

Practical Data Mining Using Microsoft Excel:

From Data Preparation to Knowledge Discovery

Politeknik Seberang Perai (PSP) | 29 Julai 2026



Pusat Pengajian Pengkomputeran
School of Computing
Universiti Utara Malaysia

INSTRUCTOR
Wan Hussain Wan Ishak
School of Computing,
Universiti Utara Malaysia

Modules

01 MODULE 1

INTRODUCTION TO DATA MINING

- What is Data Mining?
- Why Organizations Use Data Mining
- Data Warehouse vs. Data Mining
- The KDD Process
- Real-world Applications

02 MODULE 2

DATA PREPROCESSING USING EXCEL

- Data Quality Assessment
- Missing Values
- Outliers
- Normalization
- Attribute Selection
- Data Transformation
- Prepare Data for Mining

03 MODULE 3

CLASSIFICATION AND CLUSTERING

- Supervised vs. Unsupervised Learning
- Decision Tree (J48)
- Naïve Bayes
- K-Means Clustering
- Build Classifiers in Excel
- Model Evaluation
- Interpret Results

04 MODULE 4

ASSOCIATION RULE MINING



- Market Basket Analysis
- Association Rules
- Support
- Confidence
- Lift
- Apriori Algorithm
- Interpret Rules for Decision Making

MODULE 1

INTRODUCTION TO DATA MINING

Discover patterns. Extract knowledge.
Drive better decisions.



LEARNING OUTCOMES

At the end of this session, participants should be able to:

- **Define** data mining.
- **Explain** why organizations use data mining.
- **Differentiate** between Data Warehouse and Data Mining.
- **Describe** the Knowledge Discovery in Databases (KDD) process.
- **Identify** common applications of data mining.

Transform data into knowledge and knowledge into **action**.

Discover Patterns | Extract Knowledge | Support Decisions | Create Value

WHAT IS DATA MINING?

DEFINITION

Data Mining is the process of **discovering** useful **patterns**, **relationships**, **trends**, and **knowledge** from **large volumes of data** using **statistical methods**, **machine learning**, and **database technologies**.



THE DATA MINING PROCESS

DATA



Raw data is collected from various sources.

PREPROCESSING



Data is cleaned, integrated and prepared.

DATA MINING



Advanced techniques are applied to discover patterns and relationships.

PATTERNS & MODELS



Useful patterns, trends and models are identified.

KNOWLEDGE



Actionable knowledge supports better decision-making.



IN SIMPLE TERMS:

Data Mining turns raw data into **valuable knowledge** that helps organizations make **smarter decisions**, **solve problems**, and **gain a competitive advantage**.



DATA → INFORMATION → KNOWLEDGE

Turning data into better decisions

Data on its own has no meaning.

Through processing and analysis, data is transformed into information, then into knowledge, which leads to better decisions and actions.

RAW DATA

Facts and figures collected from various sources in their original, unorganized form.

7 2 9 9 6 4 3
5 8 1 9 2 0 6
3 4 7 8 5 1 2
9 6 3 2 7 4 8



PROCESSED DATA

Data that has been cleaned, sorted, organized, and formatted for analysis.

ID	Product	Region	Sales
101	Laptop	North	1250
102	Phone	South	980
103	Tablet	East	760
...	...	...	...

INFORMATION

Data with meaning and context that answers the questions "what", 'who", 'when", 'where".



- Top product: Laptop
- Highest sales: North region
- Total sales increased by 15%

KNOWLEDGE

Understanding and insights derived from information that reveal patterns and relationships.

Laptops sell best in the North. Marketing focus and inventory should be prioritized there.

DECISION

Applying knowledge to choose the best course of action.

Increase laptop stock in North region
Run targeted promotions
Optimize marketing budget

The goal: Convert data into knowledge, and knowledge into wise decisions.

WHY IS DATA MINING IMPORTANT?

Organizations collect **massive amounts of data** every day, including:

CUSTOMER TRANSACTIONS



Purchase history, payments, returns and more.

EMPLOYEE RECORDS



Personal details, attendance, performance and more.

FINANCIAL DATA



Income statements, expenses, budgets and more.

WEBSITE VISITS



Page views, clicks, time spent, and user behavior.

SENSOR AND IoT DATA



Machine readings, device status, temperature, location and more.

SOCIAL MEDIA INTERACTIONS



Likes, shares, comments, mentions and more.

WITHOUT ANALYSIS, THESE DATA HAVE LITTLE VALUE.

They are raw, unorganized, and difficult to understand or use for decision-making.





DATA MINING CONVERTS THIS DATA INTO ACTIONABLE INSIGHTS.

- Discover hidden patterns
- Improve decision making
- Increase operational efficiency
- Gain competitive advantage



WHY ORGANIZATIONS NEED DATA MINING



Most organizations ask questions such as:

- Who are our best customers?
- Which products sell together?
- Which students are likely to fail?
- Which patients are at high risk?
- Which employees are likely to resign?
- How can we increase sales?



DATA MINING HELPS ANSWER THESE QUESTIONS AND SUPPORTS BETTER DECISIONS.



DATA MINING IN ACTION

A supermarket has **one million** transaction records.

WITHOUT DATA MINING

- **Impossible** to manually identify buying patterns.



1,000,000 transaction records

WITH DATA MINING

The system discovers

Customers who buy **Bread** also frequently buy **Milk**.





The supermarket then **places Bread and Milk nearby** or creates **promotional bundles.**



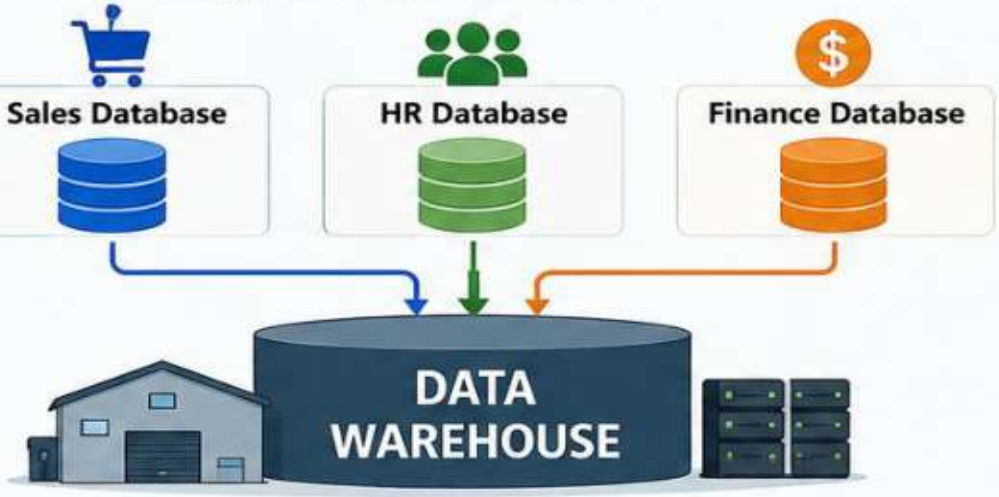
OR



DATA MINING vs DATA WAREHOUSE

DATA WAREHOUSE

Purpose:
Store large amounts of integrated historical data.



Think of it as **a large library.**



Key Characteristics

- Stores historical data
- Integrated from multiple sources
- Subject-oriented
- Supports reporting and analysis
- Stable and time-variant

DATA MINING

Purpose:
Extract hidden knowledge from stored data.



Think of it as **a researcher** exploring the library to discover new insights.



Key Characteristics

- Discovers hidden patterns
- Predictive and descriptive
- Supports decision making
- Uses techniques like classification, clustering, association, prediction

KNOWLEDGE DISCOVERY IN DATABASES (KDD) PROCESS

Data Mining is only one step in the KDD process.

1 DATA COLLECTION



Gather data from various sources such as databases, files, sensors, websites, and other systems.

2 DATA CLEANING



Remove errors, duplicates, missing values, and inconsistencies to improve data quality.

3 DATA INTEGRATION



Combine data from multiple sources into a consistent and unified format.

4 DATA SELECTION



Select relevant data that is useful for the analysis and discard unnecessary data.

5 DATA TRANSFORMATION



Convert data into the appropriate format or structure for mining, such as normalization or aggregation.

6 DATA MINING



Apply data mining techniques and algorithms to discover patterns, relationships, and trends.

This is the Data Mining Step

7 PATTERN EVALUATION



Evaluate and assess the discovered patterns to identify interesting, valid, and useful knowledge.

8 KNOWLEDGE PRESENTATION



Present the discovered knowledge in a visual or understandable format to support decision-making.



THE GOAL OF KDD:

To extract useful, understandable, and actionable knowledge from large amounts of data.



REAL-WORLD APPLICATIONS

Data Mining turns data into decisions that improve lives and drive business success.

EDUCATION



- Predict students at risk
- Improve graduation rates
- Personalize learning

HEALTHCARE



- Disease prediction
- Patient monitoring
- Drug discovery

BANKING



- Fraud detection
- Credit scoring
- Customer segmentation

RETAIL



- Product recommendations
- Market basket analysis
- Inventory forecasting

MANUFACTURING



- Predictive maintenance
- Quality control
- Supply chain optimization

GOVERNMENT



- Crime pattern analysis
- Traffic management
- Tax fraud detection
- Public health surveillance

SUMMARY



DATA

Raw facts and figures



ANALYSIS

Explore patterns and relationships



KNOWLEDGE

Discover meaningful insights



BETTER DECISIONS

Apply insights to support better outcomes

Data mining is not just about analyzing data—
it is about discovering knowledge that supports better decisions.

MODULE 2

DATA PREPROCESSING USING MICROSOFT EXCEL

Clean. Prepare. Transform.
Build quality data for reliable insights.



LEARNING OUTCOMES

At the end of this module, participants should be able to:

01 **Explain** the importance of data preprocessing.

02 **Assess** the quality of a dataset.

03 **Identify** and handle missing values.

04 **Detect** outliers.

05 **Normalize** numerical data using Excel.

06 **Select** relevant attributes for analysis.

07 **Transform** data into a suitable format for data mining.

08 **Use** Microsoft Excel to prepare data before applying data mining techniques.

Well-prepared data leads to **better** results and more **meaningful insights.**

- Improve Data Quality
- Enhance Model Performance
- Ensure Reliable Results
- Support Better Decisions

WHY DATA PREPROCESSING?

"If your data are inaccurate or incomplete, can you expect accurate analysis?"

The answer is **NO.**





"Garbage In, Garbage Out (GIGO)."

No matter how advanced the analysis or machine learning algorithm, poor-quality data will produce poor-quality results.





Research indicates that

60–80%

of a data analyst's time **is spent preparing data** rather than analyzing it.





Good data in → Good insights out.
Quality preprocessing is the foundation of successful data mining.

OVERVIEW OF DATA PREPROCESSING

RAW DATA
Data collected from various sources in its original, unprocessed form.

↓

DATA CLEANING
Detect and correct errors, handle missing values, remove duplicates and inconsistencies.

↓

DATA INTEGRATION
Combine data from multiple sources into a consistent and unified format.

↓

DATA TRANSFORMATION
Convert data into appropriate formats or structures suitable for analysis.

↓

DATA REDUCTION
Reduce the volume of data while preserving important information.

↓

PREPARED DATA
High-quality, consistent, and relevant data ready for analysis and modeling.



In this workshop, we will focus on the key data preprocessing tasks using **Microsoft Excel.**

- Practical
- Simple
- Hands-on

UNDERSTANDING DATA

Exploring and Examining Data Before Analysis

WHAT IS UNDERSTANDING DATA?

Understanding Data is the process of exploring and examining a dataset to understand its structure, content, quality, and characteristics before performing any analysis or data mining. It is the first step in data preprocessing and helps identify potential issues that may affect the accuracy of the results.

KEY IDEA

Before analysing data, you must first understand what the data represent, how they are organised, and whether they are suitable for the intended analysis.

WHY IS UNDERSTANDING DATA IMPORTANT?

- Identify the available attributes (columns).
- Determine the type of each attribute.
- Understand the meaning of each variable.
- Detect missing, inconsistent, or invalid values.
- Identify potential outliers.
- Observe data distribution.
- Select appropriate preprocessing and analysis techniques.

QUESTIONS TO ASK BEFORE ANALYSIS

- What problem am I trying to solve?
- What does each attribute represent?
- Which attributes are numerical?
- Which attributes are categorical?
- Are there missing or inconsistent values?
- Are there any duplicate records?
- Which attributes are relevant to the analysis?

WHAT SHOULD YOU EXAMINE?

Aspect	Description	Example
Number of Records	Total number of observations (rows).	100 students
Number of Attributes	Total number of variables (columns).	10 attributes
Attribute Types	Identify data types (Numeric, Nominal, Ordinal, etc.).	Attendance = Numeric
Value Ranges	Check minimum and maximum values.	Attendance: 45 – 98
Missing Values	Identify blank or null values.	Missing Attendance
Data Distribution	Observe how values are distributed.	Marks follow a normal distribution
Duplicates	Detect repeated records.	Duplicate StudentID

TWO MAIN CATEGORIES of data types

1. QUALITATIVE (CATEGORICAL) DATA

Describes qualities or categories.

1.1 NOMINAL

Categories with no **natural order.**

Examples

- Gender (Male, Female)
- Blood Type (A, B, AB, O)
- Country (Malaysia, Japan)
- Colour (Red, Blue, Green)

1.2 ORDINAL

Categories with a **natural order or ranking.**

Examples

- Education Level (Primary, Secondary, Tertiary)
- Satisfaction (Low, Medium, High)
- Grade (A, B, C, D, F)
- Ranking (1st, 2nd, 3rd)

2. QUANTITATIVE (NUMERICAL) DATA

Represents numbers that can be measured or counted.

2.1 DISCRETE

Numeric values that are **countable.**

Examples

- Number of students in a class
- Number of cars in a parking lot
- Number of siblings
- Number of transactions

2.2 CONTINUOUS

Numeric values that **can take any value within a range.**

Examples

- Height (e.g., 165.5 cm)
- Weight (e.g., 60.75 kg)
- Temperature (e.g., 36.7 °C)
- Time (e.g., 2.35 hours)

HOW THEY DIFFER

Aspect	Nominal	Ordinal	Discrete	Continuous
Nature	Categories, no order	Categories, with order	Numbers, countable	Numbers, measurable
Example Values	Male, Female	Low, Medium, High	0, 1, 2, 3, …	10.5, 10.51, 10.52, …
Arithmetic Operations	✗ Not meaningful	✗ Not meaningful	✓ Meaningful	✓ Meaningful
Order/Ranking	✗ No order	✓ Has order	✗ No natural order	✗ No natural order
Typical Charts	• Bar Chart • Pie Chart	• Bar Chart • Pareto Chart	• Bar Chart • Column Chart	• Histogram • Line Chart

QUICK EXAMPLES

- Gender → Nominal
- Service Quality (Poor, Fair, Good, Excellent) → Ordinal
- Number of books borrowed → Discrete
- Body Temperature → Continuous

ACTIVITY 1 Load and Explore the Dataset

USE: Data1_Student_Performance.xlsx

OBSERVE THE DATASET.

- How many records are there? — 100 records
- How many attributes (columns)? — 10 attributes
- What information is stored?
- Which columns are numerical?
- Which columns are categorical?

File Home Insert Page Layout Formulas Data Review View Help Tell me

	A	B	C	D	E	F	G	I	J	K
1	Student ID	Gender	Programme	Attendance (%)	Assignment (%)	Quiz (%)	Final Exam (%)	Total (%)	Grade	Status
2	1001	Female	BBA	95	85	90	88	89.5	A	Active
3	1002	Male	BIT	80	70	65	70	71.3	B	Active
4	1003	Female	BCS	90	92	88	94	91.0	A+	Active
5	1004	Male	BBA	75	60	70	70	68.8	B	Probation
6	1005	Female	BIT	98	88	92	90	92.0	A	Active
6	...	...	...	...	...	...	...	...	...	...

ATTRIBUTE TYPES

Attribute	Type	Explanation
Student ID	Identifier	Unique identifier for each student.
Gender	Nominal	Categories with no inherent order (e.g., Male, Female).
Programme	Nominal	Categories with no inherent order (e.g., BBA, BIT, BCS).
Attendance	Numeric	Percentage value (0–100).
Assignment	Numeric	Percentage value (0–100).
Quiz	Numeric	Percentage value (0–100).
Final Exam	Numeric	Percentage value (0–100).
Total	Numeric	Calculated total percentage (0–100).
Grade	Ordinal	Ordered categories (e.g., A+, A, B+, B, C).
Status	Nominal	Categories with no inherent order (e.g., Active, Probation).

NOTE: Understanding attribute types helps determine the appropriate preprocessing and analysis techniques.

EXCEL FEATURES USED

FILTER
Filter data to view specific values or conditions.

SORT
Sort data in ascending or descending order to identify patterns.

FREEZE PANES
Freeze rows/columns to keep headers visible while scrolling.

TABLE FORMAT
Use Table Format for better data organization and quick analysis.

DESCRIPTIVE ANALYSIS OF STUDENT PERFORMANCE DATASET

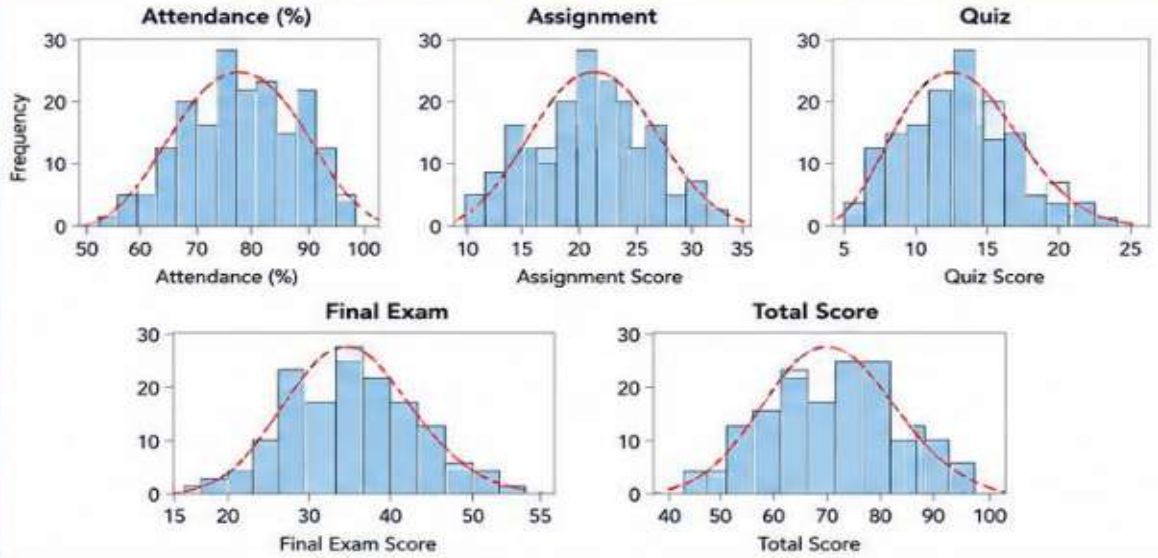
Basic Analysis: Symmetry, Distribution Curve, Skewness, and Kurtosis

1. DESCRIPTIVE STATISTICS

Variable	N	Mean	Median	Std. Dev.	Min	Max	Skewness	Kurtosis
Attendance (%)	100	77.19	78.50	13.23	55	98	-0.069	-1.323
Assignment	100	22.91	23.00	4.76	12	30	-0.187	-0.597
Quiz	100	14.90	15.00	3.15	7	20	-0.239	-0.548
Final Exam	100	37.98	38.00	8.81	19	50	-0.283	-0.927
Total Score	100	75.79	75.00	14.73	46	100	-0.187	-0.992

Skewness: -ve = Left Skewed, +ve = Right Skewed, 0 = Symmetric | Kurtosis: < 0 = Platykurtic, 0 = Mesokurtic (Normal), > 0 = Leptokurtic

2. DISTRIBUTION CURVES (HISTOGRAMS WITH NORMAL CURVE)



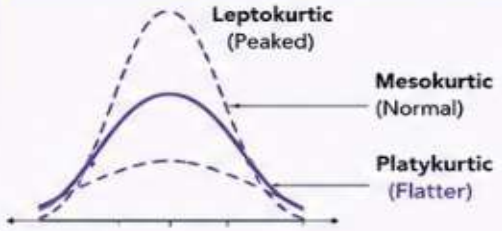
3. SYMMETRY (SKEWNESS) ANALYSIS



- ✓ All skewness values are between -0.283 and -0.069.
- ✓ Values are close to 0, which indicates approximately symmetric distributions.
- ✓ Slight negative skewness shows a small tail on the left side (more values on the higher end).
- ✓ No variable shows strong left or right skewness.

Conclusion: The dataset is approximately symmetric.

4. KURTOSIS ANALYSIS



- All kurtosis values are negative.
- This indicates platykurtic distributions.
- Distributions are flatter than normal.
- Fewer extreme values (outliers).
- Scores are more evenly spread.

5. OVERALL DISTRIBUTION SUMMARY

- Mean and median are very close for all variables.
- Skewness values are close to zero (approximately symmetric).
- Kurtosis values are negative (platykurtic distributions).
- No indication of severe outliers from descriptive measures.
- Dataset is well balanced and suitable for analysis.

6. KEY TAKEAWAYS

- The dataset (100 students) shows balanced performance across all components.
- Attendance, Assignment, Quiz, Final Exam and Total Score are all approximately symmetric.
- Distributions are flatter than normal with fewer extreme values.
- The data is suitable for statistical modeling, classification, clustering and data mining techniques.
- No data transformation is required for normality assumptions in most cases.

DESCRIPTIVE ANALYSIS OF STUDENT PERFORMANCE DATASET

Basic Analysis: Symmetry, Distribution Curve, Skewness, and Kurtosis

ANALYSIS	DEFINITION	INTERPRETATION GUIDE	HOW TO DO IN EXCEL

1. CENTRAL TENDENCY
(Mean, Median, Mode)



Measures the typical or central value of the data.

DEFINITION

- **Mean** : The average of all values.
- **Median** : The middle value when data are sorted.
- **Mode** : The most frequently occurring value.

INTERPRETATION GUIDE

- Used to understand the "center" of the data.
- If mean ≈ median → data is symmetric.
- If mean > median → right skewed.
- If mean < median → left skewed.
- Mode shows the most common score.

HOW TO DO IN EXCEL

Assume data in range B2:B101

- **Mean** : =AVERAGE(B2:B101)
- **Median** : =MEDIAN(B2:B101)
- **Mode** : =MODE.SNGL(B2:B101)

(Use Data Analysis ToolPak for Mode in older Excel)

2. DISPERSION (SPREAD)
(Standard Deviation, Min, Max, Range)



Measures how spread out the data values are.

DEFINITION

- **Standard Deviation (SD):** Average distance of data from the mean.
- **Min** : Smallest value.
- **Max** : Largest value.
- **Range** : Max – Min.

INTERPRETATION GUIDE

- Small SD → values are close to the mean (consistent).
- Large SD → values are more spread out (less consistent).
- Range gives quick idea of data spread.

HOW TO DO IN EXCEL

Assume data in range B2:B101

- **Std. Dev. (Sample)** : =STDEV.S(B2:B101)
- **Min** : =MIN(B2:B101)
- **Max** : =MAX(B2:B101)
- **Range** : =MAX(B2:B101)-MIN(B2:B101)

3. DISTRIBUTION CURVE
(Histogram with Normal Curve)



Shows the shape of data distribution and how it compares to a normal curve.

DEFINITION

- **Histogram:** Bars show frequency of values in intervals (bins).
- **Normal Curve:** Bell-shaped curve representing a normal distribution.
- Used to visualize symmetry, skewness, and outliers.

INTERPRETATION GUIDE

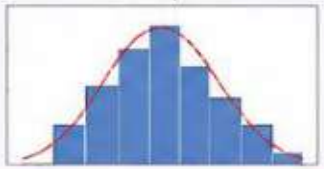
- Bell shape → data is approximately normal.
- Left tail longer → left skewed (more low values).
- Right tail longer → right skewed (more high values).
- Multiple peaks → possible subgroups.

HOW TO DO IN EXCEL

Steps:

1. Select data.
2. Insert → Statistical Chart → Histogram.
3. Right-click X-axis → Format Axis → set Bin Width (e.g., 5 or 10).
4. (Optional) Add Normal Curve:
   - Insert → Scatter with Smooth Line.
   - Calculate normal curve values using mean & SD (see below) and plot over histogram.

Example



Normal Curve Formula (for overlay)

y = NORM.DIST(x, mean, sd, FALSE) × n × bin_width

x = midpoint of each bin
mean = =AVERAGE(range)
sd = =STDEV.S(range)
n = =COUNT(range)
bin_width = chosen bin width

4. SKEWNESS (SYMMETRY)



Measures the asymmetry of the distribution.

DEFINITION

- Skewness indicates the direction and degree of asymmetry.
- Skewness = 0 → symmetric.
- Negative → left skewed.
- Positive → right skewed.

INTERPRETATION GUIDE

Guideline

Skewness Value	Interpretation
< -1	Highly Left Skewed
-1 to -0.5	Moderately Left Skewed
-0.5 to +0.5	Approximately Symmetric
+0.5 to +1	Moderately Right Skewed
> +1	Highly Right Skewed

HOW TO DO IN EXCEL

Assume data in range B2:B101

- **Skewness** : =SKEW(B2:B101) (Sample)

(Use Data Analysis ToolPak → Descriptive Statistics to get skewness automatically)

Tip: If mean ≈ median and skewness is close to 0, data is approximately symmetric.

5. KURTOSIS (PEAKNESS)



Measures the peakedness or tailedness of the distribution.

DEFINITION

- Kurtosis compares data distribution with normal distribution.
  - = 0 → **Mesokurtic** (normal)
  - > 0 → **Leptokurtic** (more peaked, heavier tails)
  - < 0 → **Platykurtic** (flatter, lighter tails)

INTERPRETATION GUIDE

Guideline

Kurtosis Value	Interpretation
< 0	Platykurtic (flatter than normal)
= 0	Mesokurtic (normal)
> 0	Leptokurtic (more peaked)

Implication:

Negative kurtosis → fewer extreme values (outliers).
Positive kurtosis → more extreme values (outliers).

HOW TO DO IN EXCEL

Assume data in range B2:B101

- **Kurtosis** : =KURT(B2:B101) (Excess Kurtosis)

(Use Data Analysis ToolPak → Descriptive Statistics to get kurtosis automatically)

Tip: Excess kurtosis of 0 means normal (mesokurtic). Excel returns excess kurtosis.

FREQUENCY ANALYSIS OF CATEGORICAL DATA

Using Student Performance Dataset (100 Records)

Dataset Overview
100 student records with 4 categorical variables: Gender, Programme, Grade, Status

WHAT IS FREQUENCY ANALYSIS?

Frequency Analysis is the process of counting how many times each category appears in a dataset.

It helps us understand the distribution and composition of categorical (nominal or ordinal) variables.

OBJECTIVES

- Identify how often each category occurs.
- Calculate the percentage of each category.
- Determine the most common or dominant category.
- Provide a simple summary of categorical data.

1. FREQUENCY TABLE RESULTS (Based on Your Dataset – 100 Records)

A. GENDER

Gender	Frequency	Percentage
Male	53	53%
Female	47	47%
Total	**100**	**100%**

Male students are 53% while female students are 47%.

B. PROGRAMME

Programme	Frequency	Percentage
Information Technology	29	29%
Data Science	25	25%
Computer Science	24	24%
Software Engineering	22	22%
Total	**100**	**100%**

Information Technology has the highest representation (29%).

C. GRADE

Grade	Frequency	Percentage
A	34	34%
B	33	33%
C	13	13%
D	16	16%
F	4	4%
Total	**100**	**100%**

Most students achieved Grade A or B (67% in total).

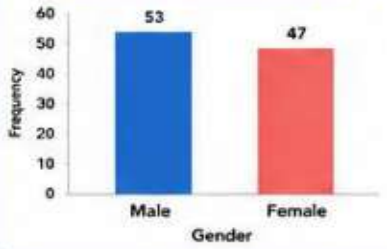
D. STATUS

Status	Frequency	Percentage
Pass	96	96%
Fail	4	4%
Total	**100**	**100%**

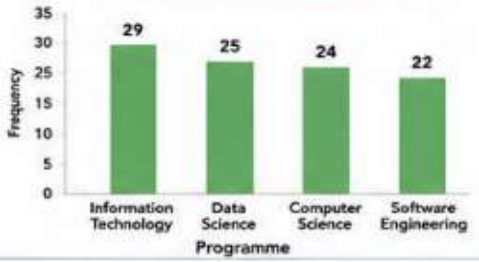
96% of students passed and 4% failed.

2. VISUALIZATION OF FREQUENCY RESULTS

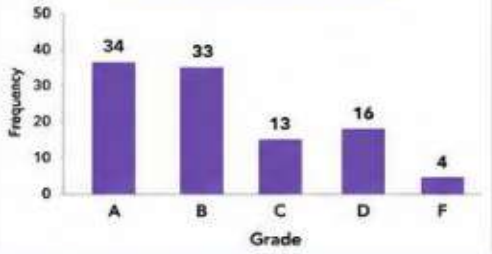
A. Gender Distribution (Bar Chart)



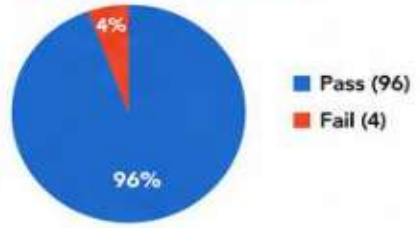
B. Programme Distribution (Bar Chart)



C. Grade Distribution (Bar Chart)



D. Status Distribution (Pie Chart)



3. HOW TO PERFORM FREQUENCY ANALYSIS IN MICROSOFT EXCEL

METHOD 1: USING PIVOT TABLE (RECOMMENDED)

1 Select your dataset.

Gender	Programme	Grade	Status
Male	Data Science	A	Pass
Female	Computer Science	B	Pass
Male	Information Technology	A	Pass
Female	Data Science	C	Pass
Male	Software Engineering	B	Fail
...	...	...	...

2 Go to Insert → PivotTable.



3 In PivotTable Fields:

- Drag the categorical variable (e.g., Gender) to **Rows**.
- Drag the same field to **Values**.



4 Excel will automatically count the frequency.

Row Labels	Count of Gender
Male	53
Female	47
Grand Total	**100**

Tip: Repeat the same steps for Programme, Grade and Status variables.

METHOD 2: USING COUNTIF FUNCTION

Assume your data is in columns: **Gender (B), Programme (C), Grade (D), Status (E)** from **row 2 to 101**.

A. Gender

Category	Formula	Frequency	%
Male	=COUNTIF(B2:B101,"Male")	53	53%
Female	=COUNTIF(B2:B101,"Female")	47	47%
Total	=SUM(F2:F3)	**100**	**100%**

Percentage formula:

=F2/SUM($F$2:$F$3) → Copy down and format as Percentage.

Other Variables

B. Programme (Column C)

=COUNTIF(C2:C101,"Information Technology")
=COUNTIF(C2:C101,"Data Science")
=COUNTIF(C2:C101,"Computer Science")
=COUNTIF(C2:C101,"Software Engineering")

C. Grade (Column D)

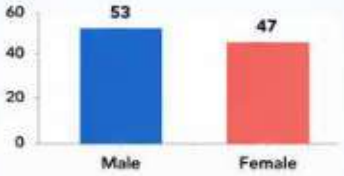
=COUNTIF(D2:D101,"A")
=COUNTIF(D2:D101,"B")
=COUNTIF(D2:D101,"C")
=COUNTIF(D2:D101,"D")
=COUNTIF(D2:D101,"F")

D. Status (Column E)

=COUNTIF(E2:E101,"Pass")
=COUNTIF(E2:E101,"Fail")

Example Output (Gender)

Category	Frequency	Percentage
Male	53	53%
Female	47	47%
Total	**100**	**100%**



DATA QUALITY

Data quality refers to whether data are **suitable for analysis and decision-making.**

A HIGH-QUALITY DATASET SHOULD BE:

ACCURATE	Data correctly reflects the real-world values.
COMPLETE	All required data are present with no missing parts.
CONSISTENT	Data is uniform across different sources and time.
VALID	Data follows the rules, formats, and constraints.
UNIQUE	No duplicate records or repeated data.
TIMELY	Data is up-to-date and available when needed.

COMMON DATA QUALITY PROBLEMS

Problem	Example	
Missing values	StudentID: 1001 / Name: Ali / Attendance (%):	Attendance is blank
Duplicate records	StudentID: 1002 / Name: Aina / Programme: Data Science (×2)	Same student appears twice
Incorrect values	StudentID: 1003 / Attendance (%): 150	Attendance = 150
Inconsistent values	StudentID: 1004 / Gender: Male StudentID: 1005 / Gender: male StudentID: 1006 / Gender: M	Male, male, M
Outliers	StudentID: 1007 / Total Marks: 250	Marks = 250

TIP: Good data quality improves the reliability of analysis, strengthens decisions, and builds trust in your results.

Activity 2 – Data Quality Inspection

Use: Data1_Student_Performance.xlsx

Inspect the dataset.

Questions:

- Are there any blank cells?
- Any impossible values?
- Any duplicate Student IDs?
- Are all programmes spelled consistently?

Based on the **Student_Performance_Dataset_100.xlsx**, the dataset inspection results are:

Inspection Question	Result	Remarks
Are there any blank cells?	✅ No	All 100 records have complete values. No missing data were found.
Any impossible values?	✅ No	All numerical values are within their expected ranges (e.g., Attendance = 0–100%, Assignment = 0–30, Quiz = 0–20, Final Exam = 0–50, Total = 0–100).
Any duplicate Student IDs?	✅ No	Every StudentID is unique. No duplicate records were detected.
Are all programmes spelled consistently?	✅ Yes	The programme names are consistently recorded as: **Computer Science**, **Data Science**, **Information Technology**, and **Software Engineering**.

Check carefully for data quality issues before analysis.

Excel Tools

Filter
Filter data to view specific values or conditions.

Sort
Sort data in ascending or descending order.

Find
Find specific text or values in the dataset.

Conditional Formatting
Highlight cells that meet certain rules or conditions.

DATA CLEANING

Improving Data Quality for Better Analysis and Decisions

Data Cleaning is the process of identifying and correcting errors, inconsistencies, and inaccuracies in a dataset to improve its quality before analysis.
It ensures that the data are accurate, complete, consistent, and reliable, leading to more meaningful insights and better decision-making.

WHY IT MATTERS

- Improves data quality and reliability
- Reduces errors in analysis
- Enhances model performance
- Supports confident decision-making

COMMON DATA CLEANING TASKS

1. REMOVE DUPLICATES

Identify and remove duplicate records to avoid redundancy.

EXAMPLE
Remove repeated StudentID or identical records.

2. HANDLE MISSING VALUES

Detect missing or blank values and decide how to handle them.

EXAMPLE
Fill missing Attendance with mean/median or remove the record.

3. CORRECT INVALID OR INCONSISTENT DATA

Fix incorrect, inconsistent, or out-of-range values.

EXAMPLE
Attendance cannot be 120%; correct to a valid range (0–100%).

4. DETECT AND TREAT OUTLIERS

Identify outliers that may distort analysis and decide how to treat them.

EXAMPLE
A Total Exam score of 5 when most scores are between 40–90.

5. STANDARDISE DATA FORMATS

Ensure consistent formats for dates, text, categories, and codes.

EXAMPLE
Convert dates to DD/MM/YYYY; use consistent Grade codes (A, B, C, D, F).

ACTIVITY 3 – HANDLING MISSING VALUES

MISSING VALUES OCCUR BECAUSE:

- Data entry mistakes
- Equipment failure
- Incomplete forms
- Lost records

METHODS TO HANDLE MISSING VALUES

Method	Suitable When
Delete record	Very few missing values
Replace with mean	Numerical attributes
Replace with median	Skewed data
Replace with mode	Categorical data
Estimate using nearby values	Advanced analysis

EXCEL DEMONSTRATION

Use: Data2_Student_Performance.xlsx

1 Count Missing Values

=COUNTBLANK(D2:D101)

	D
1	Attendance (%)
2	95
3	---
4	80
5	---

Returns the number of blank cells in D2:D101. (Example: 2)

2 Highlight Missing Values

Home
→ Conditional Formatting
→ Highlight Cell Rules
→ Blanks



3 Replace Missing Attendance with Average

=AVERAGE(D2:D101)

	D
1	Attendance (%)
2	95
3	---
4	80
5	70

→

	D
1	Attendance (%)
2	95
3	78.25
4	80
5	70

Average value is used to replace blank cells.

Notes:

- Replace only when appropriate.
- Avoid changing too much information.
- Always document what method you used.

DISCUSSION

Should we always replace missing values with the average?

- Consider the data distribution.
- The average can be affected by extreme values (outliers).
- Replacing may reduce data variability.
- Sometimes deletion or advanced estimation is better.

Think about the impact of each method on your analysis and final results.

ACTIVITY 4 – DETECTING OUTLIERS

What is an Outlier?

An outlier is a value that is much higher or lower than the rest of the data.

Example:

Attendance	85	90	88	87	20

The value 20 may be an outlier.

Why Outliers Matter

Outliers can:

- Distort averages
- Affect predictions
- Bias machine learning models

Dataset Preview (Modified) – Focus: Final Exam

	A	B	C	D	F	G	H	I	J	J
1	StudentID	Gender	Programme	Attendance (%)	Assignment	Quiz	Final Exam	Total	Grade	Status
2	1001	Female	BBA	95	28	18	45	91	A	Pass
3	1002	Male	BIT	80	20	15	38	73	B	Pass
4	1003	Female	BCS	90	27	17	44	88	A	Pass
...	...	...	...	...	...	...	...	...	...	...
45	1044	Male	BCS	82	22	15	85	100	A	Pass
73	1072	Male	BIT	79	21	14	92	100	A	Pass
96	1095	Female	BBA	88	25	18	100	100	A	Pass
...	...	...	...	...	...	...	...	...	...	...
100	1099	Male	BCS	76	16	11	37	64	C	Pass

Modified dataset contains extreme Final Exam scores: 85, 92, 100

EXCEL DEMONSTRATION

Use: Data2_Student_Performance.xlsx

1 Sort Final Exam

Identify unusually high or low values.

	A	B	C	D
1	StudentID	Gender	Programme	Final Exam
2	1050	Male	BIT	19
3	1089	Female	BCS	20
4	1026	Male	BBA	21
...	...	...	...	...
45	1044	Male	BCS	85
73	1072	Male	BIT	92
96	1095	Female	BBA	100

Lowest values

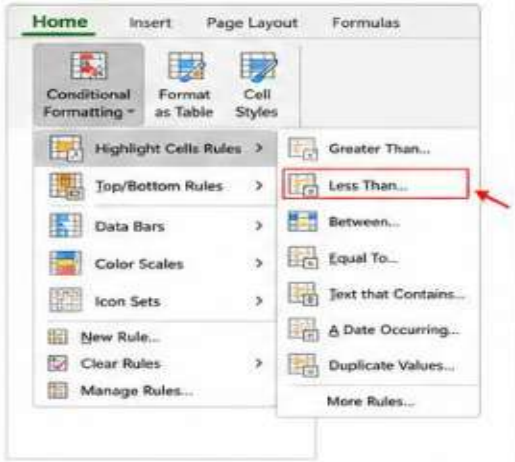
Highest values (possible outliers)

Sort from Smallest to Largest (or Largest to Smallest) to quickly spot unusual values.

2 Conditional Formatting

Highlight values below 60.

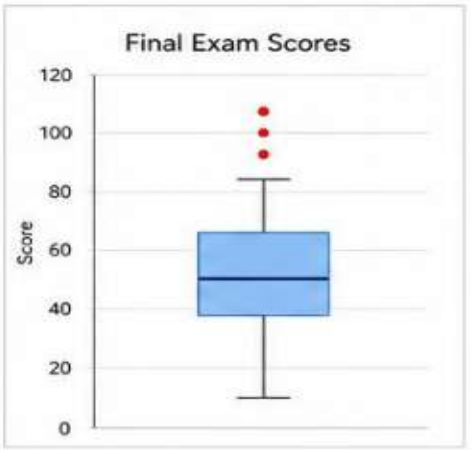
	D
1	Final Exam
2	45
3	38
4	44
5	32
6	61
7	70
8	58
9	85
10	92
11	100



Values < 60 are highlighted.

3 Box Plot (Excel 2016+)

Insert → Statistical Chart → Box and Whisker



- **Median** — Middle value of the dataset.
- **Quartiles** — Q1 (25%) and Q3 (75%) show the middle 50% of the data.
- **Whiskers** — Show the most extreme values that are not outliers.
- **Outliers** — Points outside the whiskers are outliers.

Outliers (red dots) are extreme values that are much higher than the rest of the data.

DISCUSSION

Should we remove every outlier?

- Not always.
- Some outliers are valid observations (e.g., exceptional students).
- Investigate the cause before deciding to remove or adjust.

- Check data entry errors
- Understand the context
- Decide: Keep, Adjust, or Remove

Remove only when the outlier is a clear error or not relevant to the analysis.

DATA INTEGRATION

Combining Data from Multiple Sources for a Unified View

Data Integration is the process of combining data from multiple sources into a single, consistent, and unified dataset for analysis.

It ensures that information from different systems can be used together without duplication or inconsistency.

WHEN IS DATA INTEGRATION NEEDED?

Data integration is required when data are stored in different files, databases, departments, or organisations. Common situations include:

- Combining student records from multiple campuses.
- Integrating academic, attendance, and examination data.
- Merging sales data from different branches.
- Combining customer information from CRM, websites, and online transactions.

WHY IS DATA INTEGRATION IMPORTANT?

Data integration enables organisations to:

- Obtain a complete and unified view of the data.
- Eliminate duplicate and inconsistent records.
- Improve data quality and consistency.
- Support comprehensive reporting and analysis.
- Enhance the accuracy of data mining and machine learning models.

FROM MULTIPLE SOURCES

Source 1
Student Information System

StudentID	Name	Programme
S001	Ali	BIT
S002	Sara	BCS
S003	Ahmad	BIT
...	...	...

Source 2
Attendance System

StudentID	Attendance %
S001	92
S002	85
S003	88
...	...

Source 3
Examination System

StudentID	Total Marks	Grade
S001	88	A
S002	79	B
S003	91	A
...	...	...

INTEGRATION PROCESS

- Collect data from different sources
- Match and merge common keys (e.g., StudentID)
- Resolve inconsistencies and remove duplicates
- Create a unified, consistent dataset

INTEGRATED UNIFIED DATASET

StudentID	Name	Programme	Attendance %	Total Marks	Grade
S001	Ali	BIT	92	88	A
S002	Sara	BCS	85	79	B
S003	Ahmad	BIT	88	91	A
...	...	...	...	...	...

A single, reliable, and consistent dataset ready for reporting, analysis, and data mining.

ACTIVITY 5 – DATA INTEGRATION

COMMON METHODS FOR DATA INTEGRATION

Method	Description	Example
Append (Union)	Combines datasets with the same structure by adding records vertically.	Merge student records from Semester 1 and Semester 2.
Merge (Join)	Combines datasets using a common key or identifier.	Merge Attendance.xlsx and Examination.xlsx using StudentID.
Lookup	Retrieves information from another table based on a matching key.	Retrieve Programme using StudentID with XLOOKUP() or VLOOKUP().
Data Consolidation	Summarises data from multiple worksheets into one report.	Combine marks from several classes into a single summary.
Database Import / Power Query	Imports and combines data automatically from multiple files or databases.	Imports several Excel files into one table using Power Query.

CHALLENGES

- Different formats and data types
- Inconsistent values or spelling
- Missing or duplicate records
- Different column names

EXCEL DEMONSTRATION – MERGE / LOOKUP (COMBINE COLUMNS)

1 SOURCE TABLES (SEPARATE)

Data stored in different tables.

Table 1: Student_Info

StudentID	Name	Gender	Programme
1001	Aiman	Male	BCS
1002	Siti	Female	BIT
1003	Daniel	Male	BBA
1004	Nurul	Female	BCS
1005	Hafiz	Male	BIT
...	...	...	...

Table 2: Performance

StudentID	Attendance (%)	Assignment (30)	Final Exam (50)	Total (100)
1001	95	28	45	91
1002	80	20	38	73
1003	90	27	44	88
1004	75	18	32	62
1005	98	30	48	95
...	...	...	...	...

We need to combine these tables using the common field: StudentID.

2 MERGE USING VLOOKUP / XLOOKUP

Bring performance data into Student_Info table.

Example Formula (in E2):

```
=XLOOKUP(A2, Performance!A:A, Performance!B:E, "Not Found")
```

StudentID	Name	Gender	Programme	Attendance (%)	Assignment (30)	Final Exam (50)	Total (100)
1001	Aiman	Male	BCS	95	28	45	91
1002	Siti	Female	BIT	80	20	38	73
1003	Daniel	Male	BBA	90	27	44	88
1004	Nurul	Female	BCS	75	18	32	62
1005	Hafiz	Male	BIT	98	30	48	95
...	...	...	...	...	...	...	...

XLOOKUP matches StudentID and returns the related values.

3 INTEGRATED DATASET (COMBINED)

Now all data is in one table for analysis.

StudentID	Name	Gender	Programme	Attendance (%)	Assignment (30)	Final Exam (50)	Total (100)	Status
1001	Aiman	Male	BCS	95	28	45	91	Pass
1002	Siti	Female	BIT	80	20	38	73	Pass
1003	Daniel	Male	BBA	90	27	44	88	Pass
1004	Nurul	Female	BCS	75	18	32	62	Pass
1005	Hafiz	Male	BIT	98	30	48	95	Pass
...	...	...	...	...	...	...	...	...

Integrated data enables better analysis, reporting, and decision making.

BEST PRACTICES

- Ensure key fields (e.g., StudentID) are unique.
- Standardize formats (dates, text, numbers) before integrating.
- Remove duplicates and handle missing values.
- Document the data sources and integration steps.
- Validate the integrated data for accuracy.

DISCUSSION

Points to Consider:

- What happens if a StudentID exists in one table but not another?
- How can inconsistent values affect integration?
- When would you use Append vs. Merge?
- How can Power Query make integration easier?

KEY TAKEAWAY

Data integration combines information from multiple sources into a single, consistent dataset, enabling complete and meaningful analysis.

Remember: Good data integration leads to reliable insights and better decisions.

DATA REDUCTION

Reducing Data Volume, Retaining Important Information

WHAT IS DATA REDUCTION?

Data Reduction is the process of obtaining a reduced representation of the data set that is much smaller in volume, yet produces the same (or almost the same) analytical results. It helps to reduce storage requirements, speed up data processing, and improve the efficiency of data mining algorithms.

GOALS

- Reduce data volume and storage cost
- Improve data processing speed
- Maintain important patterns and information
- Enhance the performance of data mining

COMMON DATA REDUCTION TECHNIQUES

1. DATA CUBE AGGREGATION

Summarizes data in a data cube at a higher level of abstraction by aggregating values along dimensions.

EXAMPLE

Sales data aggregated from detailed daily level to monthly, quarterly or yearly level.

Detailed Data (Daily)

Date	Product	Sales (RM)
01/05	A	120
02/05	A	150
03/05	A	130
...	...	...

Aggregated Data (Monthly)

Month	Product	Sales (RM)
May	A	4,200
May	B	3,150
May	C	2,890
...	...	...

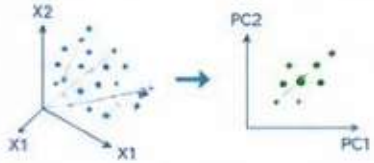
Result: Fewer records, faster analysis, same trend information.

2. DIMENSIONALITY REDUCTION

Reduces the number of dimensions (attributes/features) while preserving most of the important information.

EXAMPLE

Use Principal Component Analysis (PCA) to reduce 10 variables to 2 principal components for visualization.



Result: Fewer dimensions, easier visualization, less computation.

3. DATA COMPRESSION

Encodes data using fewer bits without significant loss of information (lossless or lossy).

EXAMPLE

Compress text or numeric data using ZIP, Run-Length Encoding, or other compression methods.

Original Data	Compressed Data
File Size (Original)	100 MB
File Size (Compressed)	25 MB
Compression Ratio	75%

Result: Less storage needed, faster data transfer.

4. NUMERICITY REDUCTION

Represents data using fewer numerical precision or smaller data types while preserving usefulness.

EXAMPLE

Store values with 2 decimal places instead of 6, or use Integer instead of Double where possible.

Original (6 decimals)	Reduced (2 decimals)
123.456789	123.46
98.765432	98.77
56.123456	56.12
...	...

Result: Smaller data size, lower storage and memory.

WHY IS DATA REDUCTION IMPORTANT?

- Handles large-scale data efficiently
- Reduces storage and memory requirements
- Speeds up data mining and analytics
- Maintains essential information
- Improves scalability and performance



Large Data (Slow, Expensive) → Reduced Data (Fast, Efficient)

COMPARISON OF TECHNIQUES

Technique	Main Idea	Applied On	Benefit	Example
Data Cube Aggregation	Summarize data at higher levels	Multidimensional data	Fewer records, faster analysis	Daily → Monthly sales
Dimensionality Reduction	Reduce number of attributes	High-dimensional data	Easier analysis, less computation	PCA, Feature Selection
Data Compression	Encode data using fewer bits	Text, numeric, multimedia	Less storage, faster transfer	ZIP, RLE
Numericity Reduction	Reduce precision or data type size	Numeric data	Lower storage & memory usage	Double → Float, 6 dp → 2 dp

DATA REDUCTION

Reducing Data Volume, Retaining Important Information

Data1_Student_Performance.xlsx

Attributes (10):
StudentID, Gender, Programme, Attendance, Assignment, Quiz, Final Exam, Total, Grade, Status

Records:
100 Students

1 DATA CUBE AGGREGATION

Summarize detailed records into higher-level information by aggregating values along dimensions.

Original Data (Detailed Records)

StudentID	Programme	Attendance	Total	Grade
S001	BIT	92	88	A
S002	BIT	85	79	B
S003	BCS	90	91	A
S004	BIT	70	68	C
...	...	...	...	...

Aggregated Data by Programme (Example)

Programme	No. Students	Avg. Attendance	Avg. Total
BIT	45	82.6	76.8
BCS	35	84.1	79.5
BIS	20	80.9	74.2

Reduce 100 detailed records to 3 programme summaries.

2 DIMENSIONALITY REDUCTION

Reduce the number of attributes (features) while keeping the most relevant ones for analysis or prediction.

Original Attributes (10)

- StudentID
- Gender
- Programme
- Attendance
- Assignment
- Quiz
- Final Exam
- Total
- Grade
- Status

Reduced Attributes (5)

- Attendance
- Assignment
- Quiz
- Final Exam
- Status

Removed 4 less relevant attributes. Focus on the 5 most useful attributes.

Helps improve model performance and reduce complexity.

3 DATA COMPRESSION

Store the data using fewer bits or compressed format without changing the information.



Original File
Student_Performance.xlsx
120 KB

Compressed File
Student_Performance.zip
32 KB

Reduced file size by ~73%. Less storage space and faster transfer.

Data remains the same, only the storage size is reduced.

4 NUMEROSITY REDUCTION

Represent the data using fewer records or summary statistics while preserving important characteristics.

Summary Statistics (Attendance)

Statistic	Value
Mean	82.5
Median	83
Minimum	48
Maximum	99
Std. Dev.	10.8

Grade Distribution (Example)

Grade	No. Students
A	18
B	30
C	28
D	16
F	8
Total	100

Replaced 100 detailed records with compact statistical summaries or distributions.

Maintains key information with much smaller data size.

ACTIVITY 6 – DATA/ATTRIBUTE SELECTION

NOT EVERY ATTRIBUTE CONTRIBUTES

Not every attribute contributes to solving the problem.

Example: Predicting student performance.

Useful attributes	Less useful attributes
Attendance	Student ID
Assignment	Phone Number
Quiz	Email Address
Final Exam	

BENEFITS

- Faster analysis
- Simpler models
- Improved prediction accuracy
- Reduced redundancy

WHY EXCLUDE THESE VARIABLES?

Student ID
Only an identifier.
Does not affect performance.

Phone Number
Not related to academic performance.
Contains personal information.

Email Address
Not related to academic performance.
Contains personal information.

EXCEL DEMONSTRATION – ATTRIBUTE SELECTION

1 ORIGINAL DATASET

Many attributes (includes unnecessary ones).

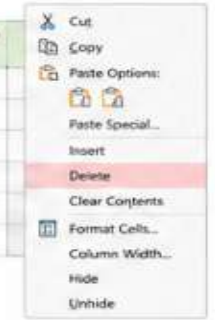
	StudentID	Name	Gender	Age	Attendance (%)	Assignment (30)	Quiz (20)	Final Exam (50)	Total (100)	Phone Number	Email Address
2	1001	Aiman	Male	20	95	28	18	45	91	012-3456789	aiman@gmail.com
3	1002	Siti	Female	21	80	20	15	38	73	013-2345678	siti@gmail.com
4	1003	Daniel	Male	20	90	27	17	44	88	017-1234567	daniel@gmail.com
5	1004	Nurul	Female	22	75	18	12	32	62	016-9876543	nurul@gmail.com
6	1005	Hafiz	Male	21	98	30	19	48	95	011-2233445	hafiz@gmail.com
	...	...	...	...	...	...	...	...	...	...	...

Unnecessary attributes (highlighted in red) do not help in predicting performance. They increase data size and may cause privacy issues.

2 REMOVE OR HIDE UNNECESSARY COLUMNS

Delete or hide columns that are not useful.

	StudentID	Name	Gender	Age	Attendance (%)
2	1001	Aiman	Male	20	95
3	1002	Siti	Female	21	80
4	1003	Daniel	Male	20	90
5	1004	Nurul	Female	22	75
6	1005	Hafiz	Male	21	98
	...	...	...	...	...



Use 'Delete' to permanently remove or 'Hide' to keep for reference.

3 SELECTED ATTRIBUTES

Dataset now contains only useful attributes.

	StudentID	Name	Gender	Age	Attendance (%)	Assignment (30)	Quiz (20)	Final Exam (50)	Total (100)
2	1001	Aiman	Male	20	95	28	18	45	91
3	1002	Siti	Female	21	80	20	15	38	73
4	1003	Daniel	Male	20	90	27	17	44	88
5	1004	Nurul	Female	22	75	18	12	32	62
6	1005	Hafiz	Male	21	98	30	19	48	95
	...	...	...	...	...	...	...	...	...

The dataset is now smaller, cleaner, and more relevant for building accurate and efficient models.

DISCUSSION

Why do we exclude certain variables?

- They do not help in predicting the target.
- They increase complexity and computational cost.
- They may introduce noise or bias.

Always focus on attributes that provide useful information for solving the problem.

DATA TRANSFORMATION

Converting Raw Data into Meaningful and Useful Information

WHAT IS DATA TRANSFORMATION?

Data Transformation is the process of converting raw data into a more appropriate format for analysis or data mining.
It improves data quality, consistency, and usability, allowing algorithms to discover more meaningful patterns and produce more accurate results.

GOAL:

Transform data into a format that is easier to analyse and more suitable for machine learning and data mining.

COMMON DATA TRANSFORMATION TECHNIQUES

1. SMOOTHING

Removes noise or random variations in the data to improve data quality.

EXAMPLE

Replace daily sales with a 7-day moving average.



2. AGGREGATION

Combines multiple records into a summary or higher-level representation.

EXAMPLE

Calculate monthly sales from daily sales records.

Month	Total Sales (RM)
Jan	25,430
Feb	28,950
Mar	30,210
Apr	27,860
...	...

3. GENERALIZATION

Replaces detailed values with higher-level concepts using a concept hierarchy.

EXAMPLE

Age 23 → Young Adult
Kuala Lumpur → Malaysia

Original Data	Generalized Data
Age 23	Young Adult
Age 45	Adult
Kuala Lumpur	Malaysia
Johor Bahru	Johor
...	...

4. NORMALIZATION

Scales numerical values into a common range to eliminate differences in measurement scales.

EXAMPLE

Marks (0–100) → Values between 0 and 1.

Original Marks	Normalized (0–1)
35	0.35
60	0.60
85	0.85
95	0.95
...	...

5. ATTRIBUTE / FEATURE CONSTRUCTION

Creates new attributes from existing data to improve analysis or prediction.

EXAMPLE

Total Marks = Assignment + Quiz + Final Exam
BMI = Weight / Height²

Assignment	Quiz	Final Exam	Total Marks
20	15	60	95
18	12	55	85
15	10	50	75
...	...	...	...

THE ROLE OF DATA TRANSFORMATION

RAW DATA → **DATA TRANSFORMATION** → **PREPARED DATA** → **ANALYSIS / DATA MINING** → **KNOWLEDGE & DECISIONS**

- **RAW DATA**: Incomplete, noisy, inconsistent
- **DATA TRANSFORMATION**: Clean, integrate, transform
- **PREPARED DATA**: High quality, meaningful
- **ANALYSIS / DATA MINING**: Discover patterns, build models
- **KNOWLEDGE & DECISIONS**: Better insights, better outcomes

WHY IS DATA TRANSFORMATION IMPORTANT?

- Improves data quality and consistency.
- Reduces noise and redundancy.
- Creates meaningful features for analysis.
- Ensures numerical attributes are comparable.
- Enhances the performance and accuracy of data mining and machine learning models.

DATA NORMALIZATION

WHY NORMALIZATION?

- Different attributes often have different scales.
- **Example:**

Attribute	Range
Attendance	0 – 100
Assignment	0 – 30
Final Exam	0 – 50

- Algorithms may be influenced by variables with larger ranges.
- Normalization brings values onto a comparable scale.

METHOD 1 MIN–MAX NORMALIZATION

Formula

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Where:
- x = original value
- x_{min} = minimum value of the attribute
- x_{max} = maximum value of the attribute

Result:

All values range between 0 and 1.
$(0 \leq x_{norm} \leq 1)$

METHOD 2 Z–SCORE STANDARDIZATION

Formula

$$z = \frac{x - \mu}{\sigma}$$

Where:
- x = original value
- μ = mean (average) of the attribute
- σ = standard deviation of the attribute

Result:
- Mean = 0
- Standard deviation = 1

METHOD 3 DECIMAL SCALING

Formula

$$x_{norm} = \frac{x}{10^j}$$

Where:
- x = original value
- j = smallest integer such that $\max(|x_{norm}|) < 1$

Result:

All values lie between −1 and 1
$(-1 < x_{norm} < 1)$

COMPARISON OF NORMALIZATION METHODS

Aspect	Min–Max Normalization	Z–Score Standardization	Decimal Scaling
Equation	$\frac{x - x_{min}}{x_{max} - x_{min}}$	$\frac{x - \mu}{\sigma}$	$\frac{x}{10^j}$
Output Range	0 to 1	Typically −3 to +3	−1 to 1
Mean	Not fixed (depends on data)	0	Not fixed (close to 0)
Standard Deviation	Not fixed	1	Not fixed
Sensitive to Outliers	Yes	Yes	No (less sensitive)
When to Use	When data has a fixed range (e.g., 0–1 required)	When data is approximately normally distributed	When you want to preserve sign and relative magnitude
Common Applications	Neural networks (sigmoid/tanh), distance-based methods	Statistical analysis, linear models, machine learning algorithms	Large datasets, data with varying magnitudes

KEY TAKEAWAYS

- Normalization ensures all attributes contribute fairly to the analysis.
- **Min–Max** is useful for bounded ranges (0 to 1).
- **Z–Score** centers the data around 0 with unit variance.
- **Decimal Scaling** is simple and effective for large or wide-ranging data.
- Choose the method based on the data distribution and the algorithm you plan to use.

Remember:
Normalization is a key preprocessing step that **improves the performance and stability** of many data mining and machine learning algorithms.

Always explore your data first, understand its distribution, and then choose the right normalization technique.

ACTIVITY 7 – NORMALIZATION

OBJECTIVE: Apply three normalization methods to the same dataset and compare the results.

Student_Performance_Dataset_100.xlsx

1 SELECT ONE FIELD TO NORMALIZE

We will normalize the field:

Final Exam (Column G)

	A	B	C	G
1	StudentID	Gender	Programme	Final Exam
2	1001	Female	BBA	45
3	1002	Male	BIT	38
4	1003	Female	BCS	44
5	1004	Male	BBA	32
...	...	...	...	...
98	1097	Female	BIT	40
99	1098	Male	BCS	46
100	1099	Female	BIT	28
101	1100	Male	BBA	42

We will use rows 2 to 101 (100 records).

2 NORMALIZATION METHODS, EQUATIONS AND EXCEL FORMULAS

Method	Equation	Excel Formula
1 **Min-Max Normalization (0 – 1)**	$x' = \frac{x - \text{Min}(X)}{\text{Max}(X) - \text{Min}(X)}$ **Result:** All values range between 0 and 1.	=(G2-MIN($G$2:$G$101))/ (MAX($G$2:$G$101)-MIN($G$2:$G$101)) Copy down to G101.
2 **Z-Score Standardization**	$z = \frac{x - \text{Mean}(X)}{\text{StDev}(X)}$ **Result:** Mean = 0, Standard Deviation = 1.	=(G2-AVERAGE($G$2:$G$101))/ STDEV.S($G$2:$G$101) Copy down to H101.
3 **Decimal Scaling**	$x' = x / 10^j$ where j = smallest integer such that max(|x'|) < 1. **Result:** Values scaled between -1 and 1.	=G2/10^2 (Use 10^2 because Max = 100, j = 2) Copy down to I101.

3 INSTRUCTIONS

1. Go to the "Normalization" worksheet.
2. The original Final Exam values (Column A) are provided.
3. Calculate the normalized values using the three methods:
   - **Min-Max (0–1)**
   - **Z-Score Standardization**
   - **Decimal Scaling**
4. Compare the original values with the normalized values.

Observe how the scale changes while the relative order of the data remains the same.

4 RESULT – COMPARISON OF ORIGINAL AND NORMALIZED VALUES (FINAL EXAM)

Original Values (Final Exam)			Min-Max Normalization (0 – 1)		Z-Score Standardization		Decimal Scaling	
Row	StudentID	Final Exam (G)	Formula Used	Normalized Value	Formula Used	Normalized Value	Formula Used	Normalized Value
2	1001	45	=(G2-MIN($G$2:$G$101))/ (MAX($G$2:$G$101)- MIN($G$2:$G$101))	0.2830	=(G2-AVERAGE($G$2:$G$101))/ STDEV.S($G$2:$G$101)	0.0082	=G2/10^2 (j = 2)	0.45
3	1002	38		0.1415		-0.7246		0.38
4	1003	44		0.2642		-0.0805		0.44
5	1004	32		0.0566		-1.3660		0.32
...	...	...		...		...		...
98	1097	40		0.1698		-0.4026		0.40
99	1098	46		0.3019		0.1068		0.46
100	1099	28		0.0000		-1.7688		0.28
101	1100	42		0.2264		-0.2412		0.42

Key Observation: Although the scale of the data changes with each method, the relative order of the values remains the same.

Discussion:
Which method would be most suitable for machine learning algorithms and why?

SUMMARY

The **quality** of your analysis depends on the **quality** of your data.

You should now understand that **data preprocessing** involves more than simply cleaning data. It includes exploring the dataset, assessing data quality, handling missing values and outliers, normalizing numerical attributes, selecting relevant variables, and transforming data into a form suitable for analysis.

KEY ACTIVITIES IN DATA PREPROCESSING

1. **Explore the Dataset**
   Understand the data and its characteristics.
2. **Assess Data Quality**
   Check for accuracy, completeness, consistency, and validity.
3. **Handle Missing Values**
   Identify and treat missing data appropriately.
4. **Detect and Handle Outliers**
   Identify unusual values and decide how to treat them.
5. **Normalize Numerical Attributes**
   Scale data to ensure fair comparison and better performance.
6. **Select Relevant Variables**
   Remove irrelevant or redundant attributes.
7. **Transform Data**
   Convert data into the appropriate format for analysis or modeling.

Good preprocessing leads to **better models**, **accurate results**, and **meaningful insights**.

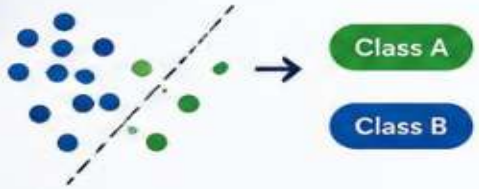
MODULE 3

CLASSIFICATION AND CLUSTERING

Turning data into predictions and patterns for smarter decisions

CLASSIFICATION

Predict known classes

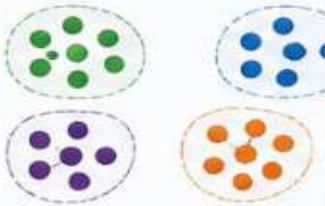


Decision Tree (J48)

Naïve Bayes

CLUSTERING

Discover unknown groups



K-Means Clustering

LEARNING OUTCOMES

At the end of this session, participants should be able to:

01 **Differentiate** between Supervised Learning and Unsupervised Learning.

02 **Explain** the basic concepts of Decision Tree (J48), Naïve Bayes, and K-Means Clustering.

03 **Build** a simple classifier using logical rules in Excel.

04 **Evaluate** classification performance using a Confusion Matrix.

05 **Perform** simple clustering using Excel.

06 **Interpret** the results for decision making.

KEY TAKEAWAYS

- Make Predictions with Classification
- Discover Patterns with Clustering
- Better Insights Better Decisions
- Stronger Data Stronger Future

INTRODUCTION

Although both techniques analyze data, they solve **different problems.**

CLASSIFICATION

"Suppose you have data about 100 students. Can you **predict** which students are **likely to pass** the course?"

This is called **Classification.**

Student Data (with known labels)

Attendance	GPA	Study Hours	Result
90	3.6	15	Pass
75	2.8	8	Fail
85	3.2	12	Pass
60	2.1	5	Fail
...	...	...	...

(100 students)

Predict
New Student

Attendance : 82
GPA : 3.0
Study Hours : 10

Prediction:
Pass

We use past data with **known results** to **predict** future or unknown results.

CLUSTERING

"Without knowing whether students pass or fail, can you automatically **group students** with **similar characteristics**?"

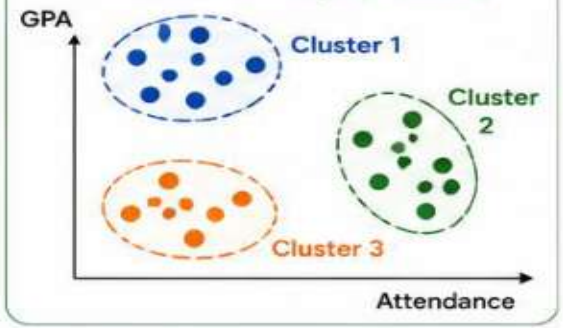
This is called **Clustering.**

Student Data (no labels)

Attendance	GPA	Study Hours
90	3.6	15
75	2.8	8
85	3.2	12
60	2.1	5
...	...	...

(100 students)

Automatic Groups (Clusters)

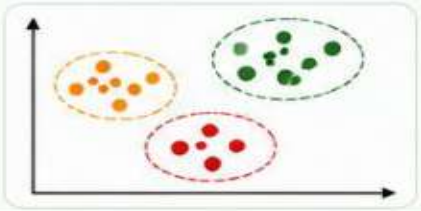


We **group similar data** together without using known results.

In summary: Classification **predicts.** | Clustering **discovers patterns.**

CLASSIFICATION vs CLUSTERING

Both techniques analyze data, but they answer **different questions** and are used in **different ways**.

CLASSIFICATION	CLUSTERING
Predicts a **known category**	Groups **similar records**
Uses **labelled data**	Uses **unlabelled data**
Supervised Learning	**Unsupervised Learning**
Target variable **exists**	**No** target variable
Example: **Pass/Fail** 	Example: **High**, **Medium**, **Low** performers 

In summary: Classification **predicts** what will happen. | Clustering **discovers** natural groupings.

SUPERVISED vs UNSUPERVISED LEARNING

Key Differences at a Glance

	SUPERVISED LEARNING *Learn from labeled data*	UNSUPERVISED LEARNING *Find patterns in unlabeled data*
DEFINITION	The algorithm learns from input data that is **labeled** with the correct output.	The algorithm explores input data **without labeled output** and finds hidden patterns or structures.
DATA	Uses **labeled** data (input + output/target).	Uses **unlabeled** data (input only).
GOAL	**Predict or classify** output for new, unseen data.	**Discover patterns**, groupings, or relationships in the data.
EXAMPLE TASKS	• Classification (e.g., spam vs. not spam) • Regression (e.g., predicting house price)	• Clustering (e.g., customer segmentation) • Association (e.g., market basket analysis) • Dimensionality Reduction (e.g., PCA)
EXAMPLE ALGORITHM(S)	• Linear Regression • Logistic Regression • Decision Tree, Random Forest • Support Vector Machine (SVM) • k-Nearest Neighbors (k-NN)	• K-Means Clustering • Hierarchical Clustering • Apriori Algorithm • Principal Component Analysis (PCA) • DBSCAN
VISUAL REPRESENTATION	 Learns the mapping from input to known output	 Discovers natural structure or patterns in the data
KEY CHARACTERISTIC	Requires a **"teacher"** (labeled examples) to learn the mapping.	No **"teacher"** needed; the algorithm finds the structure on its own.

SUPERVISED LEARNING EXAMPLE

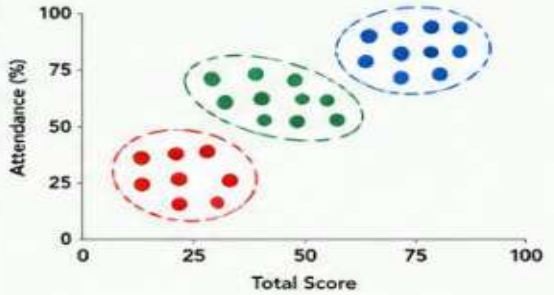
Predict Student Pass/Fail

Attendance (%)	Assignment Score	Final Exam Score	Pass/Fail (Label)
90	85	88	Pass
70	60	65	Pass
40	50	45	Fail
60	55	50	Fail
...	...	...	...

The model learns from past data with known outcomes to predict new students' Pass/Fail.

UNSUPERVISED LEARNING EXAMPLE

Group Students by Similar Behavior



The algorithm groups students into clusters without using any labels.

KEY TAKEAWAYS

- Supervised Learning uses **labeled** data to predict or classify outcomes.
- Unsupervised Learning uses **unlabeled** data to discover patterns or groupings.
- Choose **supervised** when you have target labels and want to make predictions.
- Choose **unsupervised** when you don't have labels and want to explore the data.

COMMON USE CASES

- Credit Scoring (Supervised)
- Spam Detection (Supervised)
- Sales Forecasting (Supervised)
- Customer Segmentation (Unsupervised)
- Market Basket Analysis (Unsupervised)
- Feature Extraction (Unsupervised)

ACTIVITY 1 – SUPERVISED vs UNSUPERVISED LEARNING

SUPERVISED LEARNING

The dataset **already contains the correct answer.**

Example:

Attendance	Total Marks	Status
95	88	Pass
82	76	Pass
60	45	Fail

The objective is to learn a model that **predicts Status.**

Explain:

Features (Predictors)

- Attendance
- Total Marks

These are the input variables used to make predictions.

Target (Class Label)

- Status

This is the output the model tries to predict.

Examples:

- Pass / Fail (Student Performance)
- Fraud / Genuine (Credit Card Fraud)
- Sick / Healthy (Healthcare)
- Spam / Not Spam (Email Filtering)

UNSUPERVISED LEARNING

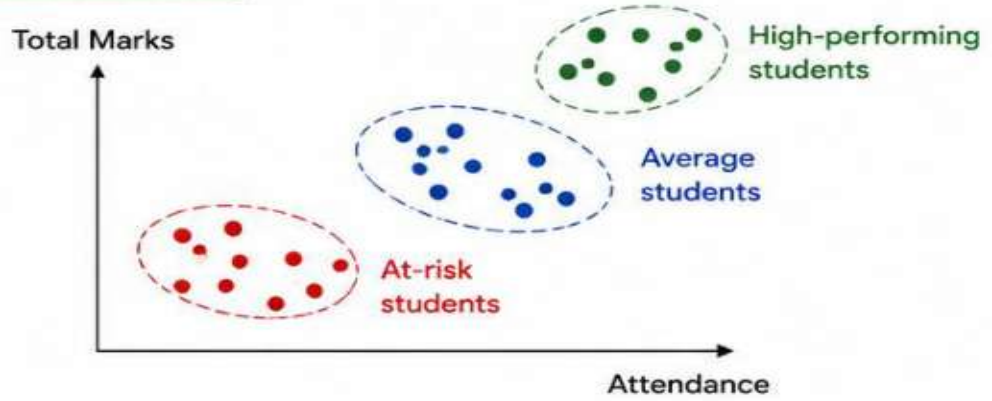
There is **no class label.**

Example:

Attendance	Total Marks
90	85
88	82
58	50
55	48
...	...

The objective is to **discover natural groups.**

Possible clusters:



DISCUSSION

If we remove the Status column, can the computer still group similar students?

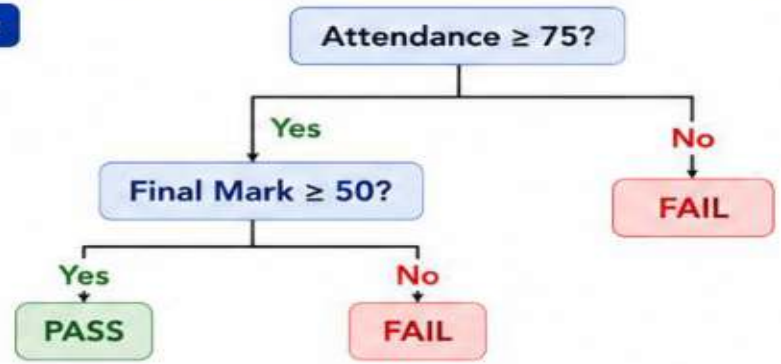
Yes.
That is **Clustering.**

ACTIVITY 2 – DECISION TREE (J48)

WHAT IS A DECISION TREE?

A Decision Tree predicts an outcome using a series of questions.

Example:



J48 is Weka's implementation of the C4.5 Decision Tree algorithm.

It automatically determines the best attribute to split the data.

ADVANTAGES

- Easy to understand
- Easy to explain
- Produces interpretable rules
- Suitable for classification

EXCEL DEMONSTRATION

Use the Student Performance dataset.

A No.	B Student ID	C Name	D Attendance (%)	E Assignment (30)	F Quiz (20)	G Final Exam (50)	H Total Marks (100)	I Actual Status (Yes/No)	J Predicted Status (Using Rule)
1	1001	Aiman	95	28	18	45	91	Pass	Pass
2	1002	Siti	80	20	15	38	73	Pass	Pass
3	1003	Daniel	60	18	12	32	62	Pass	Fail
4	1004	Nurul	55	16	10	25	51	Fail	Fail
5	1005	Hafiz	40	14	8	20	42	Fail	Fail
...	...	...	...	...	...	...	...	...	...

Create a new column named **Predicted Status.**

Use a simple rule:

```
=IF(AND(D2>=75,H2>=50),"Pass","Fail")
```

Where:

- D = Attendance
- H = Total Marks

Copy the formula down for all records.

DISCUSS

1. **Why were these thresholds chosen?**
   Discuss how the values (75 and 50) affect the classification results.

2. **Could different rules produce different results?**
   Explain how changing the rules or using other attributes might change the predictions.

HOW A DECISION TREE IS CONSTRUCTED (TRAINING PROCESS)

1 START WITH THE DATASET

We have data of students with features and class label (Status).

ID	Attendance (%)	Total Marks	Status
1	95	88	Pass
2	90	85	Pass
3	82	73	Pass
4	60	45	Fail
5	55	40	Fail
6	48	35	Fail
7	76	52	Pass
8	70	49	Fail

Pass (4) Fail (4)

2 EVALUATE EACH ATTRIBUTE (FIND THE BEST SPLIT)

Use Information Gain (IG) to find which attribute separates the data best.

Attribute	Information Gain (IG)
Attendance (%)	0.811
Total Marks	0.420
Quiz Score	0.251

Attendance has the highest IG (0.811). Choose Attendance as the root node.

3 SPLIT THE DATA

Split the dataset using the best attribute (Attendance ≥ 75%).

Attendance ≥ 75%?

Yes (≥ 75%) 4 records

ID	Att.(%)	Total	Status
1	95	88	Pass
2	90	85	Pass
3	82	73	Pass
7	76	52	Pass

All are Pass → Make this a leaf node.

No (< 75%) 4 records

ID	Att.(%)	Total	Status
4	60	45	Fail
5	55	40	Fail
6	48	35	Fail
8	70	49	Fail

Not all the same. Need another split.

4 REPEAT FOR EACH BRANCH

For the right branch (Attendance < 75%), find the next best split.

Best split: Total Marks ≥ 50

Total Marks ≥ 50?

Yes (≥ 50) 1 record

ID	Att.	Total	Status
8	70	49	Fail

All are Fail → Leaf node.

No (< 50) 3 records

ID	Att.	Total	Status
4	60	45	Fail
5	55	40	Fail
6	48	35	Fail

All are Fail → Leaf node.

5 FINAL DECISION TREE

The tree is now ready.

Attendance ≥ 75%?

Yes (≥ 75%) → PASS

No (< 75%) → Total Marks ≥ 50?

Yes (≥ 50) → PAIL

No (< 50) → FAIL

This tree will be used to predict new students.

HOW THE DECISION TREE WORKS (PREDICTION PROCESS)

NEW STUDENT RECORDS

We want to predict the Status of new students.

Student	Attendance (%)	Total Marks
A	85	78
B	65	55
C	60	40
D	45	30

DECISION TREE (From Training)

Attendance ≥ 75%?

Yes (≥ 75%) → PASS

No (< 75%) → Total Marks ≥ 50?

Yes (≥ 50) → FAIL

No (< 50) → FAIL

PREDICTION – TRAVERSE THE TREE

A 85, 78
Attendance ≥ 75%? 85 ≥ 75 → Yes → PASS
Predicted Status: PASS

B 65, 55
Attendance ≥ 75%? 65 ≥ 75 → No → Total Marks ≥ 50? 55 ≥ 50 → Yes → FAIL
Predicted Status: FAIL

C 60, 40
Attendance ≥ 75%? 60 ≥ 75 → No → Total Marks ≥ 50? 40 ≥ 50 → No → FAIL
Predicted Status: FAIL

D 45, 30
Attendance ≥ 75%? 45 ≥ 75 → No → Total Marks ≥ 50? 30 ≥ 50 → No → FAIL
Predicted Status: FAIL

SUMMARY

- The tree was built automatically from historical data.
- It asks simple questions.
- It follows the path based on the answers.
- It predicts the class at the leaf node.

KEY POINTS

- The tree starts at the root and splits the data.
- It keeps splitting to create pure or homogeneous groups.
- During prediction, follow the path based on the answers.
- The final leaf node gives the predicted class.

ACTIVITY 3 – NAÏVE BAYES

1. WHAT IS NAÏVE BAYES?

- Naïve Bayes is a **probabilistic classifier**.
- It predicts the **most likely class** based on probabilities.

Example:

Condition	Probability of Passing
Attendance > 80	95%
Attendance < 60	30%

Bayes' Theorem

$$P(C|X) = \frac{P(X|C)\,P(C)}{P(X)}$$

Choose the class (C) with the **highest probability** given the data (X).

The algorithm combines these probabilities to estimate the **most likely outcome**.

2. WHY IS IT CALLED "NAÏVE"?

- It assumes that all predictor variables are **independent**, even though this is rarely true in practice.

Example Assumption



Naïve Bayes assumes these variables **do not influence** one another.

Although this assumption is "naïve", the algorithm often **performs surprisingly well.**

3. ADVANTAGES

Fast
Training and prediction are very quick.

Simple
Easy to understand and implement.

Works well with large datasets
Handles high-dimensional data efficiently.

Effective for text classification and spam detection
Widely used in real-world applications.

4. EXCEL DEMONSTRATION (USING FREQUENCY TABLE)

Instead of calculating full probabilities, we explain the concept using a **frequency table.**

Example:

Attendance	Students (Total)	Pass (Count)	Fail (Count)
Above 80	40	38	2
Below 80	60	35	25
Total	**100**	**73**	**27**

Probability Calculation:

$$P(\text{"Pass"} \mid \text{"Attendance"} > 80) = \frac{38}{40} = 0.95 = 95\%$$

There is a 95% chance of passing when attendance is above 80%.

$$P(\text{"Pass"} \mid \text{"Attendance"} < 80) = \frac{35}{60} \approx 0.583 = 58.3\%$$

There is a 58.3% chance of passing when attendance is below 80%.

CLASSIFICATION EVALUATION METRICS

What They Mean, How They Differ & When to Use Them

These metrics help us understand how well a classification model performs.
Different metrics answer different questions and are useful in different situations.

- Positive Class = 1 (e.g., Pass)
- Negative Class = 0 (e.g., Fail)

	1. CONFUSION MATRIX (TP, FP, TN, FN)	2. ACCURACY	3. PRECISION	4. RECALL (SENSITIVITY)	5. F1-SCORE
WHAT IS IT?	A table that shows the comparison between actual and predicted classes.	Overall percentage of correct predictions (both classes).	Of all instances predicted as Positive, how many are actually Positive?	Of all actual Positive instances, how many did the model correctly find?	The harmonic mean of Precision and Recall. Balances both metrics.
FORMULA	N/A (It is a summary table)	$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$	$\text{Precision} = \frac{TP}{TP + FP}$	$\text{Recall} = \frac{TP}{TP + FN}$	$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$
WHAT DOES IT TELL US?	Shows the four types of outcomes: • TP: Correctly predicted Positive • FP: Incorrectly predicted Positive • FN: Actual Positive but missed • TN: Correctly predicted Negative	Out of all predictions, how many were correct? Overall correctness	Measures the quality of Positive predictions. Focus on avoiding False Positives (Type I error)	Measures the ability to find all Positive cases. Focus on avoiding False Negatives (Type II error)	Gives a single score that balances Precision and Recall.
RANGE	Counts (0 to total number of samples)	0 to 1 (or 0% to 100%) Higher is better	0 to 1 (or 0% to 100%) Higher is better	0 to 1 (or 0% to 100%) Higher is better	0 to 1 (or 0% to 100%) Higher is better
WHEN TO USE?	**Always use it as the FIRST step.** It is the foundation for all other metrics.	Use when classes are balanced and FP and FN have similar cost/impact.	Use when the cost of False Positives is high. (Example: Flagging a good student as at-risk)	Use when the cost of False Negatives is high. (Example: Missing a truly at-risk student)	Use when you need a single metric that balances both FP and FN. (Example: Imbalanced classes)
KEY TAKEAWAY	It gives the complete picture of model performance.	Easy to understand, but can be misleading when classes are imbalanced.	High Precision means most predicted Positive are truly Positive.	High Recall means few actual Positives are missed.	Best choice when you want a balance between catching Positives and avoiding False alarms.

1. CONFUSION MATRIX (TP, FP, TN, FN)

Predicted Class \ Actual Class	Positive (1)	Negative (0)
Positive (1)	TP True Positive	FP False Positive
Negative (0)	FN False Negative	TN True Negative

CONFUSION MATRIX AND ACCURACY

Understanding with an Example (100 Students)

1. CONFUSION MATRIX

Compare:
Actual Status
vs
Predicted Status
Create a Confusion Matrix.

Predicted Status	Actual Pass	Actual Fail
Predicted Pass	TP	FP
Predicted Fail	FN	TN

where

- **TP** — **True Positive**: Model predicts Pass and student actually Passes.
- **FP** — **False Positive**: Model predicts Pass but student actually Fails.
- **FN** — **False Negative**: Model predicts Fail but student actually Passes.
- **TN** — **True Negative**: Model predicts Fail and student actually Fails.

2. EXCEL FORMULA – CHECK CORRECTNESS

Check whether the prediction is correct:

```
=IF(I2=J2,"Correct","Incorrect")
```

Where:

- I = Actual Status
- J = Predicted Status

No.	Student ID	Attendance (%)	Total Marks (100)	Actual Status	Predicted Status	Correctness
1	1001	85	78	Pass	Pass	Correct
2	1002	65	55	Pass	Fail	Incorrect
3	1003	60	40	Fail	Fail	Correct
4	1004	45	30	Fail	Fail	Correct
5	1005	80	70	Pass	Pass	Correct
...	...	...	...	...	...	...

3. CALCULATE ACCURACY

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Records}}$$

Example:
95 correct out of 100 students ➡ Accuracy $= \frac{95}{100} = 95\%$

No.	Student ID	Attendance (%)	Total Marks (100)	Actual Status (I)	Predicted Status (J)	Correctness
1	1001	85	78	Pass	Pass	Correct
2	1002	65	55	Pass	Fail	Incorrect
3	1003	60	40	Fail	Fail	Correct
4	1004	45	30	Fail	Fail	Correct
5	1005	80	70	Pass	Pass	Correct
...	...	...	...	...	...	...

Correct Predictions = **95** Total Records = **100** Accuracy $= \frac{95}{100} = 95\%$

4. DISCUSSION – LIMITATION OF ACCURACY

High accuracy can be misleading when classes are **imbalanced**.

- **Example: Imbalanced Dataset**

Status	Actual Count	Predicted Count	Correct
Pass (Majority)	950	940	940
Fail (Minority)	50	0	0
Total	1000	940	940

Accuracy $= \frac{940}{1000} = 0.940 = 94\%$

But the model failed to detect any **Fail** cases!

Therefore, also consider:

- ✅ **Precision** = TP / (TP + FP) (Of predicted positives, how many are correct?)
- ✅ **Recall** (Sensitivity) = TP / (TP + FN) (Of actual positives, how many are detected?)
- ✅ **F1-Score** = 2 × (Precision × Recall) / (Precision + Recall)
- ✅ **Confusion Matrix** (Shows the types of errors clearly)

PRECISION, RECALL, F1-SCORE

Understanding with an Example (100 Students)

SCENARIO

- Total students = 100
- Actual results:
  - ✓ 80 students actually PASS (Positive)
  - ✓ 20 students actually FAIL (Negative)
- The classifier predicts:
  - ✓ 90 students as PASS
  - ✓ 10 students as FAIL

Of the 90 predicted PASS, 80 are actually PASS (10 are actually FAIL).

Key Idea:
The model predicted more students as PASS, but some of them were actually FAIL.

CONFUSION MATRIX

Predicted Class	Actual Class: Actual Pass (Positive)	Actual Class: Actual Fail (Negative)	Total
Predicted Pass	80 (TP) True Positive	10 (FP) False Positive	90
Predicted Fail	0 (FN) False Negative	10 (TN) True Negative	10
Total	80	20	100

LEGEND

- TP (True Positive) : Correctly predicted PASS
- FP (False Positive) : Incorrectly predicted PASS
- FN (False Negative) : Actual PASS but predicted FAIL
- TN (True Negative) : Correctly predicted FAIL

Positive Class = PASS (1)
Negative Class = FAIL (0)

1. PRECISION

Of the students predicted to PASS, how many actually passed?

Formula

$$\text{Precision} = \frac{TP}{TP + FP}$$

Calculation

$$\text{Precision} = \frac{80}{80 + 10} = \frac{80}{90} = 0.889 = 88.9\%$$

Interpretation

Among the 90 students predicted to PASS, 80 actually passed.
So, 88.9% of positive predictions were correct.

2. RECALL (SENSITIVITY)

Of all students who actually passed, how many were correctly identified?

Formula

$$\text{Recall} = \frac{TP}{TP + FN}$$

Calculation

$$\text{Recall} = \frac{80}{80 + 0} = \frac{80}{80} = 1.000 = 100\%$$

Interpretation

The model successfully identified all students who actually passed.
No passing students were missed.

3. F1-SCORE

The harmonic mean of Precision and Recall. It balances both metrics in a single score.

Formula

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Calculation

$$\text{F1-Score} = 2 \times \frac{0.889 \times 1.000}{0.889 + 1.000} = 2 \times \frac{0.889}{1.889} = 0.941$$

Interpretation

F1-Score provides a balance between Precision and Recall.
In this case, the model performs very well.

4. ACCURACY (FOR REFERENCE)

Overall percentage of correct predictions (both classes).

Formula

$$\text{Accuracy} = \frac{TP + TN}{\text{Total}}$$

Calculation

$$\text{Accuracy} = \frac{80 + 10}{100} = \frac{90}{100} = 90\%$$

Interpretation

90% of all predictions (PASS and FAIL) were correct.

KEY TAKEAWAYS

- Precision focuses on the quality of positive predictions (How many predicted PASS were correct).
- Recall focuses on the coverage of actual positives (How many actual PASS were found).
- F1-Score balances both Precision and Recall.
- Accuracy looks at overall correctness (both PASS and FAIL).

ANOTHER EXAMPLE (MORE REALISTIC)

A case with both False Positives and False Negatives

	Actual Pass	Actual Fail	Total
Predicted Pass	72 (TP)	18 (FP)	90
Predicted Fail	8 (FN)	2 (TN)	10
Total	80	20	100

- Accuracy = (72 + 2)/100 = 74%
- Precision = 72/90 = 80%
- Recall = 72/80 = 90%
- F1-Score = 2×(0.80×0.90)/(0.80+0.90) = 84.7%

ACTIVITY 5 – K-MEANS CLUSTERING

1. WHAT IS CLUSTERING?

Clustering groups **similar** records **together.**

Unlike classification, there is **no correct answer.**

The algorithm discovers patterns **automatically.**

2. K-MEANS

K-Means divides data into K groups.

Example: K = 3

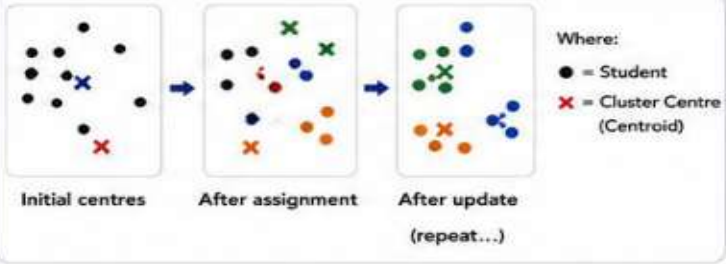
Cluster 1
High-performing students

Cluster 2
Average students

Cluster 3
Students needing intervention

3. HOW K-MEANS WORKS

1. Select K.
2. Choose initial cluster centres.
3. Assign each record to the **nearest centre.**
4. **Recalculate** the centre of each cluster.
5. Repeat until the clusters stabilize.



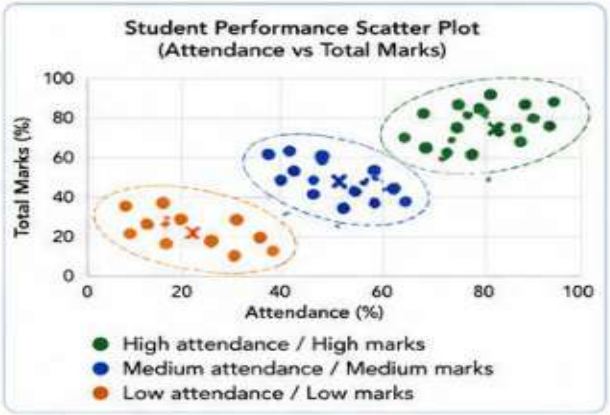
4. EXCEL DEMONSTRATION

Using the Student Performance dataset:

Select:

- Attendance
- Total Marks

Insert: Scatter Plot



K-Means **automates** this grouping mathematically.

5. DISCUSSION

1 How many clusters can you identify?

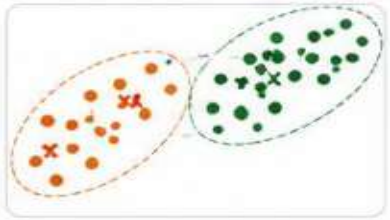
Look at the scatter plot. How many natural groups do you see?

2 Why might different values of K produce different results?

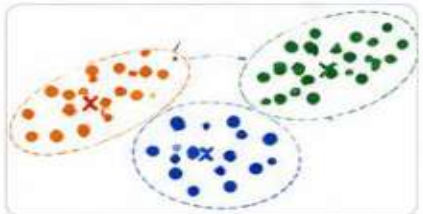
Trying K = 2, 3, 4 … will create different groupings.
There is no single "best" K, it depends on the problem.

EXAMPLE: EFFECT OF DIFFERENT K VALUES

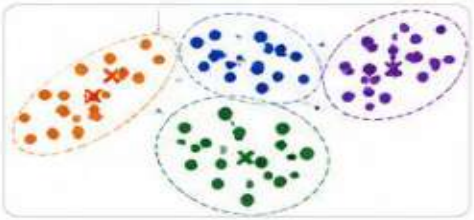
K = 2
Two broad groups



K = 3
Three meaningful groups



K = 4
More detailed groups



KEY TAKEAWAY:

K-Means clustering helps us discover **natural groups** in data without predefined labels.
The **choice of K** affects the results, so understanding the data is important.

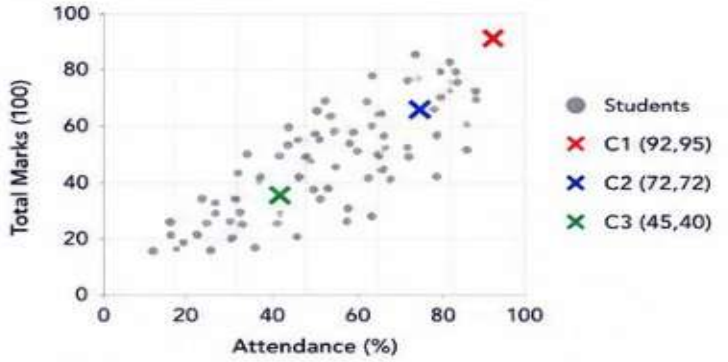
K-MEANS CLUSTERING (K = 3) – STEPS 2, 3 & 4

2 CHOOSE INITIAL CLUSTER CENTRES

Select K = 3 initial centroids from existing data points (randomly).

Cluster	Attendance (%)	Total Marks (100)
C1 (Red)	92	95
C2 (Blue)	72	72
C3 (Green)	45	40

Plot the initial centres.



3 ASSIGN EACH RECORD TO THE NEAREST CENTRE

Calculate distance from each record to all 3 centres using Euclidean distance and assign to the nearest.

Euclidean Distance
between point (x,y) and centre (a,b):

$$d = \sqrt{(x-a)^2 + (y-b)^2}$$

Example for Student ID 1

Attendance = 95, Total Marks = 88

d to C1 = $\sqrt{(95-92)^2 + (88-95)^2} = \sqrt{58} = 7.62$

d to C2 = $\sqrt{(95-72)^2 + (88-72)^2} = \sqrt{785} = 28.02$

d to C3 = $\sqrt{(95-45)^2 + (88-40)^2} = \sqrt{4909} = 70.06$

Nearest centre = C1 (Red)

Do this for all students.

ID	Attendance (%)	Total Marks (100)	Distance to Centres			Assigned Cluster
			C1 (92,95)	C2 (72,72)	C3 (45,40)	
1	95	88	7.62	28.02	70.06	1
2	90	85	11.18	23.26	64.03	1
3	85	82	16.28	15.81	56.28	2
...	...	...	...	...	...	...

Cluster 1 (Red)	Cluster 2 (Blue)	Cluster 3 (Green)
17 students	41 students	42 students

4 RECALCULATE THE CENTRE OF EACH CLUSTER

Find the new centre of each cluster by taking the mean of Attendance and Total Marks of all students in that cluster.

Cluster 1 (Red) – 17 students

New Attendance = $\frac{\sum \text{Attendance}}{17} = \frac{92.529 = 37}{17} = 54.43$

New Total Marks = $\frac{\sum \text{Total Marks}}{17} = \frac{846.941 = 17}{17} = 49.82$

Cluster 2 (Blue) – 41 students

New Attendance = $\frac{\sum \text{Attendance}}{41} = \frac{3066.471 = 31}{41} = 74.79$

New Total Marks = $\frac{\sum \text{Total Marks}}{41} = \frac{2977.471 = 32}{41} = 72.62$

Cluster 3 (Green) – 42 students

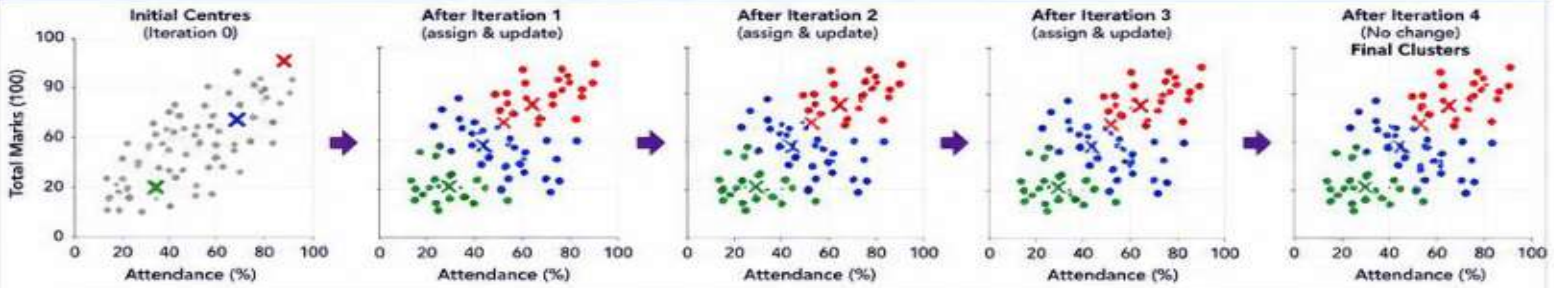
New Attendance = $\frac{\sum \text{Attendance}}{42} = \frac{2064.000 = 62}{42} = 49.14$

New Total Marks = $\frac{\sum \text{Total Marks}}{42} = \frac{1950.588 = 32}{42} = 46.44$

These become the new centroids for the next iteration.

Cluster	New Attendance (%)	New Total Marks (100)
C1 (Red)	54.43	49.82
C2 (Blue)	74.79	72.62
C3 (Green)	49.14	46.44

REPEAT UNTIL CLUSTERS STABILIZE



Red = Cluster 1 (At-risk) | Blue = Cluster 2 (Average) | Green = Cluster 3 (High Performing) | × = Centroid

FINAL DATASET WITH CLUSTER LABELS

ID	Attendance (%)	Total Marks	Final Cluster
1	95	88	3 (Green)
2	90	85	3 (Green)
3	85	82	3 (Green)
4	60	45	1 (Red)
5	55	40	1 (Red)
...	...	...	...

Cluster Interpretation

Cluster 1 (Red):
At-risk / Low performance
Lower attendance & Lower marks)

Cluster 2 (Blue):
Average performance

Cluster 3 (Green):
High performance

SUMMARY OF FORMULAS

Euclidean Distance

$$d = \sqrt{(x-a)^2 + (y-b)^2}$$

(x, y) = record point
(a, b) = centroid

New Centroid (Mean)

$$\bar{x} = \frac{\sum x_i}{n}, \quad \bar{y} = \frac{\sum y_i}{n}$$

n = number of records in the cluster

Stop Condition

When cluster assignments no longer change between iterations.

Goal

Minimize within-cluster variance (SSE):

$$SSE = \sum \sum d^2$$

KEY TAKEAWAY

K-Means groups students into K clusters by repeatedly assigning to the nearest centre and updating centres until stable.

SUMMARY

Classification	Clustering
Predict known classes	Discover unknown groups
Requires labelled data	Requires unlabelled data
Supervised Learning	Unsupervised Learning
Decision Tree, Naïve Bayes	K-Means

Classification is about **predicting** a known outcome.
Clustering is about **discovering** natural patterns in data.

MODULE 4

ASSOCIATION RULE MINING

Discover hidden relationships in transactional data and turn them into actionable insights.



KEY MEASURES

SUPPORT
How frequently the rule occurs

CONFIDENCE
How reliable the rule is

LIFT
How strong the association is compared to random chance

LEARNING OUTCOMES

At the end of this session, participants should be able to:

01 **Explain** the concept of Market Basket Analysis.

02 **Understand** Association Rules.

03 **Calculate** Support, Confidence, and Lift using Excel.

04 **Explain** how the Apriori algorithm works.

05 **Interpret** association rules for business decision-making.

APRIORI PROCESS

Transaction Data → Find Frequent Itemsets → Generate Rules → Apply Support & Confidence → Strong Rules

Uncover patterns.
Increase sales.
Make smarter decisions.

- Identify product combinations
- Drive targeted promotions
- Improve customer satisfaction
- Boost sales and revenue

Introduction to Market Basket Analysis

"When you go to a supermarket, have you noticed that bread, butter, and jam are often placed close together?"



WHY SUPERMARKETS ARRANGE PRODUCTS THIS WAY

- It is **not random.**
- Retailers **analyse millions of customer transactions** to discover which products are frequently purchased together.
- This analysis is known as **Market Basket Analysis.**

WHAT IS MARKET BASKET ANALYSIS?

Market Basket Analysis is a data mining technique that identifies **associations and patterns** between items based on customer purchase behaviour.

Key Characteristics

- Analyses **transaction data** (shopping baskets).
- Discovers items that are **frequently bought together.**
- Helps in better **decision-making and planning.**
- Improves **store layout, promotions,** and **recommendations.**

EXAMPLE

From thousands of transactions, the analysis may find rules like:

Bread	+	Butter	→	Often bought together
Bread	+	Jam	→	Often bought together
Butter	+	Jam	→	Often bought together

These patterns help retailers **understand customer behaviour** and **increase sales.**

HOW MARKET BASKET ANALYSIS WORKS

1. **Collect Transaction Data**
   Gather historical sales data where each record represents a shopping basket.
2. **Find Frequent Itemsets**
   Identify items or item combinations that appear in many transactions (minimum support).
3. **Generate Association Rules**
   Create rules of the form A → B (If a customer buys A, they are likely to buy B).
4. **Evaluate Rules**
   Measure the strength of the rules using Support, Confidence and Lift.
5. **Apply Insights**
   Use the patterns for store layout, promotions, cross-selling, and inventory management.

KEY TAKEAWAY: Market Basket Analysis turns large amounts of sales data into valuable insights about what customers buy together. → The right insights lead to the right actions, happy customers, and higher sales!

What is Association Rule Mining?

Association Rule Mining is a technique used to discover **relationships between items** in a dataset.

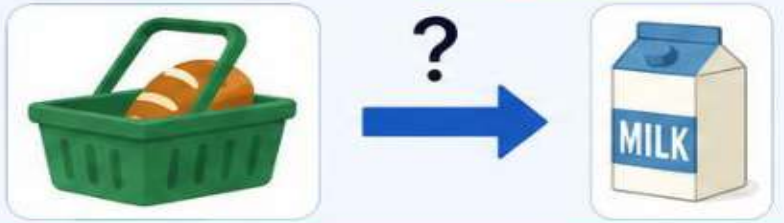
Instead of predicting, it **answers questions** such as

- Which products are frequently purchased together?
- Which courses are commonly taken together?
- Which symptoms usually occur together?
- Which websites are frequently visited together?

EXAMPLE

Transaction	Items Purchased
T1	Bread, Milk
T2	Bread, Butter
T3	Milk, Eggs
T4	Bread, Milk, Eggs
T5	Bread, Milk, Butter

QUESTION



Do customers who buy **Bread** also buy **Milk**?

Association Rule Mining helps answer this by analysing transaction data and discovering hidden patterns.

KEY CONCEPT

Finds items that **occur together** frequently. → % Measures the **strength** of the relationship using Support, Confidence and Lift. → Helps businesses and organizations make better decisions.

IN SIMPLE TERMS

It finds **"IF–THEN"** patterns.

Example: IF a customer buys **Bread**, **THEN** they are likely to buy **Milk**.

GOAL

Discover **meaningful relationships** in data to improve customer experience, increase sales, plan promotions, and make **data-driven decisions**.

ACTIVITY 1 – ASSOCIATION RULE (SUPPORT, CONFIDENCE AND LIFT)

Use the provided **Supermarket_Transactions_100.xlsx** dataset.

DATASET OVERVIEW

Total Transactions	Transactions with Bread	Transactions with Milk	Transactions with Bread and Milk
100	15	29	11

1 SUPPORT

Support measures how frequently an itemset appears.

Formula

$$\text{Support}(A,B) = \frac{\text{Transactions containing A and B}}{\text{Total Transactions}}$$

Example (Bread → Milk)

Total Transactions	100
Bread and Milk purchased together	11

Calculation

$$\text{Support}(\text{Bread, Milk}) = \frac{11}{100} = 0.11$$

Result: 11%

Interpretation

11% of all customers purchased Bread and Milk together.

2 CONFIDENCE

Confidence measures how often B is purchased when A is purchased.

Formula

$$\text{Confidence} = \frac{\text{Support}(A,B)}{\text{Support}(A)}$$

Example (Bread → Milk)

Bread purchased	15
Bread and Milk together	11

Calculation

$$\text{Confidence} = \frac{11}{15} = 0.733$$

Result: 73.3%

Interpretation

73.3% of customers who purchased Bread also purchased Milk.

3 LIFT

Lift measures the strength of an association.

Formula

$$\text{Lift} = \frac{\text{Confidence}}{\text{Support}(B)}$$

Example (Bread → Milk)

Confidence	73.3%
Support(Milk)	29%
Calculation	0.733 / 0.29 = 2.53

Result: 2.53

Interpretation

- Lift > 1 → Positive relationship
- Lift = 1 → Independent (no association)
- Lift < 1 → Negative relationship

Since **Lift = 2.53 > 1**, customers who buy **Bread** are **2.53 times** more likely to buy **Milk** than an average customer.

SUMMARY

Measure	Purpose
Support 11%	**Frequency** How often the items appear together in all transactions.
Confidence 73.3%	**Reliability** How often B is purchased when A is purchased.
Lift 2.53	**Strength of Relationship** How much more likely the items occur together compared to random chance.

Key Takeaway

Bread and Milk have a strong positive association. This insight can help supermarkets place them near each other or run joint promotions.

KEY MESSAGE

Association Rule Mining helps us discover hidden patterns in transaction data.
It reveals what customers buy together, helping businesses make smarter decisions.



APRIORI ALGORITHM

1. WHAT IS APRIORI?

Apriori is one of the most popular algorithms for Association Rule Mining.

It identifies:

1. Frequent Itemsets
2. Association Rules



2. BASIC PRINCIPLE – APRIORI PROPERTY

If an itemset is frequent, then all of its subsets must also be frequent.

Example:

If

Bread + Milk + Butter

is frequent,

then

Bread + Milk

must also be frequent.

This is called the **Apriori Property.**

EXAMPLE TRANSACTION DATA

TID	Items Purchased
T1	Bread, Milk, Butter
T2	Bread, Milk
T3	Bread, Butter
T4	Milk, Butter, Eggs
T5	Bread, Milk, Eggs

Suppose the minimum support (min_sup) = 2 transactions (40%)

3. STEPS OF APRIORI

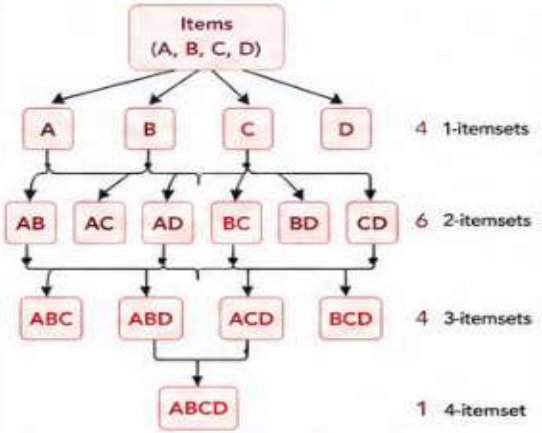
1. **Transaction Data**
   Input the transaction dataset.
2. **Count Item Frequency**
   Count how many times each item appears.
3. **Remove Infrequent Items**
   Remove items with support < min_sup.
4. **Generate Candidate Itemsets**
   Join frequent itemsets to form larger candidate itemsets (k-itemsets).
5. **Calculate Support**
   Scan the database and count support for each candidate itemset.
6. **Generate Association Rules**
   From frequent itemsets, generate rules $X \rightarrow Y$ (where $X \cap Y = \emptyset$).
7. **Calculate Confidence**
   Calculate confidence of each rule:
   $$\text{Confidence} = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)}$$
8. **Evaluate Lift**
   Evaluate lift to measure the strength of the rule.
   $$\text{Lift} = \frac{\text{Confidence}}{\text{Support}(Y)}$$

4. HOW APRIORI REDUCES COMPUTATION

Apriori reduces computation by eliminating combinations that cannot possibly be frequent.

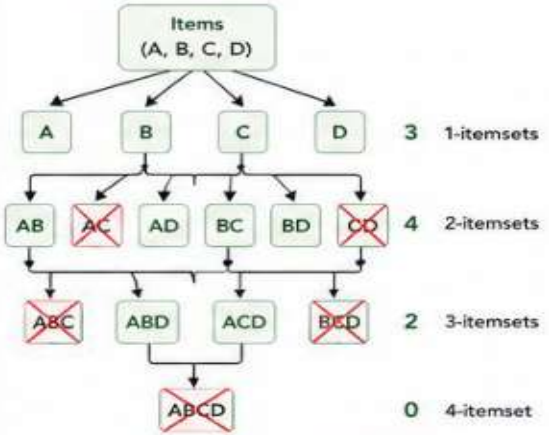
Example:

Without Apriori Property (Brute Force)
All possible itemsets must be generated and counted.



Total candidates = 4 + 6 + 4 + 1 = 15

With Apriori Property
Infrequent itemsets are pruned early. Fewer candidates are generated.



Total candidates = 3 + 4 + 2 = 9

By applying the Apriori Property, we avoid generating and counting itemsets that contain an infrequent subset (e.g., AC or CD).

SUMMARY

Step	Purpose	Key Benefit
1–3	Filter out infrequent items early	Reduces the search space
4–5	Generate and evaluate only promising candidates	Saves time and memory
6–8	Produce strong, meaningful rules	Helpful for decision making

KEY TAKEAWAY
Apriori uses the "if frequent – then all subsets are frequent" property to efficiently find strong association rules with less computation.

ACTIVITY 2 – EXCEL DEMONSTRATION

Open: Supermarket_Transactions_100.xlsx

1 STEP 1 – REVIEW THE TRANSACTION LIST

Review the transaction list.

Example:

Transaction	Items
T001	Bread, Milk
T002	Coffee, Sugar
T003	Bread, Butter
...	...

	A	D
1	Transaction ID	Items
2	T001	Bread, Milk
3	T002	Coffee, Sugar
4	T003	Bread, Butter
5	T004	Milk, Butter, Eggs
6	T005	Bread, Milk, Butter
7	...	...

2 STEP 2 – COUNT ITEM FREQUENCY

Use Pivot Table or COUNTIF to count how many times each item appears.

A. Pivot Table Example

Row Labels	Count of Items
Bread	45
Milk	55
Butter	38
Coffee	32
Sugar	31
Grand Total	**100**

B. Using COUNTIF Function

Assume items are in column D.

Bread	=COUNTIF(D:D,"*Bread*")
Milk	=COUNTIF(D:D,"*Milk*")
Butter	=COUNTIF(D:D,"*Butter*")
Coffee	=COUNTIF(D:D,"*Coffee*")
Sugar	=COUNTIF(D:D,"*Sugar*")

Repeat the same for all items.

3 STEP 3 – CALCULATE SUPPORT

Support measures how frequently an item (or itemset) appears in all transactions.

Example:

Bread appears 45 times

Total transactions 100 times

Support(Bread) = 45 / 100 = 0.45 = **45%**

General Formula:

$$\text{Support}(X) = \frac{\text{Count of } X}{\text{Total Transactions}}$$

4 STEP 4 – CALCULATE CONFIDENCE

Confidence measures how often item B is purchased when item A is purchased.

Example: Rule Bread → Milk

Bread appears 45 times

Bread and Milk appear 35 times

$$\text{Confidence}(\text{Bread} \rightarrow \text{Milk}) = \frac{35}{45} = 0.78 = 78\%$$

General Formula:

$$\text{Confidence}(A \rightarrow B) = \frac{\text{Support}(A, B)}{\text{Support}(A)}$$

5 STEP 5 – CALCULATE LIFT

Lift measures the strength of an association.

Example: Rule Bread → Milk

Milk Support 55%

Confidence (Bread → Milk) 78%

Lift = 0.78 / 0.55 = **1.42**

Interpretation

- Lift > 1 **Positive association**
- Lift = 1 Independent
- Lift < 1 **Negative association**

General Formula:

$$\text{Lift}(A \rightarrow B) = \frac{\text{Confidence}(A \rightarrow B)}{\text{Support}(B)}$$

6 STEP 6 – DISCUSS THE RESULTS

Calculate Support, Confidence and Lift for important rules and discuss.



Rule	Support	Confidence	Lift	Interpretation
Bread → Milk	35%	78%	1.42	Positive association
Coffee → Sugar	24%	81%	1.65	Strong association
Tea → Biscuits	20%	76%	1.48	Positive association

- Which rule is strongest?
- Which rule would you use for product promotion?
- Which products should be displayed together?

Real-world Applications

Industry	Example
Retail	**Cross-selling products** (e.g., customers who buy bread also buy milk).
Healthcare	**Frequently co-occurring symptoms** (e.g., fever and cough often occur together).
Banking	**Customers likely to purchase additional financial products** (e.g., a credit card with a savings account).
Education	**Courses commonly enrolled together** (e.g., students who take Data Mining also take Machine Learning).
Telecommunications	**Frequently bundled services** (e.g., customers who subscribe to internet also subscribe to TV service).
E-commerce	**Online recommendation systems** (e.g., "Customers who bought this item also bought...").

Summary



Association Rule Mining discovers hidden relationships within transactional data.

The three key measures are:

- **Support** – How frequently the rule occurs.
- **Confidence** – How reliable the rule is.
- **Lift** – How strong the association is compared with random chance.

Although Excel cannot execute the full Apriori algorithm automatically,
it is an excellent tool for **illustrating** the concepts and **manually calculating** these measures.

SUMMARY & CONCLUSION

From Data to Decisions: The Data Mining Journey

You have learned the essential steps to transform raw data into meaningful knowledge and actionable insights using Microsoft Excel.

MODULE 1

INTRODUCTION TO DATA MINING



- Defined data mining and its importance.
- Differentiated between Data Warehouse and Data Mining.
- Understood the KDD process.
- Identified real-world applications.

MODULE 2

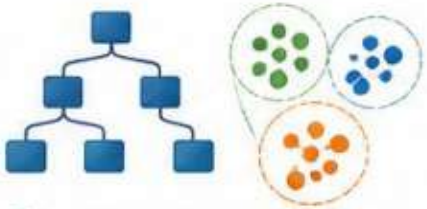
DATA PREPROCESSING USING EXCEL



- Assessed data quality.
- Handled missing values and outliers.
- Normalized and transformed data.
- Selected relevant attributes.
- Prepared data for analysis.

MODULE 3

CLASSIFICATION AND CLUSTERING



- Understood supervised vs. unsupervised learning.
- Explored Decision Tree (J48), Naïve Bayes, and K-Means.
- Built simple classifiers using logical rules in Excel.
- Evaluated models with Confusion Matrix.
- Performed clustering and interpreted results.

MODULE 4

ASSOCIATION RULE MINING



- Understood Market Basket Analysis and Association Rules.
- Calculated Support, Confidence, and Lift in Excel.
- Explained the Apriori algorithm.
- Interpreted rules to support business decisions.

OUTCOME

TURNING DATA INTO VALUE



- You can now apply data mining techniques using Excel.
- You are ready to extract patterns, make predictions, and uncover insights.
- You are equipped to make better, data-driven decisions.

KEY TAKEAWAYS

- Data has value.
- Good data leads to good results.
- The right techniques reveal hidden patterns.
- Insights drive smarter decisions.
- Practice and apply what you have learned!