



Studybeesofficial

Where Minds Pollinate

Class 12
Computer 

Chapter 4
Data & Analysis
notes

Follow Our Socials



studybeesofficial



studybeesofficialpak



studybeesofficialpak

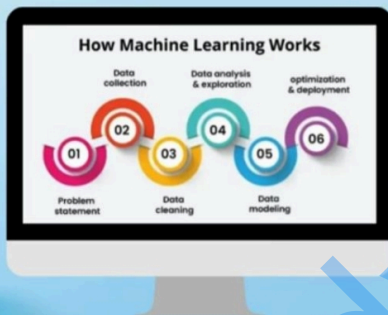


Chapter 4

Data & Analysis



Introduction to Machine Learning



- **Machine Learning** enables computers to learn from existing data.
- Computers **use predefined rules** during training.
- Raw data → transformed into models.

These models:

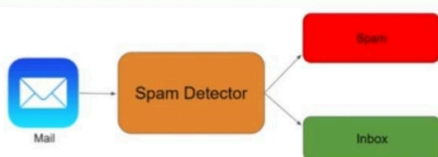
- Are **evaluated** on different environments.
- **Learn patterns.**
- **Predict future** values based on past patterns.
- Helps understand **cause-effect relationships** in data.
- **Key Use:** Forecasting based on data insights.

Real-World Example – Email Spam Filter

Spam filters use ML to classify emails.

How it works:

- Learns from labeled spam/non-spam emails.
- Detects patterns/keywords.
- Marks similar future emails as spam.



Data Types Overview

Data is raw facts and figures.
In ML, data is categorized into types:

Data Type	Description	Example
Numerical	Quantitative values	Age, Salary, Marks
Categorical	Discrete values (Nominal & Ordinal)	Gender, Rank
Boolean	Binary values	True/False, Yes/No
Text	Sequence of characters	Review, Feedback
Time Series	Collected over time	Stock price, Weather data
Image	Pixel values	X-Ray, Photograph
Audio	Sound waves	Music, Speech
Structured	Row-column format	Spreadsheets, Databases
Semi-Structured	Some structure, not fixed schema	XML, JSON

Feature	Nominal Data	Ordinal Data
Definition	Categories with no order	Categories with a meaningful order
Order	Not important	Important
Examples	Colors, Gender, Blood Group	Education Level, Satisfaction Ratings
Math Operations	Only counting, mode	Counting, ordering, median

Introduction to Machine Learning (Basics)

Key Concepts before diving deeper:

- How ML models are trained.
- How are they evaluated?
- Metrics to assess performance.
- Understanding these is essential to grasp ML techniques.

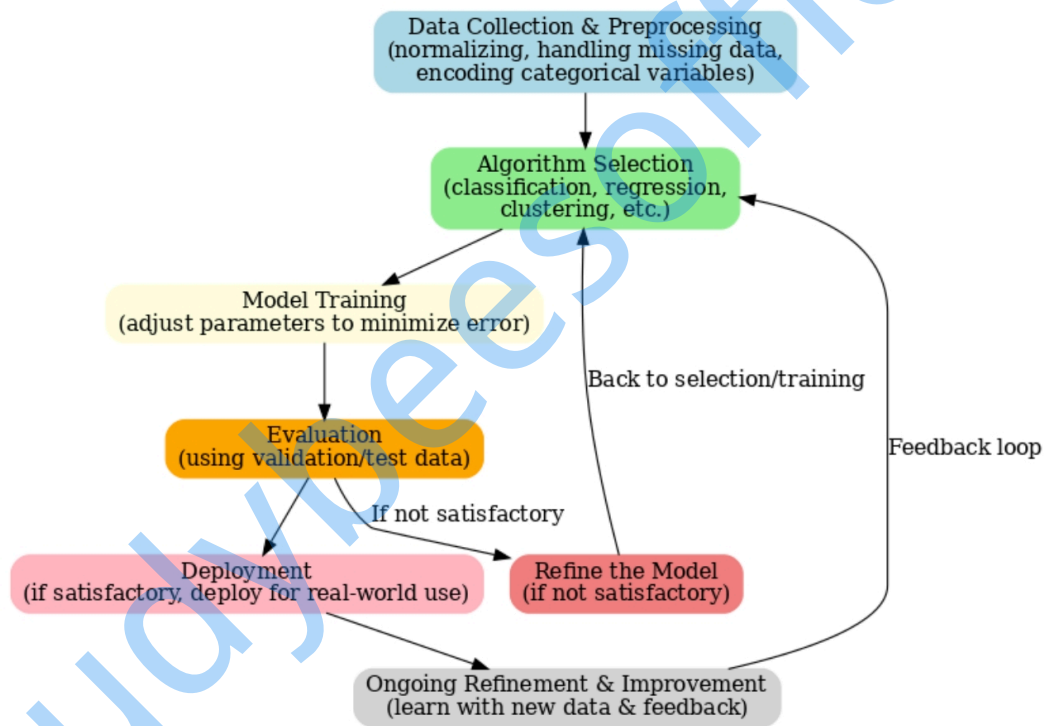
Feature	Rule-Based Algorithm	Machine Learning
Definition	A program that follows explicit rules written by humans.	A program that learns patterns automatically from data.
Who makes the decision rules?	Human programmer writes “if-else” conditions.	The computer finds the rules from data during training.
Adaptability	Doesn't change unless the programmer edits the rules.	Continuously improves as more data is provided.
Example	Loan approval based on fixed conditions.	Loan approval learned from past customer data.

What is Model Building in ML?

A step-by-step process to teach machines how to learn from data and make predictions.

Involves:

- **Data collection and preprocessing** (normalizing values, handling missing data, and encoding categorical variables)
- **Algorithm selection** (classification, regression, clustering, or another type)
- **Model training** (adjusting its parameters to minimize error and improve accuracy)
- **Evaluation** (Evaluate performance using validation/test data)
- **Deployment** If satisfactory, deploy for real-world use
- Otherwise, **refine the model**
- The model keeps learning with new data and feedback
- **Ongoing refinement and improvement ensure accuracy**



Dataset: Loans were approved or rejected

Applicant	Age	Income	Credit Score	Loan Approved
A	25	30,000	600	No
B	40	80,000	750	Yes
C	30	50,000	720	Yes
D	22	20,000	580	No

Learning Process

- **The computer looks for *patterns* in this data.**
For example, it might learn that:
- **Higher income** → more likely to approve
- **Higher credit score** → more likely to approve
- It builds a **model** — basically, a set of rules or a mathematical function that can predict whether a new applicant should get the loan or not.
- You can illustrate like this:
- **Model (simplified rule):**
If Income > 50,000 and Credit Score > 700 → Approve
Else → Reject



Test the Model

- Now, we test the model with *new data* (someone who didn't exist in the training examples). This is called **testing or prediction**.

Applicant	Age	Income	Credit Score	Predicted Decision
E	28	60,000	710	 Approve
F	24	25,000	590	 Reject

Feedback and Improvement

If the model makes mistakes — for example, approves someone who doesn't repay — we show it the correct answers again and let it learn better.

This is called **training again** or **model improvement**.

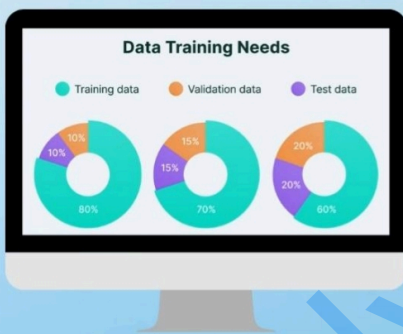
Key Terms in Model Building

To understand the **entire process of model building** using **Machine Learning**, some key terms are necessary to understand.

Let's look at an example:

- ➔ Building a model to classify images of dogs, cats, and horses.

Train-Test Split



Dividing the dataset into training and testing portions:

- **Training Data:** 70–80%
- **Test Data:** 20–30%
- Too much training data → **Overfitting**
- Too little training data → **Underfitting**

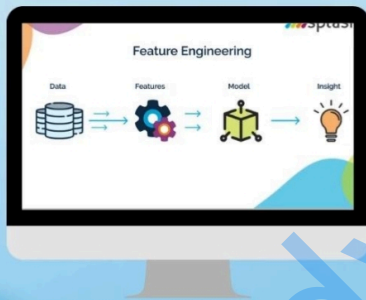
Validation Dataset -> **Purpose:** Used **during training** to tune the model.

Feature Engineering

Process of selecting and transforming important features to improve model accuracy.

Example features:

- Shape of ears, eyes, and nose
- Size, fur texture, and body posture
- Prevents confusing a sitting horse with a standing dog.



Overfitting & Underfitting in ML Models

- **Overfitting**
- Occurs when the model learns the training data **too well**, including noise and irrelevant details.
- It **performs well on training data** but **poorly on new data**.
- **Underfitting**
- Occurs when the model is **too simple** to capture meaningful patterns in the data.
- **Performs poorly on both training and new data.**

Overfitting:

- Definition:** When a model learns the **training data too well**, including noise and irrelevant patterns.

- Problem:** The model performs **very well on training data** but **poorly on new/unseen data**.

- Cause:** Model is **too complex** (too many parameters, deep trees, large neural networks, etc.).

- Example:**

A student memorizes answers to practice questions word-for-word.

In the exam, if the question is slightly different, they struggle.

Underfitting:

Definition: When a model is **too simple** and fails to learn the important patterns in the training data.

Problem: The model performs **poorly on both training and test data**.

Cause: Model is **too simple** (too few parameters, not enough training, wrong features).

Example:

A student only skims the book without learning details.

They cannot answer even the practice questions properly.

Model Training in Machine Learning

- **Model training is the key feature in Machine Learning.**
- It helps the model make accurate and meaningful decisions.
- Without proper training, the model cannot handle new data.

Types of Model:

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Types of Model Training:

1. Supervised Learning

- Learns from labeled data (input-output pairs).

Example: Linear Regression predicts continuous values, such as prices.

2. Unsupervised Learning

- Learns from unlabeled data to identify hidden patterns.

Example: K-Means Clustering is used for grouping similar data points.

3. Reinforcement Learning

- Learns by interacting with an environment using rewards and penalties for feedback.

Example: Q-Learning applied in games AI and robotic navigation. The model learns to make decisions to maximize its rewards.

Supervised Learning

Models learn from **labeled** data.

Training data

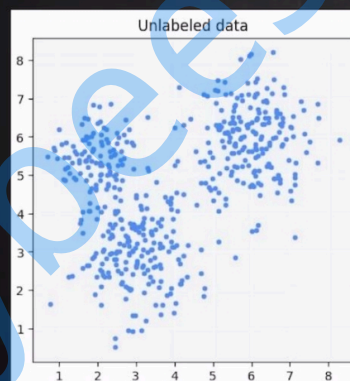
Email Text	Sender Email	Link	Exclamations	Label
"Win a brand new iPhone now! Click the link to claim your prize!"	promo@fakesite.com	yes	2	spam
"Dear team, please find the attached report for last quarter."	manager@company.in	no	0	not spam
"Cheap meds available!!! Order today and save big."	sale@pharmacy.in	yes	3	spam

Handwritten digits dataset

[illegible]

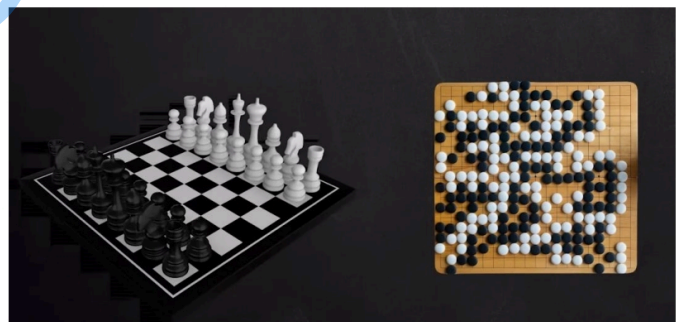
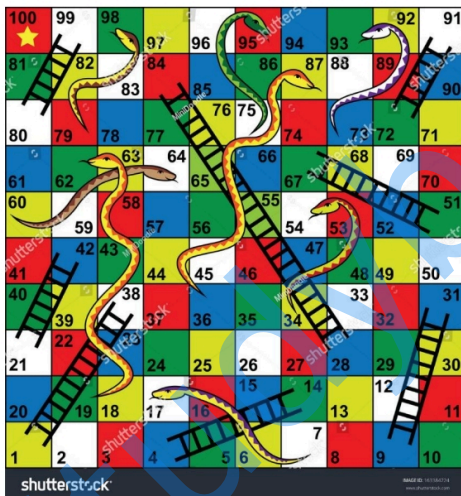
Unsupervised Learning

Models learn from **unlabeled** data.



Reinforcement Learning

In RL, an agent learns to make decisions by interacting with an environment, receiving **rewards** or **penalties** for its actions.



Reinforcement Learning (RL)

- Reinforcement Learning (RL) is a type of **machine learning** where an agent (like a computer program or robot) learns by **interacting with an environment**.
- Instead of being given correct answers (like in supervised learning), the agent **learns through trial and error** by receiving **rewards** (positive points) or **penalties** (negative points). The goal is to maximize total rewards over time.
- **Example:**
- **Game AI:** A computer playing chess learns which moves lead to winning (reward) or losing (penalty).
- **Self-driving car:** Learns to follow traffic rules by getting rewards for safe driving and penalties for accidents.

Model Assessment & Evaluation parameters

Model assessment evaluates how well a machine learning model performs using unseen data.

Parameters like accuracy, precision, recall, and F1 score help determine its effectiveness.

Key Evaluation Parameters:

1. **Accuracy:** The percentage of total correct predictions.

- Measures how often the model is correct.

Formula:

- $\text{Accuracy} = \frac{\text{Correct Prediction}}{\text{Total Predictions}}$

Example: 80 correct out of 100 images → 80% accuracy.



Truth



Prediction

Dog

Dog

Dog

No
Dog

Dog

No
Dog

Dog

No
Dog

Dog

Dog

Truth										
Prediction	Dog	Dog	Dog	No Dog	Dog	No Dog	Dog	No Dog	Dog	Dog
	✓	✗	✓		✓		✗		✓	✗

Truth



Prediction



True Positive = 4

False Positive = 3

Truth



Prediction



True **Negative** = 1

False **Negative** = 2

Truth										
Prediction	Dog	Dog	Dog	No Dog	Dog	No Dog	Dog	No Dog	Dog	Dog
	✓	✗	✓	✗	✓	✓	✗	✗	✓	✗

How many we got right? → 5

Accuracy → $5/10 \rightarrow 0.5$

Truth										
Prediction	Dog	Dog	Dog	No Dog	Dog	No Dog	Dog	No Dog	Dog	Dog
	✓	✗	✓		✓		✗		✓	✗

True Positive = 4

False Positive = 3

Precision is out of all dog predictions how many you got it right?

$$\text{Precision} = 4 / 7 = 0.57$$

Truth										
Prediction	Dog	Dog	Dog	No Dog	Dog	No Dog	Dog	No Dog	Dog	Dog
	✓	✗	✓		✓		✗		✓	✗

Recall is out of all **dog truth** how many you got it right?

Total Dog truth samples = 6

True **Positive** = 4

Recall = 4 / 6 = **0.67**

$Recall = TP / (TP + FN)$

For **precision**, think about **predictions** as
your base

For **recall**, think about **truth** as your base

Key Evaluation Terms

True Positive (TP) — Correct yes.

- The model says 'Cat', and it really is a cat.

False Positive (FP) — Wrong yes.

- The model says 'Cat', but it's actually a dog or a horse.

True Negative (TN) — Correct no.

- The model says 'Not Cat', and it's really not a cat.

False Negative (FN) — Wrong no.

- The model says 'Not Cat', but it's actually a cat.

Key Evaluation Metrics

2. Precision: How correct are model YES answers?"

- Measures how many of the predicted positive cases are actually correct.
- **Formula:**

Precision = True Positive / True Positives + False Positives

Example: Model predicts **50** images as cats. Out of these, **40 are actually cats** and 10 are dogs/horses. **Precision = 40/50 = 80%.**

3. Recall: Did the model find all the YES cases?

- Measures how well the model identifies all actual positive cases.
- It shows the model's ability to find all the relevant instances.
- **Formula:**

Recall = True Positives / True Positives + False Negatives

Example: There are **60 real cat images**. **Model correctly finds 50**, but misses 10 cats (labels them as dogs/horses). **Recall = 50/60 ≈ 83%.**

F1 Score – Balanced Evaluation

Combines **precision and recall** into one metric.

Especially useful when the dataset is **imbalanced**. Useful when both “how correct” (Precision) and “how complete” (Recall) are important.

Formula:

$$F1\text{ Score} = 2 \times (\text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall}))$$

Example (Cats):

- Precision = 80% (40 out of 50 predicted cats were correct)
- Recall = 83% (50 out of 60 real cats were found)
- **F1 Score = $2 \times (0.80 \times 0.83) / (0.80 + 0.83) \approx 81.5\%$**

Training Data

Training data is the **main part of the dataset** used to **teach the model** how to make correct decisions.

- It is given to the model during the **training phase**.
- Usually, **60–70%** of the total data is used for training.
- In our example, we give the model many labeled images of **cats, dogs, and horses** so it can learn to recognize them.
- Each image has a **label** (like “cat” or “dog”), especially when using **supervised learning**.

Test Data

Test data is a part of the dataset that is **not used** during **training**. It is used after the model is trained to **check** how well it can work **on new and unseen data**.

- Usually, **30–40%** of the data is kept for **testing**.
- It helps us see if the model can give the **right answers** for data it hasn't seen before.
- For example, we give the model new images of cats, dogs, and horses to see if it can **identify them correctly**.

Human Learning vs Machine Learning

Machines and humans both learn, but in different ways.

- **Machine Learning learns from data.**

- **For example**, it looks at many labeled images like "cat," "dog," or "horse" to understand patterns.
- The model uses these patterns to make predictions on new images.

- **Humans Learn through examples, pictures, and comparisons.**

- For example, we can recognize a cat after seeing it only a few times by remembering features like whiskers, pointed ears, and a tail.
- **Machines need data to learn, humans use experience and reasoning.**

Learning from Data

Continuous Learning:

- Keeps the model up to date with new data.
- Improves accuracy over time.

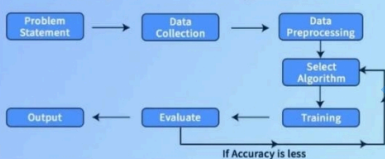
Impact of Data Quality:

- Good data → better predictions.
- Poor-quality data → poor results.

Key Idea:

- Understanding trends & patterns is essential.
- The right data helps the model keep learning and give correct answers

Machine Learning Model



Choosing the Right Algorithm

- Selecting the right algorithm is very important.
- Different algorithms are suitable for different types of data.
- **Key factors to consider:**
 - Data type (structured or unstructured)
 - Problem type (prediction, classification, grouping)
 - Data size & complexity

Data Types & Algorithms

1. Structured Data

- Organized into rows & columns (e.g., databases, spreadsheets, CSV files)
- Easy to search & quantify
- **Examples:** Student records, Bank accounts, sales report.
- **Suitable algorithms:** Regression, Classification

2. Unstructured Data

- No fixed format (e.g., text, images, videos, audio)
- **Examples:** Social media posts, emails, sensor readings
- **Suitable algorithms:** Neural Networks

Algorithm Selection by Problem Type

Supervised Learning → Uses labeled data

- **Example:** Predicting exam scores from study hours

Unsupervised Learning → Finds patterns in unlabeled data

- **Example:** Grouping customers by buying habits

Regression → Predicts continuous values (uses labeled data)

- **Example:** Forecasting house prices

Classification → Sorts data into categories (uses labeled data)

- **Example:** Email → spam or not spam

Clustering → Groups similar data points (uses unlabeled data)

- **Example:** Grouping similar images

Neural Networks → Best for large, complex patterns

- **Example:** Image recognition, speech analysis

Algorithm Selection: Data Size & Complexity

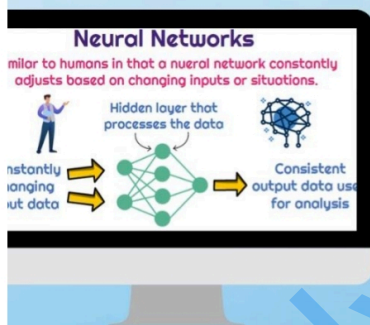
Small & Simple Data

- Use traditional algorithms (e.g., Linear Regression, Decision Trees)
- Fast to train and require less computing power
- **Example:** Predicting monthly sales from last year's records

Large & Complex Data

- Use advanced algorithms (e.g., Neural Networks, Deep Learning)
- Can process massive datasets in parallel (**using GPUs**)
- **Example:** Analyzing millions of social media posts about COVID-19

Neural Networks



- **Best for complex patterns** (e.g., colors, textures, shapes, angles in images)
- Suitable for **large datasets** (e.g., millions of social media posts)
- Can **process data in parallel** using powerful **GPUs**
- Ideal for **image recognition, speech processing, and natural language understanding**

Difference between predictive outcomes vs. causality

1. Predictive Outcomes

Using data to guess what will happen in the future.

Goal: Accuracy in prediction, not explaining "why."

Example: A model predicts that a student will score 85% in the next exam based on past marks and attendance.

Key point: It tells us what might happen, but not why it happens.

2. Causality

Understanding the cause-and-effect relationship between variables.

Goal: Explain why something happens.

Example: Finding out that regular study increases exam scores because it improves understanding and memory.

Key point: It focuses on the reason behind the outcome, not just predicting it.

Introduction to Data Visualization

Definition:

- Graphical representation of information and data.

Purpose:

- Helps identify trends, outliers, and patterns in data.

Why important:

- Makes complex data easier to understand.
- Helps present data to non-technical audiences clearly.



Visual Elements in Data Visualization

Common elements:

- Charts
- Graphs
- Maps

Benefits:

- **Accessibility:** easier for everyone to interpret data
- **Clarity:** visual format avoids confusion
- **Storytelling:** turns data into a narrative



Data Visualization Tools

SQL: A **database query language** used to **retrieve, insert, update, and manage** data stored in **relational databases**.

Python (Matplotlib): A **Python library** used to **create basic and advanced charts** such as line graphs, bar charts, and scatter plots.

Python (Seaborn): A Python library built on Matplotlib that is used for **creating attractive, statistical visualizations** with less code.

R Language: A programming language designed for **statistical analysis and data visualization**, widely used by statisticians and data scientists.

Interactive Visualization Tools: **Software or libraries** that allow users to **interactively explore, filter, and analyze data** in graphical form.

Choosing the Right Tool

Consider:

- Task type (statistical, interactive, simple chart)
- Skill level of user
- Audience (technical or non-technical)

Goal: Present findings clearly & effectively.

Common Visualization Types

KPIs



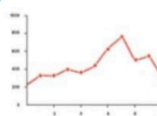
Tables

	A	B	C
X	\$40	240	48
Y	\$50	200	59
Z	\$60	310	79

Bar charts



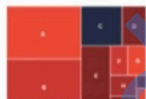
Line charts



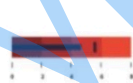
Donut charts



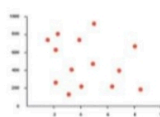
Tree Maps



Bullet Charts



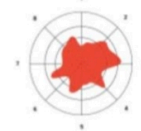
Scatter plots



Geo Maps



Radial charts



Studydrive Official

- KPIs:** Show a single important metric (e.g., revenue growth 180%) at a glance.

Example: Used on business dashboards to track sales or customer satisfaction.

- Tables:** Organize data in rows and columns for detailed comparison.

Example: Used in financial reports to list expenses, income, and balances.

- Bar Charts:** Compare categories using rectangular bars.

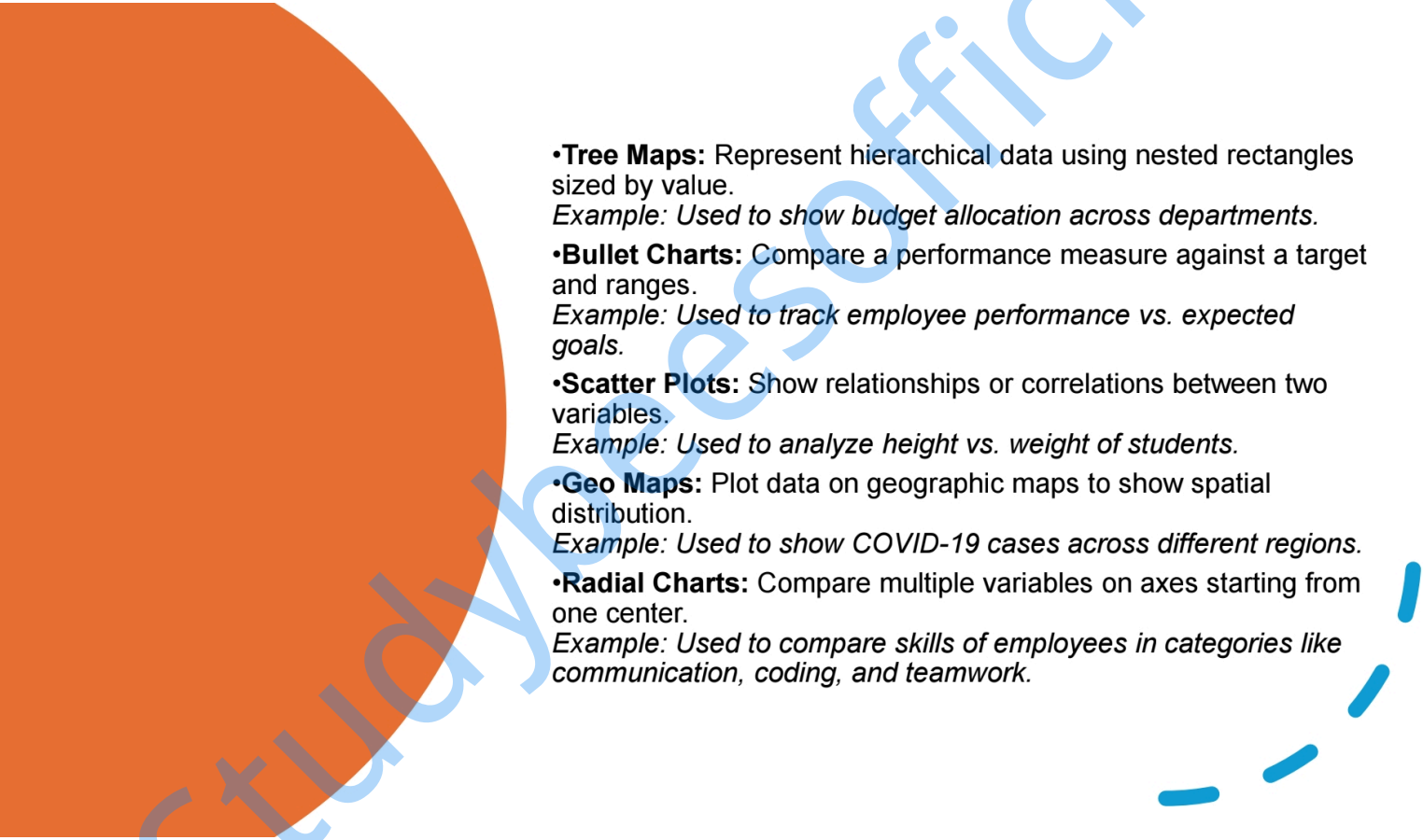
Example: Used to compare sales of products A, B, C, and D.

- Line Charts:** Display trends or changes over time with connected points.

Example: Used to show monthly temperature changes or stock price trends.

- Donut Charts:** Show proportions of a whole, like a pie chart with a center cut.

Example: Used to display market share of different companies.



•**Tree Maps:** Represent hierarchical data using nested rectangles sized by value.

Example: Used to show budget allocation across departments.

•**Bullet Charts:** Compare a performance measure against a target and ranges.

Example: Used to track employee performance vs. expected goals.

•**Scatter Plots:** Show relationships or correlations between two variables.

Example: Used to analyze height vs. weight of students.

•**Geo Maps:** Plot data on geographic maps to show spatial distribution.

Example: Used to show COVID-19 cases across different regions.

•**Radial Charts:** Compare multiple variables on axes starting from one center.

Example: Used to compare skills of employees in categories like communication, coding, and teamwork.

Data Visualization:



NumPy (np) → A library for numerical operations, especially arrays and math calculations.



Matplotlib (plt) → A plotting library to create charts, graphs, and visualizations.



sklearn.linear_model → Comes from **Scikit-Learn**, a machine learning library.



LinearRegression helps us build a **simple linear regression m**



X → Input values (study hours).



2D array `[[1],[2],...]` because Scikit-Learn expects inputs in **column form**.



Y → Output values (marks).

Data Visualization:



NP.Linspace(1, 5, 100) →
CREATES 100 EVENLY
SPACED VALUES BETWEEN 1
AND 5 (SMOOTH CURVE).



.RESHAPE(-1, 1) →
CONVERTS IT INTO A
COLUMN SHAPE (NEEDED
FOR THE MODEL).



MODEL.PREDICT(X_LINE) →
PREDICTS MARKS FOR ALL
THOSE 100 X VALUES →
HELPS US DRAW THE
REGRESSION LINE.



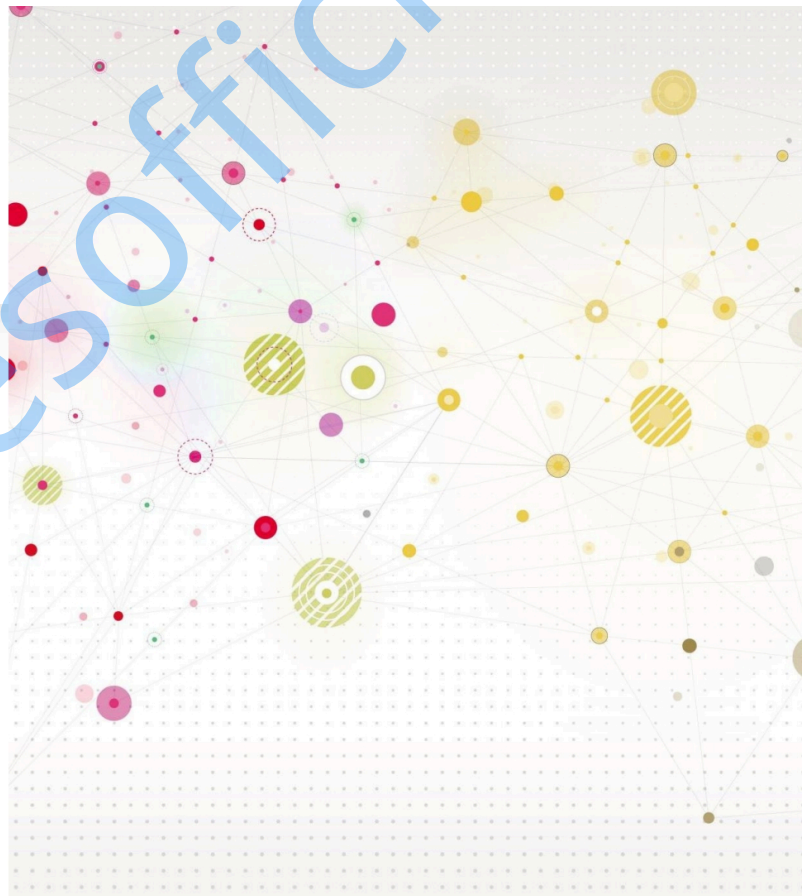
PLT.LEGEND() → DISPLAYS
COLOR LABELS ("TRAINING
DATA", "REGRESSION LINE",
ETC.).



PLT.SHOW() → FINALLY
SHOWS THE FULL GRAPH
WINDOW.

Data Storytelling

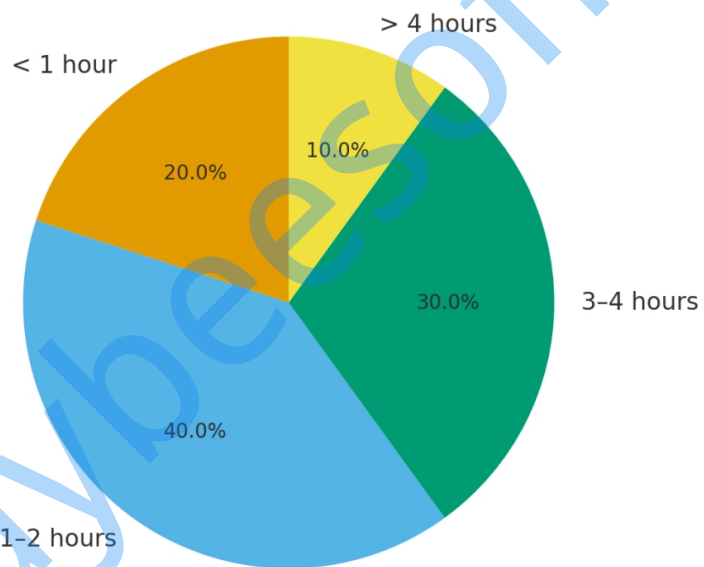
- Data storytelling is the **art of using data, visuals (like charts/graphs), and simple explanations** to tell a meaningful story.
- Instead of just showing numbers in a table, data storytelling helps people **understand patterns, trends, and insights** in a way that makes sense and influences decisions.
- **Data** gives facts.
- **Visualization** (charts, graphs) makes those facts easy to see
- **Storytelling** explains what those facts mean and why they matter.



Imagine your school conducted a survey about **students' study habits** and collected the following data:

Study Hours per Day	% of Students
Less than 1 hour	20%
1–2 hours	40%
3–4 hours	30%
More than 4 hours	10%

Study Hours per Day (Student Survey)



Data Storytelling: Example

Context: School improving student attendance

Data:

- Avg. attendance: 85%
- Highest: 9th grade (90%)
- Lowest: 12th grade (78%)
- Avg. absences: 5 days

Story:

- Meet **Jameel (12th grade)** – misses **5 days** on average
- **9th-grade** success shows that **improvement is possible**
- **Call to Action:** Apply 9th-grade strategies to 12th grade → make attendance a priority and support students' success!

Key Elements of the Story

1. **Relatable character** (Jameel)
 2. **Contextual data** (attendance rates, absences)
 3. **Insight** (9th-grade success factors)
 4. **Call to action** (apply to 12th-grade students)
- Impact:** Evokes empathy and inspires action, making complex information more engaging and memorable.

Formulating Questions

- Before we collect or analyze data, we must decide:
“What do we want to know?”
- This is called **formulating questions**.
Good questions are:
- **Clear** → easy to understand.
- **Specific** → focused on one thing.
- **Relevant** → related to the problem we are solving.

Scenario

- Your school principal wants to improve the **library services** for students.

To actually improve the library, we need **specific, useful questions**, such as:

- “How many times do you visit the library in a week?”
- “What type of books do you borrow the most (textbooks, storybooks, reference books, magazines)?”
- “What problems do you face in the library (lack of books, seating issues, opening hours, etc.)?”
- “How satisfied are you with the library staff support? (Very satisfied, Neutral, Dissatisfied)”



4.2.3 Formulating Questions

1. **Guides data exploration & analysis**
2. Identify relevant datasets
3. Evaluate the data's ability to answer the question
4. Compare new vs. previous findings
5. Refine questions after evaluation
6. **Good questions = accurate, meaningful insights**
7. Data evaluation helps in understanding and interpreting findings

Question Formulation: Example

Topic: Impact of School Breakfast Programs on Attendance
Questions:

1. Does participation improve attendance?
2. Which grade benefits most?
3. Do low-income students' attendance improve more?

Relevant Data:

- Attendance records
- Breakfast program participation
- Demographics (grade, income)

Data Evaluation

- Once we collect the data, we need to **check and analyze** it. This process is called **data evaluation**.
- In data evaluation, we:
- **Check accuracy** → Is the data correct?
- **Look for patterns** → What does the data show?
- **Draw conclusions** → What decisions can we make?

Example:

If survey data shows that **60% of students study less than 2 hours daily**, we can evaluate that:

- Many students may not be preparing enough.
- Teachers might need to arrange extra classes.

Question Formulation: Example

Data Evaluation

- Check for missing/inconsistent data
- Ensure attendance records are accurate
- Verify program participation data matches student records

Outcome:

- Focused exploration → meaningful insights → informed decisions

4.2.4 Descriptive Statistics Techniques

Purpose: Understand, summarize & describe data effectively

Core Measures:

- **Mean:** average value
- **Median:** middle value
- **Mode:** most frequent value
- **Range:** difference between the highest and the lowest
- **Standard Deviation:** spread of values
- **Variance:** average of squared differences from the mean
- **Percentiles:** threshold values
- **Histograms:** frequency distribution

Importance of Descriptive Statistics Techniques in Analysis

- Summarizes large datasets
- Identifies patterns & variations
- Provides a basis for Machine Learning models
- Supports accurate, data-driven decision-making

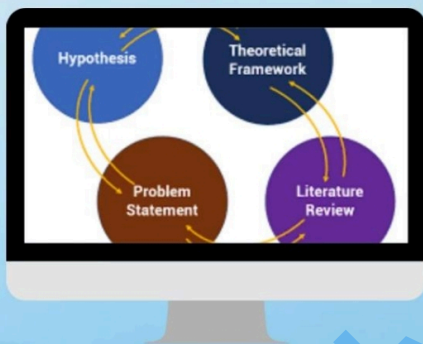
Hypothesis Formulation

- Hypothesis formation is the **process of developing a clear, testable statement** that predicts the **relationship between two or more variables**.
- Based on **existing data or prior knowledge**.
- Provides a **framework for testing relationships** in data.

Example:

"Increased study hours lead to better exam scores."

- **Independent Variable:** Study hours
- **Dependent Variable:** Exam scores
- **Cause-and-effect relationship:** One factor influences another.



Hypothesis Testing

Statistical process to check if data supports or refutes a hypothesis.

Used to validate:

- Assumptions
- Model performance
- Data patterns

There are two key hypotheses in hypothesis testing:

- Null Hypothesis
- Alternate Hypothesis
- **Main tools:** p-values and significance tests.

Two Key Hypotheses

Null Hypothesis (H_0)

1. No effect or relationship exists.
2. Any difference is due to chance.

Alternative Hypothesis (H_1)

1. A Significant effect or relationship exists.
2. What researchers aim to prove.

Importance of Both:

- Provide a clear framework for analyzing data.
- Guide the research and testing process.
- Ensure results are interpreted accurately.

Null vs. Alternative Hypothesis

Feature	Null Hypothesis (H_0)	Alternative Hypothesis (H_1)
Definition	A statistical statement that there is no effect or no relationship between the variables being studied, and any observed difference is due to random variation.	A statistical statement that there is an effect or relationship between the variables, and the observed difference is due to a real cause, not chance.
Purpose	Acts as the default assumption to be tested and possibly challenged.	Represents what the researcher wants to prove .
Interpretation of differences	Any observed differences are due to random chance .	Observed differences are due to a real effect .
Evidence needed	Accepted when there is insufficient evidence against it.	Accepted when there is strong evidence against H_0 .
Example	"Drinking coffee has no effect on cognitive function in students."	"Drinking coffee improves cognitive function in students."

P-Values

- **Measure how likely results are due to chance.**
- **Low P-value** → Strong evidence **against** the null hypothesis (H_0).
- **High P-value** → **Weak evidence against** the null hypothesis (H_0). Results may be due to random chance.
- Help in interpreting test results and ensuring conclusions are based on reliable evidence.

Example 1 – Coin Flip:

- Flip a coin 10 times → get 7 heads, 3 tails.
- Is the coin unfair, or did it happen by chance?
- P-value tells the probability of this happening randomly.

P-Values

Example 2 – Fertilizer Test:

- **Two plant groups:**
- Group A → gets fertilizer
- Group B → no fertilizer

After a month, fertilized plants grow taller.

- **Question: Did the fertilizer cause this, or was it random?**
- **P-value = 0.02** → Very unlikely by chance → Fertilizer likely effective.
- **P-value = 0.5** → Likely due to chance → No strong evidence of effect.

Significance Testing

Definition: Statistical method used to decide if a result is due to chance or statistically significant.

Statistically significant → Result is likely due to a real effect or factor.

Purpose:

- Determine if results are meaningful.
- Guide data-driven decisions.
- Validate research findings.

Process:

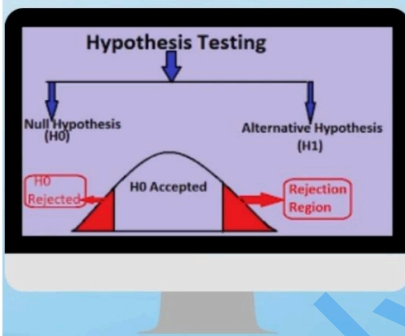
1. Compare test statistic with critical value.
2. Use a significance level (α) → threshold for deciding importance.
3. Decide whether to reject or fail to reject the null hypothesis.

Significance Testing

Example – Exercise & Blood Pressure:

- **H₀:** Exercise doesn't affect blood pressure.
- **H₁:** Exercise lowers blood pressure.
- **Data:** 6 weeks of exercise → average BP drop = 8 mmHg.
- **P-value** = 0.005 (0.05% chance result is random).
- **Decision:** **P-value** < α → Reject H₀, accept H₁.
- **Conclusion:** Exercise likely lowers blood pressure.

Hypothesis Testing: Introduction



Definition: Evaluates assumptions using statistical methods.

Test Statistic: Measures deviation from the null hypothesis (H_0).

Purpose: To draw valid conclusions from sample data.

Example Question:

- Does drinking coffee improve productivity?

Key Terms in Hypothesis Testing

- 1. Test Statistic:** A numerical value calculated from sample data to measure how far the data deviates from the null hypothesis.
- 2. Confidence Interval (CI):** A range of values within which the true population parameter is expected to lie with a given level of confidence (e.g., 95%).
- 3. Type I Error:** When a true null hypothesis is incorrectly rejected. Also called a "false positive."
- 4. Type II Error:** Failing to reject a false null hypothesis when it's actually false. Also called a "false negative."
- 5. Power of a Test:** The probability of correctly detecting a real effect when it exists. Higher power means better detection ability.
- 6. Significance Level (α):** The probability threshold (e.g., 0.05) below which results are considered statistically significant. How much evidence is enough to reject H_0 ?
- 7. Sample Size:** The number of observations in the study. Larger samples give more reliable and precise results.