GITAM
(DEEMED TO BE UNIVERSITY)
(Estd. u/s 3 of the UGC Act, 1956)
VISAKHAPATNAM • HYDERABAD • BENGALURU

Meta Developer Circles
GITAM Visakhapatnam

GITAM
Career Guidance Center
GCGC

DECODING ML 2.0

Date & Time:

15th Mar 2023 - 03:00PM to 05:00PM
16th Mar 2023 - 03:00PM to 05:00PM
17th Mar 2023 - 04:00PM to 05:00PM

AGENDA

- Introduction & Recap
- Pandas Profile
- Pandas
- Plotting
- Feature Engineering and its types

RECAP

- Introduction to ML
- Overview of types of ML & its challenges
- Machine Learning Development Life Cycle
- Data Gathering – (CSV’s , Web Scraping, API’s)
- Types of Data
- EDA - (Univariate and Bivariate/Multivariate)
- Hands on Project

PANDAS PROFILER

Pandas Profiler is an open-source Python package that generates a report with descriptive statistics of a given Pandas DataFrame.

Overview: General information about the dataset

Variables: A detailed description of each variable in the dataset, including its name, type, and data distribution.

Correlation: A correlation matrix showing the pairwise correlations between all numerical variables in the dataset**.**

Missing Values: a table that shows the number and percentage of missing values in each variable.

Sample: a table that shows a random sample of observations from the dataset.

PANDAS

- Pandas is a popular open-source Python library used for data manipulation and analysis.
- There are two types of DataStructures
    a. Series
    b. DataFrames

Id	SepalLengthCm	SepalWidthCm	PetalLengthCm
1	5.1	3.5	1.4
2	4.9	3	1.4
3	4.7	3.2	1.3
4	4.6	3.1	1.5
5	5	3.6	1.4
6	5.4	3.9	1.7
7	4.6	3.4	1.4

MACHINE LEARNING DEVELOPMENT LIFE CYCLE

1. Data Gathering
2. Data Preprocessing
3. Exploratory Data Analysis
4. Feature Engineering
5. Model Training and Testing
6. Model Deployment
7. Testing

FUNCTIONS

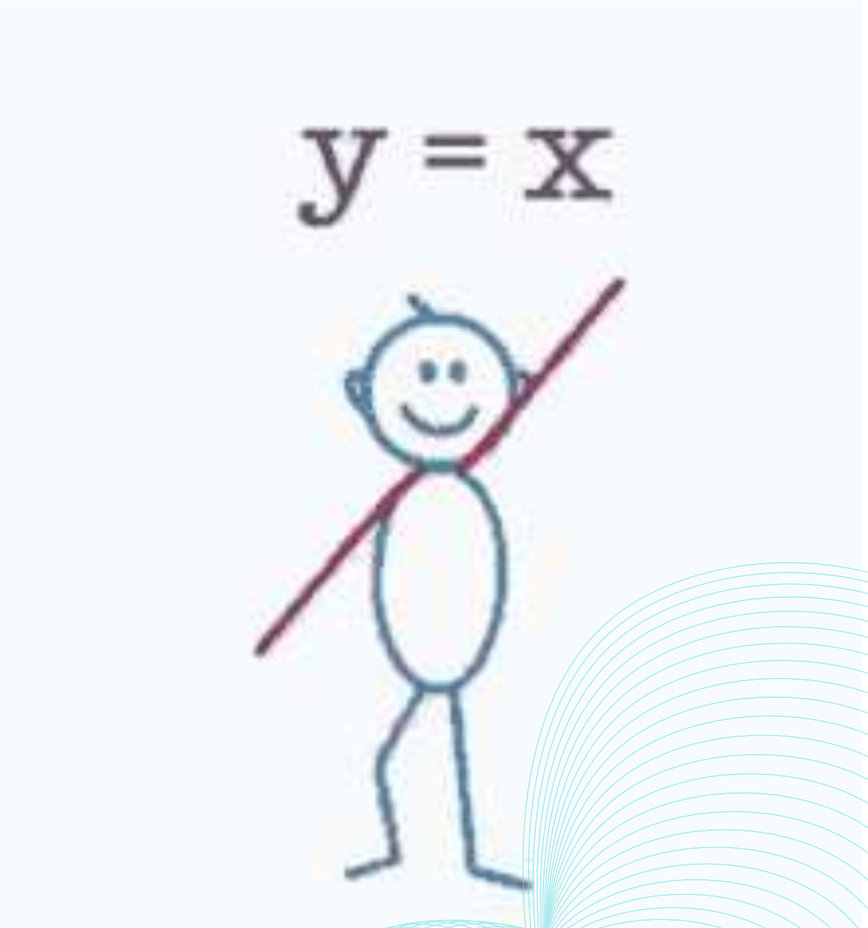
Gathering Data:
read_csv(), read_sql(), to_csv(), to_sql()

Data Preprocessing:
dropna(), fillna(), drop_duplicates() , quantile()

Data Exploration:
head(), tail()

Data Manipulation:
loc[] and iloc[], merge(), sort_values(), groupby()

Model Train and Test:
train_test_split()

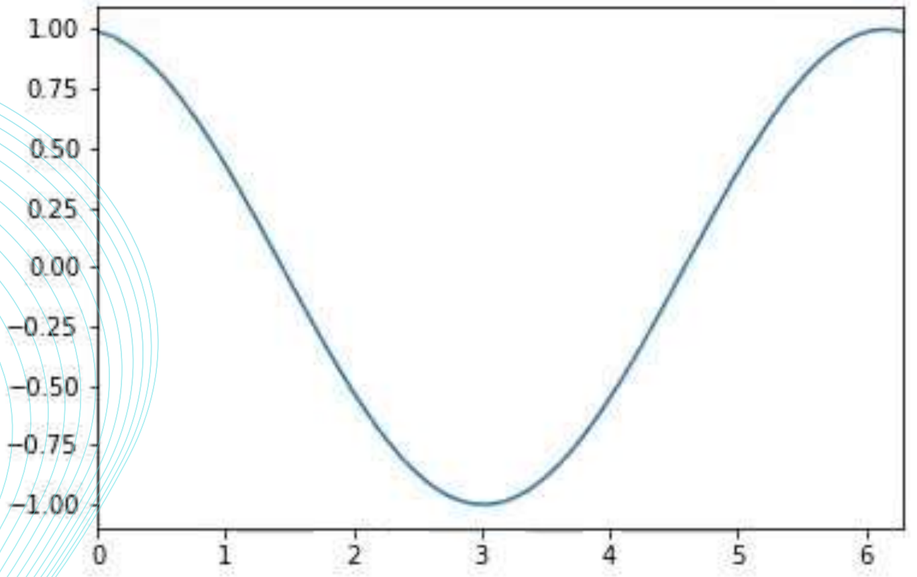


PLOTS

matplotlib

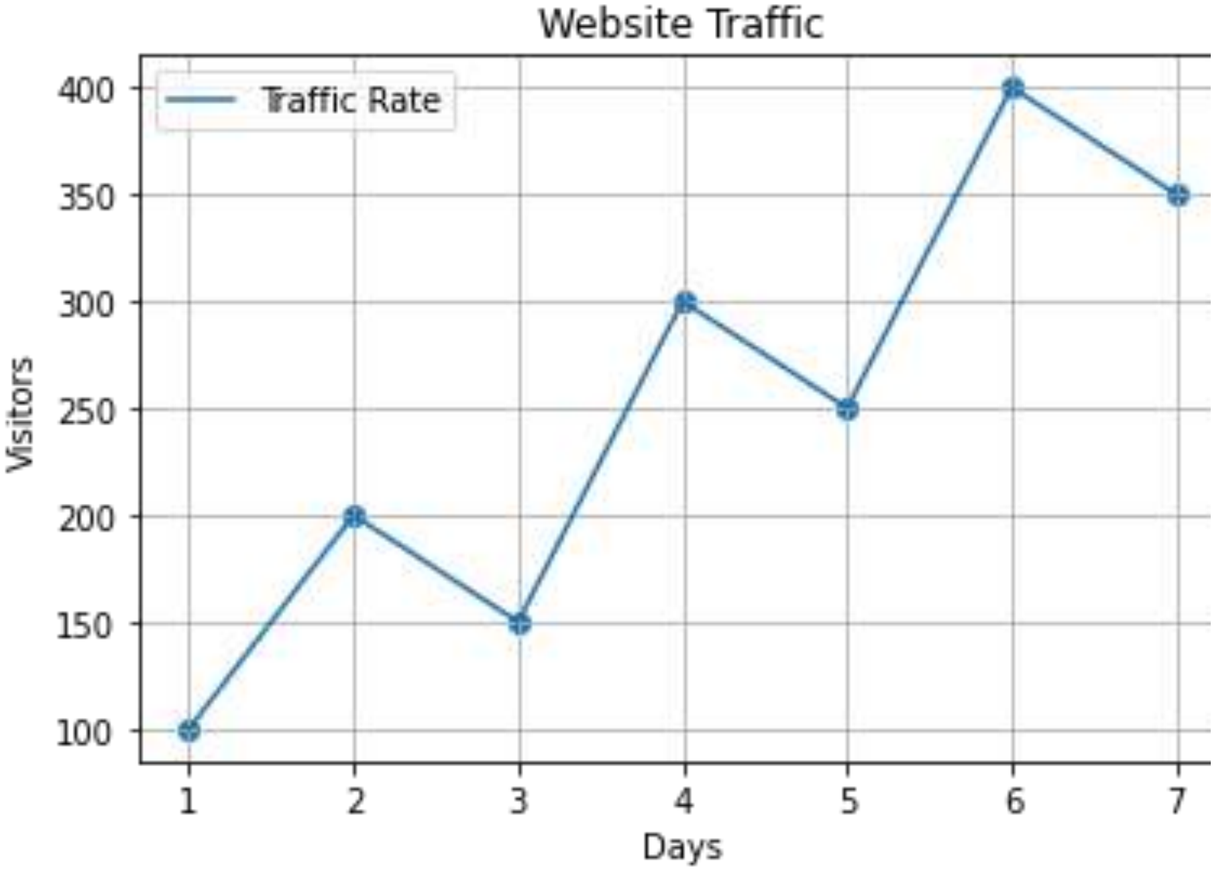
seaborn

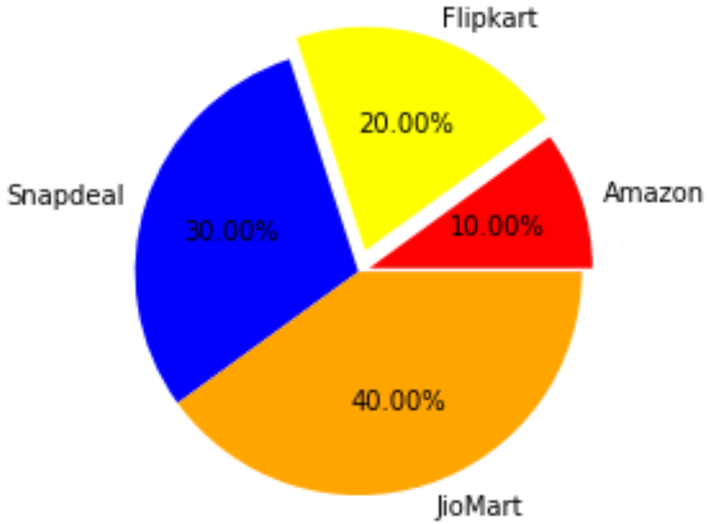
- Plotting is the process of creating a visual representation of data using graphs, charts, or other types of visual aids.
- It is used to help visualize and understand patterns, trends, and relationships in data, and to communicate this information effectively to others.



FUNCTIONS

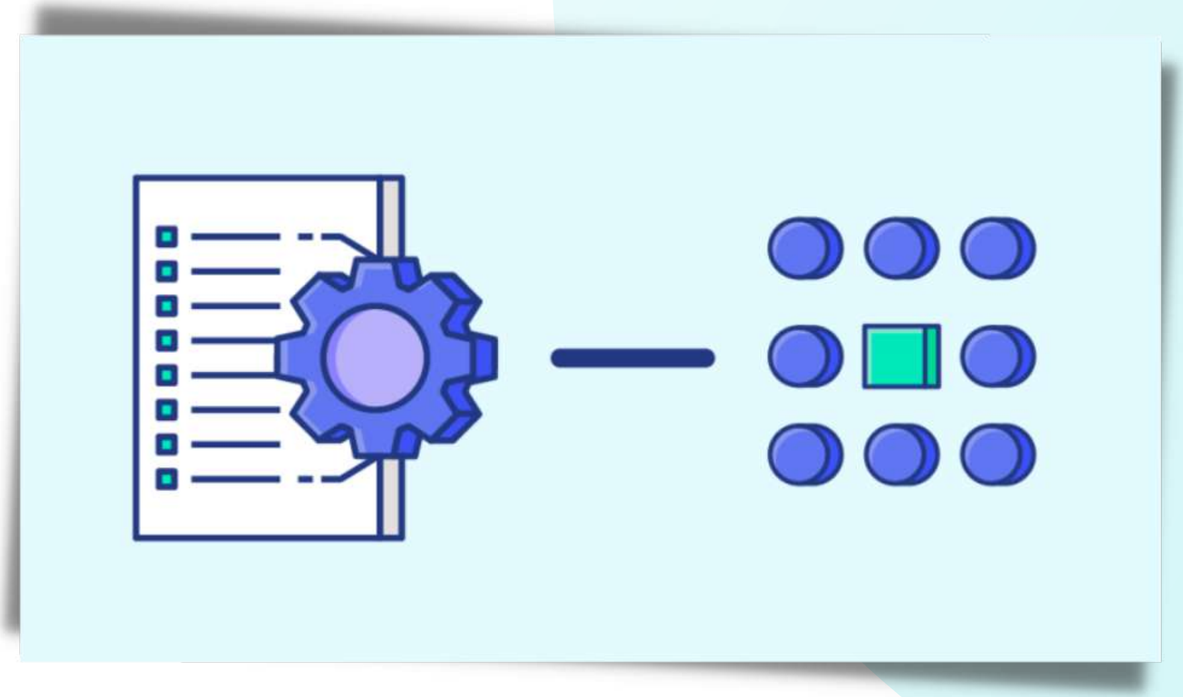
- scatter
- plot
- show
- xlabel
- ylabel
- legend
- grid
- axis
- title
- pie



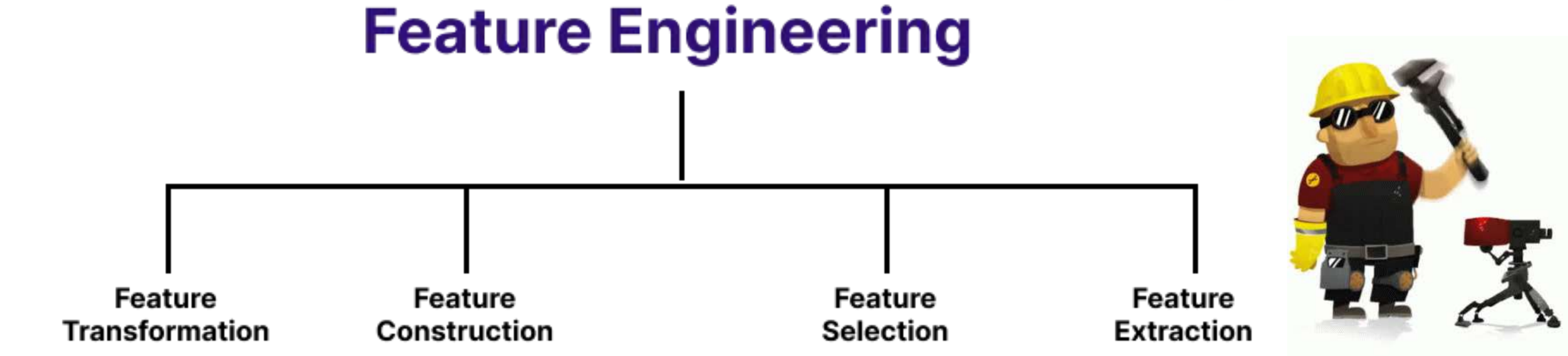


FEATURE ENGINEERING

Feature engineering is the process of using domain knowledge to extract features from raw data. These features can be used to improve the performance of machine learning model



TYPES OF FEATURE ENGINEERING



- Missing values
- Outlier detection
- Feature scaling

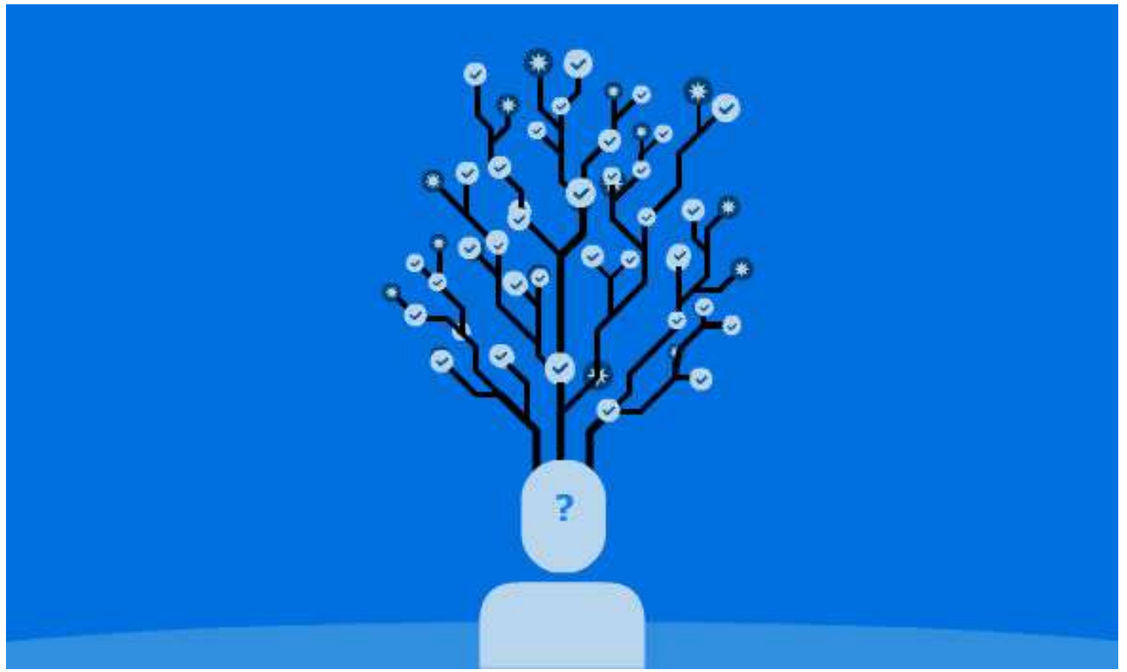
Combining two features to make single column

Selecting only the required features

Out of the given features create new features

FEATURE TRANSFORMATION

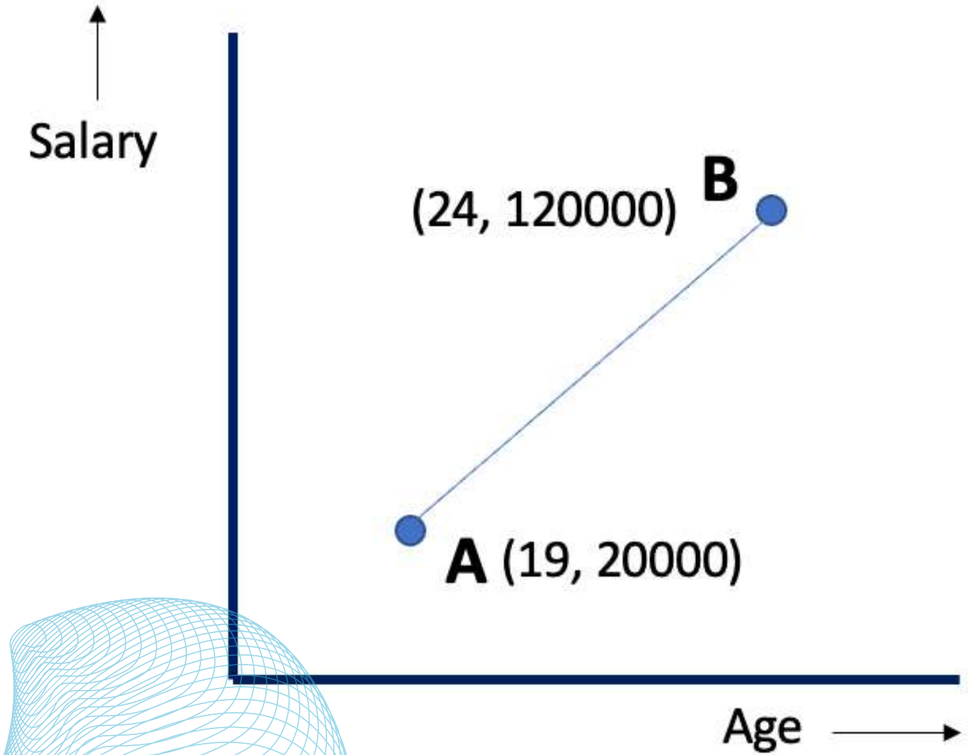
- Missing Values
- Categorical Values
- Outlier Detection
- Feature Scaling



FEATURE SCALING

Age	Salary	Purchase
19	80000	1
20	120000	1
17	20000	0
40	50000	0
12	160000	1

FEATURE SCALING



Get it to a common range

Distance = sqrt((y2-y1)^2 + (x2-x1)^2)

(120000 - 20000) ^ 2 = 10^10
(24 - 19) ^ 2 = 25

FEATURE ENCODING

Data

Numerical

Categorical

Nominal

One - Hot Encoder

Ordinal Data

Ordinal Encoder

ORDINAL ENCODING

Roll Number	Grades	Grades_new
12201020022	A+	1
12201020023	O	2
12201020024	A	0
12201020025	A+	1
12201020026	A	0
12201020027	A	0

Based on the order of the categories

O	2
A+	1
A	0

ONE-HOT ENCODING

Name	Section	Section_Encoded	
Student **P**	B13	0	1
Student **Q**	B26	1	0
Student **R**	B39	1	1
Student **S**	B13	0	1
Student **T**	B26	1	0

The categories are not related

HANDLING MISSING VALUES

Strategies :

- Deletion
- Mean/Median
- Most Frequent

SUMMARY

GITAM
(DEEMED TO BE UNIVERSITY)
(Estd. u/s 3 of the UGC Act, 1956)
VISAKHAPATNAM • HYDERABAD • BENGALURU

Meta Developer Circles
GITAM Visakhapatnam

GITAM
Career Guidance Center
GCGC

DECODING ML 2.0 Day 2

Date & Time:

15th Mar 2023 - 03:00PM to 05:00PM
16th Mar 2023 - 03:00PM to 05:00PM
17th Mar 2023 - 04:00PM to 05:00PM

AGENDA

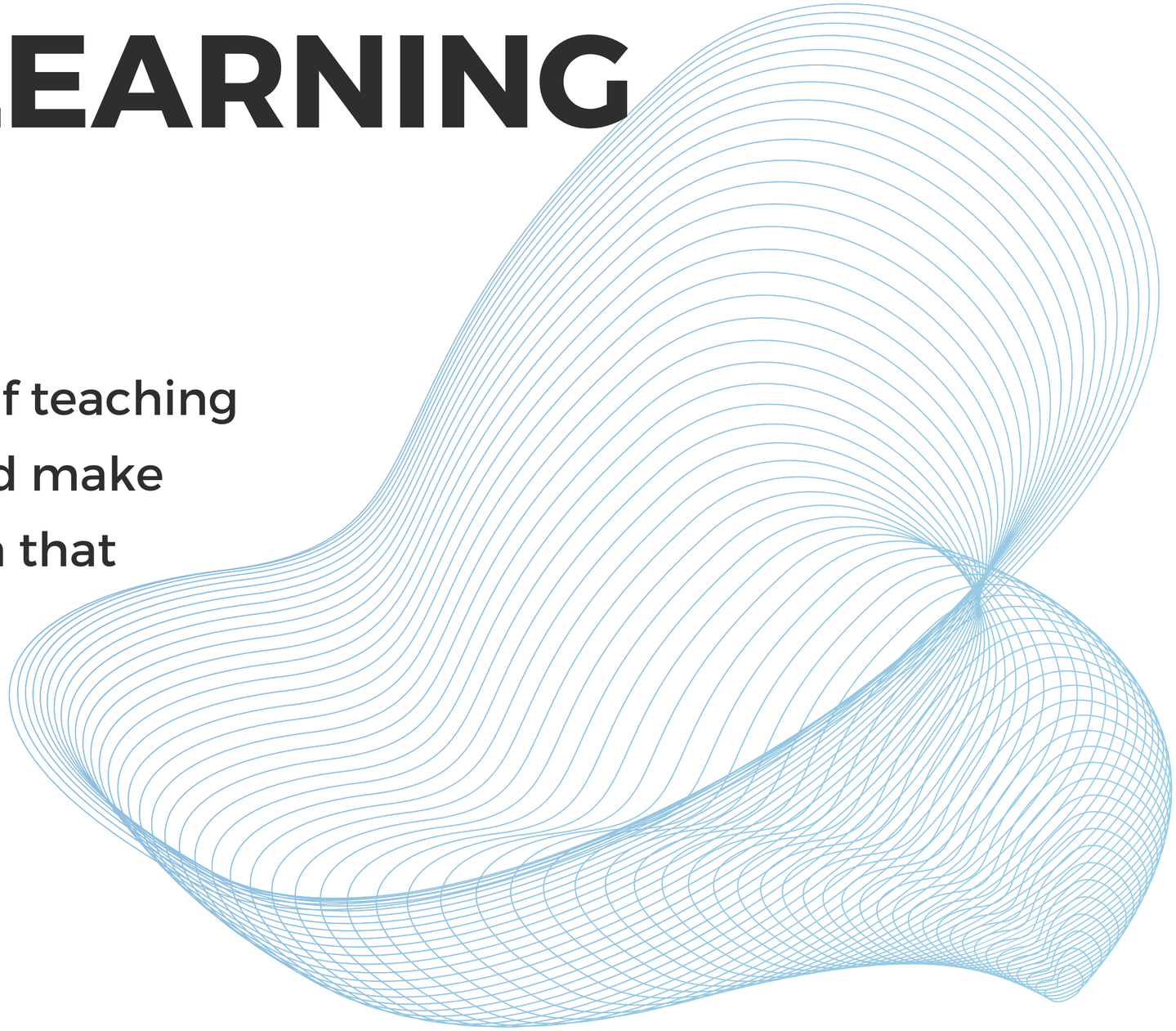
- Recap
- Types of Machine Learning
- Supervised Learning and its Types
- Un-Supervised Learning and its Types
- Hands-On Projects

RECAP

- Introduction to ML
- Evolution of ML
- Types of Machine Learning

MACHINE LEARNING

Machine learning is the process of teaching computers to learn from data and make predictions or decisions based on that learning.

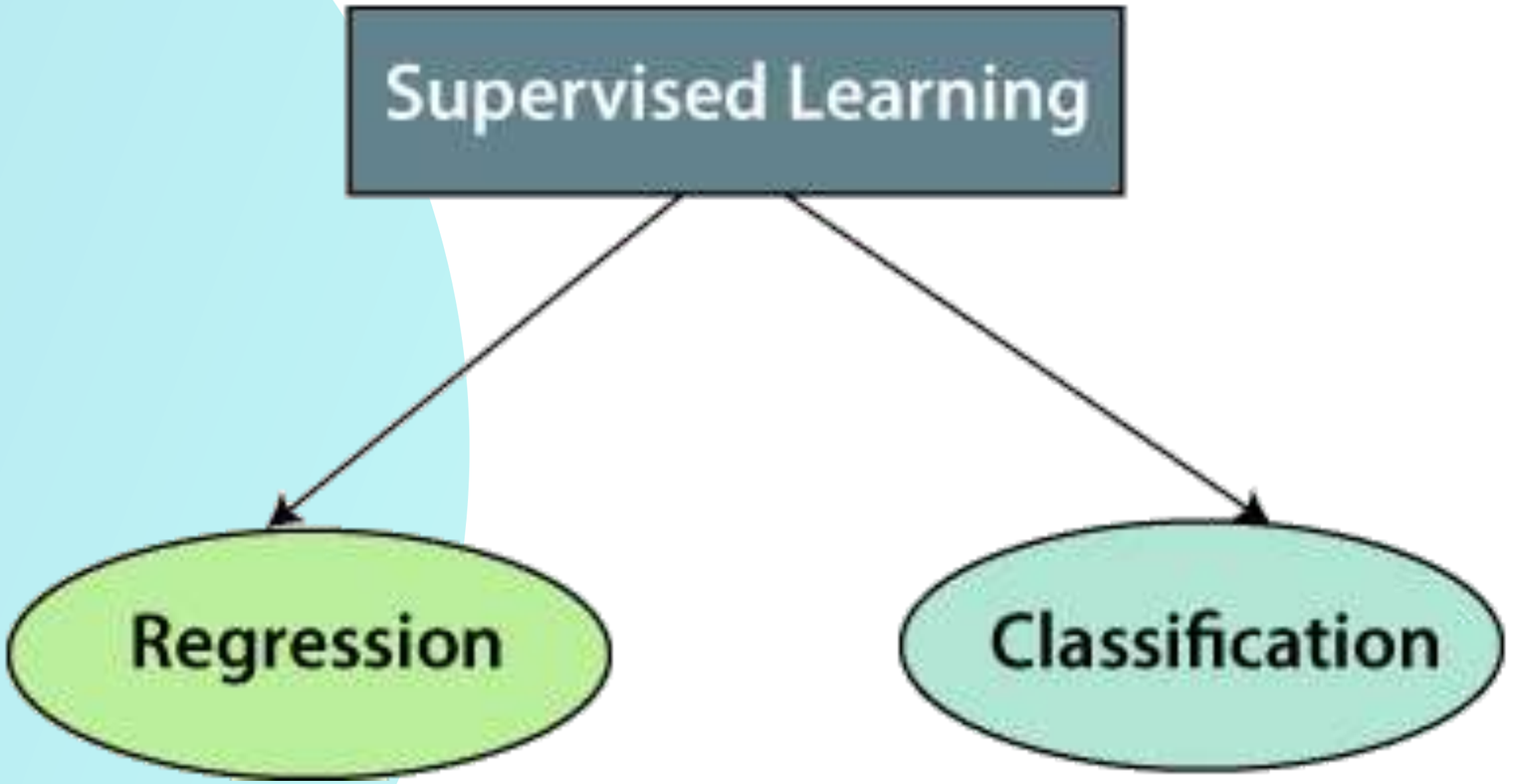


Types of Machine Learning

- Supervised Learning
- Unsupervised Learning
- Semi - Supervised Learning
- Reinforcement Learning

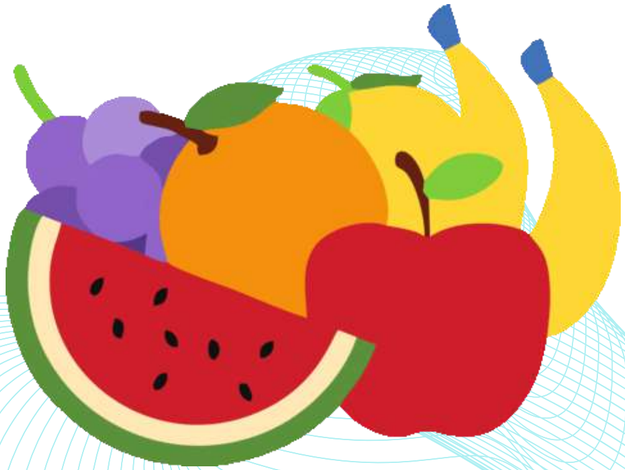
SUPERVISED LEARNING

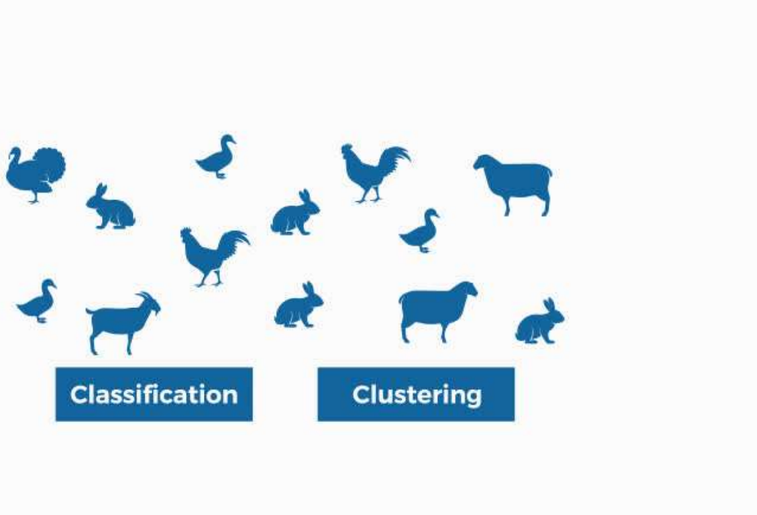
Supervised learning is a type of machine learning in which an algorithm is trained on a labeled dataset.


Supervised Learning
Regression
Classification

CLASSIFICATION

Classification is a type of supervised learning in machine learning, where the goal is to predict a discrete target variable or class label for new input data.





FEW CLASSIFICATION ALGORITHMS

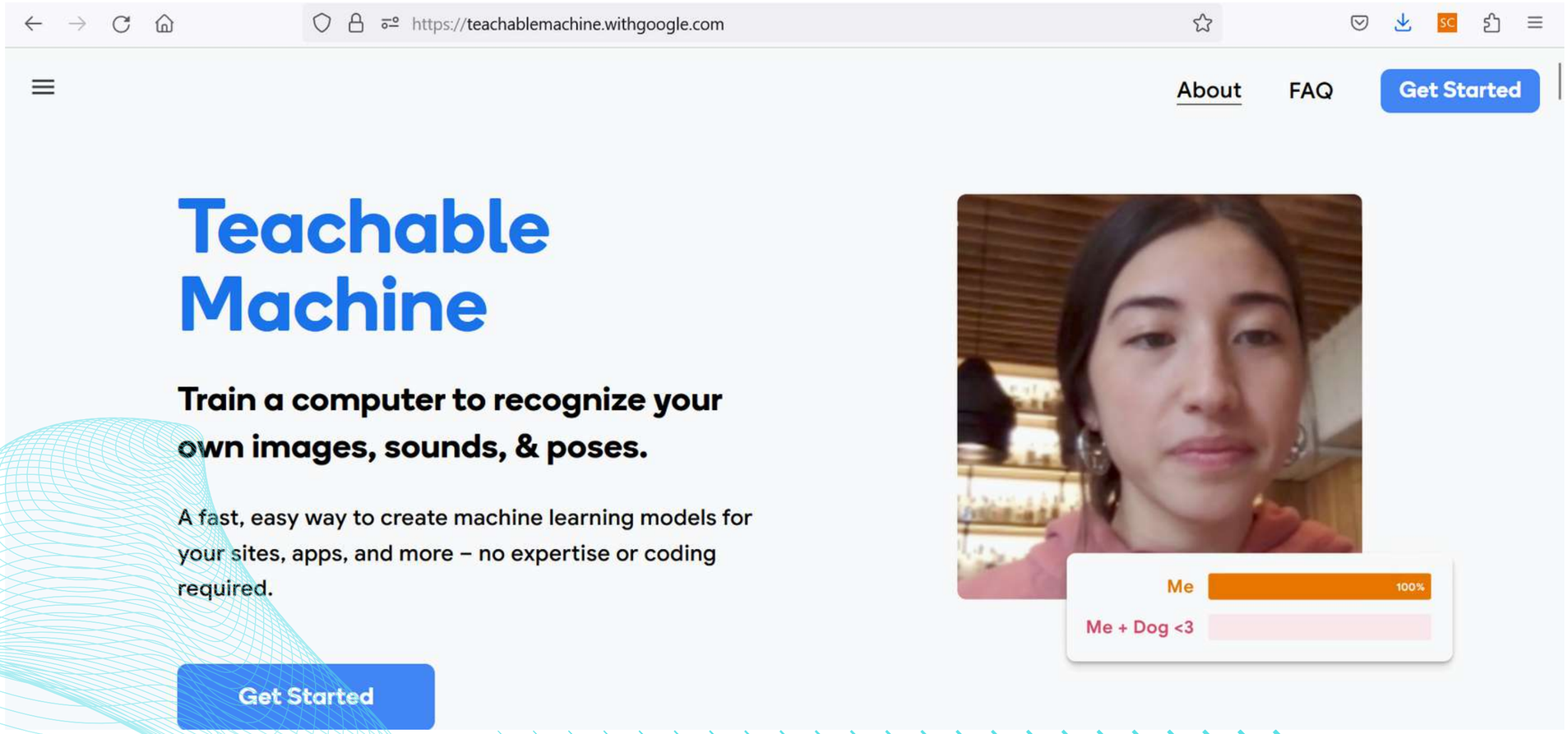
- Logistic Regression
- Decision Trees
- Random Forest
- Support Vector Machine

CLASSIFICATION IRL

1. Image Classification
2. Fraud detection
3. Product recommendation
4. Credit risk assessment
5. Email spam filtering

CLASSIFICATION

A live example



REGRESSION

Regression is a type of supervised learning in machine learning, used to predict a continuous numerical output variable for new input data.

FEW REGRESSION ALGORITHMS

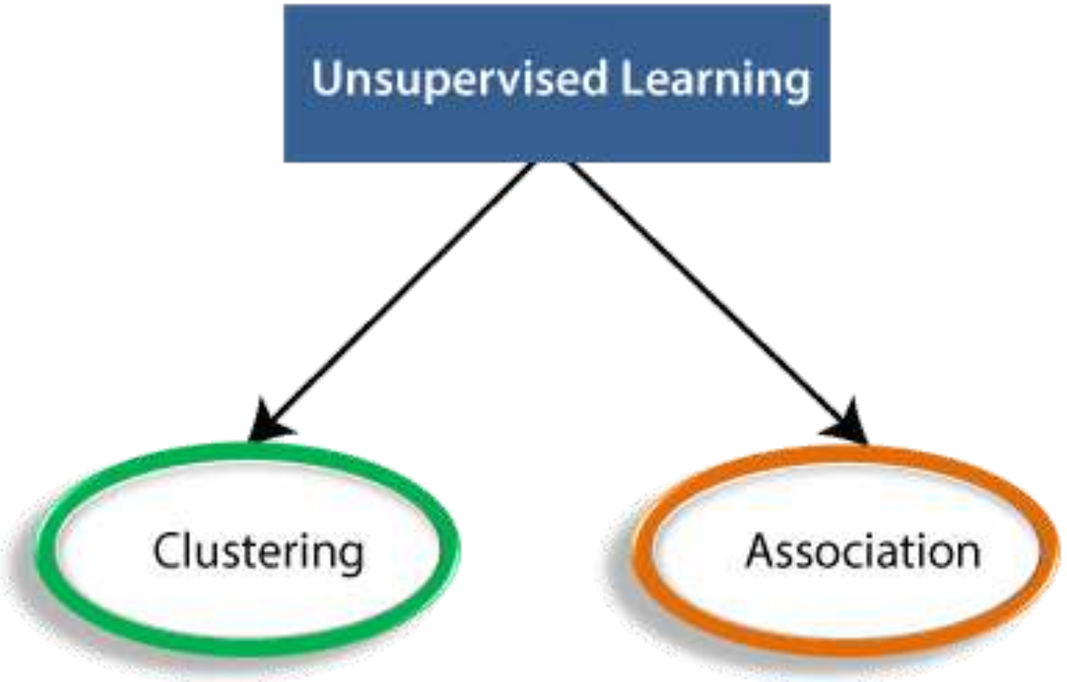
- Simple Linear Regression
- Polynomial Regression
- Multiple Linear Regression
- Random Forest

REGRESSION IRL

1. Weather Forecasting
2. Sales Forecasting
3. Customer Behavior Analysis
4. Healthcare Analysis
5. Financial Analysis

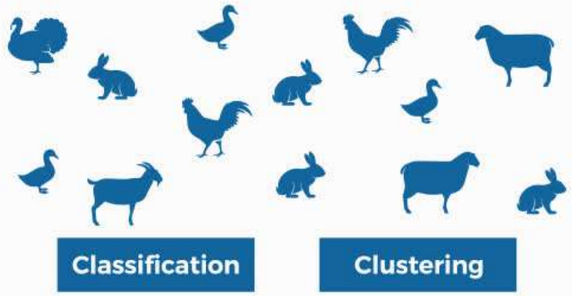
UN-SUPERVISED LEARNING

Unsupervised machine learning, uses machine learning algorithms to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention.


Unsupervised Learning
Clustering
Association

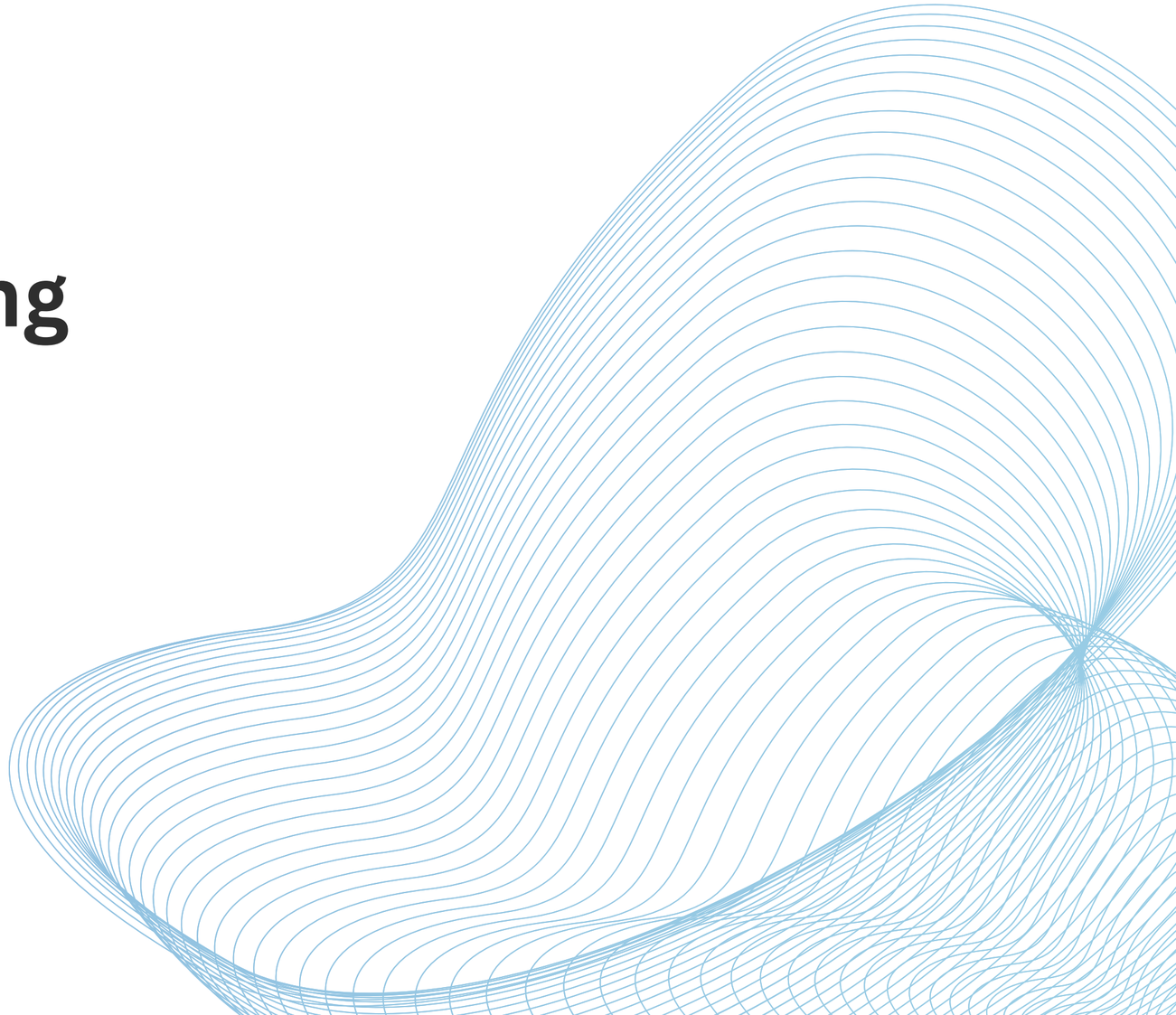
CLUSTERING

- Clustering is a technique used in machine learning to group similar data points together
- clustering is a powerful technique for discovering patterns in data



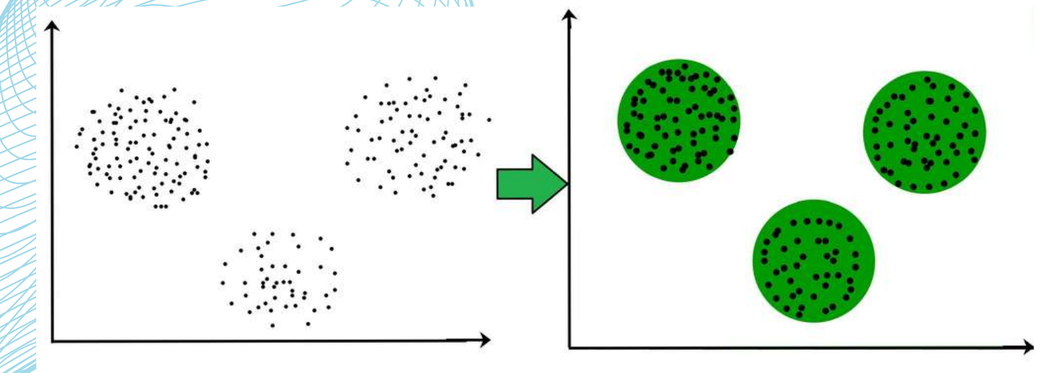
PRIMARY ALGORITHMS OF CLUSTERING

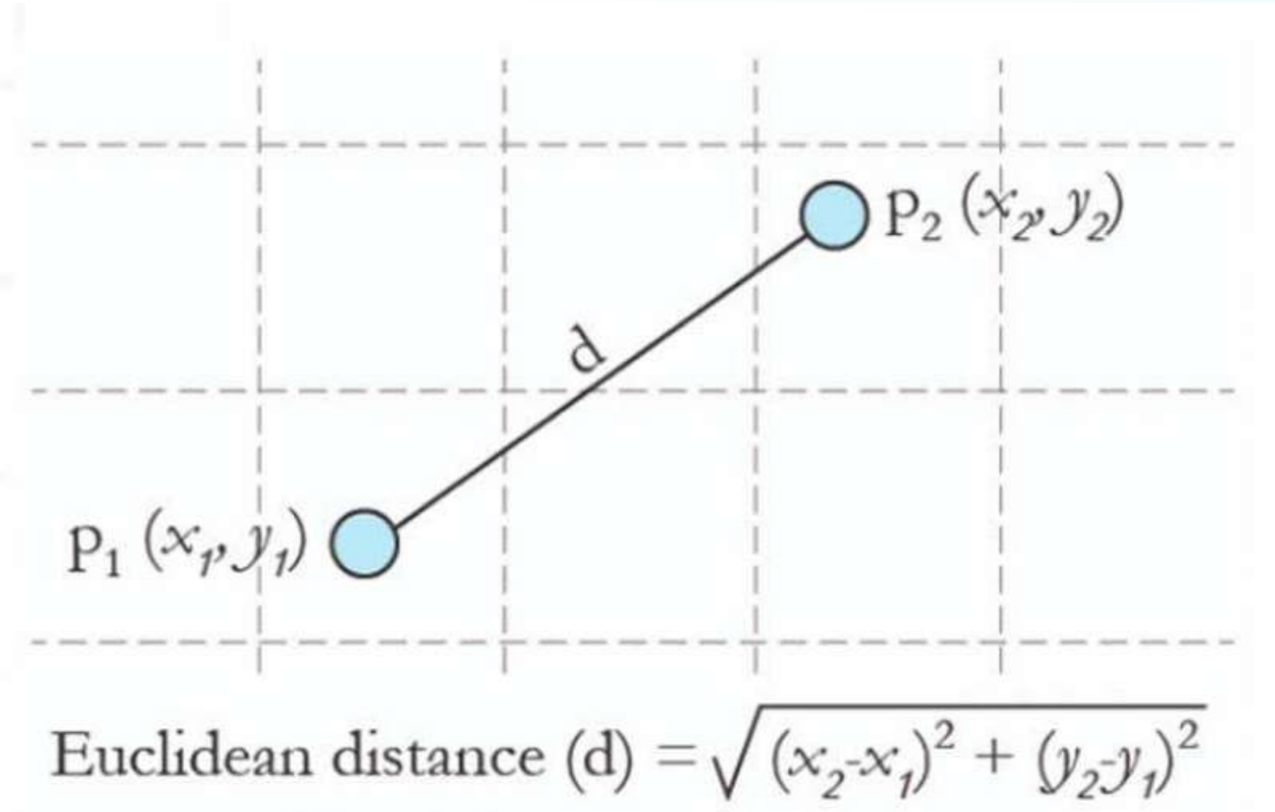
- K-means clustering
- Hierarchical clustering
- Density-based clustering



K-MEANS CLUSTERING

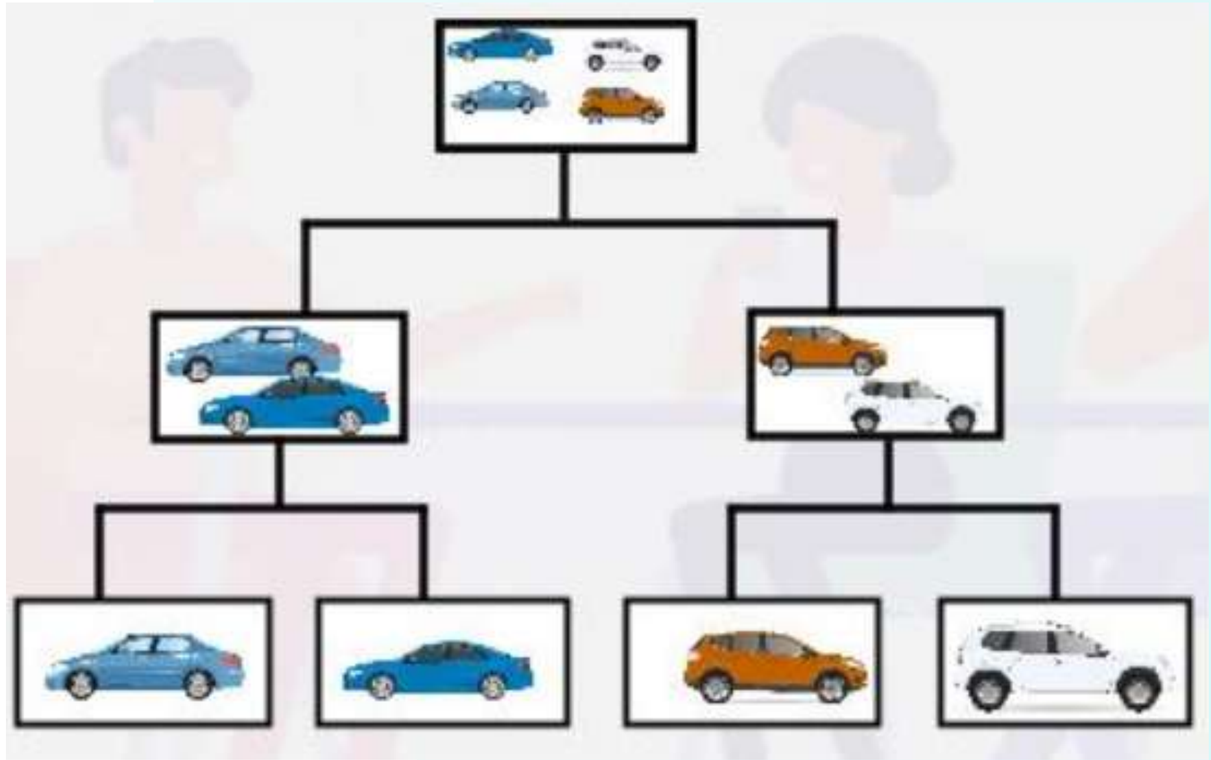
- This algorithm groups data points into k clusters based on their similarity to each other
- It works by iteratively updating the centroids of the clusters until convergence is reached.

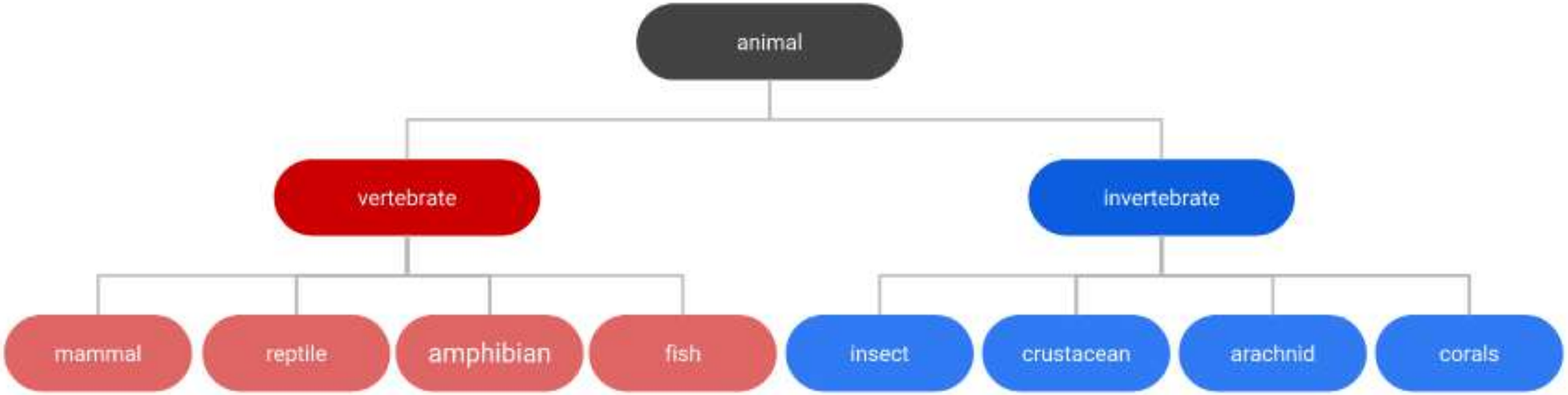


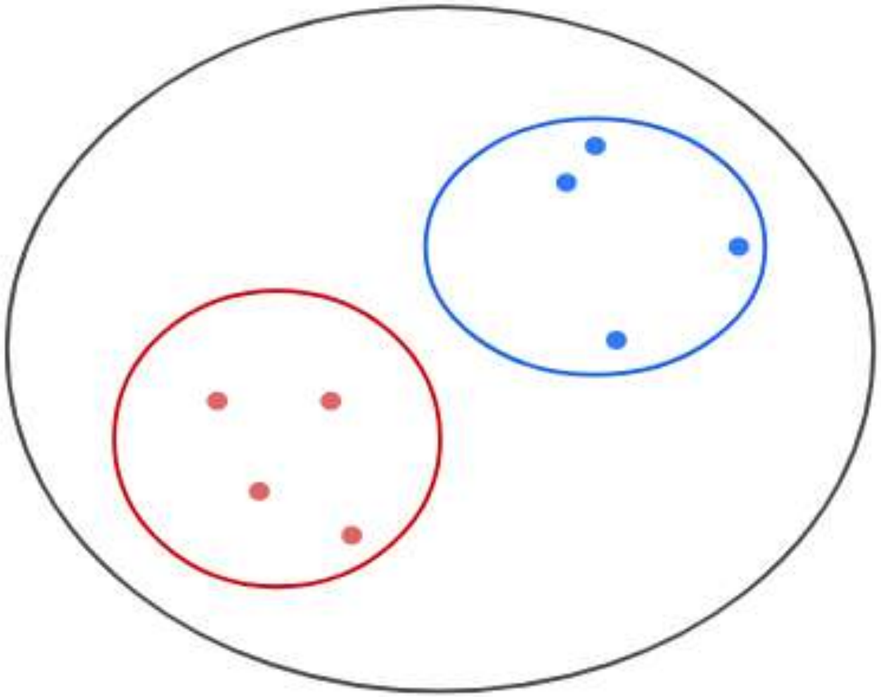


Hierarchical clustering

This algorithm builds a hierarchy of clusters, either by starting with individual data points and merging them into clusters, or by starting with all data points in one cluster and recursively splitting them

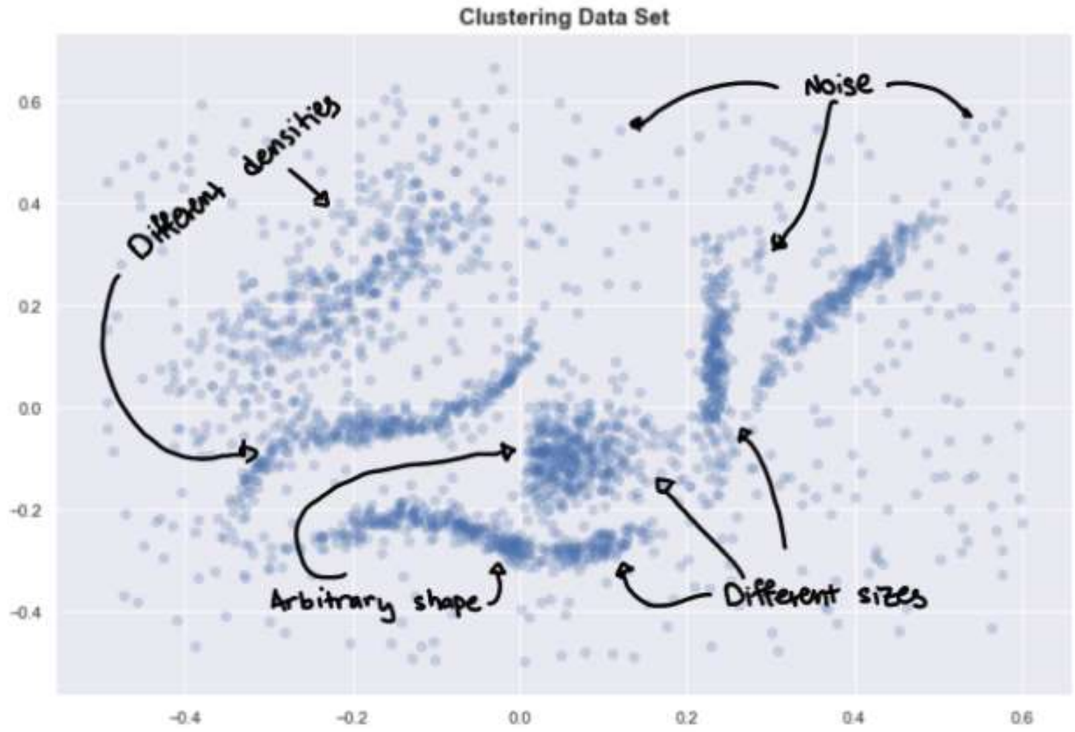



animal
vertebrate
invertebrate
mammal
reptile
amphibian
fish
insect
crustacean
arachnid
corals



Density-based clustering

This algorithm groups data points based on their proximity to each other in terms of density. It is useful for detecting clusters of arbitrary shape.

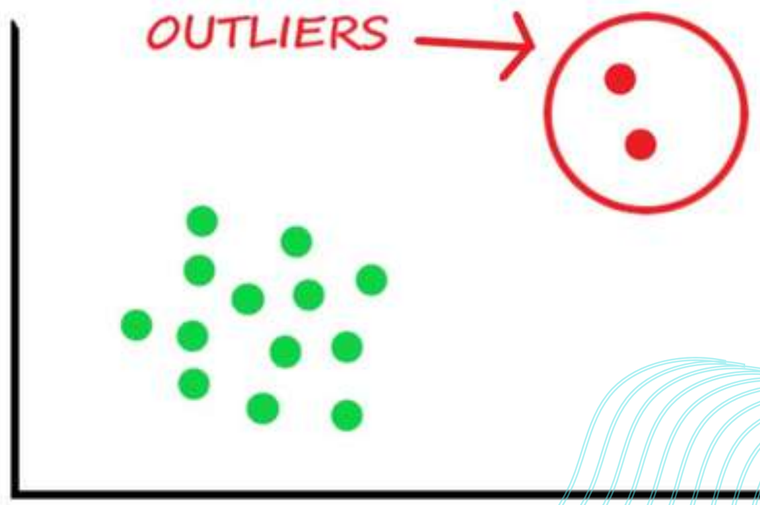


APPLICATIONS

Image segmentation: Clustering can be used to group pixels together based on their color or texture, which can then be used to segment an image into distinct regions.

APPLICATIONS

Anomaly detection: Clustering can be used to detect outliers or anomalies in a dataset, which may indicate errors or fraud.



ASSOCIATION

- Machine learning technique used to discover interesting relationships between variables in a dataset

Key measures used in association analysis

- **Support:** This measures the frequency of a particular itemset in the dataset. It represents the proportion of transactions that contain the itemset.

- **Confidence:** This measures the strength of the association between two itemsets. It represents the proportion of transactions that contain both itemsets, divided by the proportion of transactions that contain the first itemset.

GITAM
(DEEMED TO BE UNIVERSITY)
(Estd. u/s 3 of the UGC Act, 1956)
VISAKHAPATNAM • HYDERABAD • BENGALURU

Meta Developer Circles
GITAM Visakhapatnam

GITAM
Career Guidance Center
GCGC

DECODING ML 2.0 Day 3

Date & Time:

15th Mar 2023 - 03:00PM to 05:00PM
16th Mar 2023 - 03:00PM to 05:00PM
17th Mar 2023 - 04:00PM to 05:00PM

AGENDA

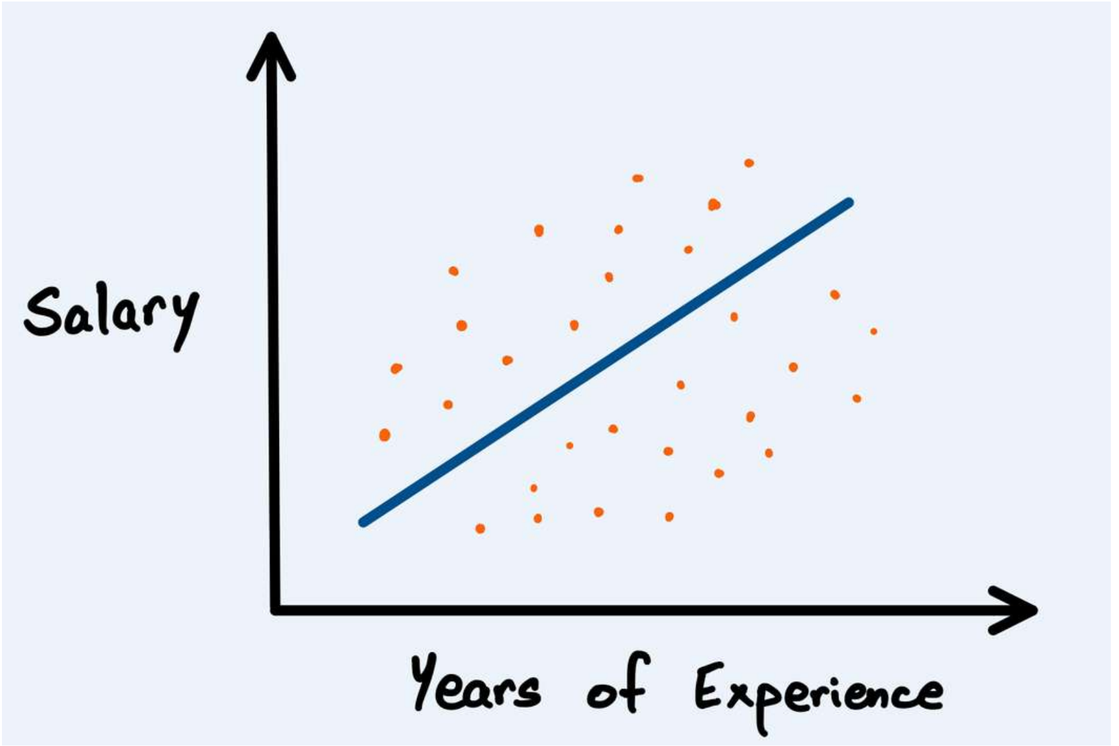
- Introduction & Recap
- Simple Linear Regression Evaluation Metrics
  - MSE
  - RMSE
  - MAE
  - R2 score
- Cross-validation
- Confusion Matrix
  - Precision and Recall
  - F1 Score

RECAP

- What is Machine Learning?
- Evolution of ML
- Types of Machine Learning
- MLDLC (ML Development Life Cycle)
- Exploratory Data Analysis (EDA)
- Feature Engineering
- Simple Linear Regression and KNN

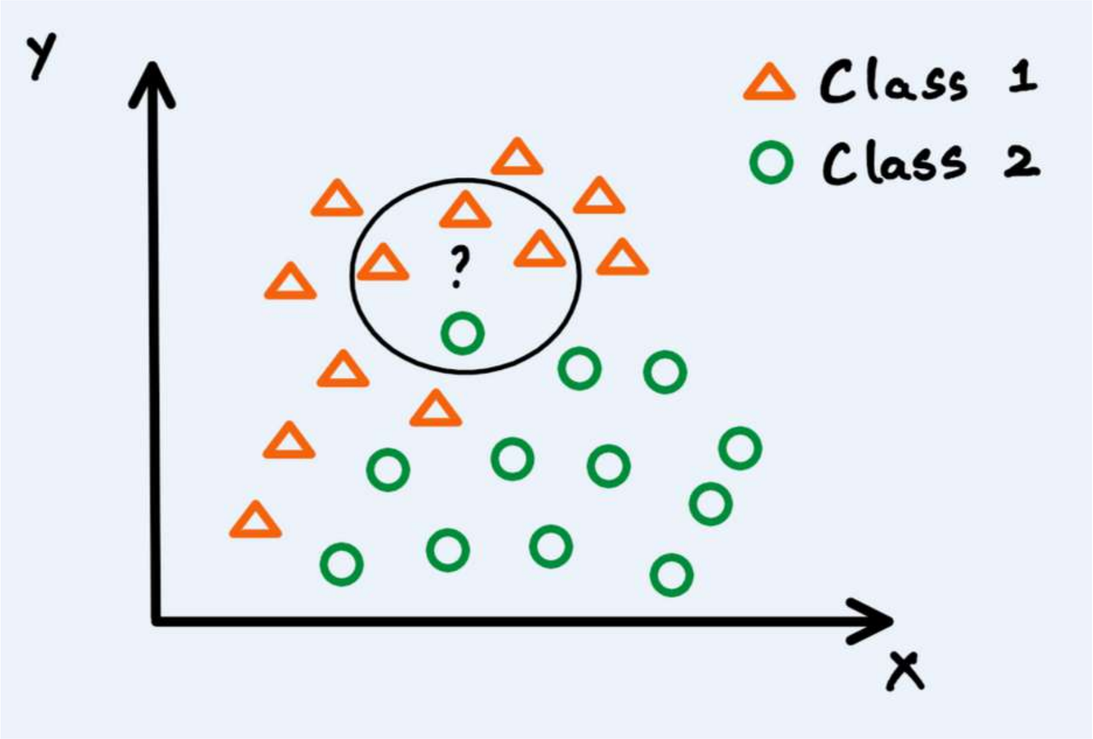
LINEAR REGRESSION

Linear regression is a type of predictive analysis that uses one or more variables to estimate the value of another variable.



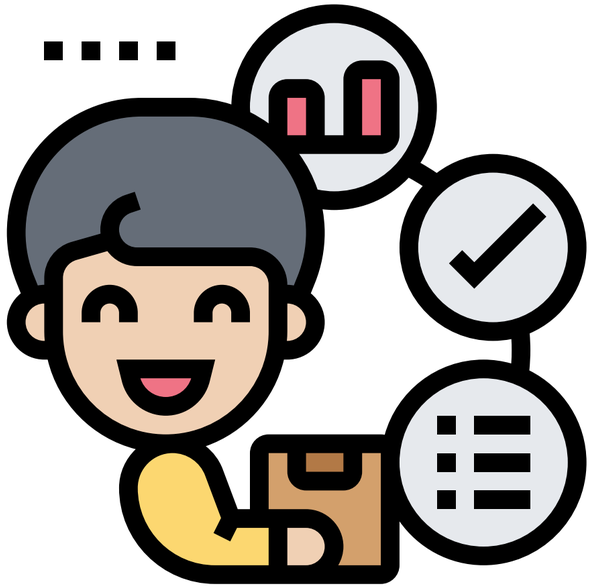
K-NEAREST NEIGHBORS (KNN)

KNN is a machine learning algorithm that uses the distance or similarity between data points to assign them to a group or class.



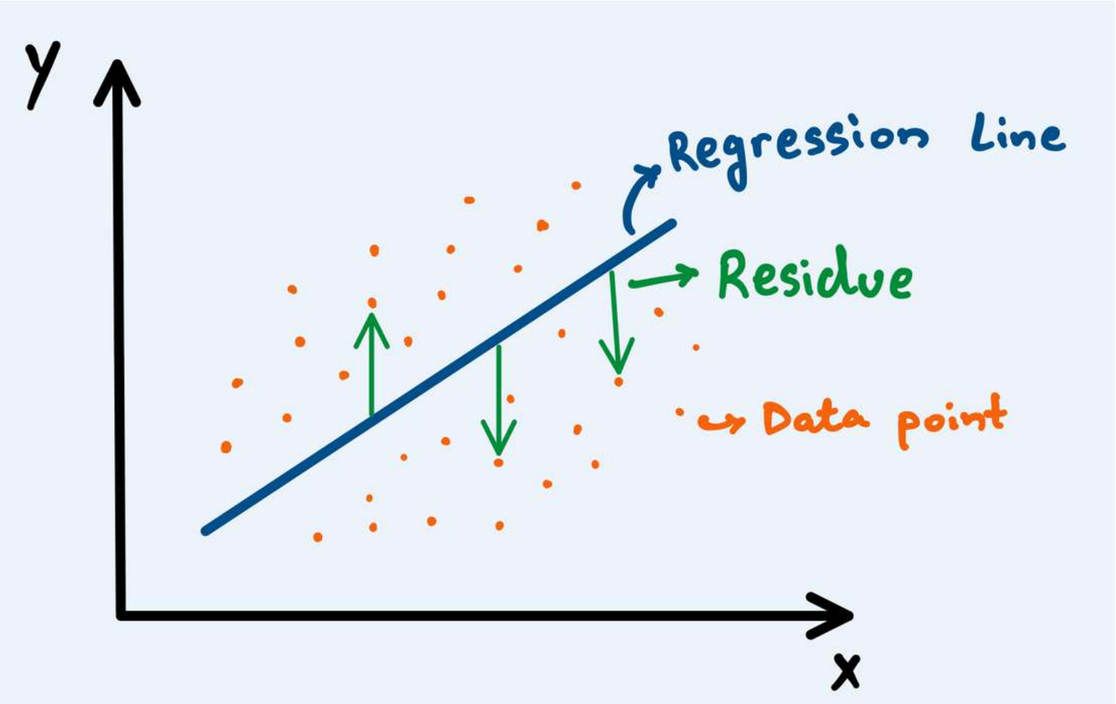
EVALUATION METRICS FOR SIMPLE LINEAR REGRESSION

- Mean Squared Error
- Root Mean Squared Error
- Mean Absolute Error
- R2 Score

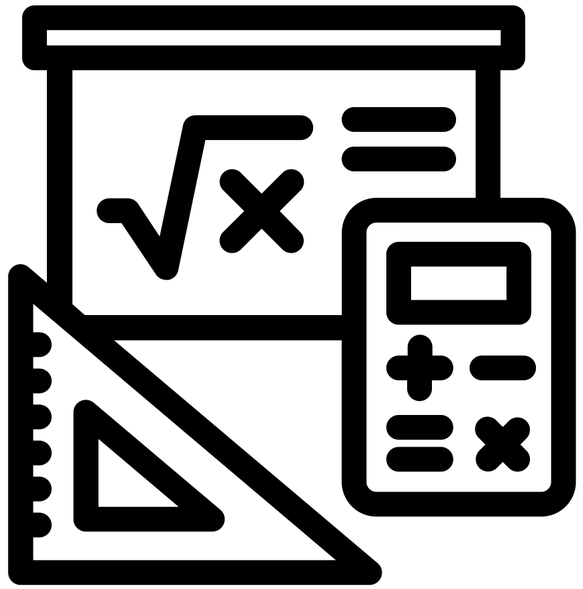


WHAT ARE RESIDUALS?

Residual is the difference between the actual target value and the fitted value.



FORMULAS



MSE

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

MAE

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

RMSE

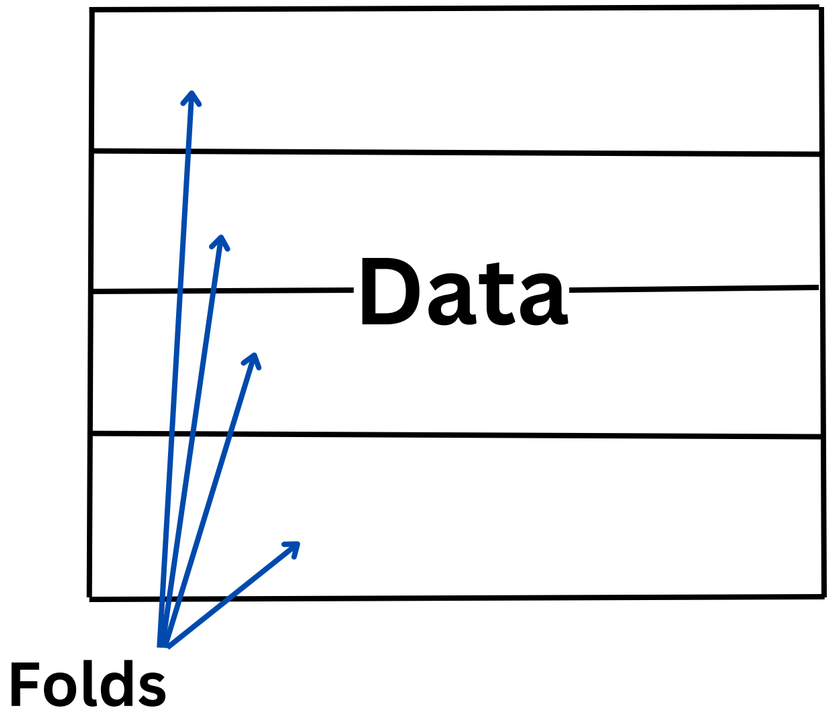
$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

R Squared Error

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

CROSS VALIDATION

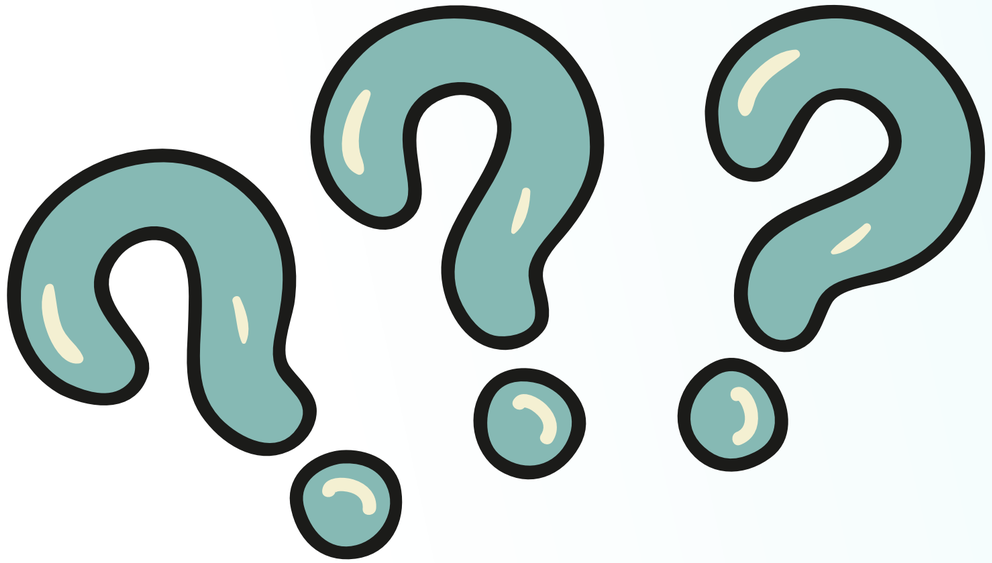
Cross-validation involves splitting the data into smaller parts, called folds, and using some to train the model and some to test the model



CONFUSION MATRIX

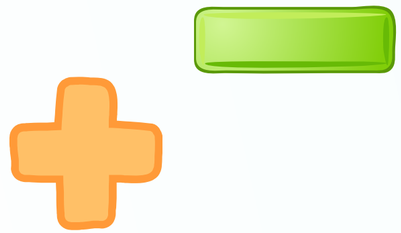
A confusion matrix is a table that evaluates a classification model's performance by comparing the actual and predicted outcomes.

- Accuracy
- Precision
- Recall
- F1 Score



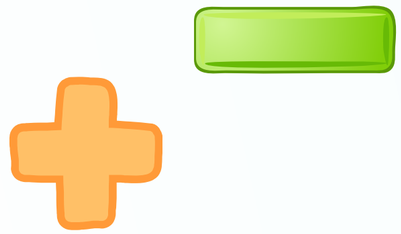
CONFUSION MATRIX

		Predicted: Positive	Predicted: Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
Actual	Negative	False Positive (FP)	True Negative (TN)



EXAMPLE

		Predicted: Cancer	Predicted: No Cancer
Actual	Cancer	21 (TP)	6 (FN)
	No Cancer	4 (FP)	49 (TN)



ACCURACY, PRECISION, RECALL AND F1 SCORE

ACCURACY

$$\frac{TP + TN}{TP + TN + FP + FN}$$

TP	FN
FP	TN

PRECISION

$$\frac{TP}{TP + FP}$$

RECALL

$$\frac{TP}{TP + FN}$$

F1 SCORE

$$\frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$