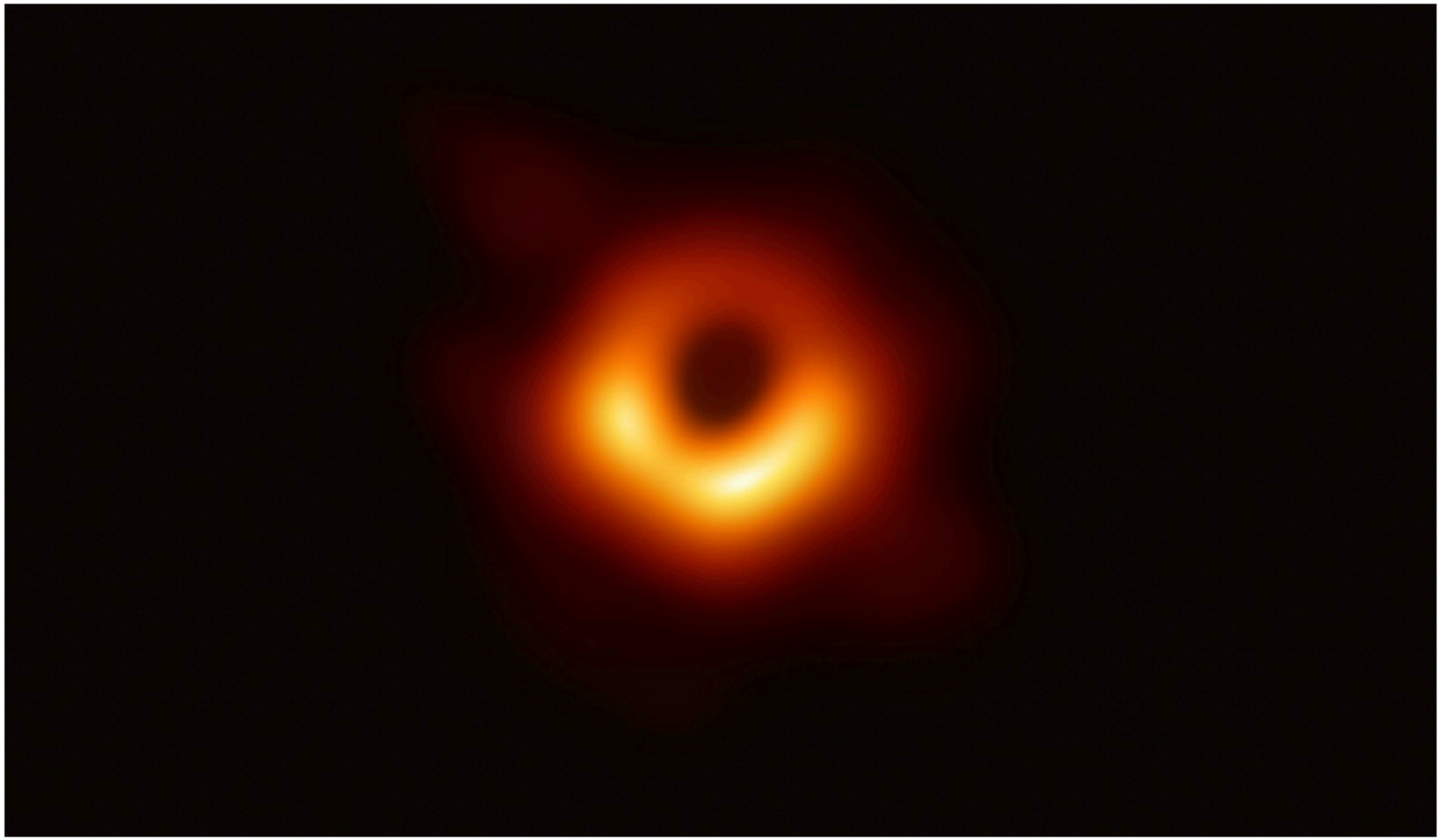
Geographic Processing

Shapely
GeoPandas
Folium

Architecture & Engineering

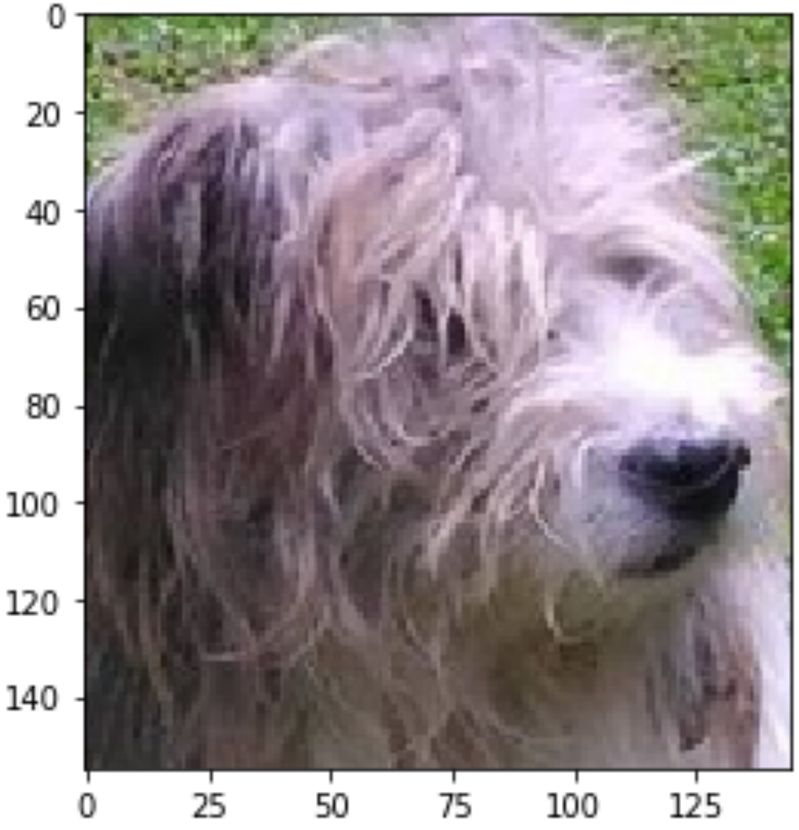
COMPAS
City Energy Analyst
Sverchok

NumPy Use Cases

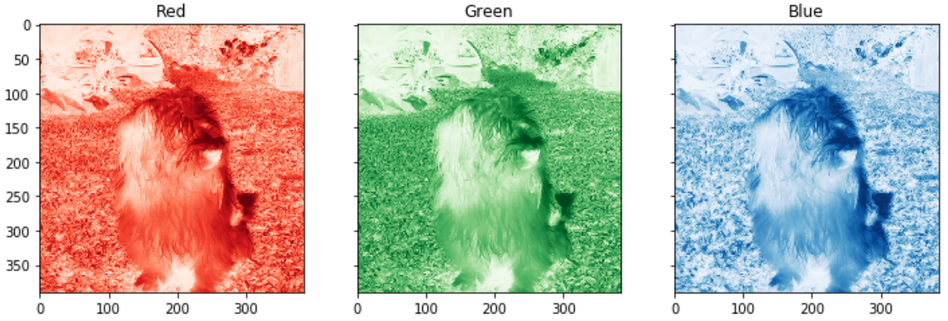


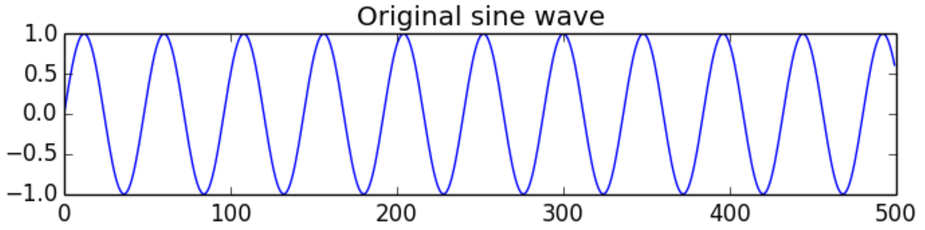
Black Hole M87 #

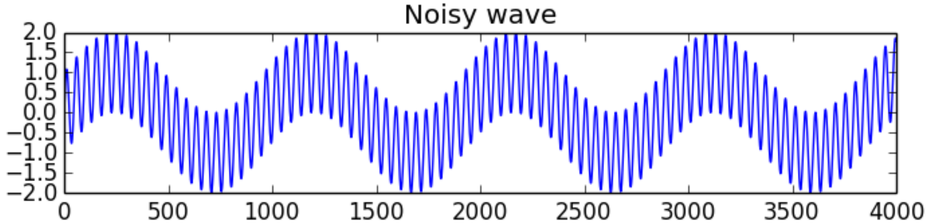
(Image Credits: Event Horizon Telescope Collaboration)

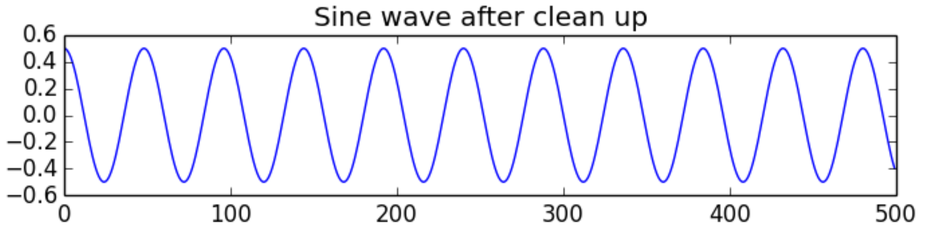


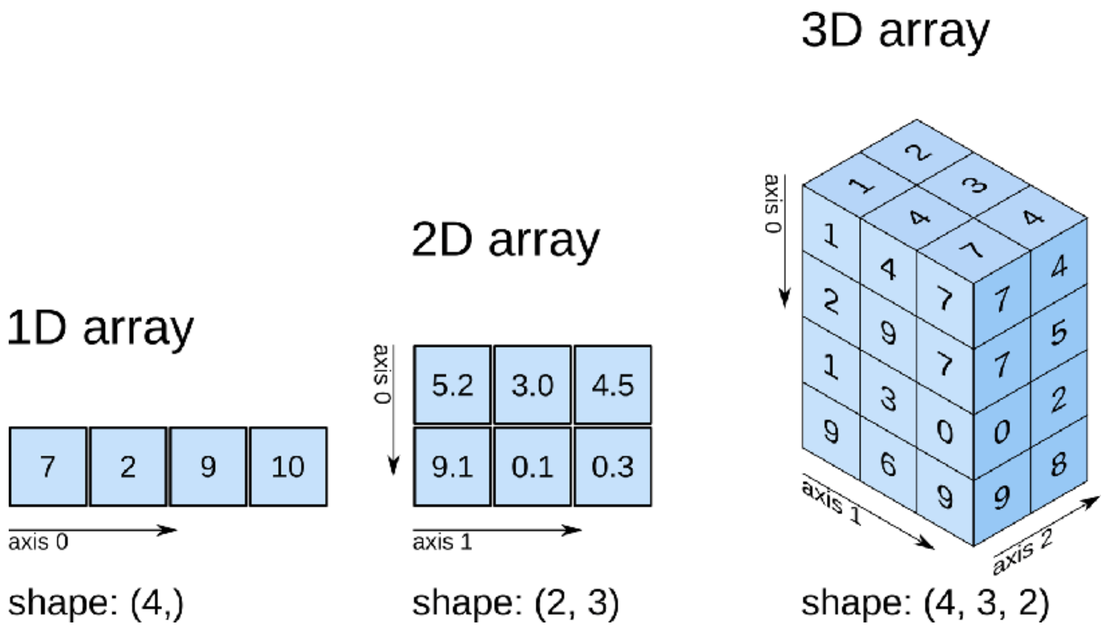

```
Array([[[ 33,  35,  30],
        [ 32,  35,  28],
        [ 33,  36,  29],
        ...,
        [ 35,  33,  35],
        [ 40,  38,  40],
        [ 43,  42,  41]],

       [[ 34,  36,  30],
        [ 33,  36,  29],
        [ 32,  35,  28],
        ...,
        [ 34,  32,  35],
        [ 37,  35,  36],
        [ 44,  43,  42]],

       [[ 34,  36,  30],
        [ 33,  36,  29],
        [ 34,  37,  30],
        ...,
        [ 40,  38,  41],
        [ 37,  35,  36],
        [ 40,  38,  38]],

       ...,

       [[149, 162, 110],
        [153, 166, 117],
        [156, 169, 122],
        ...,
        [ 99, 110,  71],
        [102, 111,  71],
        [ 85,  93,  58]],
```











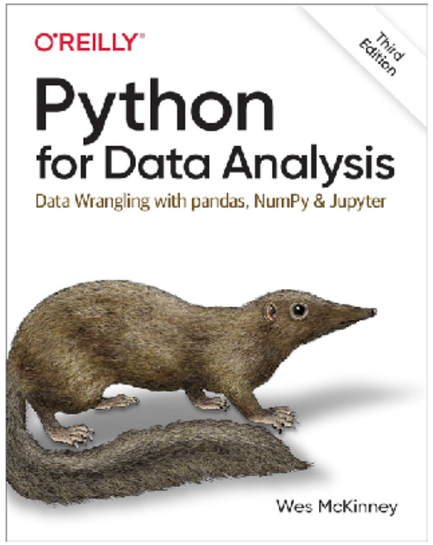
Any rectangular arrangement of number(real or complex) is called a **matrix**. If a matrix has m rows and n columns then the order of matrix is $m \times n$.

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1j} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2j} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{i1} & a_{i2} & a_{i3} & \cdots & a_{ij} & \cdots & a_{in} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mj} & \cdots & a_{mn} \end{bmatrix}$$

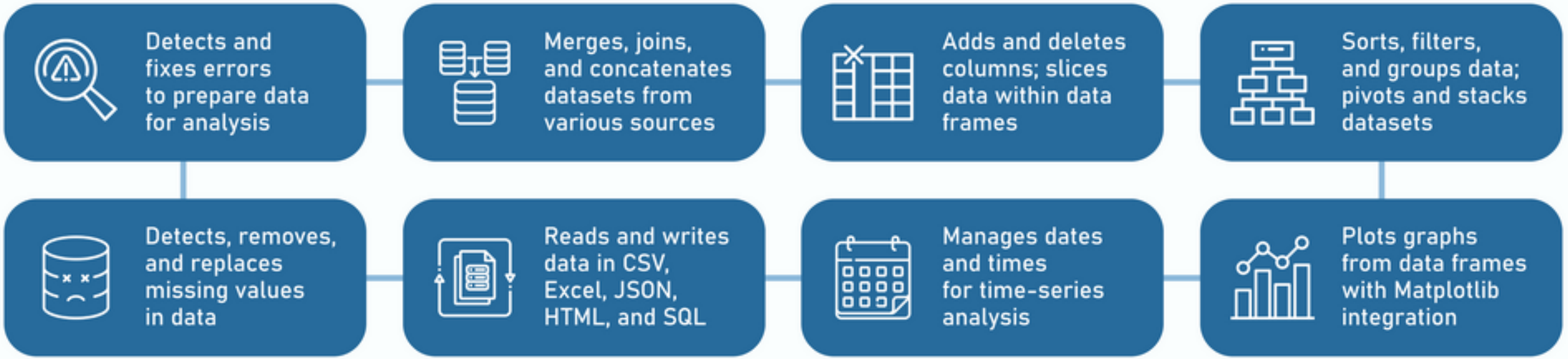
where a_{ij} denote the element of i^{th} row & j^{th} column.

Pandas

Pandas is a fast, powerful, flexible and easy to use open source **data analysis and manipulation tool**, built on top of the Python programming language.



PANDAS MAIN CAPABILITIES



altexsoft

Data Science Use Cases

"In 2012, Target (the US retail chain) famously predicted a teenage girl's pregnancy before her own family knew, just by analyzing her shopping patterns."

The 787 continuously streams data from hundreds of thousands of sensors (engines, hydraulics, electronics).

Airlines use this data to predict part failures before they happen, reducing delays and costs.

Teslas continuously send driving data to the cloud. With billions of miles analyzed, Tesla improves its AI driving models.

Hospitals analyze patient vitals and lab data to detect sepsis (a deadly infection) hours earlier than doctors might notice it. This saves thousands of lives.

Netflix saves about $1 billion every year by using data analysis to recommend shows and movies. Most people watch what Netflix suggests rather than browsing endlessly.

Every year Spotify uses listening data to create personalized "Wrapped" summaries for each user. This data-driven feature became a viral global trend

Companies like BMW and Mercedes use vehicle telematics to analyze engine vibrations, brake wear, and fuel system efficiency.

Each F1 car has 300+ sensors streaming 1.1 million data points per second during a race.

Teams analyze this data in real time to decide on pit stops, tire changes, and strategy, often deciding who wins.

Pandas

Series

pd.Series

pd.Series(data, index)

Creation of Series

Arithematic operations

Data Frame

pd.DataFrame

pd.DataFrame(data, index, columns)

pd.read_csv('file_path')

df.head()

df.tail()

df.describe()

df.info()

Create/drop the columns

iloc and loc

Conditional Filtering

pd.DataFrame(Conditional)

df[df['column_name'] <= 'value']

Multiple columns filtering -

And(&) Or (|)

df.isin()

GroupBy Operations

df.groupby()

GroupBy Operations

Item	year	Sale
tea	2010	1200
coffee	2010	1050
sugar	2010	500
tea	2011	1500
coffee	2011	1200
sugar	2011	1000
tea	2012	1230
coffee	2012	1300
sugar	2012	1420

Groupby Item

```
coffee:
     item  year  sales
1  coffee  2010   1050
4  coffee  2011   1200
7  coffee  2012   1300


sugar:
    item  year  sales
2  sugar  2010    500
5  sugar  2011   1000
8  sugar  2012   1420


tea:
  item  year  sales
0  tea  2010   1200
3  tea  2011   1500
6  tea  2012   1230
```

DATA SCIENCE
PARICHAY

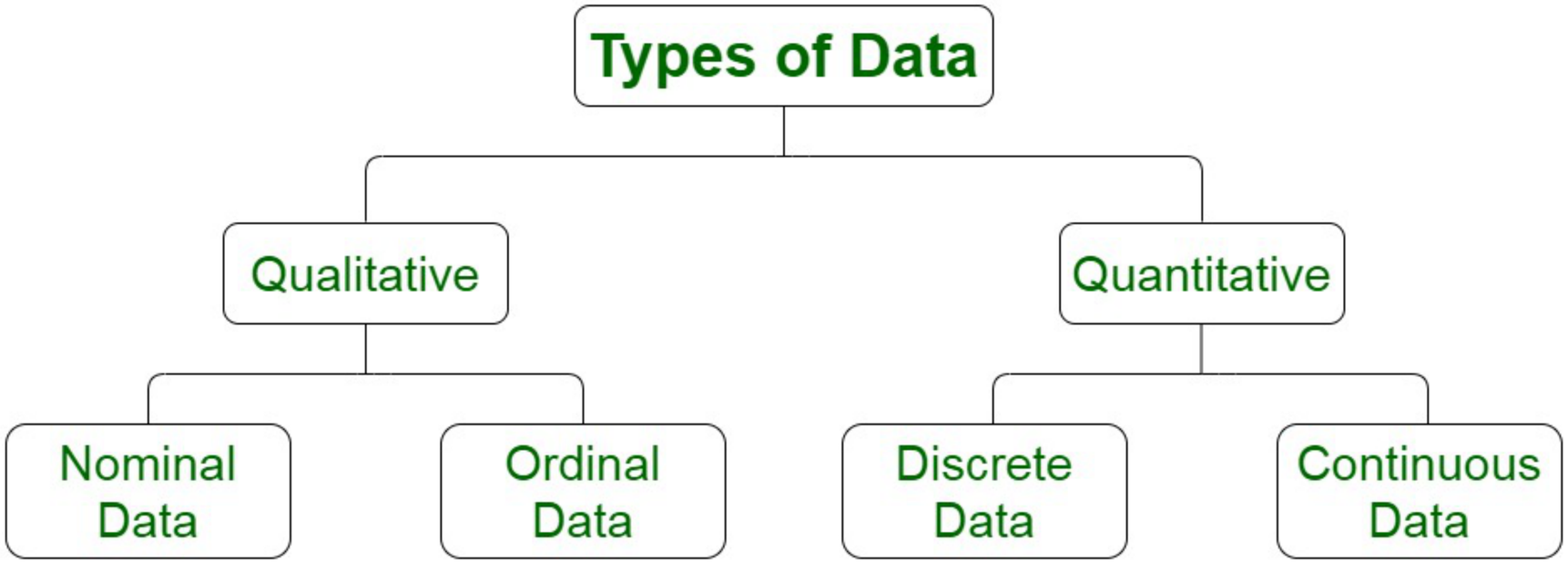
Team	Score
A	8.1
A	8.3
A	9.2
B	6.5
B	7.1
B	7.9
B	9.2

A	8.3
B	7.5

Median of each Group

Statistics Data Types

Types of Data in Statistics



Qualitative Data Type (Strings/Object)

Nominal Data

Examples of nominal data include:

- Gender (Male or female),
- Race (White, Black, Asian),
- Religion (Hinuduism, Christianity, Islam, Judaism),
- Blood type (A, B, AB, O), etc.
- Names (Rohan > Rohit < SuryaKumar)
- City : Boston > Mumbai > Cal

Ordinal Data

Examples of ordinal data include:

- Education level (Elementary, Middle, High School, College),
- Job position (Manager, Supervisor, Employee), etc.
- ratings : 5 stars > 4 stars > 3>2>1

Intern < SWE < Tech Lead < Mana< CEO

Quantitative Data Type (Numerical data)

Discrete Data(Integers)

Example of the discrete data types are,

- Count of Students in a class
- Marks of the students in a class test
- Weight of different members of a family, etc.
- age
- Children : 1 ,2 3

Continuous Data

Examples of the continuous data types are

- Temperature Range
- Salary range of Workers in a Factory, etc. 54000.12
- bmi

Mileage : 10.5 km/l

Matplotlib and Seaborn

Types of data visualization


Bar charts


Line charts


Pie charts


Scatter charts


Histograms


Heatmaps


Area charts


Box plots


Bubble charts


Tree maps


Word clouds


Pictogram charts

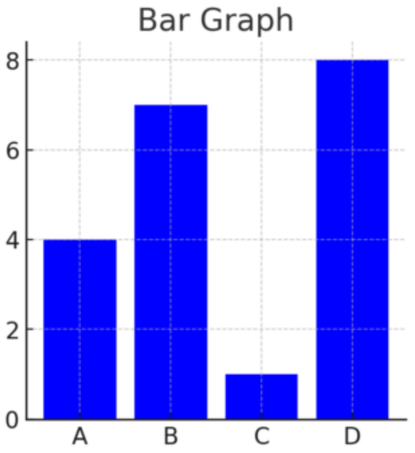

Streamgraphs

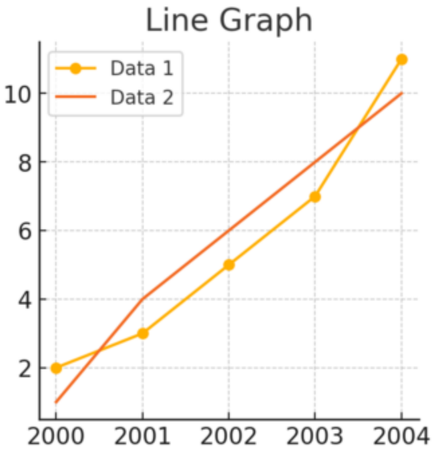

Bullet graphs

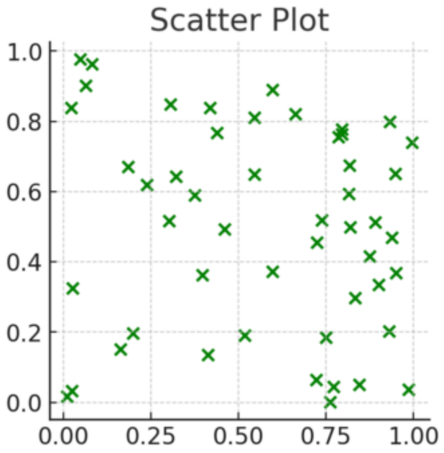

Gantt charts

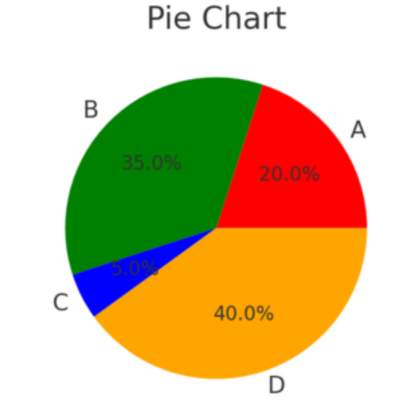

Waterfall charts

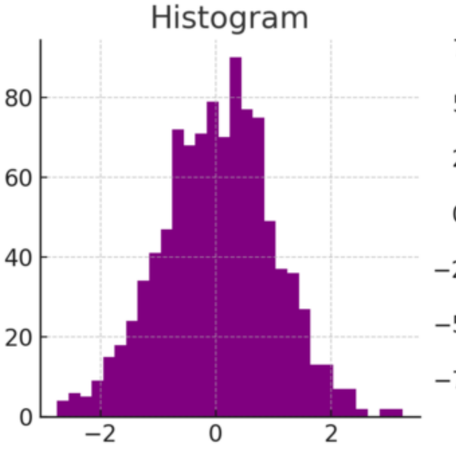
Syracuse University
School of Information Studies

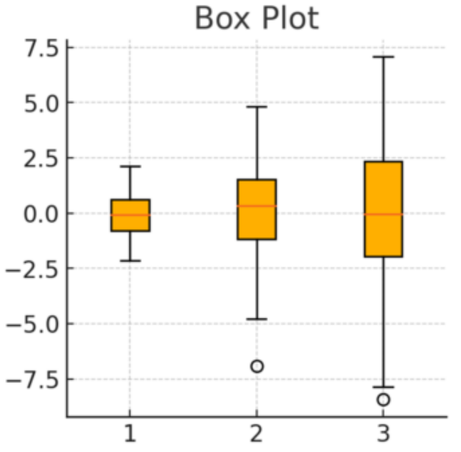

Bar Graph
A
B
C
D

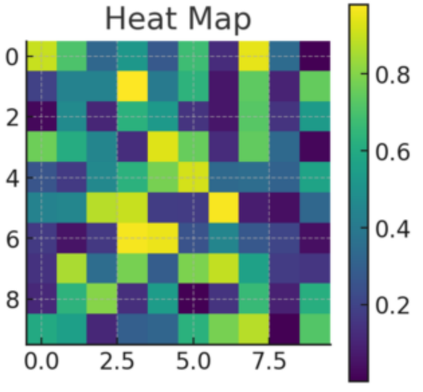

Line Graph
Data 1
Data 2

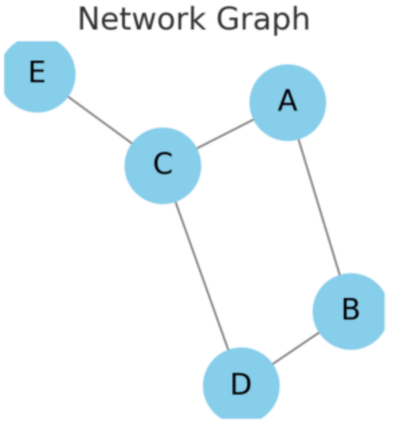

Scatter Plot

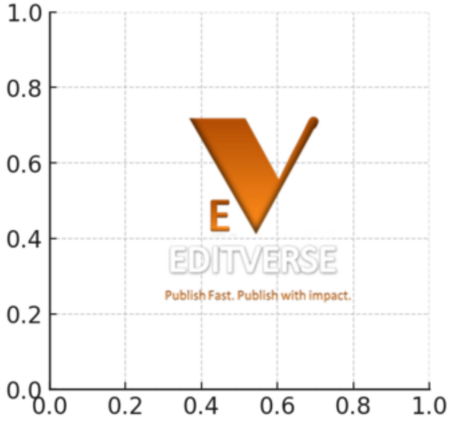

Pie Chart
A
B
C
D
20.0%
35.0%
5.0%
40.0%


Histogram


Box Plot


Heat Map


Network Graph
E
A
C
B
D


EDITVERSE
Publish Fast. Publish with impact.

Types of Plots and its Purposes

Type of	Plots	Purpose
Bar Plot	(sns.barplot)	Compare averages (or other estimates) across categories.
Count Plot	(sns.countplot)	Show frequency counts of categorical values.
Box Plot	(sns.boxplot)	Summarize distribution, spot outliers, compare categories.
Histogram	(sns.histplot)	Show distribution of a numeric variable (often with `kde=True`).
Scatter Plot	sns.scatterplot	Explore relationship between two numeric variables (with hue for categories).

MindMap for Visualisation

Machine Learning

Supervised : Regression

Supervised ML
Regression Algorithms

1) Simple Linear Regression

2) Polynomial Regression

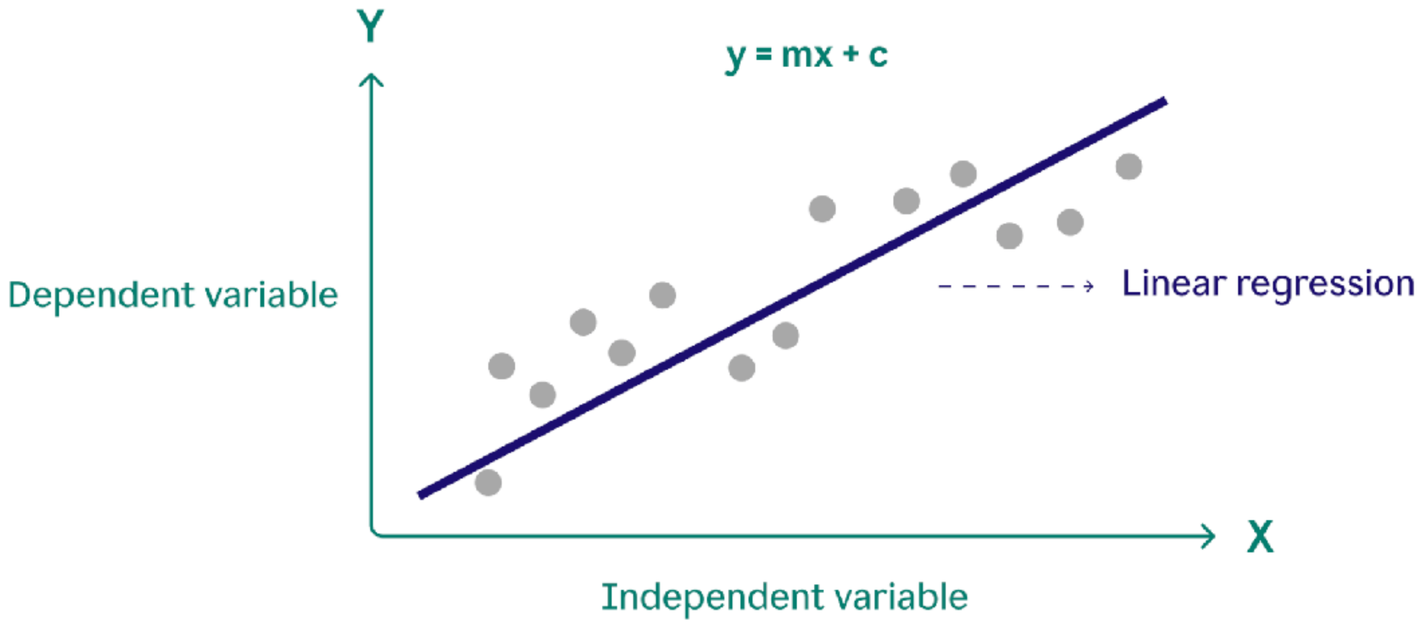
3) Tree Based : Decision Tree Regressor

4) Ensemble Methods :

Random Forest Regressor , Gradient Boosting Regressor, XGBoost Regressor.

Simple Linear Regression

Supervised Regression



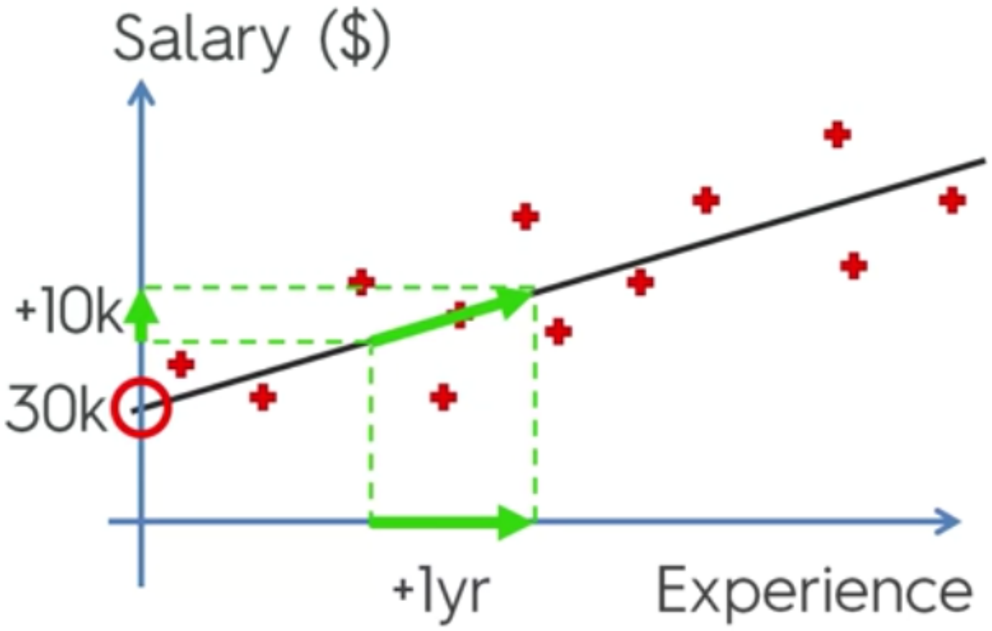
If we have one Target Variable

And One Feature Variable ONLY.

Target Variable : ***Salary***

Feature Variable : ***Experience***

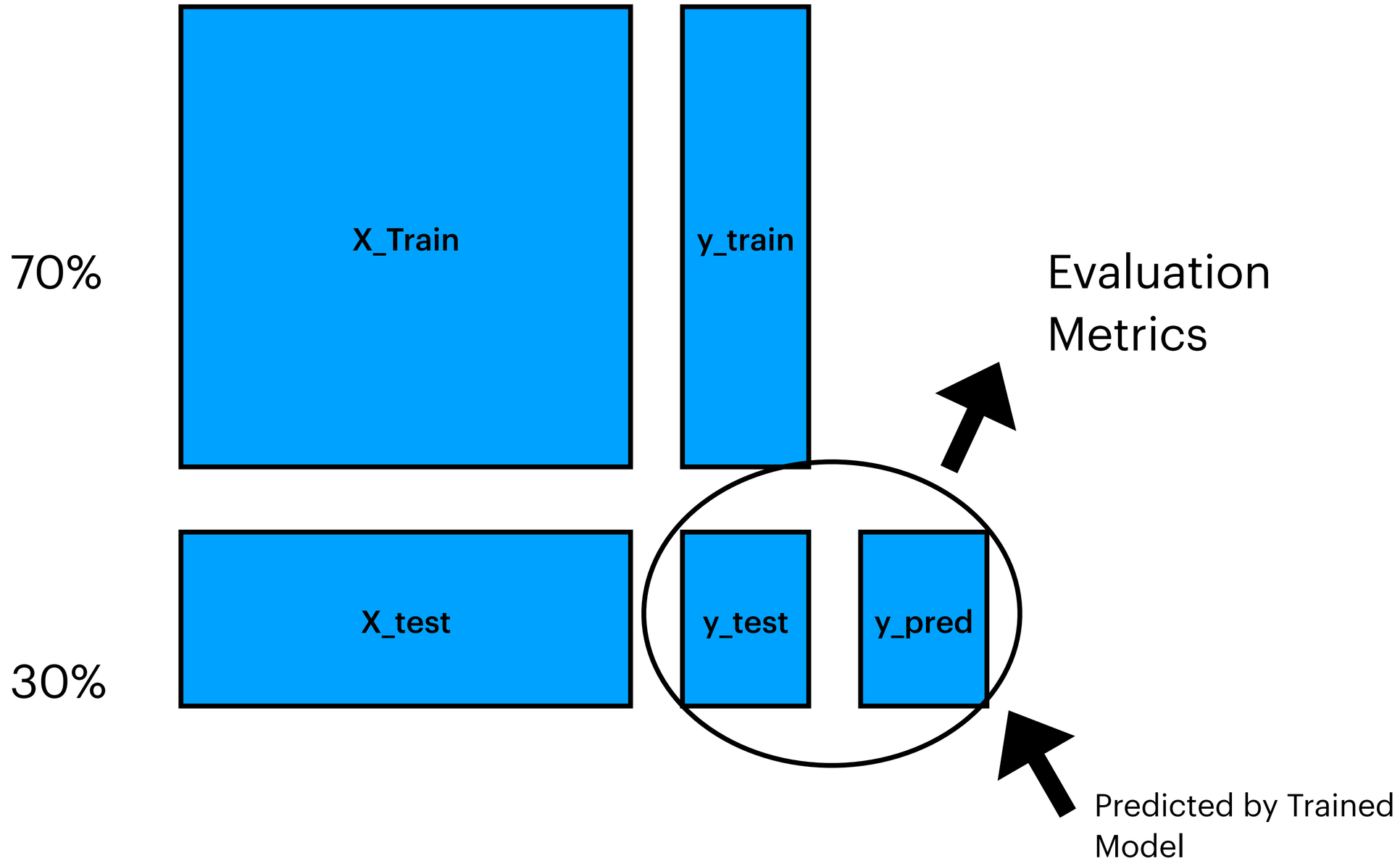
Simple Linear Regression:



$y = b_0 + b_1 * x$

Salary = b_0 + b_1 *Experience

Train Test Split



Supervised ML

Evaluation Metrics

Regression

MSE - Mean Squared Error

MAE - Mean Absolute Error

RMSE - Root Mean Squared Error

R2 - R Squared

Classification

Accuracy

Precision

Recall

F1 Score

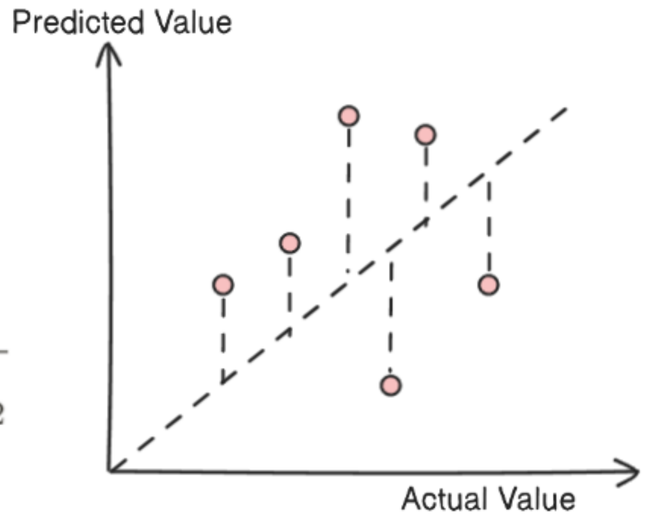
Supervised ML

Regression Metrics

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

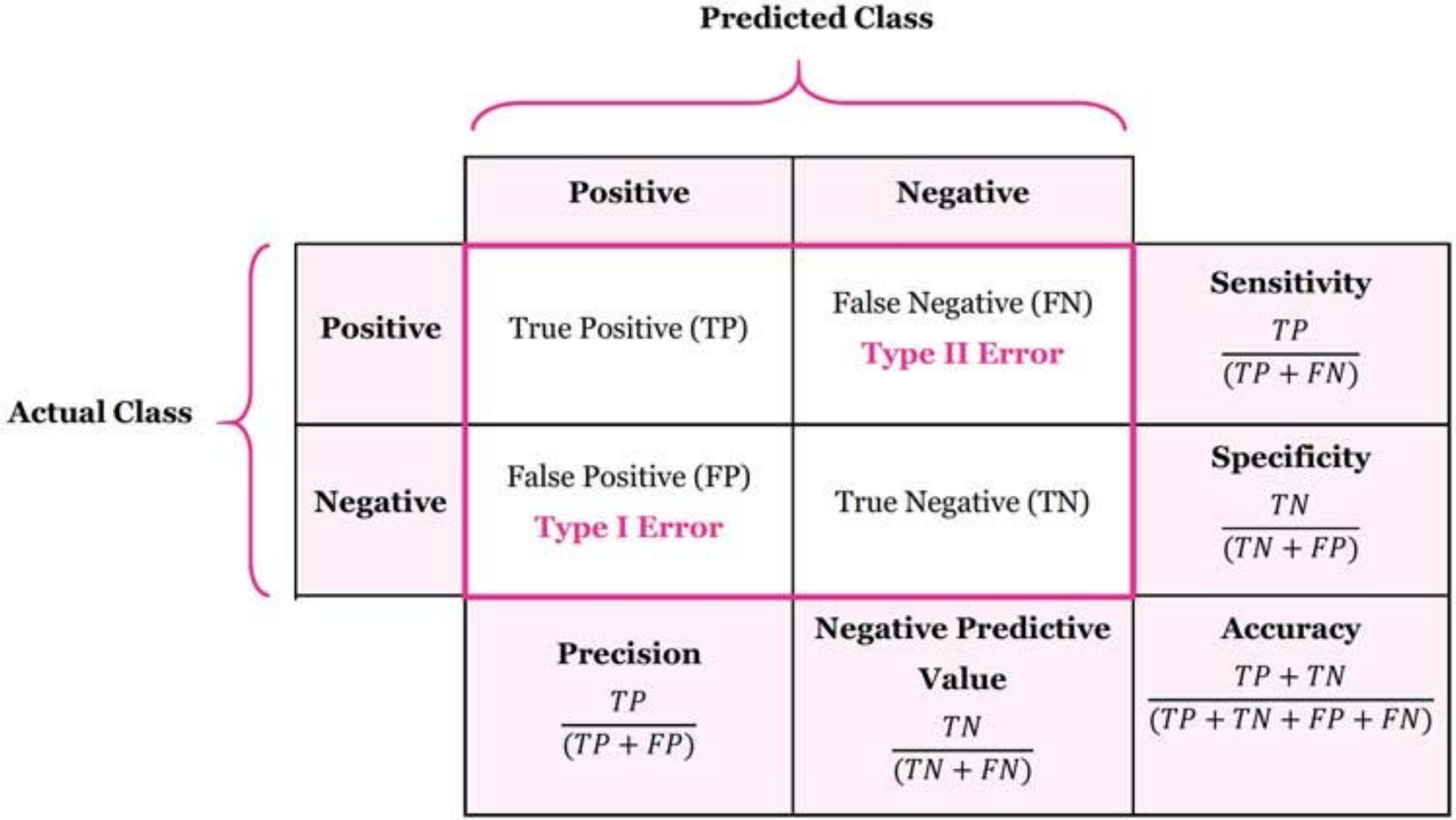


$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

Supervised ML

Classification Metrics



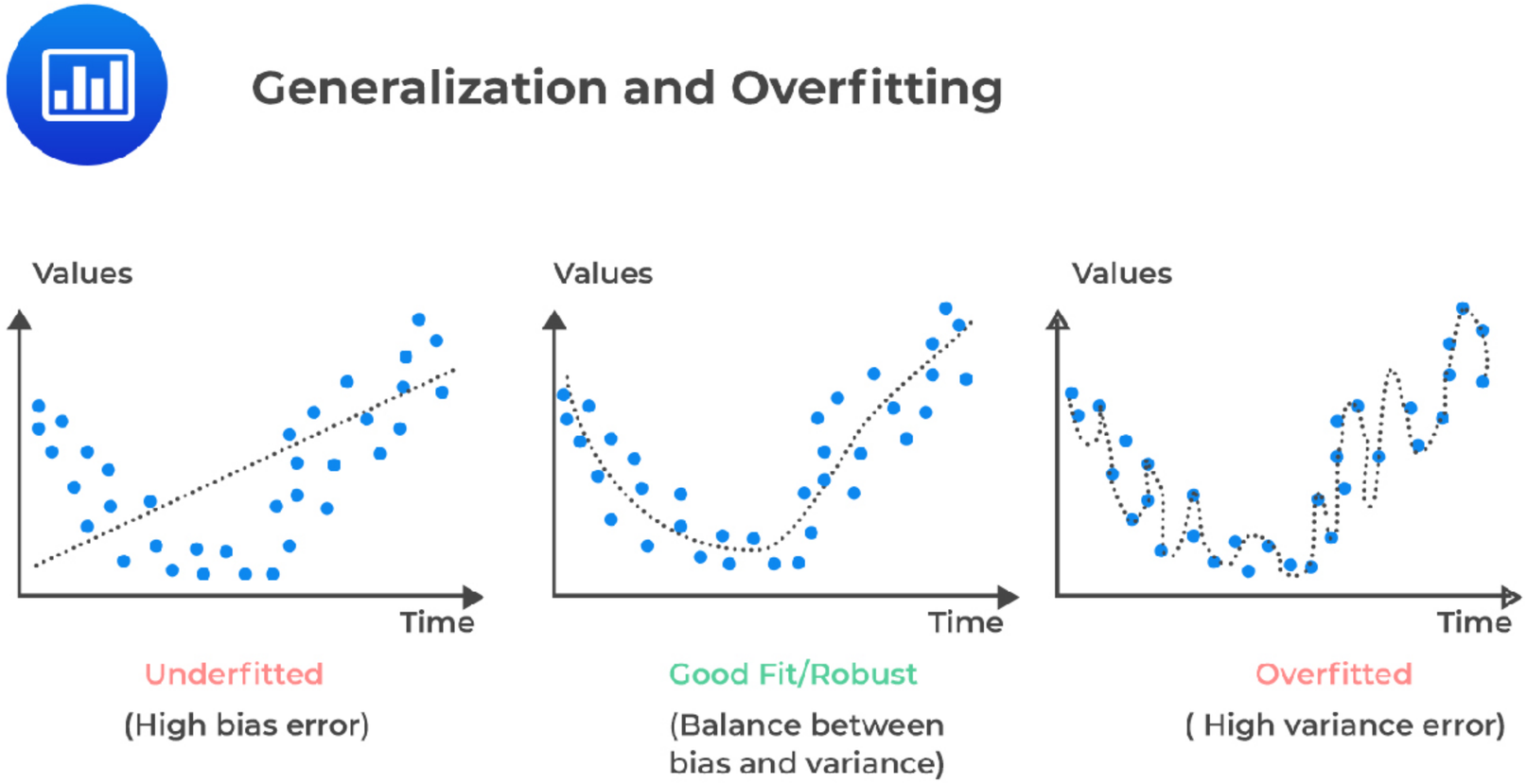
Bias and Variance Trade off

Bias and Variance Trade off

What is it ?

Bias as the model's inability to capture the true relationship between variables.

Variance as the model's ability to produce consistently good predictions across different datasets

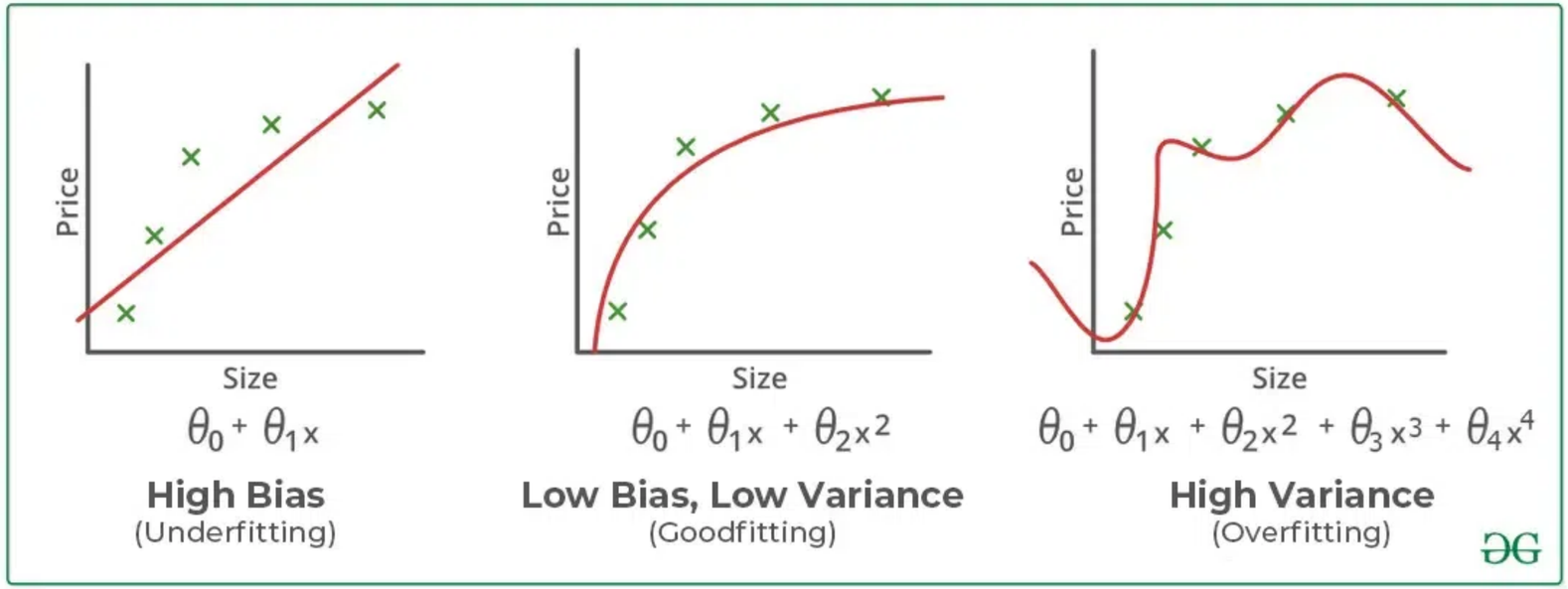


UnderFitting Vs OverFitting

Underfitting : Straight line trying to fit a curved dataset but cannot capture the data's patterns, leading to poor performance on both training and test sets.

Overfitting: A squiggly curve passing through all training points, failing to **“generalize”** performing well on training data but poorly on test data.

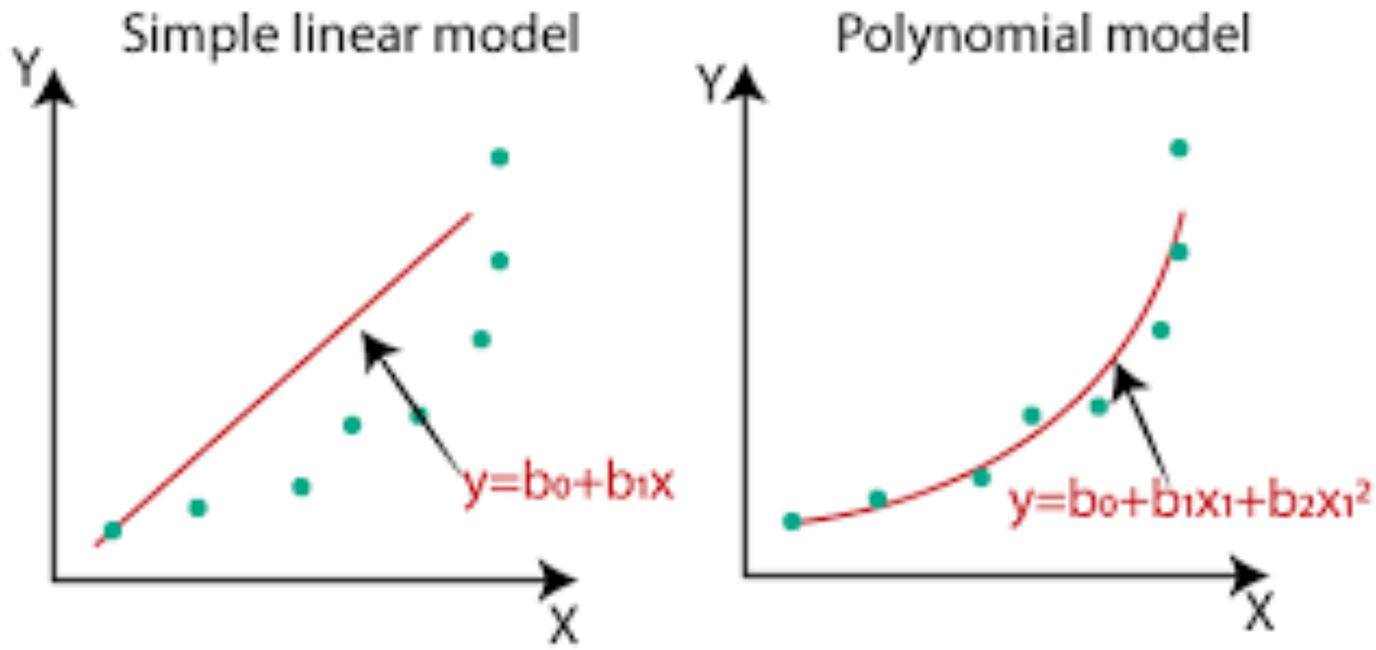
Good Fit: Curve that follows the data trend without overcomplicating to capture the true patterns in the data.



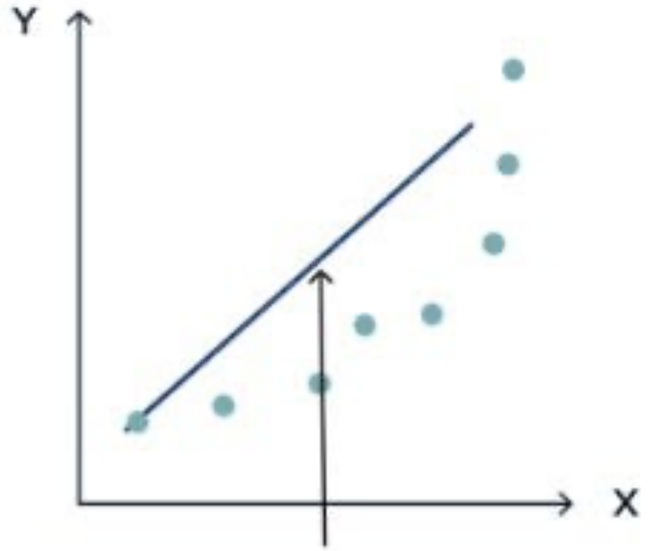
“We Aim for **Low Bias** and **Low Variance**”

Polynomial Regression

Supervised Regression

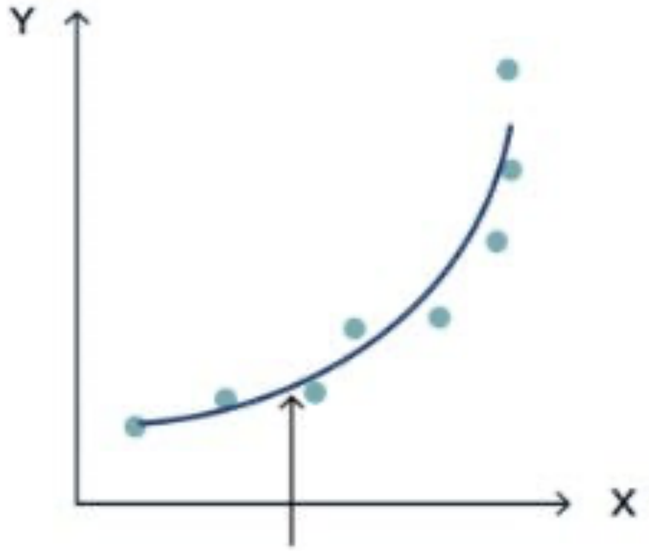


Simple linear model



$y=b_0+b_1x$

Polynomial model



$y=b_0+b_1x+b_2x_1^2$

Feature Engineering

Feature Engineering

Ways to handle the data...

Missing Values

1) Imputation with Mean , Median or Mode

2) Dropping

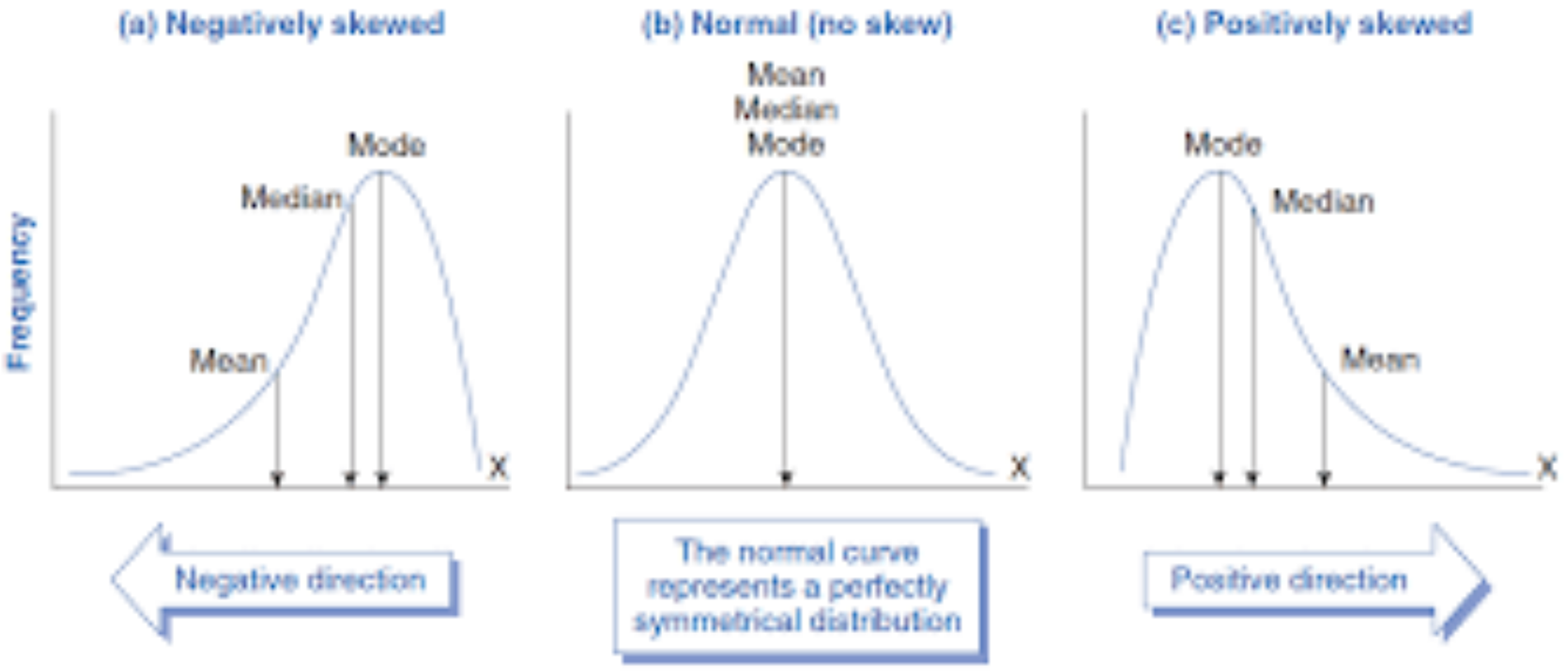
Outlier Detection

Filter the Outlier

Standarisation/ Normalisation

Standardization = (x - mean) / std

Normalisation = (x - mean/ x_max - x_min


(a) Negatively skewed
(b) Normal (no skew)
(c) Positively skewed
Frequency
Mode
Median
Mean
Mean
Median
Mode
Mode
Median
Mean
X
X
X
Negative direction
The normal curve represents a perfectly symmetrical distribution
Positive direction

Machine Learning

Supervised : Classification

Supervised ML
Classification Algorithms

1) Logistic Regression

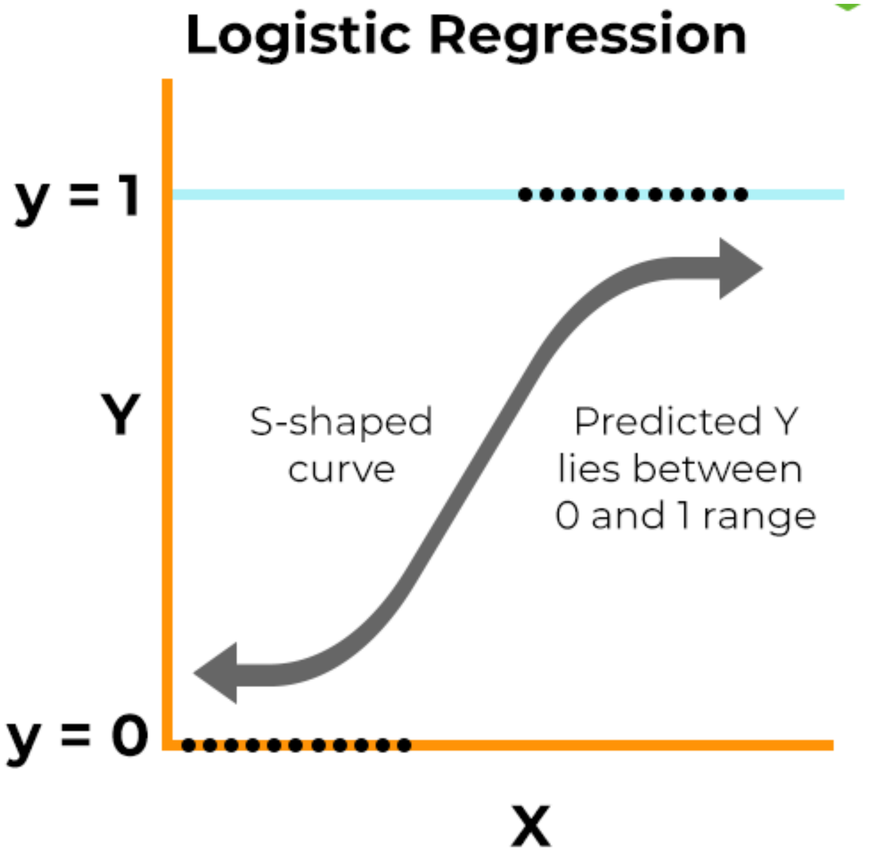
2) Tree Based : Decision Tree Classifier

3) Ensemble Methods :

Random Forest Classifier , Gradient Boosting classifiers, XGBoost Classifiers.

Logistic Regression

Supervised Classification



Logistic Regression

Mathematical Formula

1) It outputs the continuous numerical number however it is still a classification algorithm.

2) The value of “y” falls between 0 and 1.

3)If the output is greater than 0.5 then 1 else 0.

4) It is majorly used for Binary class classification.

$$y = \frac{e^{(b_0 + b_1 X)}}{1 + e^{(b_0 + b_1 X)}}$$

Logistic Regression

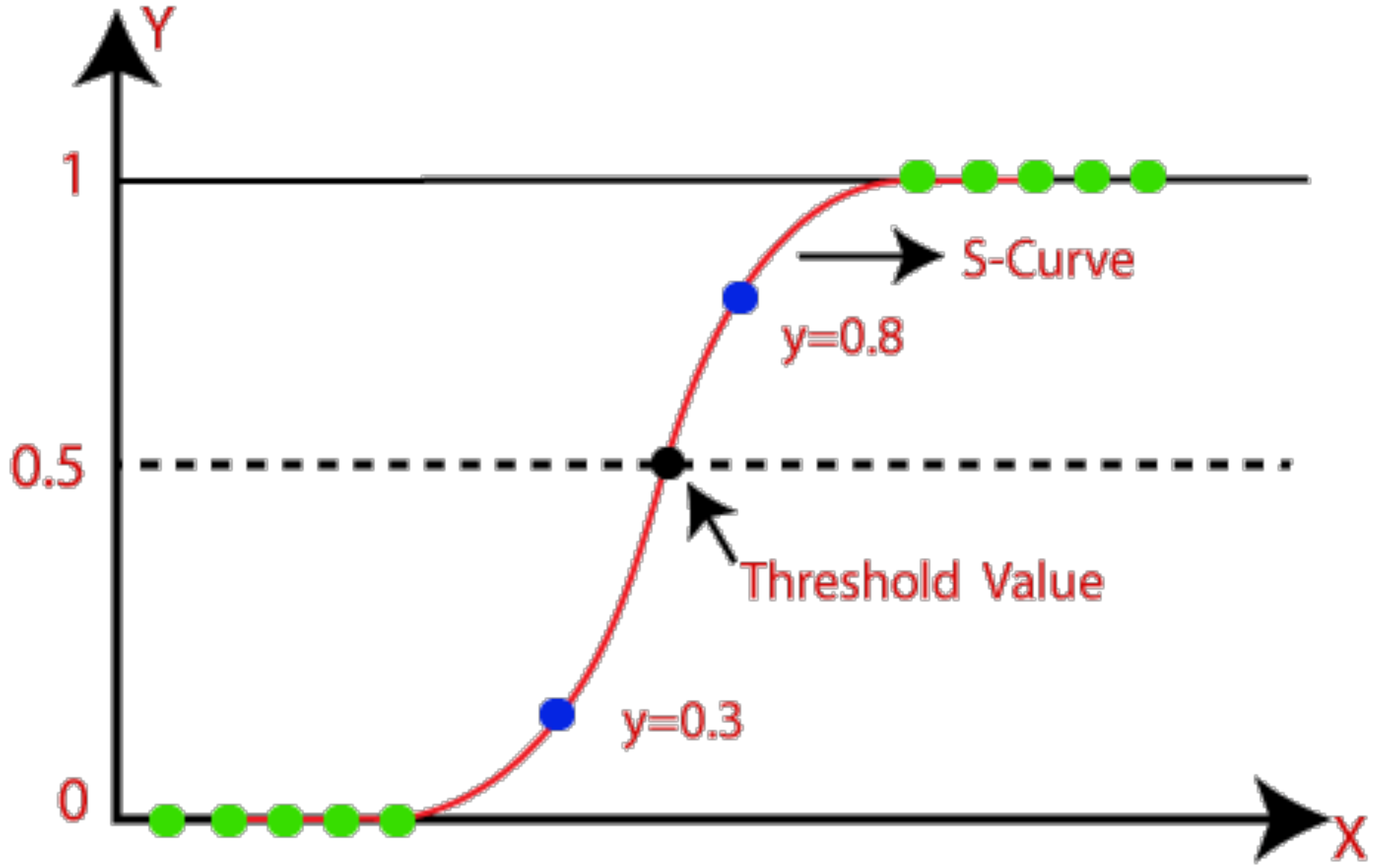
Mathematical Formula

$$y = \frac{e^{(b_0 + b_1 X)}}{1 + e^{(b_0 + b_1 X)}}$$

O.5 is the threshold

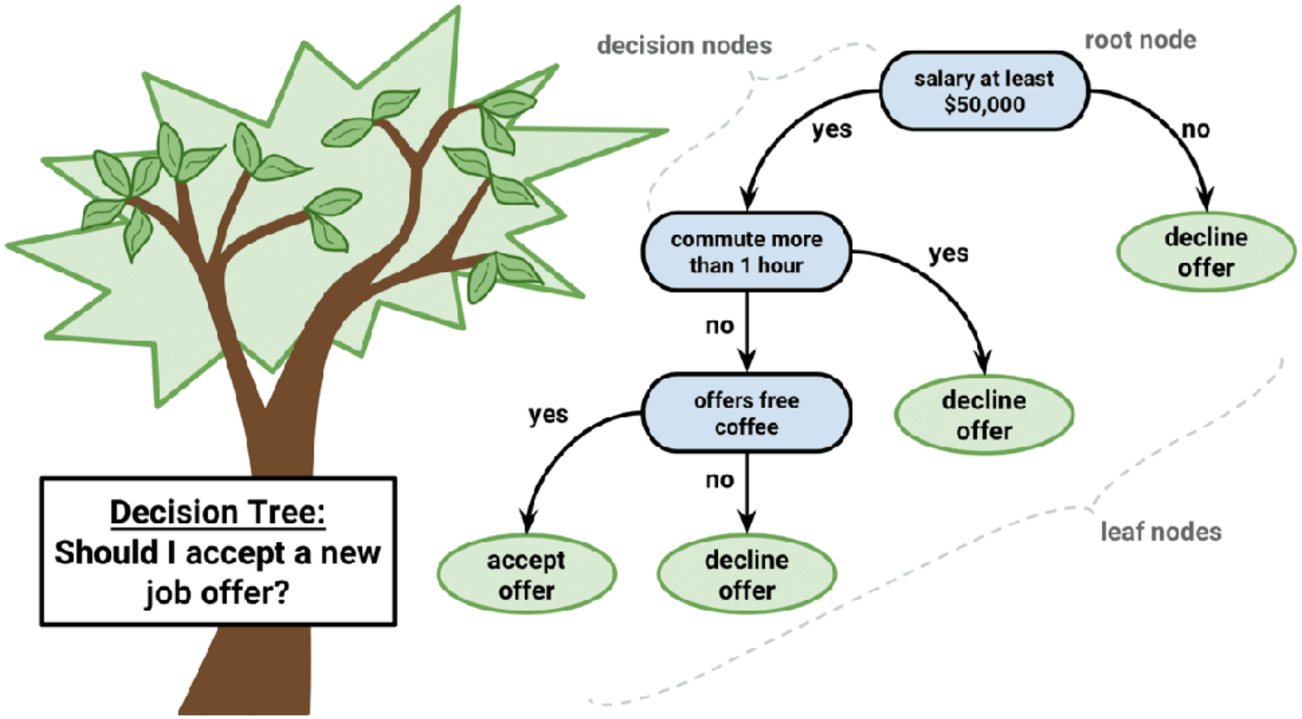
Y - 0.8, > 0.5 then y = 1

Y - 0.3, < 0.5 then y - 0



Decision Tree

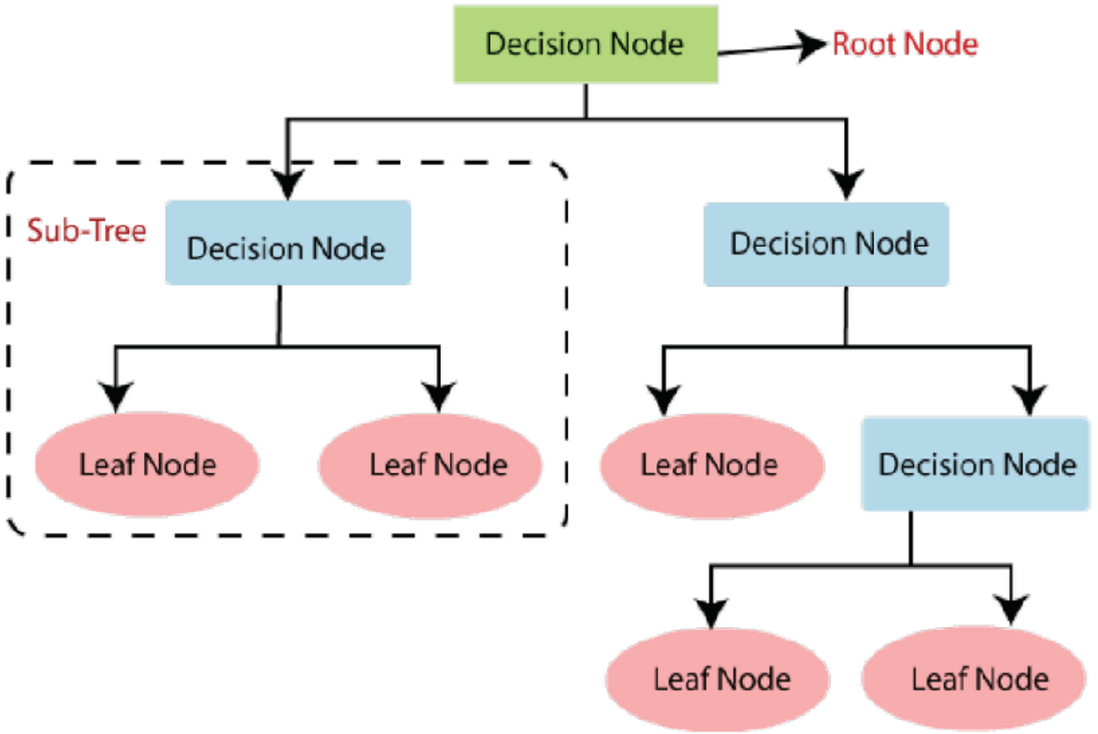
Regression | Classification



Decision Tree

(a) Node Types

- **Root Node:** First split of the dataset.
- **Internal Nodes:** Subsequent splits.
- **Leaf Nodes:** Final output (class label or regression value).



(b) Splitting Criterion

To decide where to split, the algorithm uses a measure of impurity:

- **For Classification:**
  - Gini Impurity
  - Entropy / Information Gain
- **For Regression:**
  - Mean Squared Error (MSE) or Mean Absolute Error (MAE)

Gini Impurity

Gini Impurity:

$$Gini = 1 - \sum_{i=1}^{C} p_i^2$$

where p_i = proportion of samples belonging to class i.

Entropy:

$$Entropy = -\sum_{i=1}^{C} p_i \log_2(p_i)$$

Information Gain (IG)

Step 2: Compute Information Gain (IG)

After splitting on a feature A:

$$IG(A) = Impurity_{parent} - \sum_{k} \frac{n_k}{n} \times Impurity_k$$

The feature with **maximum Information Gain** (or minimum Gini) is chosen for splitting.

Person	Weather	Buy Car?
1	Sunny	No
2	Sunny	No
3	Overcast	Yes
4	Rainy	Yes
5	Rainy	Yes
6	Rainy	No
7	Overcast	Yes
8	Sunny	No
9	Sunny	Yes
10	Rainy	Yes

🧩 Step 1: Before any split — Root Node

At the start, we have **10 samples.**

- "Yes" = 6
- "No" = 4

So proportions:

$$p(Yes) = \frac{6}{10} = 0.6, \quad p(No) = \frac{4}{10} = 0.4$$

🧱 Step 2: Compute Impurity at Root

(a) Gini Impurity

$$Gini = 1 - \sum p_i^2 = 1 - (0.6^2 + 0.4^2) = 1 - (0.36 + 0.16) = 1 - 0.52 = 0.48$$

✅ So Gini at root = **0.48**

(b) Entropy

$$Entropy = -\sum p_i \log_2(p_i)$$

$$= -(0.6 \log_2 0.6 + 0.4 \log_2 0.4)$$

$$= -(0.6 * (-0.7369) + 0.4 * (-1.3219)) = 0.9709$$

✅ So Entropy at root = **0.97 bits**

☀️ Step 3: Try Splitting on "Weather"

Split into groups:

Weather	Count	Yes	No
Sunny	4	1	3
Overcast	2	2	0
Rainy	4	3	1

🌞 Step 4: Compute Gini for Each Split

(a) Sunny:

$$p(Yes) = \frac{1}{4} = 0.25, \quad p(No) = 0.75$$

$$Gini = 1 - (0.25^2 + 0.75^2) = 1 - (0.0625 + 0.5625) = 0.375$$

(b) Overcast:

$$p(Yes) = 1.0, \quad p(No) = 0$$

$$Gini = 1 - (1^2 + 0^2) = 0$$

(c) Rainy:

$$p(Yes) = \frac{3}{4} = 0.75, \quad p(No) = 0.25$$

$$Gini = 1 - (0.75^2 + 0.25^2) = 1 - (0.5625 + 0.0625) = 0.375$$

Weighted Gini after split:

$$Gini_{split} = \frac{4}{10}(0.375) + \frac{2}{10}(0) + \frac{4}{10}(0.375) = 0.15 + 0 + 0.15 = 0.3$$

Gini Gain:

$$\text{Gini Gain} = 0.48 - 0.3 = 0.18$$

✅ The **gain = 0.18**, meaning impurity reduced by 0.18.

🔥 Step 5: Compute Entropy for Each Split

(a) Sunny:

$$Entropy = -[0.25 \log_2 0.25 + 0.75 \log_2 0.75] = -(0.25 * (-2) + 0.75 * (-0.415)) = 0.811$$

(b) Overcast:

$$Entropy = 0$$

(c) Rainy:

$$Entropy = -[0.75 \log_2 0.75 + 0.25 \log_2 0.25] = 0.811$$

Weighted Entropy after split:

$$Entropy_{split} = \frac{4}{10}(0.811) + \frac{2}{10}(0) + \frac{4}{10}(0.811) = 0.6488$$

Information Gain:

$$IG = 0.9709 - 0.6488 = 0.3221$$

✅ The **Information Gain = 0.32 bits**

📝 Step 6: Interpretation

Metric	Before Split	After Split	Gain
Gini	0.48	0.30	0.18
Entropy	0.97	0.65	0.32

Both show impurity **decreased**, which means the split on **Weather** improved class separation.

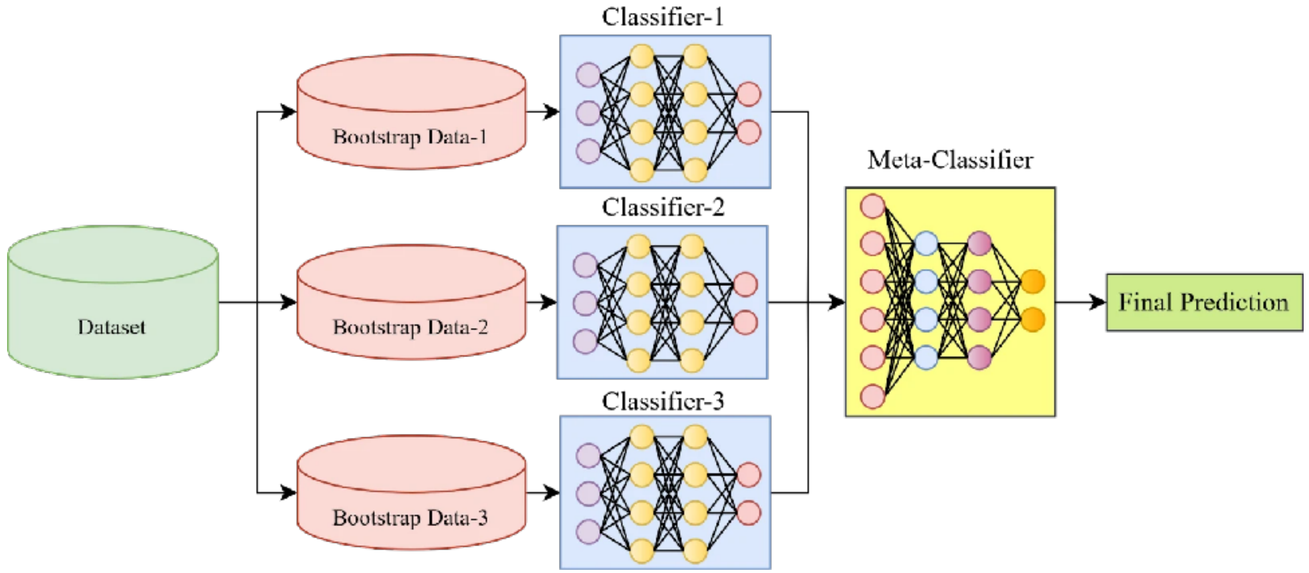
So the Decision Tree would **choose "Weather"** as the first splitting feature.

🎯 Step 7: Key Takeaways

Concept	Formula	Range	Meaning
Gini Impurity	$1 - \sum p_i^2$	[0, 0.5] (for 2 classes)	Measures how often you'd misclassify a random sample
Entropy	$-\sum p_i \log_2 p_i$	[0, 1] (for 2 classes)	Measures the uncertainty/disorder
Information Gain	$Entropy_{parent} - Entropy_{child}$	Higher = better	How much uncertainty is reduced after the split

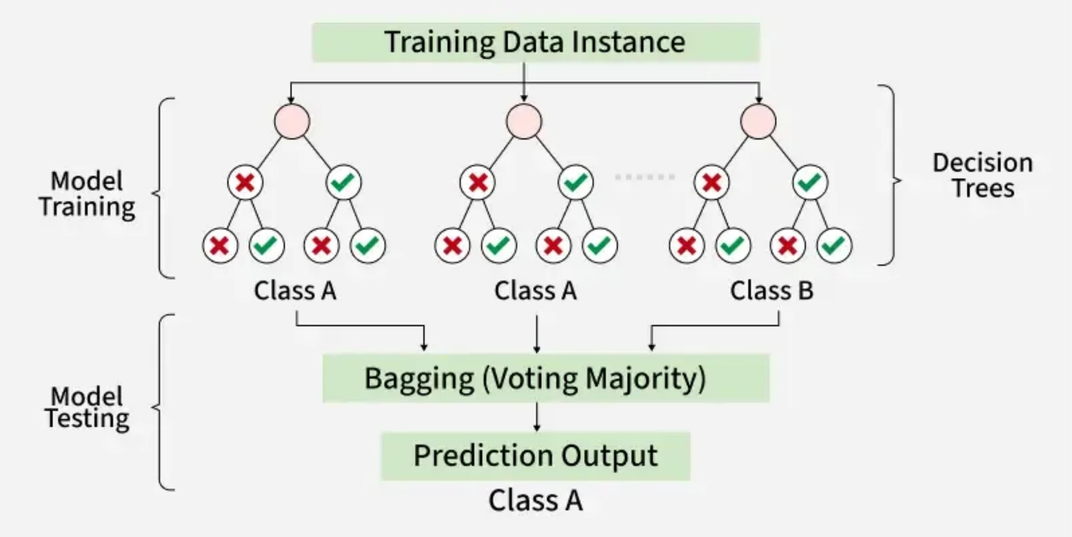
Ensemble Methods

Regression | Classification



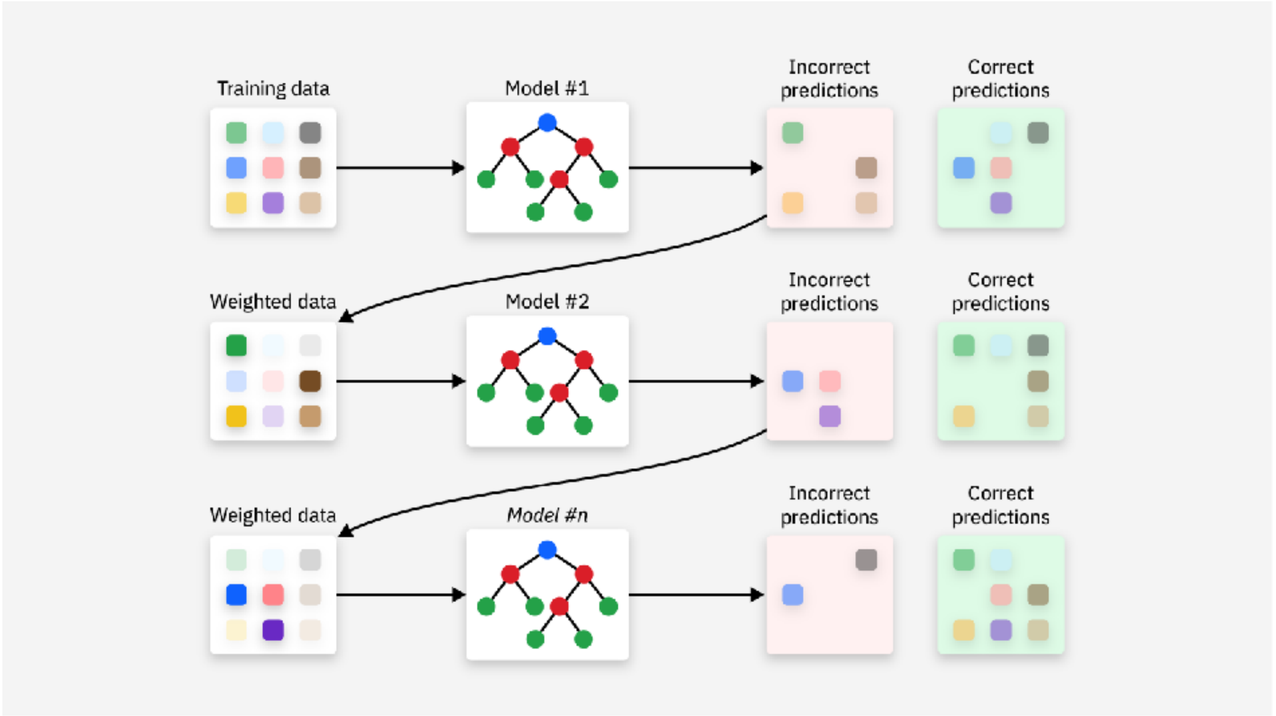
Random Forest

Regression | Classification



Gradient Boosting

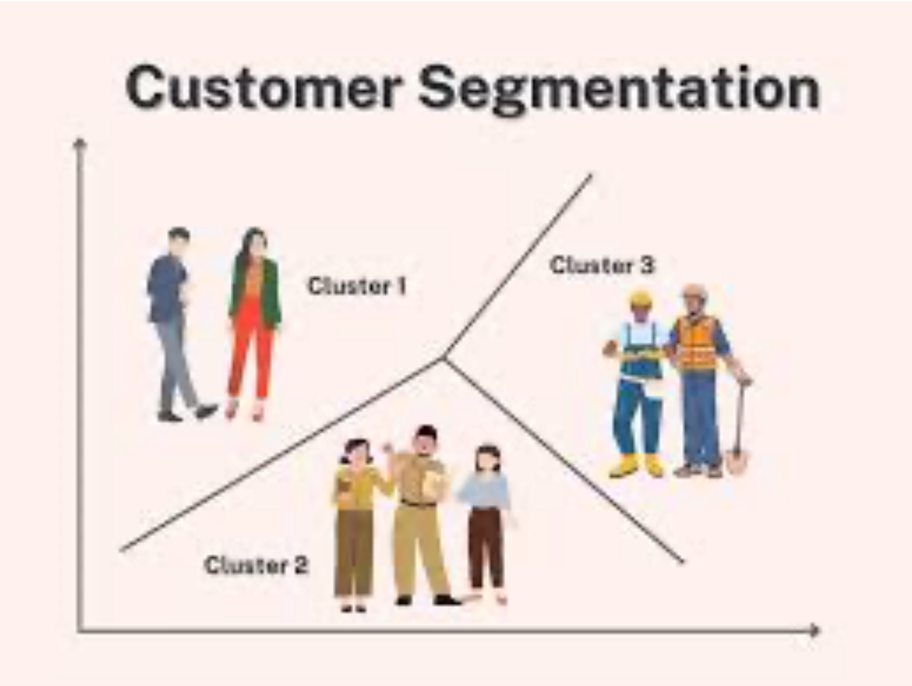
Regression | Classification



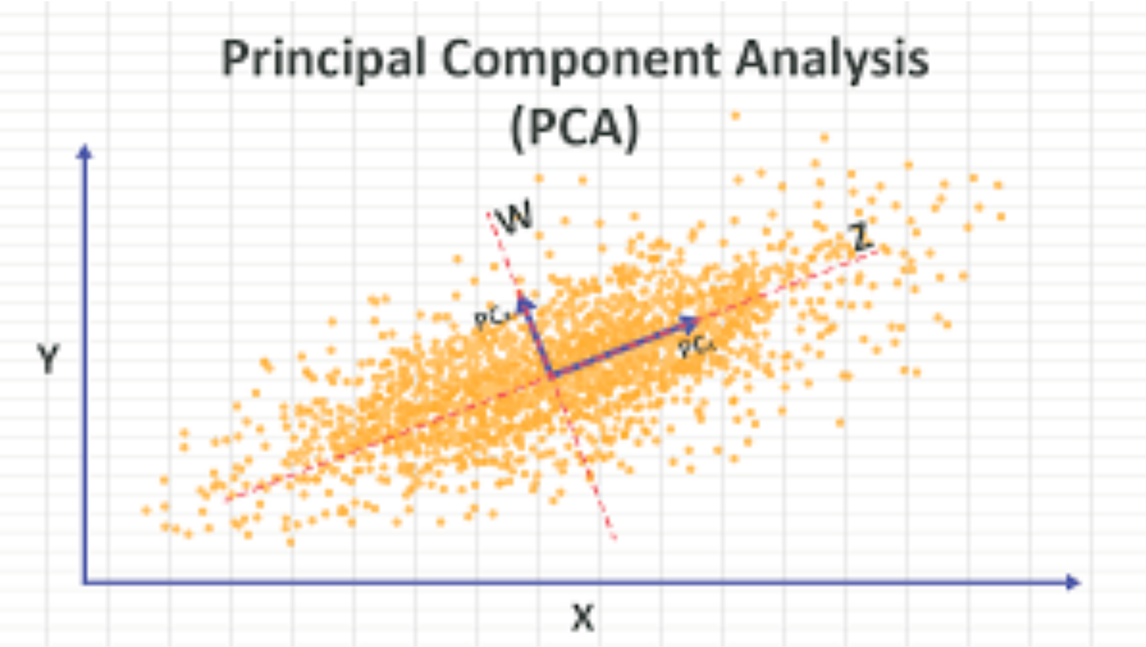
Machine Learning

Un - supervised : Clustering

Applications of Un - supervised ML



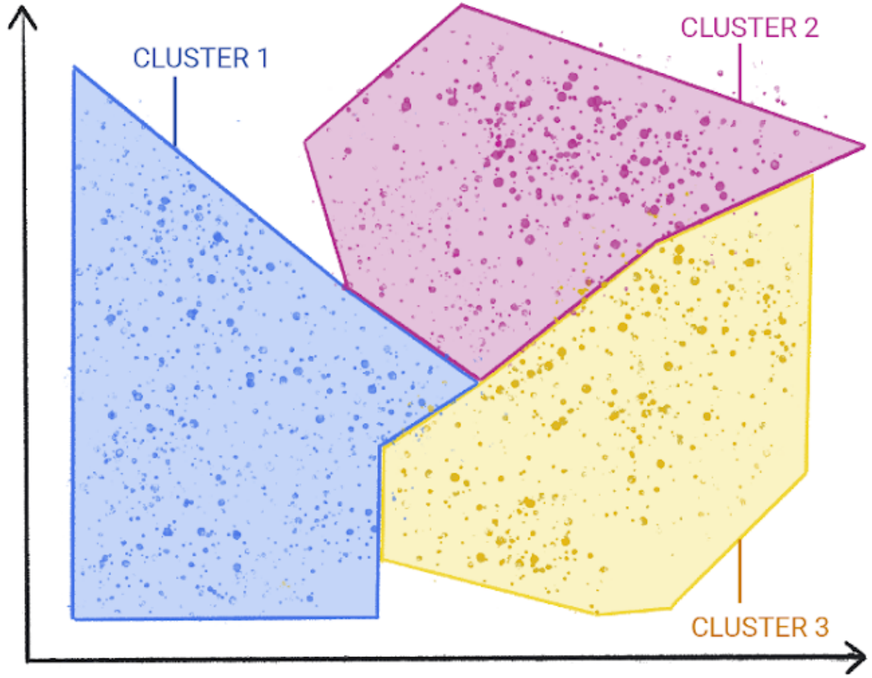
Clustering



Dimensionality Reduction

KMeans

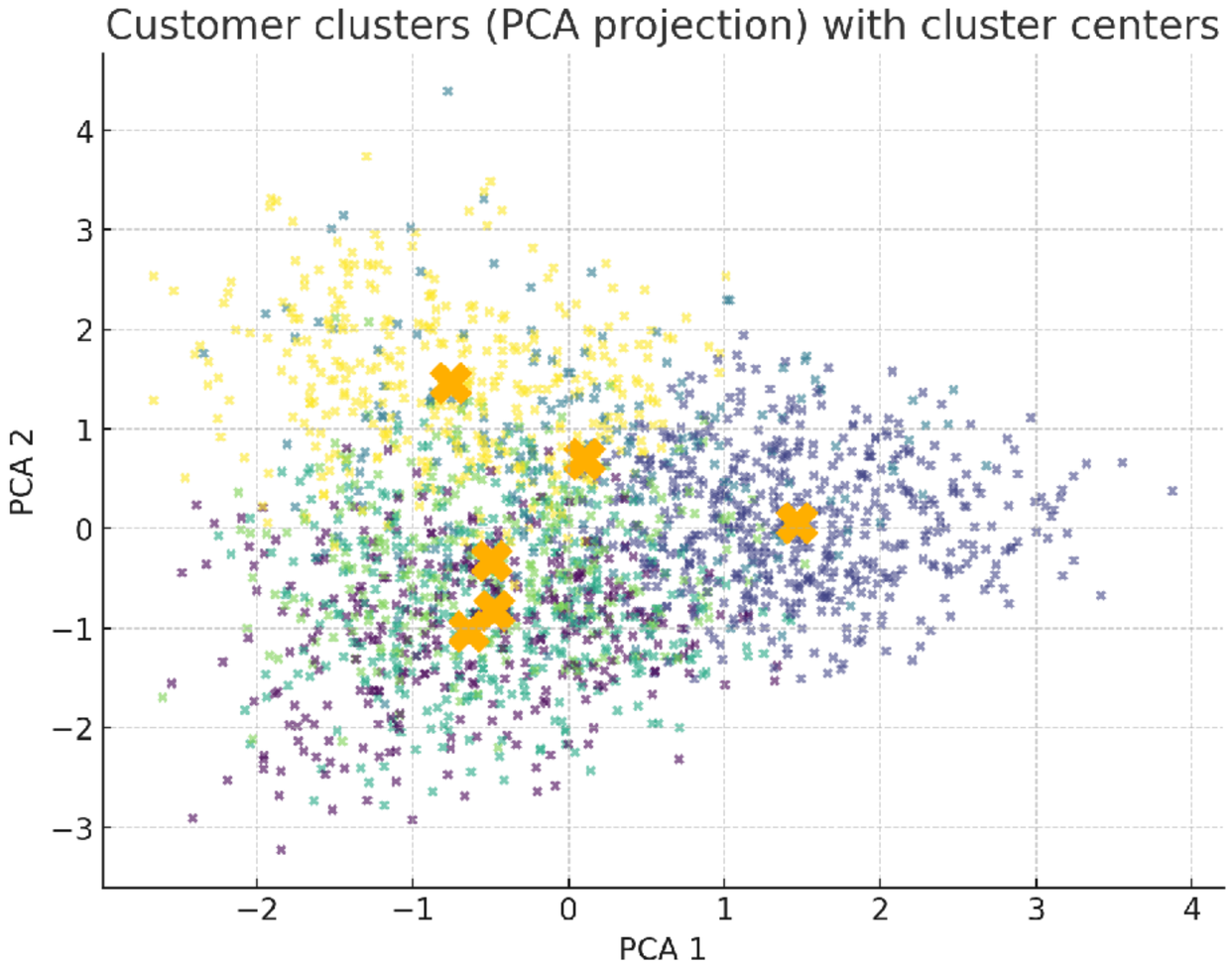
Clustering Algorithm



KMeans
Algorithm Explained

Step - 1 : Choose K

Randomly choosing the no. of clusters (K).



KMeans
Algorithm Explained

Step - 2 : Initialize centroids

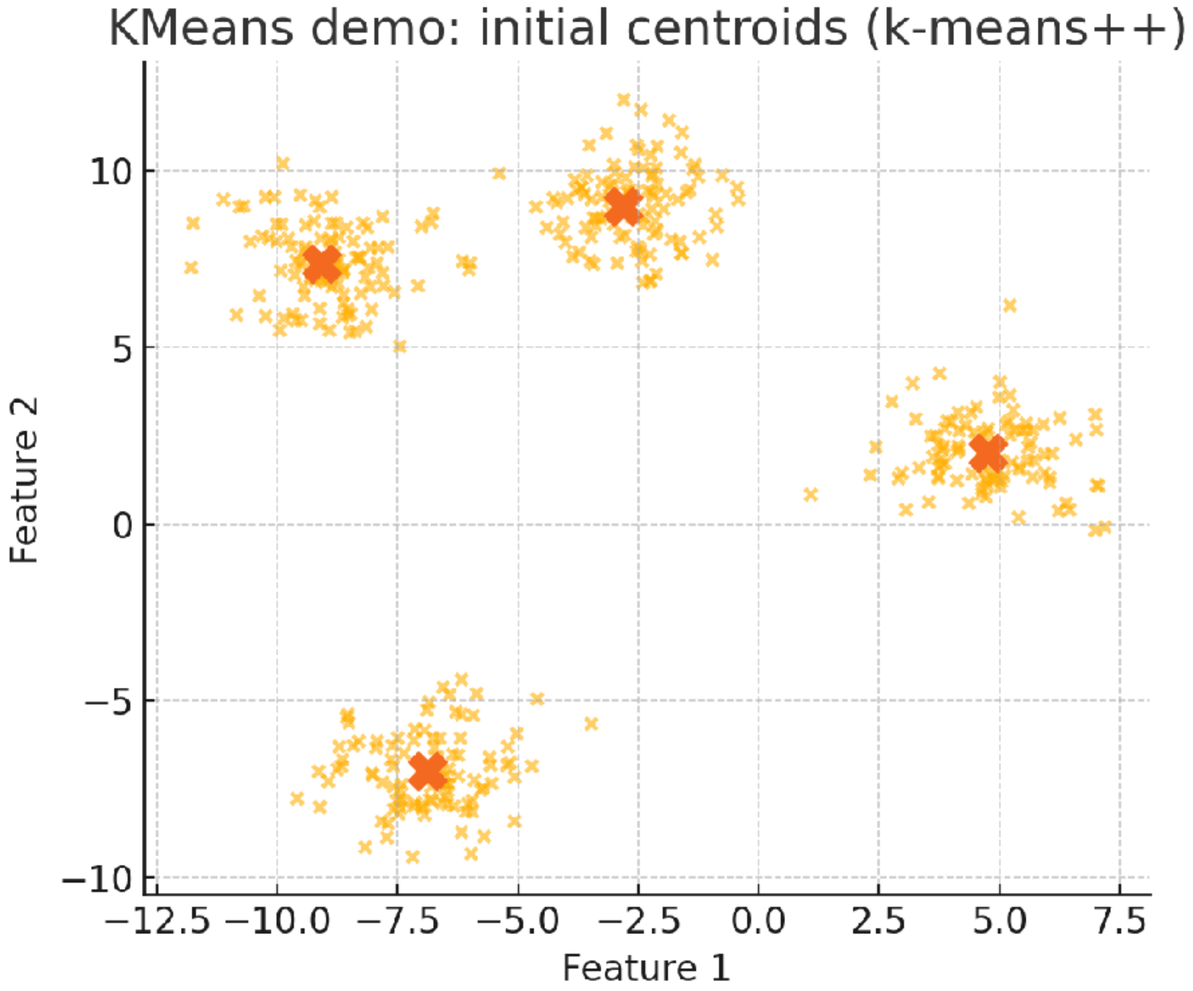
Pick K

K initial centroid vectors

- Simple option: random K points from data.
- Better option: **k-means++** — spreads initial centroids to reduce bad starts (recommended).
-

Step - 3 : Assignment step

Intuition: each point joins the cluster whose centroid is closest.



KMeans
Algorithm Explained

Step - 4 : Update step

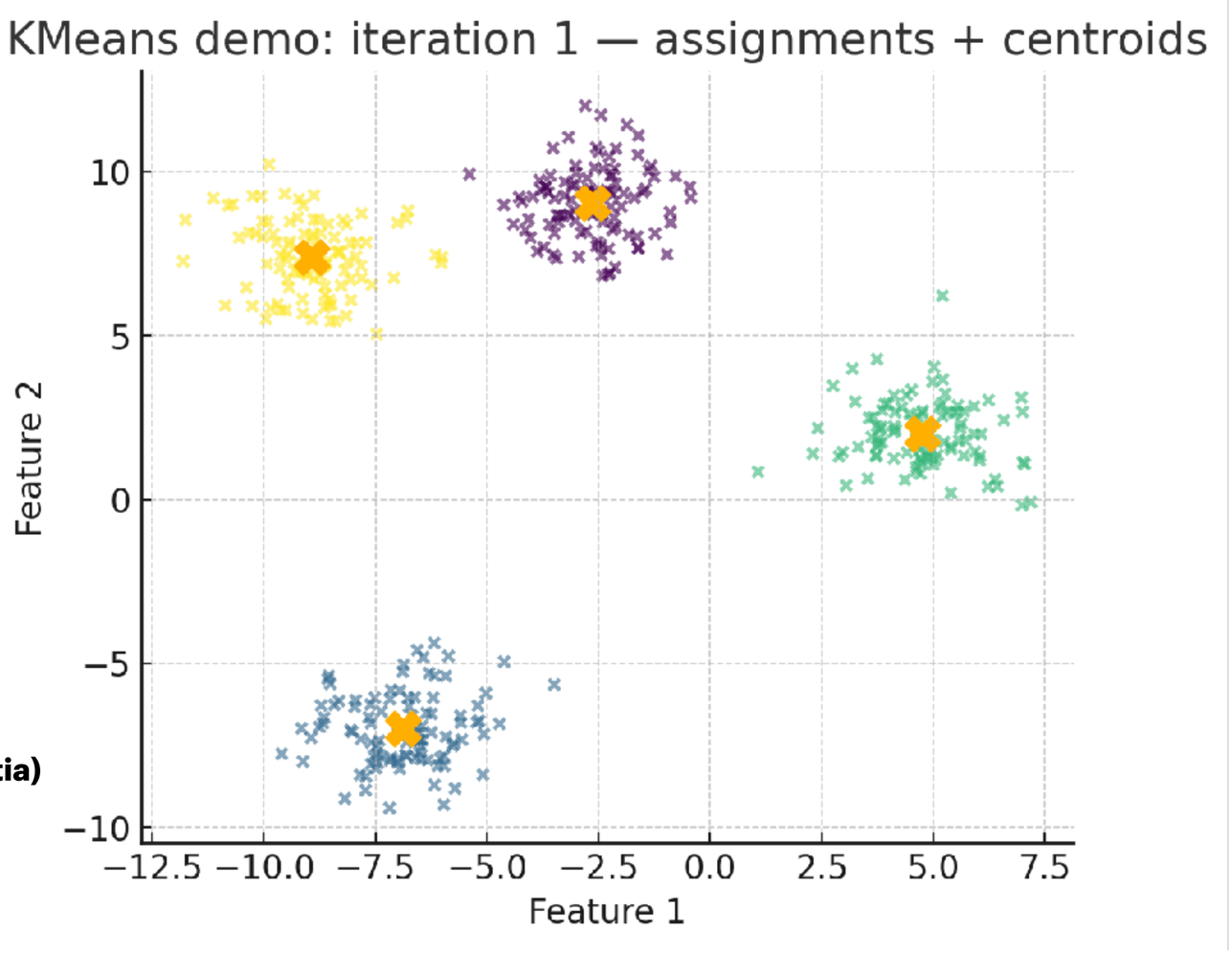
Intuition: move centroid to the center of its members.

Step - 5 : Repeat assignment + update

Assignments don’t change or centroids move very little, or a max iteration count is reached.

Step - 6 : Objective

KMeans minimizes the **within-cluster sum of squared distances (inertia)**



Machine Learning

Un - supervised : Dimensionality Reduction

Principle Component Analysis

PCA : Dimensionality Reduction