LOVELY PROFESSIONAL UNIVERSITY

NAAC GRADE A++

INT395- SUPERVISED ML

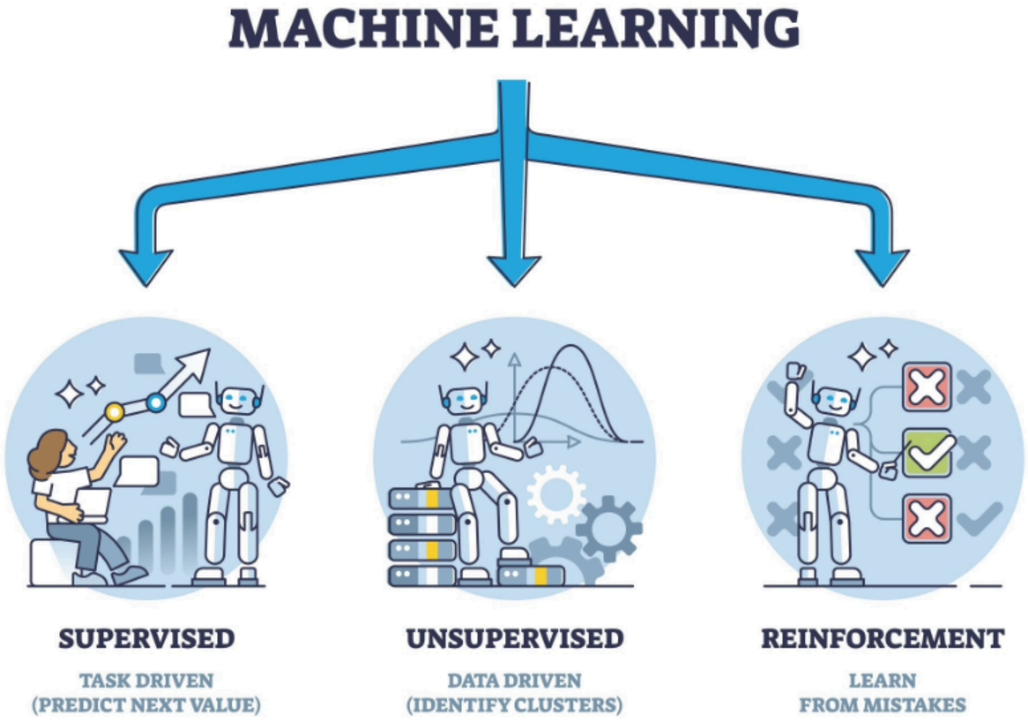
Unit 1:

- **Introduction**
- **Data Preprocessing**

Presented By: Blossom Kaler
Assistant Professor
SCSE, LPU

WHAT IS MACHINE LEARNING?

- Machine Learning is a branch of Artificial Intelligence
- it enables computers to learn from data and make predictions without being explicitly programmed
- Instead of writing specific rules, we train models to find patterns in data and make decisions on their own



LOVELY PROFESSIONAL UNIVERSITY

NAAC GRADE A++

TYPES OF MACHINE LEARNING

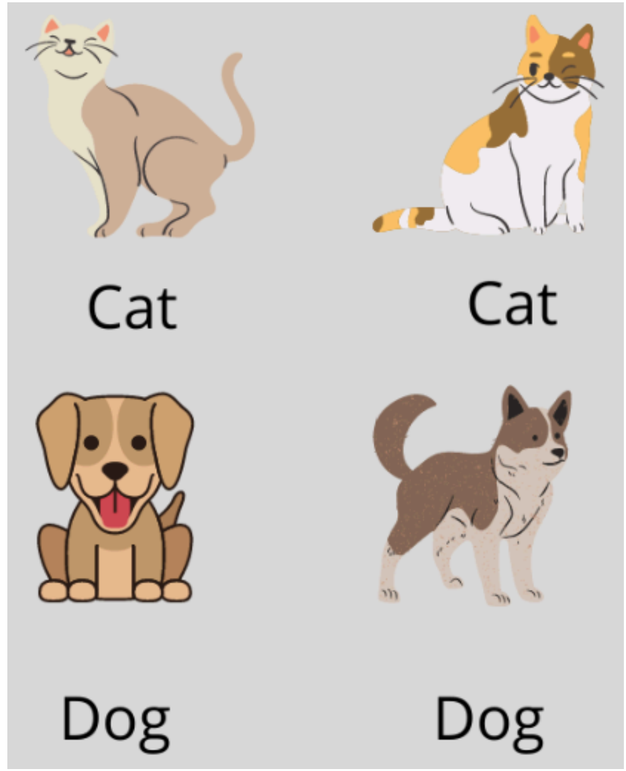
SUPERVISED

UNSUPERVISED

REINFORCEMENT

SUPERVISED ML

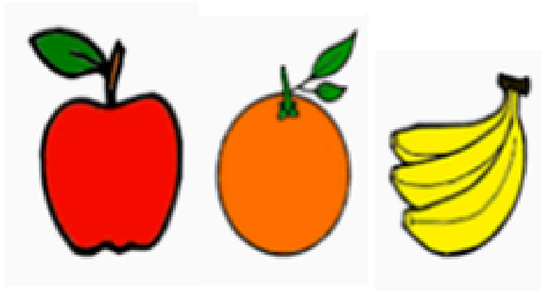
- when we train the machine using **labelled data.**
- labelled data is already tagged with the correct answer.
- After that, the machine is provided with a new set of examples (testing data) to make predictions.
- Types : Classification and Regression



labelled data

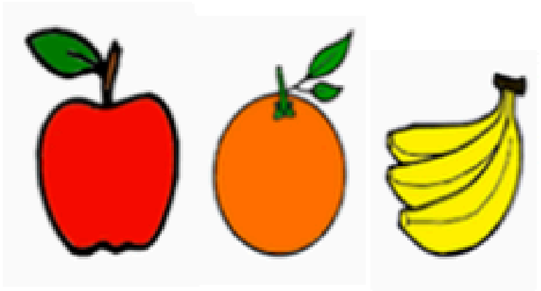
LABELLED VERSUS UNLABELLED DATA

LABELED DATA (BASED ON TASTE)



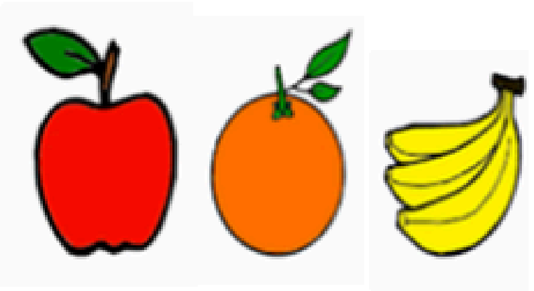
Apple Orange Banana

LABELED DATA (BASED ON COLOUR)



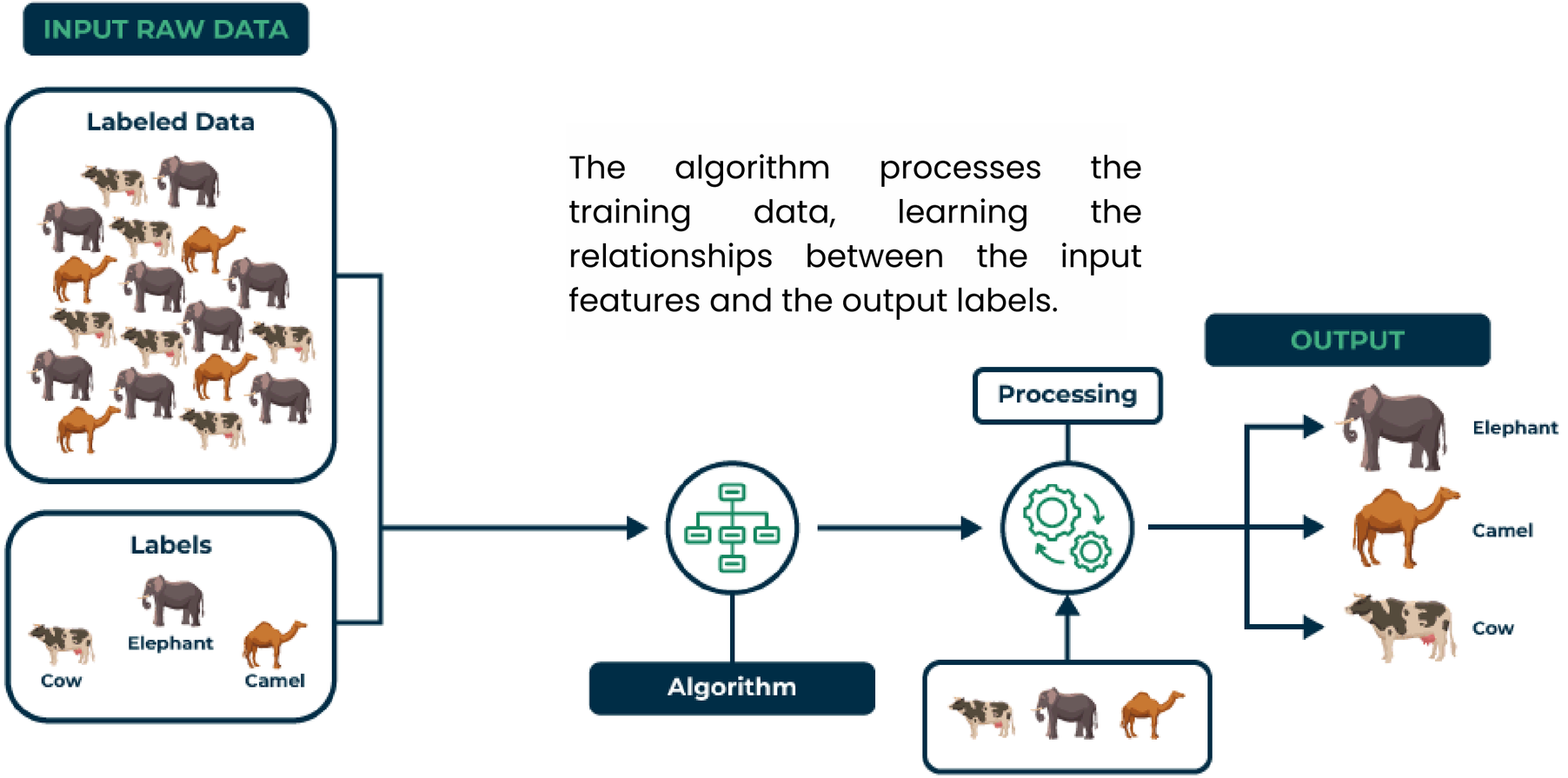
Red Orange Yellow

UNLABELED DATA



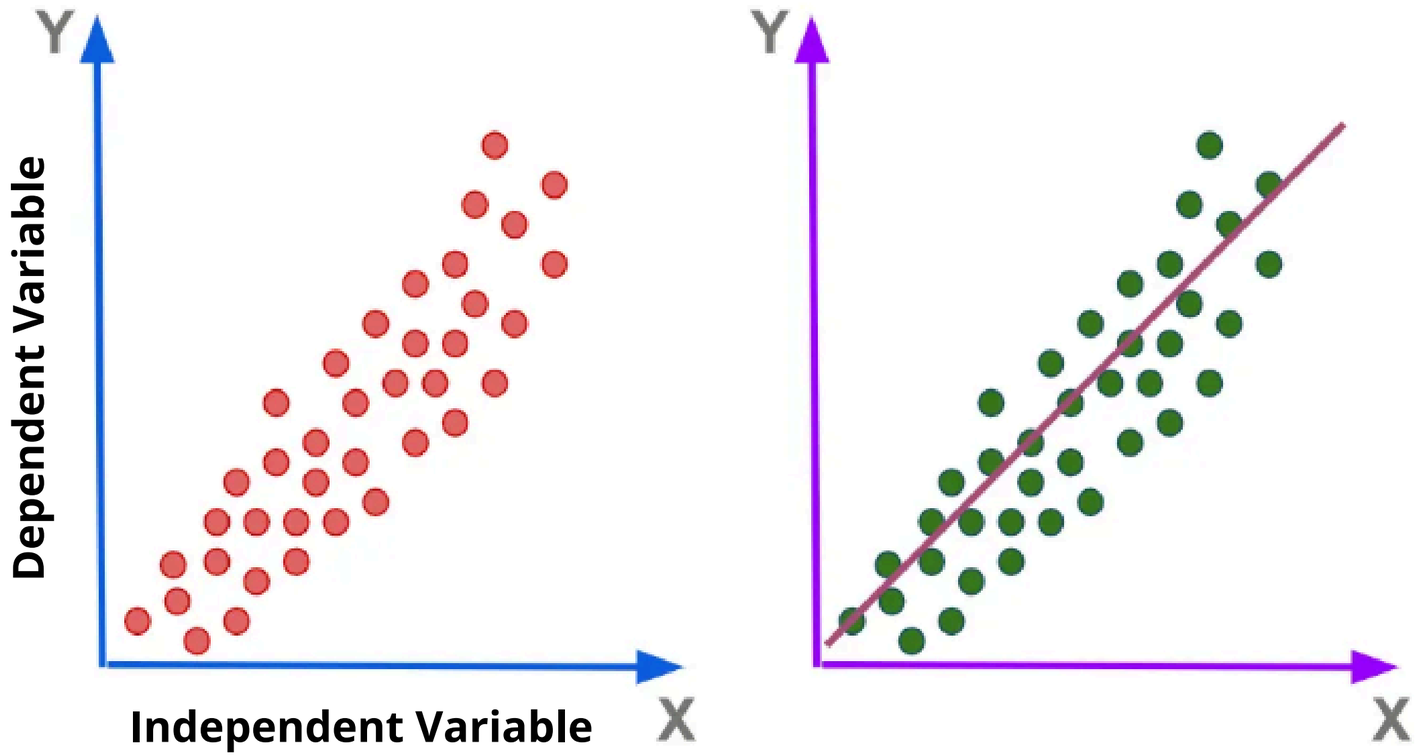
CLASSIFICATION: SUPERVISED ML

Training data



The algorithm processes the training data, learning the relationships between the input features and the output labels.

REGRESSION: SUPERVISED ML



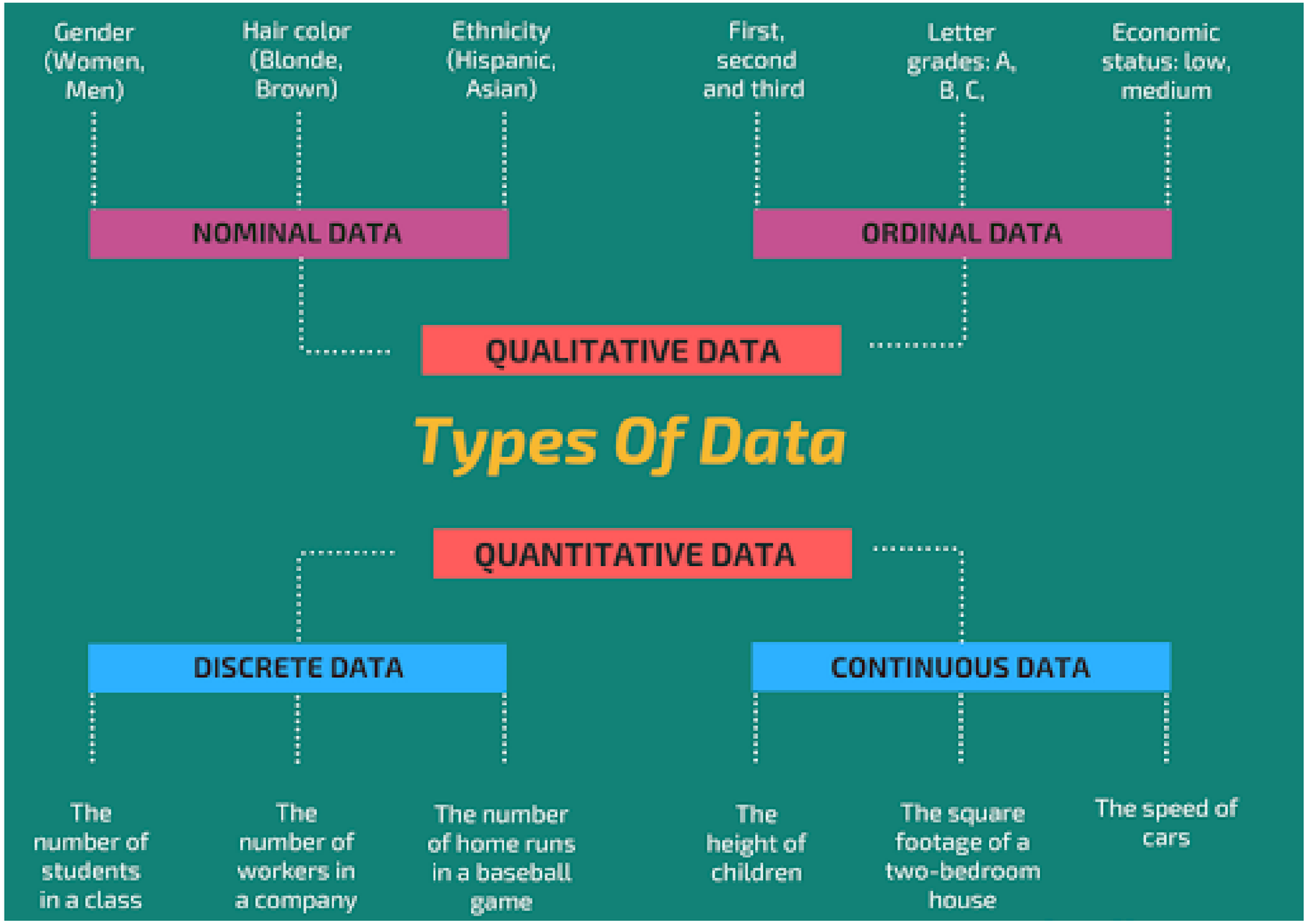
Linear Regression

REAL LIFE APPLICATIONS

Image Classification

House Price Prediction

Spam Detection


Gender (Women, Men)
Hair color (Blonde, Brown)
Ethnicity (Hispanic, Asian)
First, second and third
Letter grades: A, B, C,
Economic status: low, medium
NOMINAL DATA
ORDINAL DATA
QUALITATIVE DATA
Types Of Data
QUANTITATIVE DATA
DISCRETE DATA
CONTINUOUS DATA
The number of students in a class
The number of workers in a company
The number of home runs in a baseball game
The height of children
The square footage of a two-bedroom house
The speed of cars

COMMON DATA ISSUES

Raw data often contains :

- **missing values** due to human error or system failure
- **duplicate data**
- **inconsistent formats** (e.g., date, currency)
- **outliers** that misrepresent the overall trend
- **class imbalance** causing biased model training
- **irrelevant features**

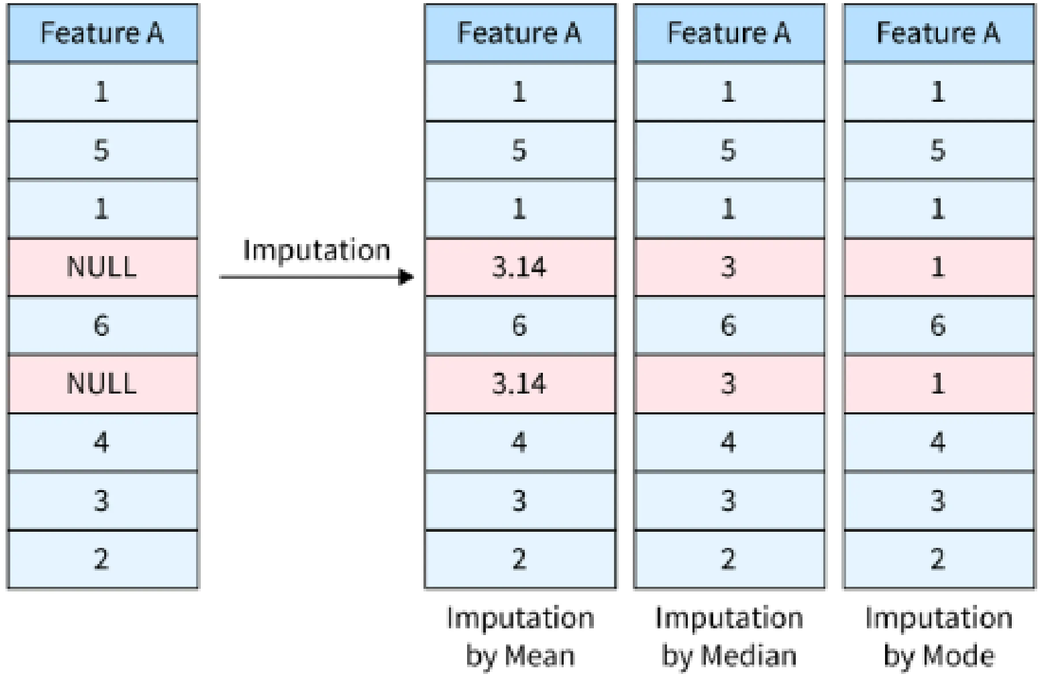
These issues can negatively impact model performance.

EXAMPLE OF DIRTY DATA

Student ID	Student Name	Age	GPA	Classification
100122014	Joseph	21	3.5	Junior
100232015	Patrick	200	3.2	Sophomore
100122012	Seller	24	3.0	Senior
100342013	Roger	23	234	Senior
100942012	Davis	2.8	3.7	Sophomore
	Travis	23	3.4	Sr
100982015	Alex	27		Sophomore
100982013	Trevor	-22	4.0	Senior
AUC2016XC	Aman	30	3.5	Jr

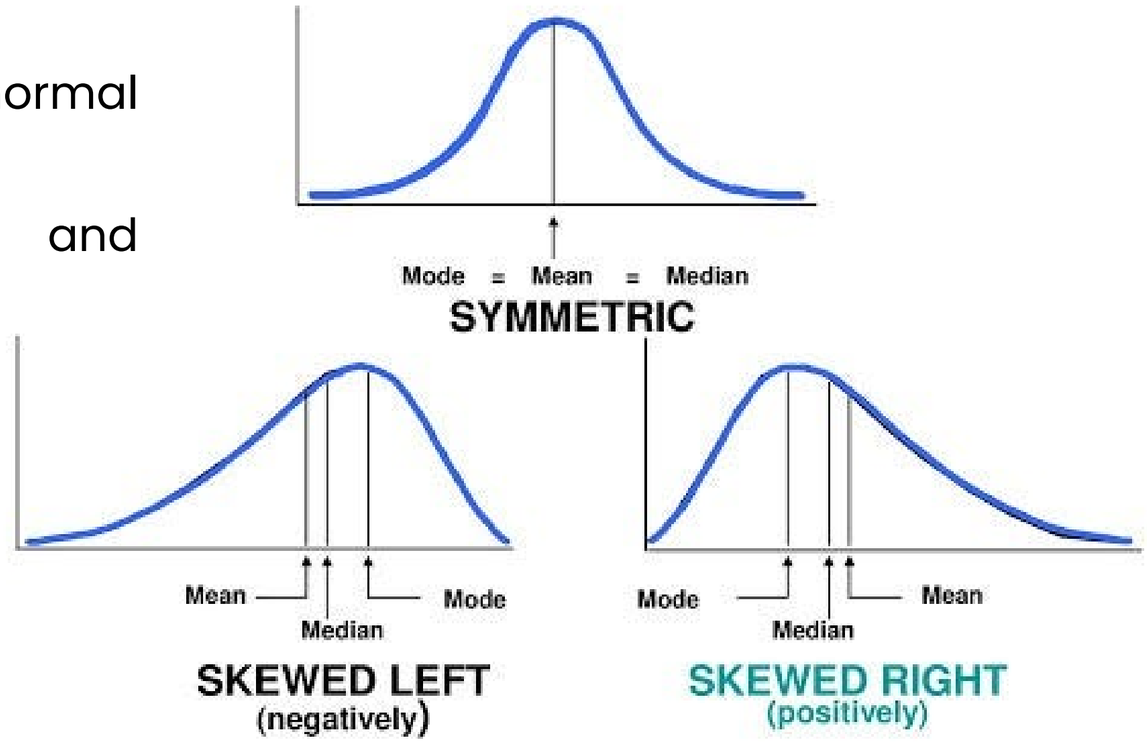
HANDLING MISSING VALUES

- **Delete** complete row (generally not recommended)
- Impute with **mean/median** for numerical features
- Use **mode** for categorical feature imputation
- Imputation methods: KNN or regression-based filling



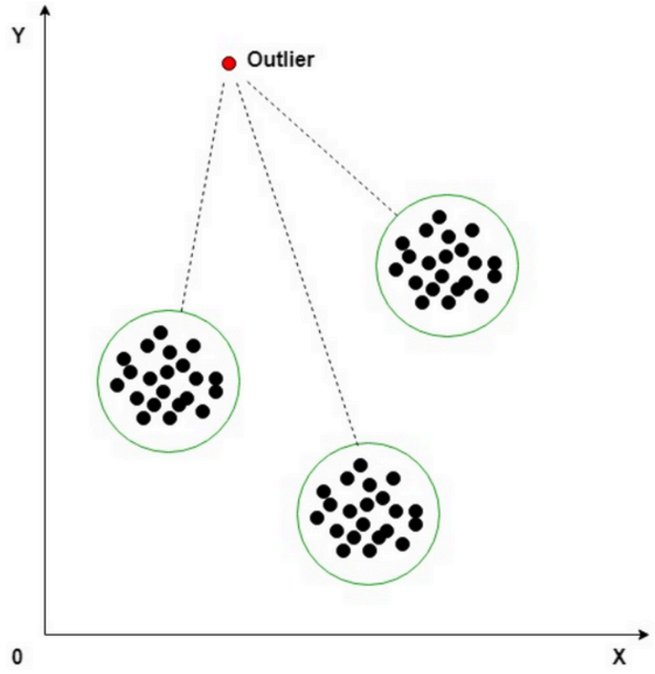
SKEWNESS

- Skewness measures asymmetry in data distribution.
- Positive skew → tail on right.
- Negative skew → tail on left.
- Skewed data may harm algorithms assuming normal distribution (e.g., linear regression).
- Helps improve model performance and interpretability.

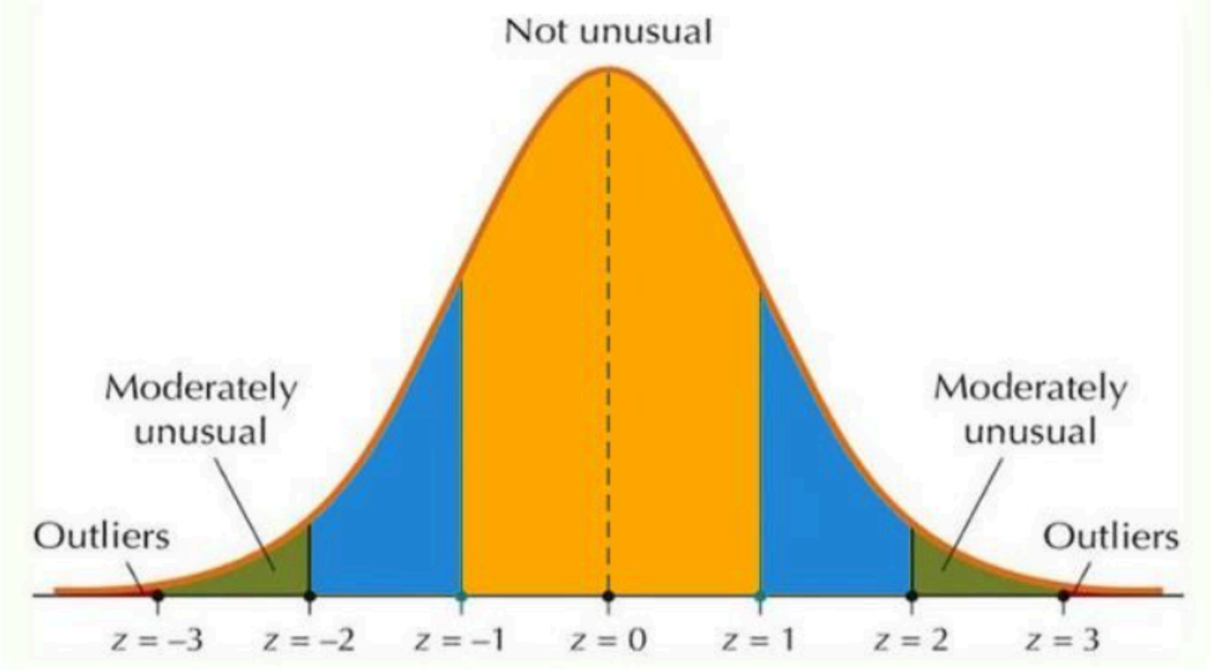


OUTLIER DETECTION

An outlier is a data point that **significantly deviates from the rest of the data**.

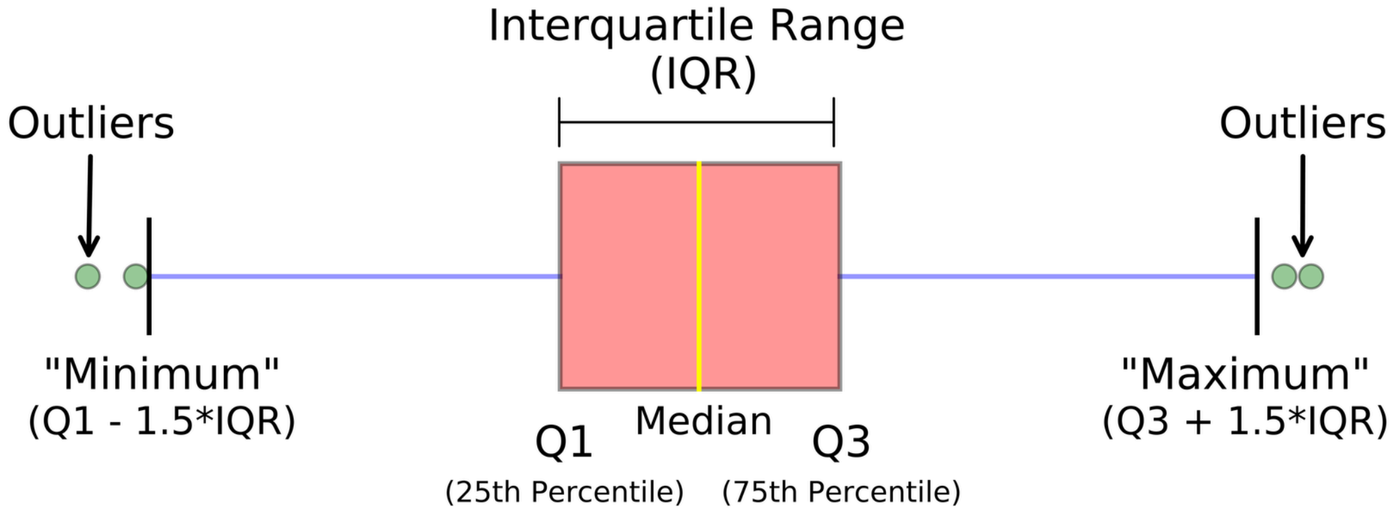


Using **Clustering**



Using **Z-score**

OUTLIER DETECTION



Using **Box Plot**

OUTLIER HANDLING

- **Remove the Outliers:** In cases where outliers are simply errors or irrelevant to your analysis, you can remove them.
- **Transform Data:** Apply transformations like logarithmic or square root transformations to reduce the impact of outliers.
- **Use Clipping (trimming):** Clipping involves setting an upper or lower limit on values.
- **Use Capping (Winsorizing)**: replaces outlier values with a defined maximum or minimum value from the dataset's distribution.
- **Use Robust Algorithms**: Certain algorithms like decision trees and random forests are less sensitive to outliers.

LOG TRANSFORMATION

- The logarithmic transformation “compresses” large values more than small ones.
- Reduces range and variability in data
- Makes skewed data more symmetric
- Works well for right-skewed, positive data

Original	log(x)
1	0
2	0.30
3	0.48
4	0.60
100	2.00

right skewed data

x	log(x)
0.1	-2.30
0.5	-0.69
1	0
10	2.30

left skewed data

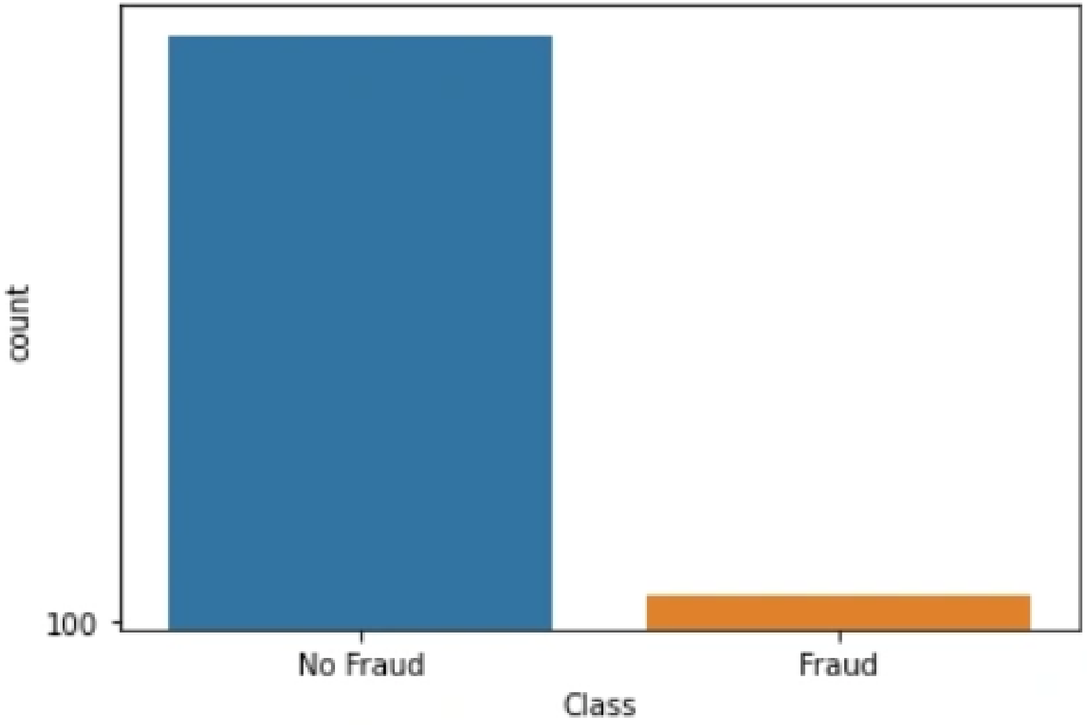
SQUARE ROOT TRANSFORMATION

- Used to reduce mild right-skewness in data
- Applies transformation y = sqrt{x}
- Compresses larger values, keeps small values stable
- Makes distribution closer to normal
- Simpler and safer than log for small datasets

Original	√x
1	1.00
4	2.00
9	3.00
100	10.00

CLASS IMBALANCE

- happens when some classes in a classification problem have significantly more instances than others
- it often leads to biased model performance.



UNDERSAMPLING

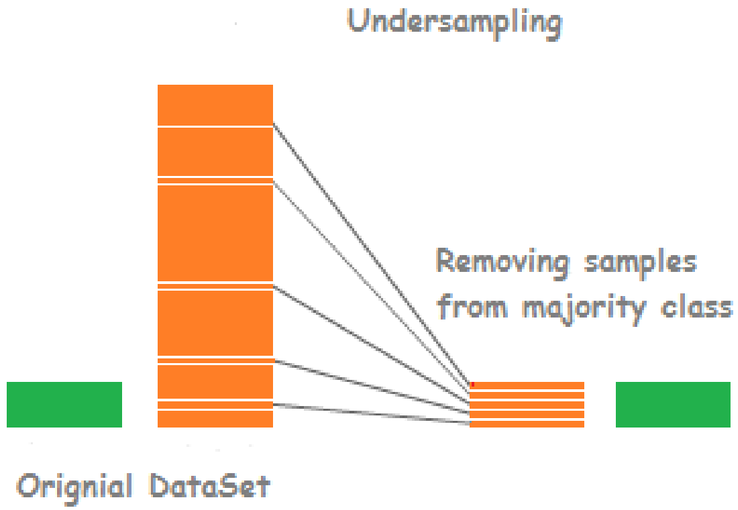
- Randomly removes samples from the majority class to balance the dataset.

Pros:

- Reduces dataset size, making training faster.
- Helps prevent overfitting to the majority class.

Cons:

- Risk of losing important data points, potentially affecting model performance.
- May not work well if the dataset is already small.
- Example: If Class 1 has 1000 samples and Class 2 has 100, RU randomly removes 900 samples from Class 1 to match Class 2.



OVERSAMPLING

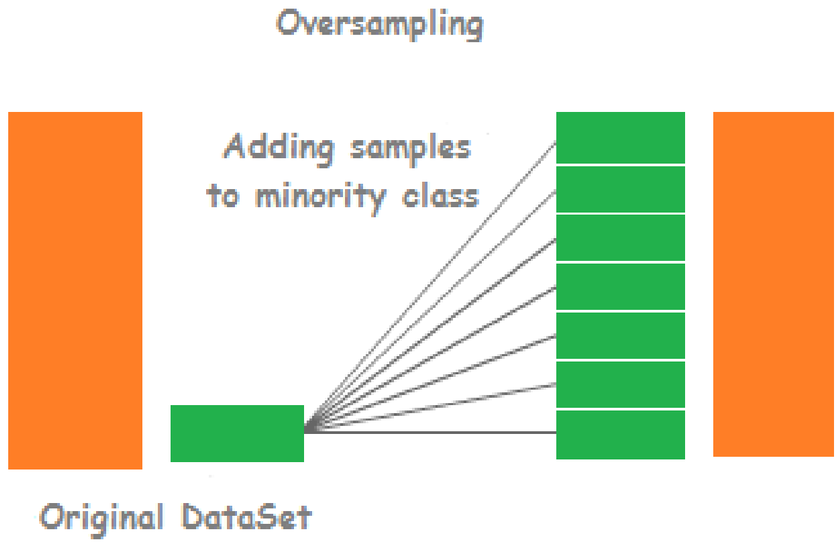
- Randomly duplicates samples from the minority class to balance the dataset.

Pros:

- Prevents loss of information.
- Helps the model learn from more balanced data.

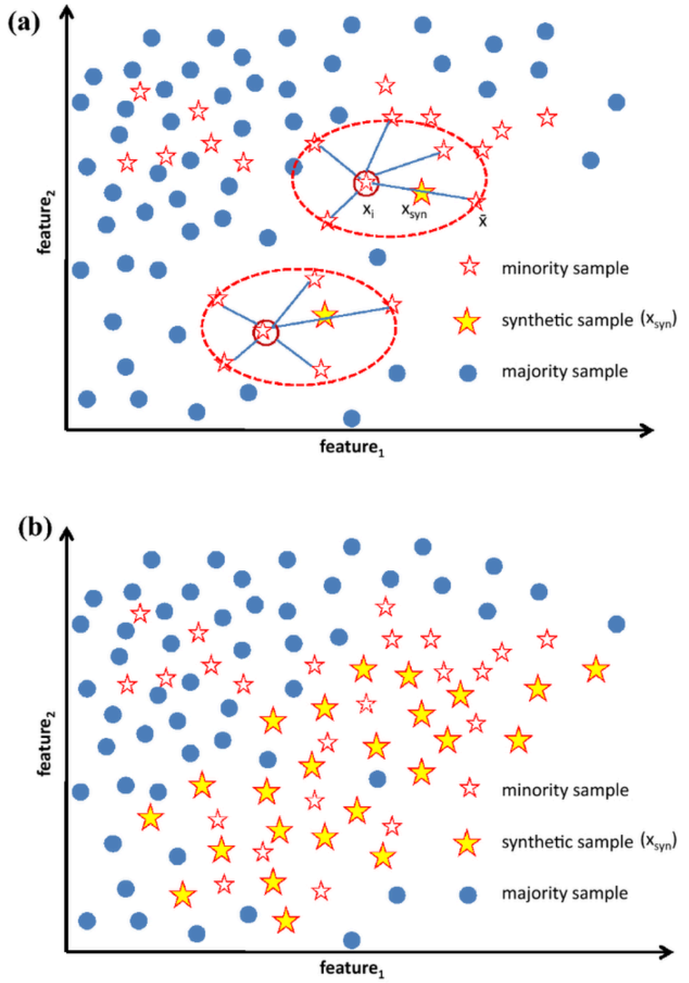
Cons:

- Can lead to overfitting as the model sees repeated data points.
- Does not create new informative examples.
- Example: If Class 1 has 1000 samples and Class 2 has 100, RO duplicates 900 samples from Class 2 to match Class 1.



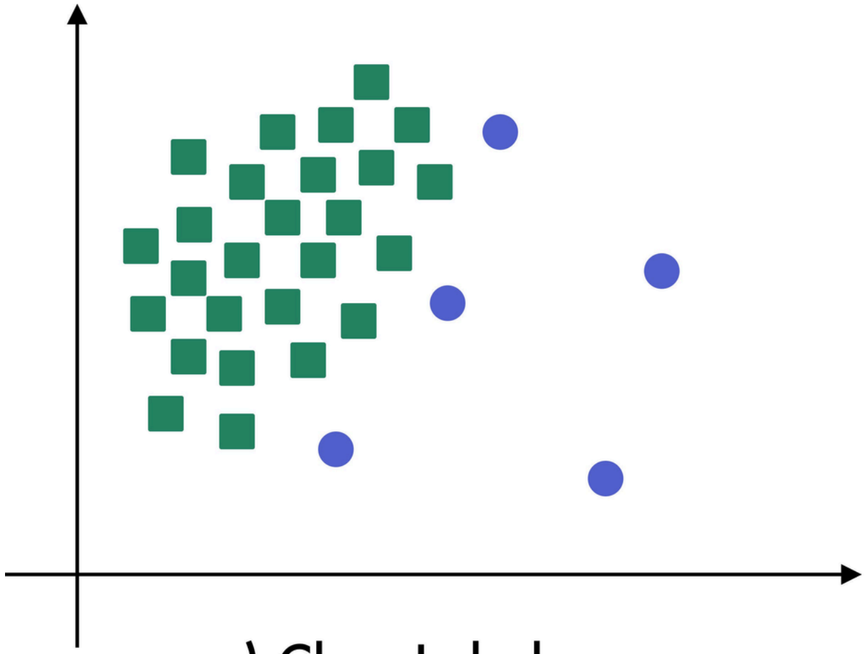
SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE)

- Instead of duplicating existing samples, SMOTE generates synthetic samples by interpolating between existing minority class instances.
- Selects a random minority class sample
- Finds its k nearest minority neighbors
- Randomly selects one of these neighbors
- Creates synthetic point between sample and neighbor
- Repeats until desired balance is achieved
- Pro:Prevents overfitting caused by simple oversampling
- Con:May generate noisy or irrelevant samples

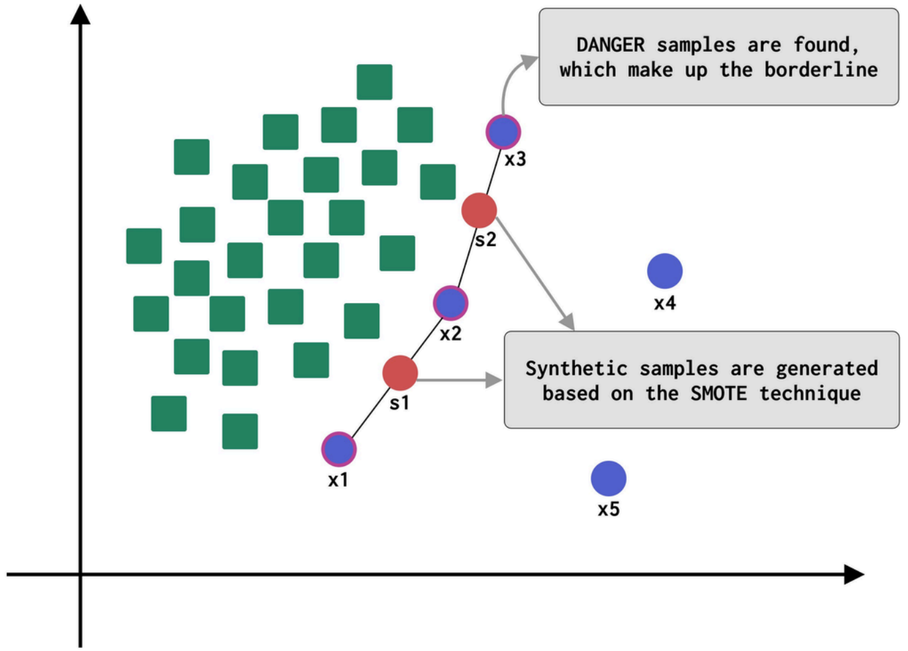


BORDERLINE SMOTE

- Focuses on samples near decision boundary
- Identifies “danger” samples close to majority class
- Generates new samples around these critical points
- Improves learning of hard-to-classify regions



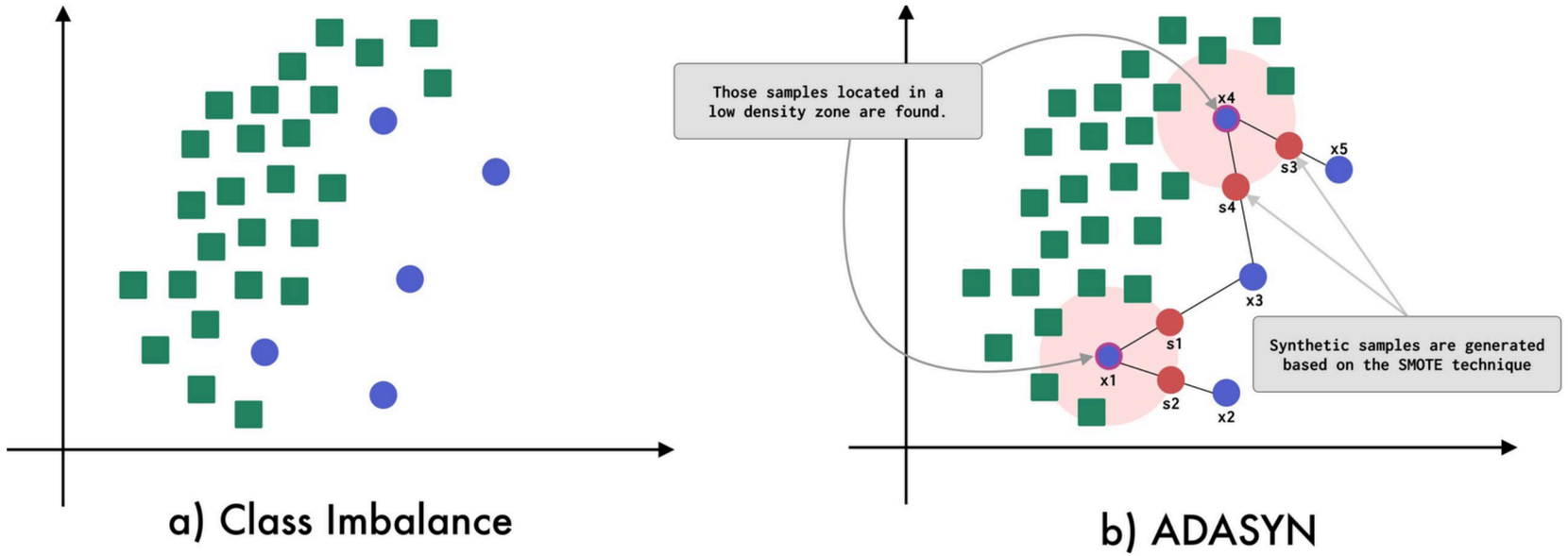
a) Class Imbalance



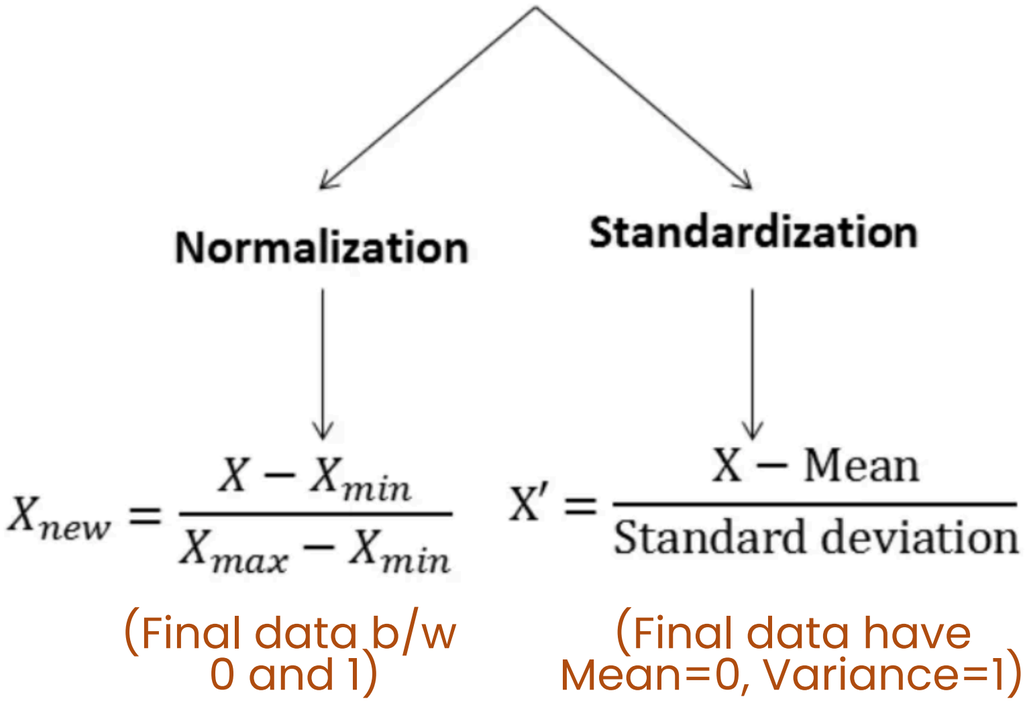
b) Borderline-SMOTE

ADAPTIVE SYNTHETIC SAMPLING (ADASYN)

- It is an extension of the SMOTE technique
- Minority samples harder to classify get more attention
- Regions with higher learning difficulty get more new points
- Assign higher weights to hard-to-learn samples
- Generate synthetic samples accordingly using nearest neighbors
- Cons: Can create noise if difficult samples are actually outliers.



FEATURE SCALING



- it is the process of standardizing or normalizing data **to bring all features to a common scale**.
- Without scaling, **features with larger magnitudes may dominate** the learning process, biasing the model towards those features.

NORMALIZATION

	fare	fare_scaled
0	7.2500	0.014151
1	71.2833	0.139136
2	7.9250	0.015469
3	53.1000	0.103644
4	8.0500	0.015713
5	8.4583	0.016510
6	51.8625	0.101229
7	21.0750	0.041136
8	11.1333	0.021731
9	30.0708	0.058694

- Also called Min–Max Scaling
- Rescales data to a fixed range, usually [0, 1]
- Preserves relative relationships among data points
- Suitable for methods using Euclidean distance (kNN, K-Means)
- Sensitive to outliers — range may distort if extremes exist

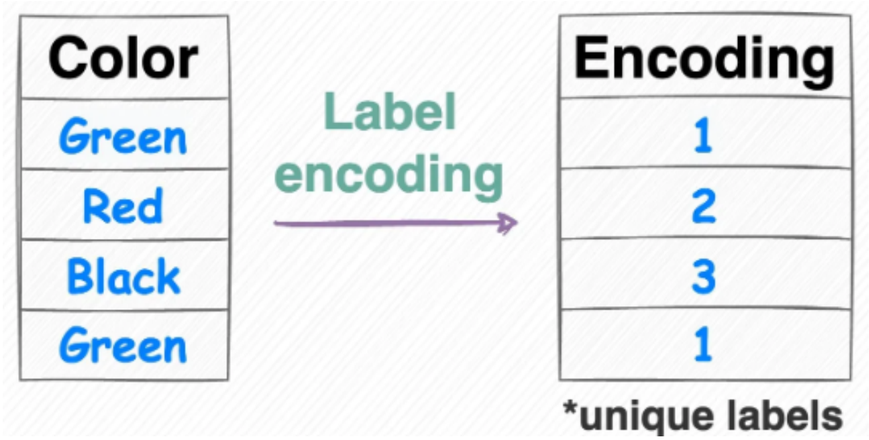
STANDARDIZATION

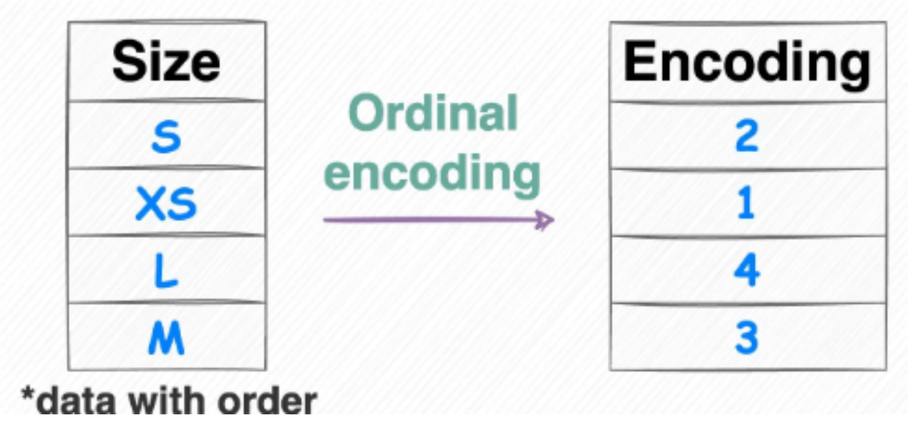
	fare	fare_scaled
0	7.2500	-0.502445
1	71.2833	0.786845
2	7.9250	-0.488854
3	53.1000	0.420730
4	8.0500	-0.486337
5	8.4583	-0.478116
6	51.8625	0.395814
7	21.0750	-0.224083
8	11.1333	-0.424256
9	30.0708	-0.042956

- Also called Z-score scaling
- Centers data around mean 0, standard deviation 1
- Does not bound values to [0, 1]
- Retains outlier influence but balances feature variance
- Preferred when outliers exist but shouldn't be clipped

FEATURE ENCODING

- conversion of categorical or text data into a numerical format
- this helps algorithms to easily understand and process the data
- techniques: One-Hot Encoding, Label Encoding, Ordinal Encoding





ONE-HOT ENCODING

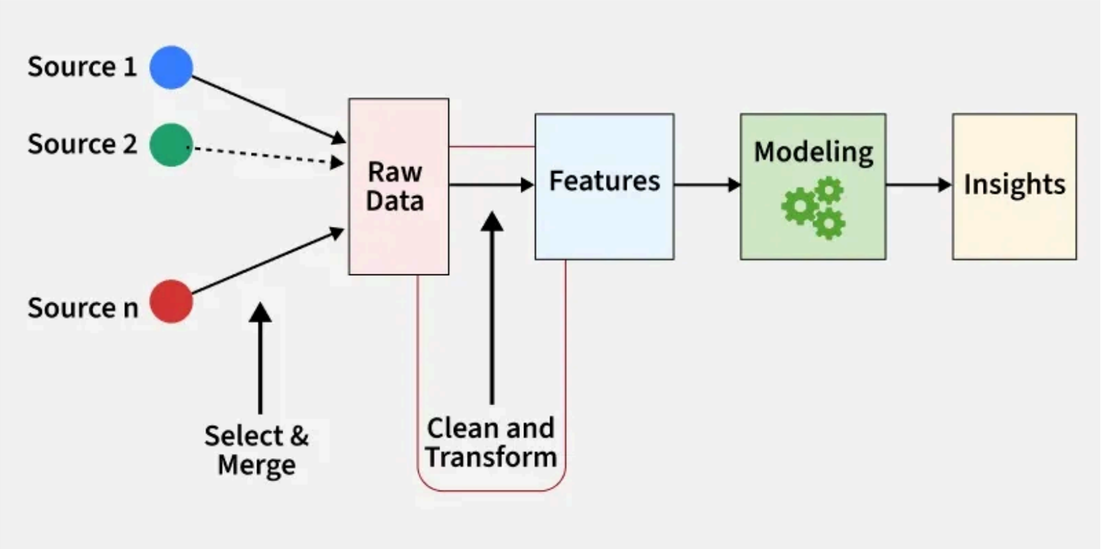
Color
Green
Red
Black
Orange

One-hot encoding →

Green	Red	Black	Orange
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

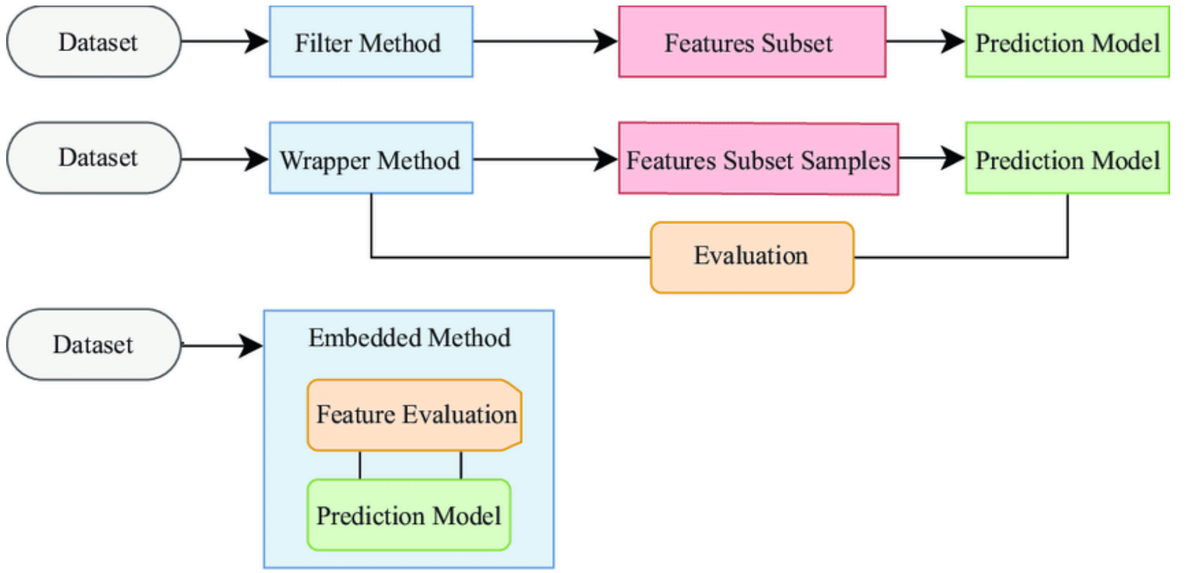
FEATURE ENGINEERING

- Feature engineering transforms raw data into meaningful input features.
- It improves model performance and learning efficiency significantly.
- Techniques include encoding, scaling, binning, and feature creation.
- Domain knowledge helps in designing impactful custom features.
- Goal: enhance predictive power and simplify model complexity.



FEATURE SELECTION

- identifying and retaining only the most important features for model training
- Helps remove irrelevant, redundant, or noisy features
- Focuses on the most informative attributes of the dataset
- reduces overfitting, improves accuracy, speeds up training
- **Example**: In medical diagnosis, only the most relevant biomarkers are selected from hundreds of tests to predict disease.



FEATURE SELECTION

Filter-based Approach

Filter-based feature selection approaches are based on data intrinsic attributes such as feature correlation or statistics.

Wrapper-based Approach

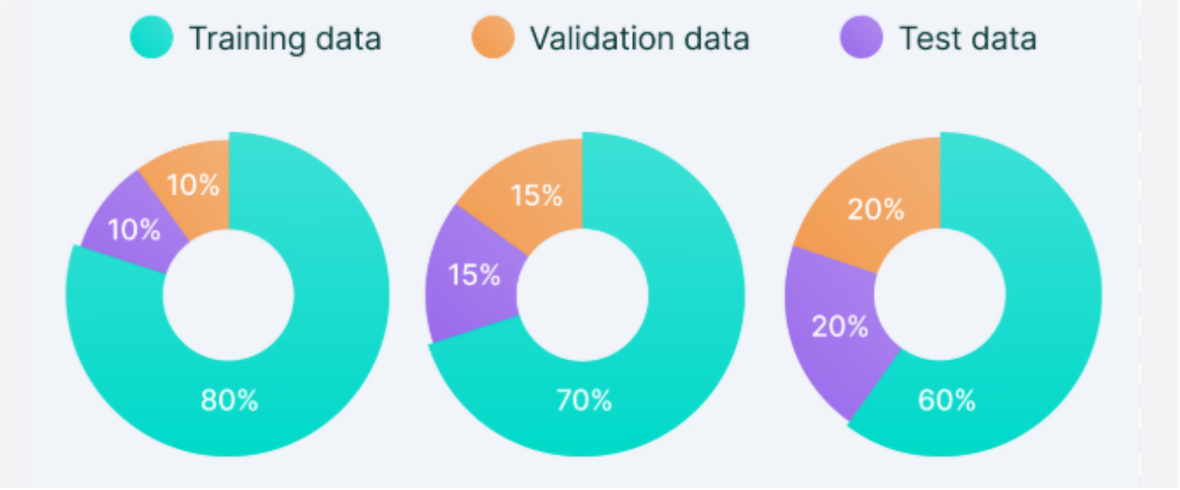
Wrapper-based feature selection approaches include assessing the importance of features using a specific machine learning algorithm.

Embedded Approach

Embedded feature selection approaches include the feature selection process as part of the learning algorithm.

DATA SPLITTING

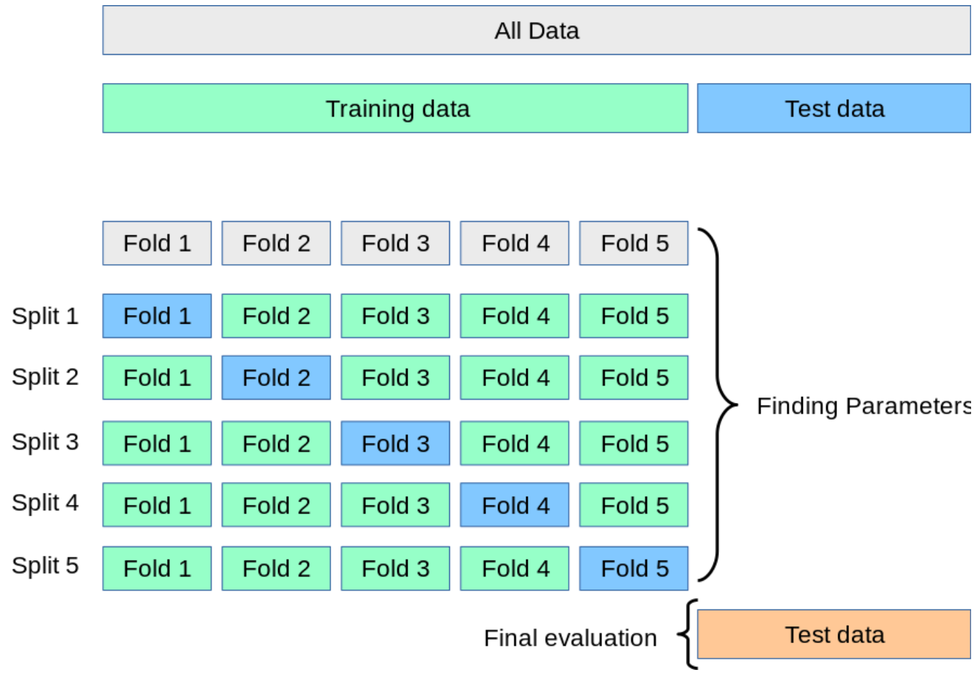
- dividing a dataset into training, validation, and testing sets to evaluate model performance.
- Split data into training and testing sets
- Use validation set for model tuning
- Ensure proportional representation in splits
- Avoid data leakage during splitting process
- Test set remains unseen during training



K-FOLD CROSS VALIDATION

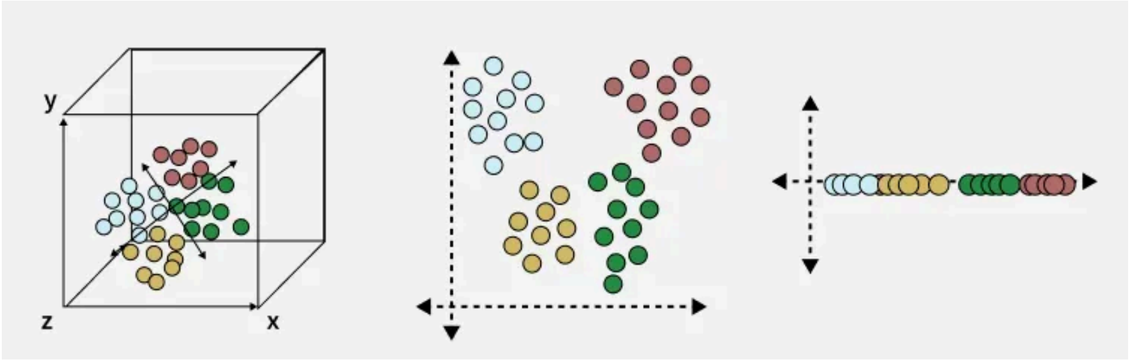
- To evaluate model performance more reliably using all available data.

  Process:
- Split dataset into k equal folds.
- For each iteration (total k):
- Train model on (k-1) folds
- Validate on the remaining 1 fold
- Repeat this k times (each fold acts as validation once).
- Average the validation results → gives final performance estimation



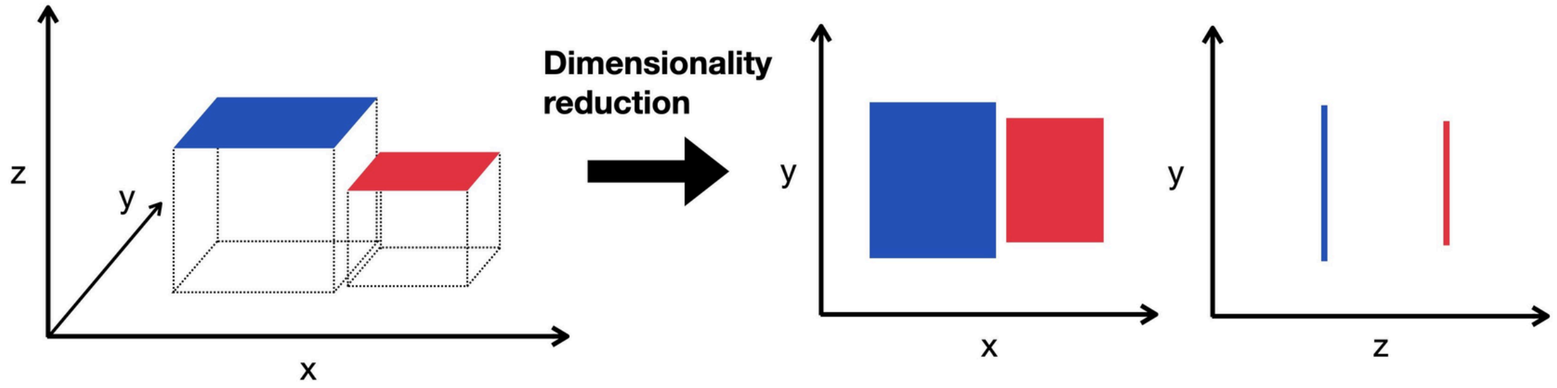
DIMENSIONALITY REDUCTION

- process of reducing the number of input variables while preserving essential information.
- Helps overcome the curse of dimensionality in large datasets.
- Improves computation time and reduces model overfitting risk.
- PCA extracts new orthogonal features maximizing variance captured.
- Simplifies models, enhances interpretability, and boosts generalization.



DIMENSIONALITY REDUCTION

- There are two components of dimensionality reduction:
- **Feature selection**: In this, we try to find a subset of the original set of variables, or features, to get a smaller subset which can be used to model the problem. It usually involves three ways: **Filter, Wrapper, Embedded**
- **Feature extraction**: This reduces the data in a high dimensional space to a lower dimension space, i.e. a space with lesser no. of dimensions. Methods of Dimensionality Reduction:
- Principal Component Analysis (**PCA**)
- Linear Discriminant Analysis (**LDA**)
- Generalized Discriminant Analysis (**GDA**)



DIMENSIONALITY REDUCTION:PCA

- Imagine you're analyzing a dataset with dozens of features, like a customer survey with age, income, purchase history, and website behavior.
- While rich, this high dimensionality can be a curse: complex models, slower training, and even irrelevant info hiding the good stuff. That's where Principal Component Analysis (PCA) comes in as your dimensionality reduction hero!
- **The Gist**: PCA takes your high-dimensional data and squishes it into a lower- dimensional space, capturing the most important information but ditching the redundancy. Think of it like summarizing a long lecture into key points – you lose some detail, but the core meaning remains.

LOVELY PROFESSIONAL UNIVERSITY

NAAC GRADE A++

THANK YOU