Redesigning Common Mind & Business Towards Excellence

3Purpose Process People Culture

SNS INSTITUTIONS
www.snsgroups.com

1st GenAI Powered Design Thinking FrameWork

Build an Entrepreneurial Mindset Through Our Design Thinking FrameWork

SNS COLLEGE OF TECHNOLOGY

Coimbatore-35
An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A++' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF COMPUTER APPLICATIONS

23CAT702 - MACHINE LEARNING

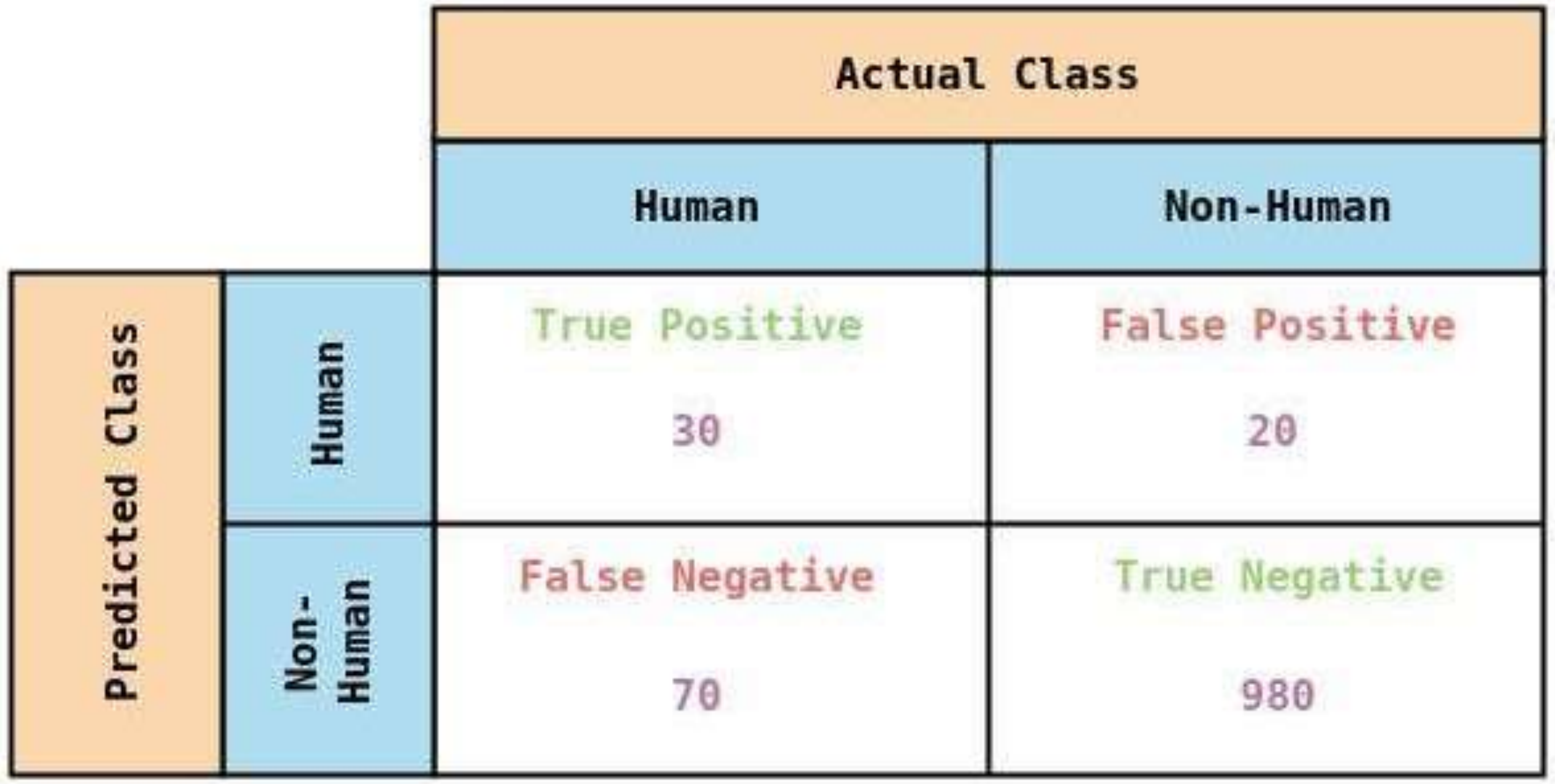
II YEAR III SEM

UNIT I – FOUNDATIONS OF LEARNING

TOPIC 6 – Error and noise - Training versus Testing

Error

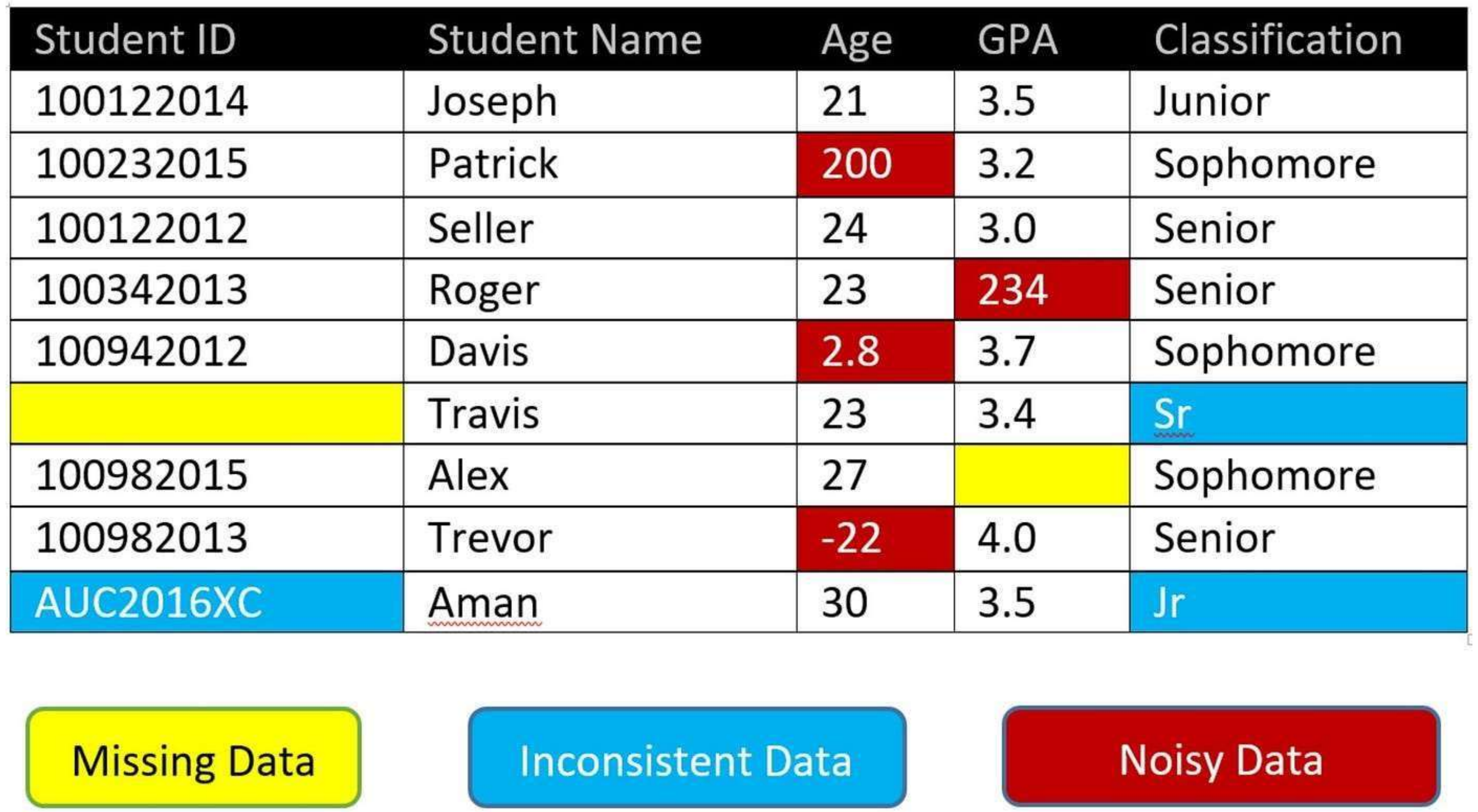
Error: Measures the average squared difference between the predicted and the actual target values within a dataset



		Actual Class	
		Human	Non-Human
Predicted Class	Human	True Positive 30	False Positive 20
	Non-Human	False Negative 70	True Negative 980

Noise

Noise: Unwanted behavior within dataset.



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

Missing Data
Inconsistent Data
Noisy Data

Causes of Noise

1. Errors in data collection.
2. Noise can also be introduced by measurement mistakes, such as inaccurate instruments or environmental conditions.
3. Another form of noise in data is inherent variability resulting from either natural fluctuations or unforeseen events.
4. Inaccurate data point labeling or annotation can introduce noise and affect the learning process.

Types of Noise in Machine Learning

1. **Feature Noise**: It refers to superfluous or irrelevant features present in the dataset that might cause confusion and impede the process of learning.
2. **Systematic Noise**: Recurring biases or mistakes in measuring or data collection procedures that cause data to be biased or incorrect.
3. **Random Noise**: Unpredictable fluctuations in data brought on by variables such as measurement errors or ambient circumstances.
4. **Background noise**: It is the information in the data that is unnecessary or irrelevant and could distract the model from the learning job.

Simple ways to Handle Noises

Identify the source of noise

Use domain knowledge, metadata, or documentation to understand how your data was collected, processed, or stored, and what are the expected values, ranges, or formats for each variable.

Filter out outliers

Outliers are extreme values that deviate significantly from the rest of the data and can distort your analysis or model. Identify and exclude outliers based on their similarity or proximity to other data points

Simple ways to Handle Noises

Handle missing values

Missing values are gaps or blanks in your data that can reduce your sample size, bias your analysis, or cause errors in your model.

Filter out outliers

Outliers are extreme values that deviate significantly from the rest of the data and can distort your analysis or model. Identify and exclude outliers based on their similarity or proximity to other data points

Remove the row or column else replacing them with a constant, a mean, a median, a mode

Simple ways to Handle Noises

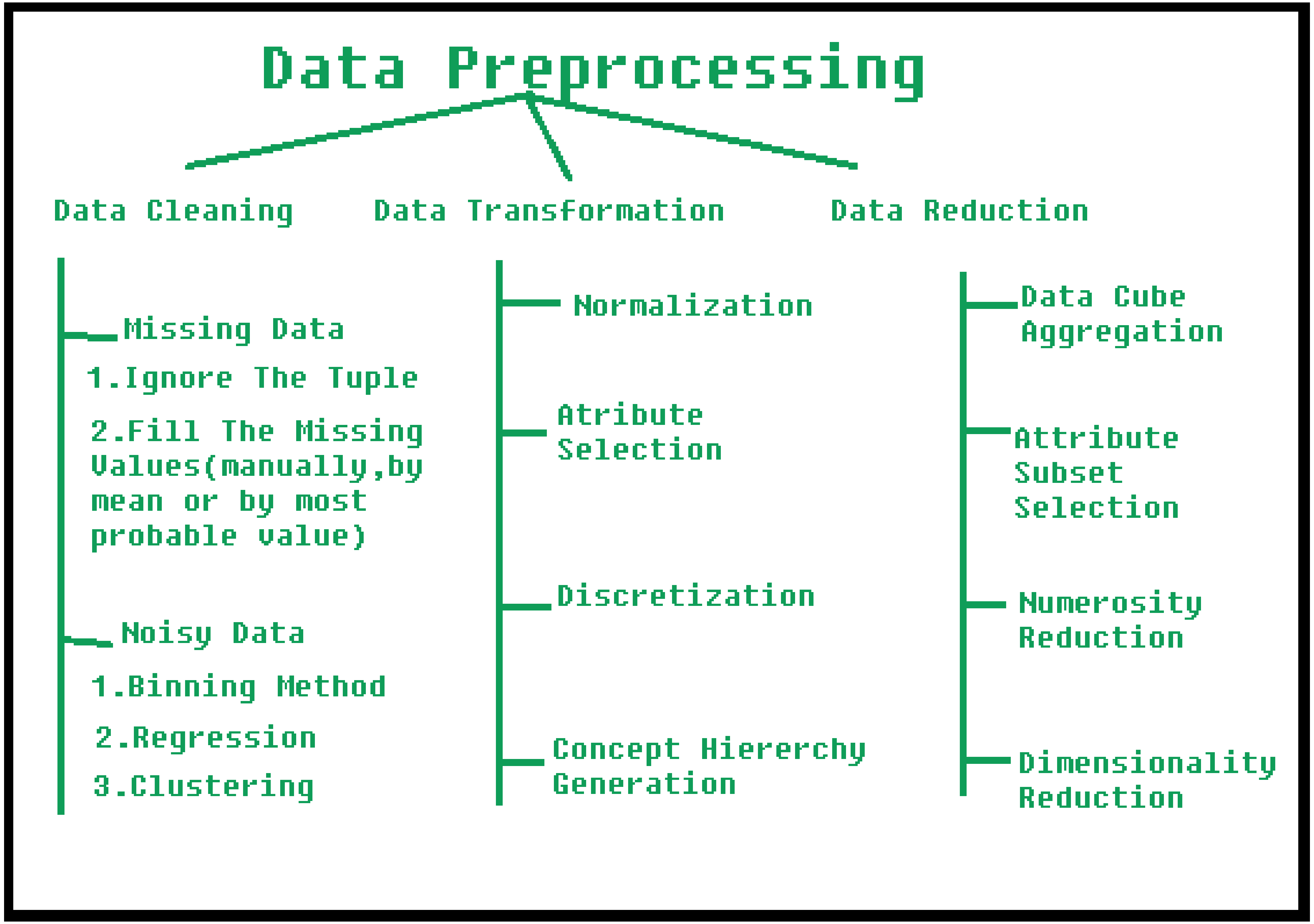
Remove duplicates

Duplicates are repeated or identical records in your data. You can also use fuzzy matching techniques to identify and remove duplicates that have slight variations or typos in their values.

Standardize formats

Formats are the way your data is represented or structured, such as dates, times, currencies, units, or categories. Inconsistent or incorrect formats can cause confusion.

For date and time use the common format YYYY-MM-DD HH:MM:SS



Data Preprocessing

Data Cleaning

- Missing Data
  1. Ignore The Tuple
  2. Fill The Missing Values(manually,by mean or by most probable value)
- Noisy Data
  1. Binning Method
  2. Regression
  3. Clustering

Data Transformation

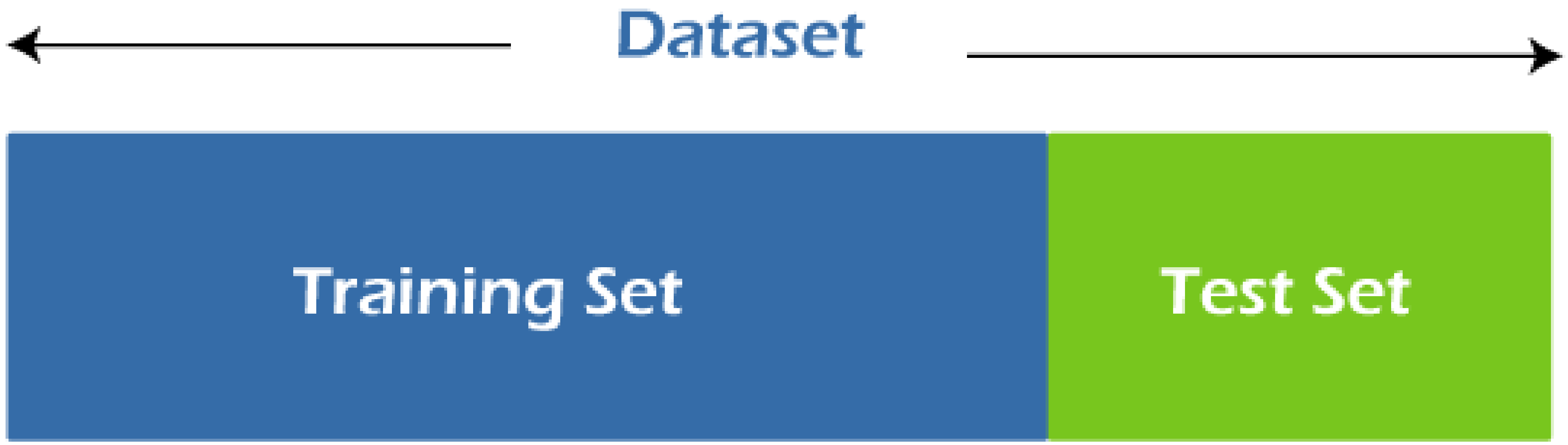
- Normalization
- Atribute Selection
- Discretization
- Concept Hiererchy Generation

Data Reduction

- Data Cube Aggregation
- Attribute Subset Selection
- Numerosity Reduction
- Dimensionality Reduction

Training Vs Testing

1. Train and test datasets are the two key concepts of machine learning, where **the training dataset is used to fit the model, and the test dataset is used to evaluate the model.**



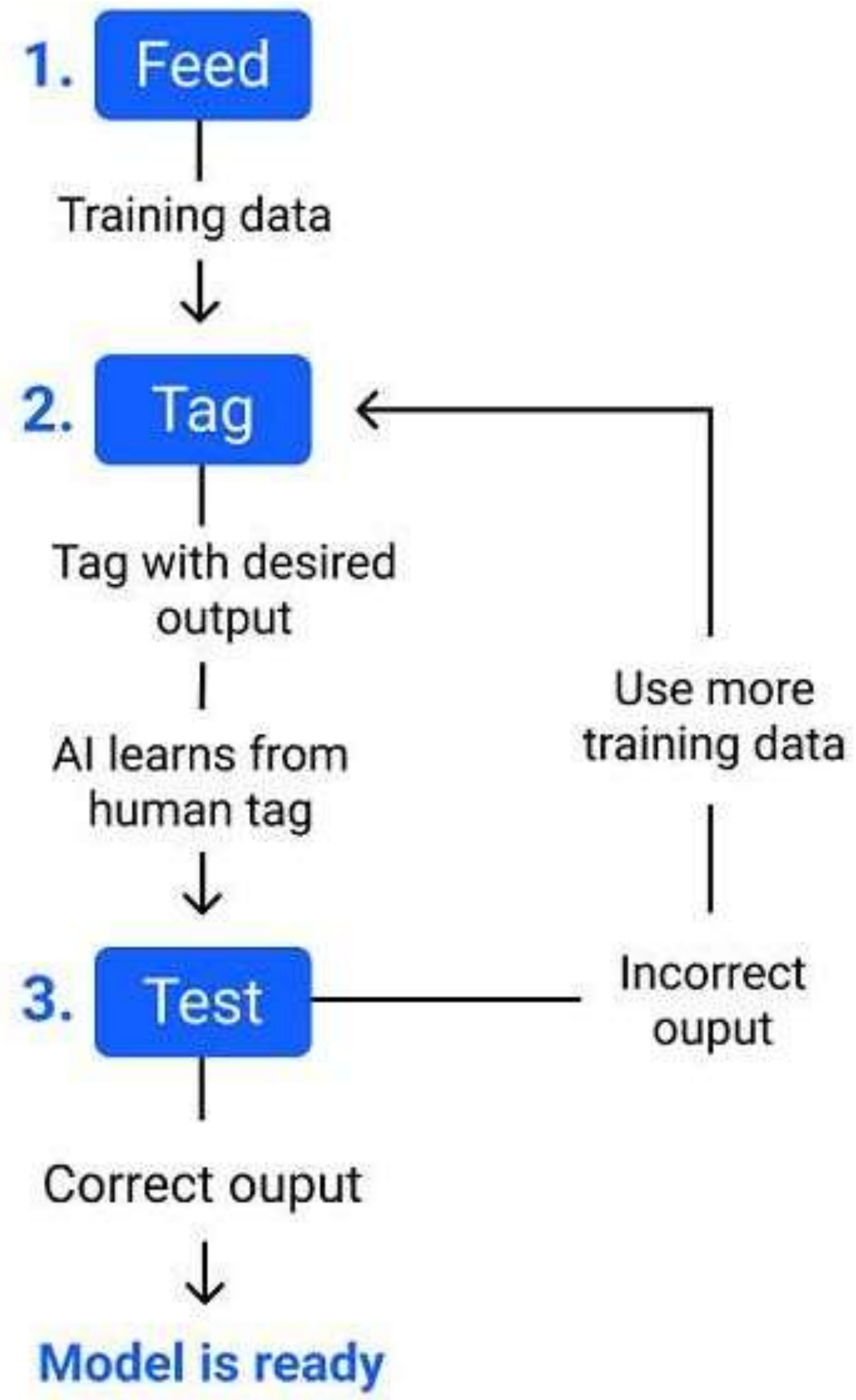
What is Training Dataset?

Training data is the biggest (in -size) subset of the original dataset, which is used to train or fit the machine learning model.

What is Testing Dataset?

This dataset evaluates the performance of the model and ensures that the model can generalize well with the new or unseen dataset

Workflow



1. Feed

Training data

2. Tag

Tag with desired output

AI learns from human tag

3. Test

Correct ouput

Model is ready

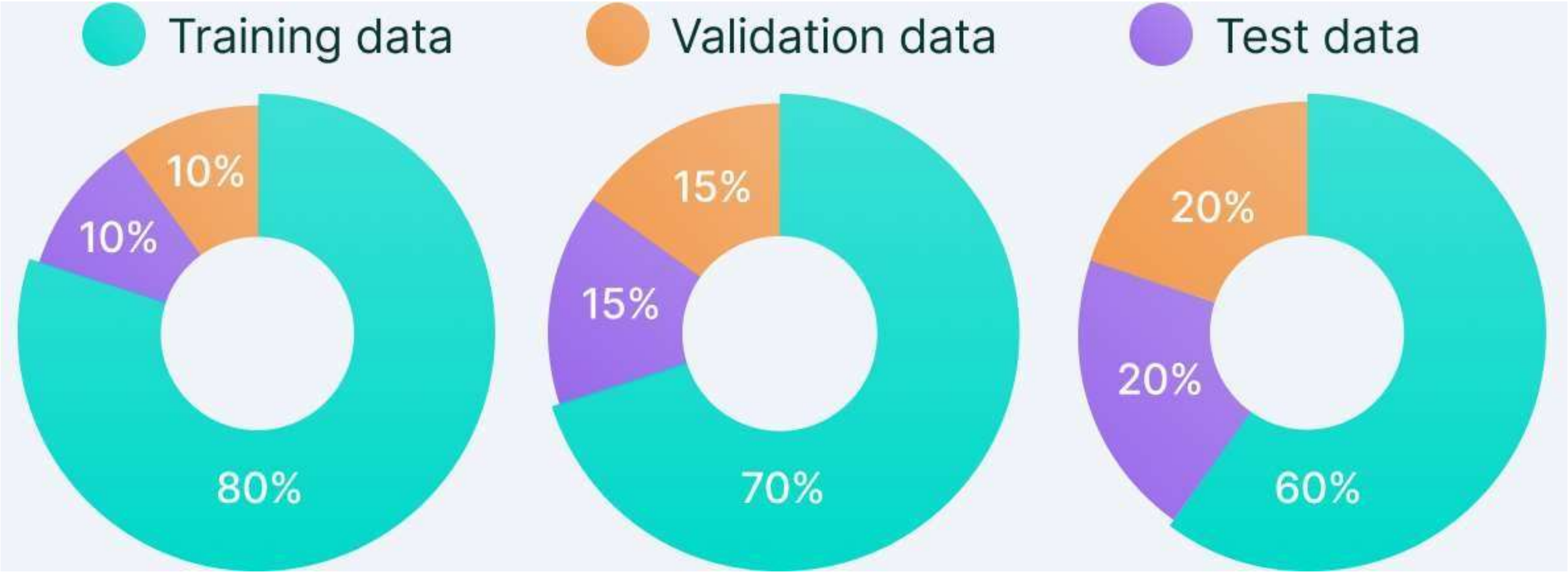
Incorrect ouput

Use more training data

Training Vs Testing

Feature	Training Data	Testing Data
Purpose	Train the machine-learning model. The more training data the model has, the better it can make predictions.	Testing Data is used to evaluate the performance of the model.
Exposure	The model can learn from the training data and improve its predictions.	The testing data is not exposed to the model before evaluation. This ensures the model cannot memorize the testing data and make perfect predictions.
Use	Training data is used to prevent overfitting.	The testing data is used to evaluate the model's performance by making predictions on it and comparing the predictions to the actual labels.
Size	The training data is typically larger. This is because the model needs to see many examples of input data to learn how to make accurate predictions.	The size of the testing data is typically smaller than the training data because the testing data is used to evaluate the performance of the model that has been trained on the training data.

Allocation of Training, validation and testing



References

Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, ―Learning from Data, AML Book Publishers, 2012.

P. Flach, ―Machine Learning: The art and science of algorithms that make sense of data‖, Cambridge University Press, 2012.

W3school.com

https://testsigma.com/blog/difference-between-training-data-and-testing-data/

https://www.v7labs.com/blog/train-validation-test-set#h1

Thank You