DMBI Assignment 1

Chapter 3: Classification

Name: Atharva Lotankar
Class: D15C ; RNum: 27
Batch B ; DMBJ

Q1. Define classification and prediction. How do they differ in data mining.

→ Classification in data mining refers to the process of organizing data into predefined categories or classes based on their attributes or features, using models trained with labeled data.

↝ It's basically a supervised data mining technique that assigns each instance in a dataset to one of several predefined categories labels based on its traits.

→ Prediction in data mining, is about estimating continuous or missing values for new observations, often predicting numerical outcomes using past data & established models.

↝ It's used to forecast all continuous values such as predicting a customer's future spending or estimating missing information in dataset.

Point Of Distinction	Classification	Prediction
1. Core Difference	It is used to assign data to predef. class labels, typically discrete outcomes.	It is used to forecast or estimate contin-uous numeric vals for unknown data, focusing on quantitative outcome.
2. Nature Of output	results in categor-ical outputs ↳ spam / notSpam ↳ safe / risky	continuous or numerical outputs, such as predicting sales amt. or custspend
3. Model Used	built by task doer – "classifier" which determines likely categories of new observations	built for predictive task – "predictor" which estimates the unknown numeric value of new data points
4. Accuracy Criteria	Accuracy is measured by how correctly the model assigns class label to new datas.	Accuracy is assessed by how close the estimated value is to actual val. for new observ.
5. Evaluation metrics	Usually evaluated using "Accuracy, Confusion Matrix, F1-Score, or ROC-AUC."	Usually evaluated using regression error metrics such as "Mean Abs Error, R-Squared"

Q2. Give real-world applications of classification prediction.

→ Applications of Classification:

C1: Email Spam Filtering
Classification is widely used to filter emails into spam & non-spam categories based on email content, sender details, & metadata. This helps users manage unwanted messages efficiently.

C2: Medical Diagnosis
Classification algorithms help categorize patients into disease groups by analyzing symptoms, medical history, & test results. For e.g. classifying whether a patient has diabetes or cancer based on medical data.

C3: Credit Risk Analysis
Banks & financial institutions classify applicants into risk categories (low, medium, high) to aid in loan approval decisions, using features like credit score, income, & credit history.

C4: Customer Segmentation
Businesses classify customers into segments based on demographics, buying behaviour

& interaction history to target marketing campaigns effectively & improve customer retention.

C5: Sentiment Analysis
It is used to categorize texts such as social media posts or product reviews into sentiments like positive, negative or neutral, aiding businesses in customer feedback analysis.

→ Applications of Prediction

Sales Forecasting

P1: It models analyze historical sales data, seasonality, & market trends to forecast future sales volume, enabling better inventory & supply chain management.

P2: Credit Scoring & Default Prediction
Beyond classification, continuous prediction models estimate the probability of loan default or creditworthiness to assess financial risk more granularly.

P3: Healthcare Risk Prediction
Predictive models forecast patient outcomes such as likelihood of developing chronic illnesses or hospital readmission risk, helping guide preventive care.

P4: Customer Churn Prediction
Predictive analytics estimates the likelihood that a customer will leave a service, allowing companies to proactively engage at-risk customers to reduce churn.

P5: Weather & Climate Forecasting
Prediction techniques model meteorological data over time to forecast temperature, precipitation, & severe weather events for planning & disaster preparedness.

Q3. What is decision tree induction? Give a real-world example.

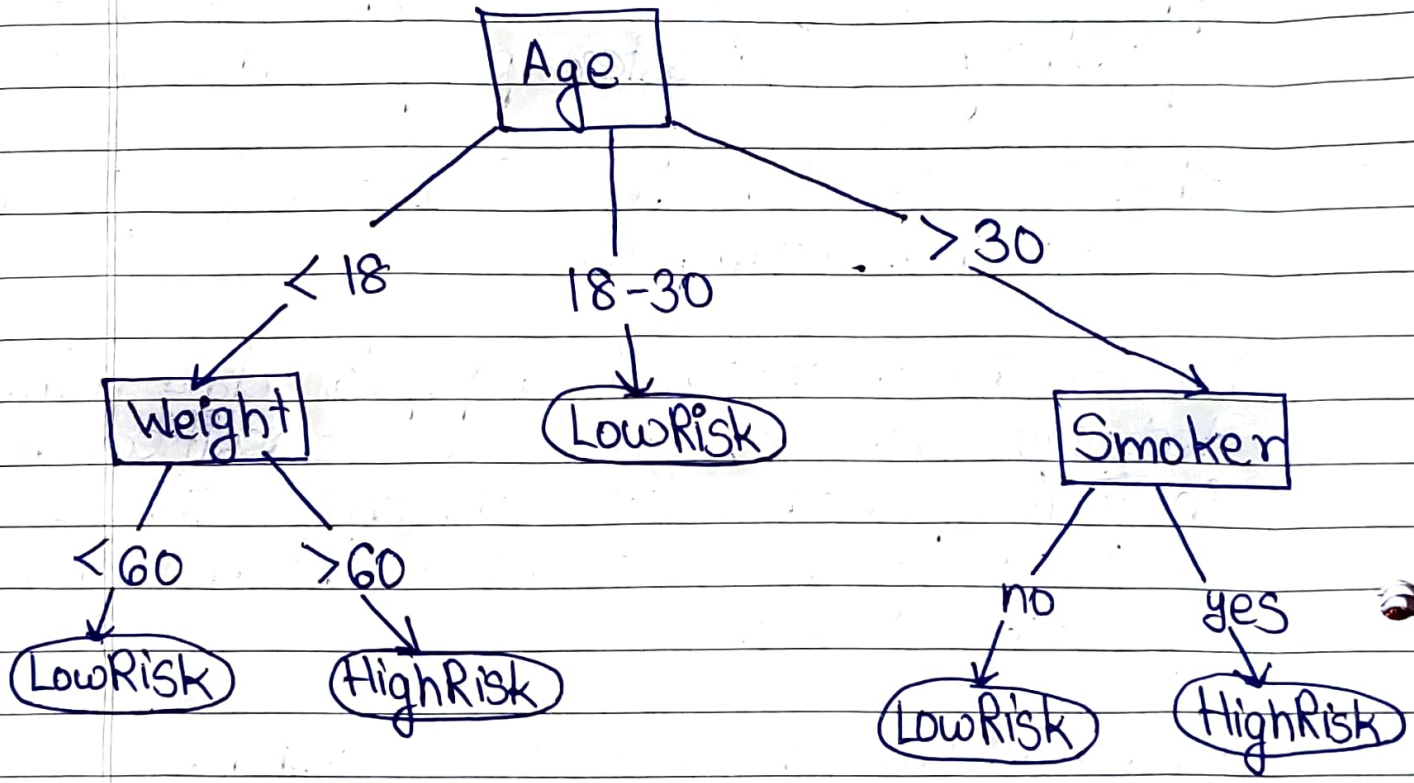
→ Decision Tree induction is a popular data mining technique used to create a classified model in form of a tree-like structure. This tree consists of a root node, internal nodes, & leaf nodes.

Each internal node tests an attribute, each branch represents an outcome of test, & each leaf node assigns a class label or decision.

The goal is to classify data by recursively splitting dataset based on attributes that best separate different classes, using measures like Gini index or info gain to select the splits.

This recursive partitioning continues until the subsets are pure or meet stopping criteria.
Decision trees are easy to interpret & apply for classification tasks.

⇒ Real-World Example:
Understanding the risks to prevent a heart attack.



Suppose a doctor uses this decision tree to quickly assess a patient's heart attack risk. By checking patient's age:

- If under 18, dr. checks wt. less than 60 kg means LowRisk, otherwise HighRisk
- If age is between 18 and 30, risk is classified as LowRisk

> If above 30, doctor asks if patient is a smoker: non-smokers are "LowRisk", smokers are HighRisk.

* This simplified pathway enables fast, data-driven risk assessment in a clinical setting.

Q4. What is tree pruning? Explain pre-pruning methods & post-pruning methods.

→ Tree pruning in decision trees is a technique used to simplify overly complex models by removing branches that do not provide significant predictive power.

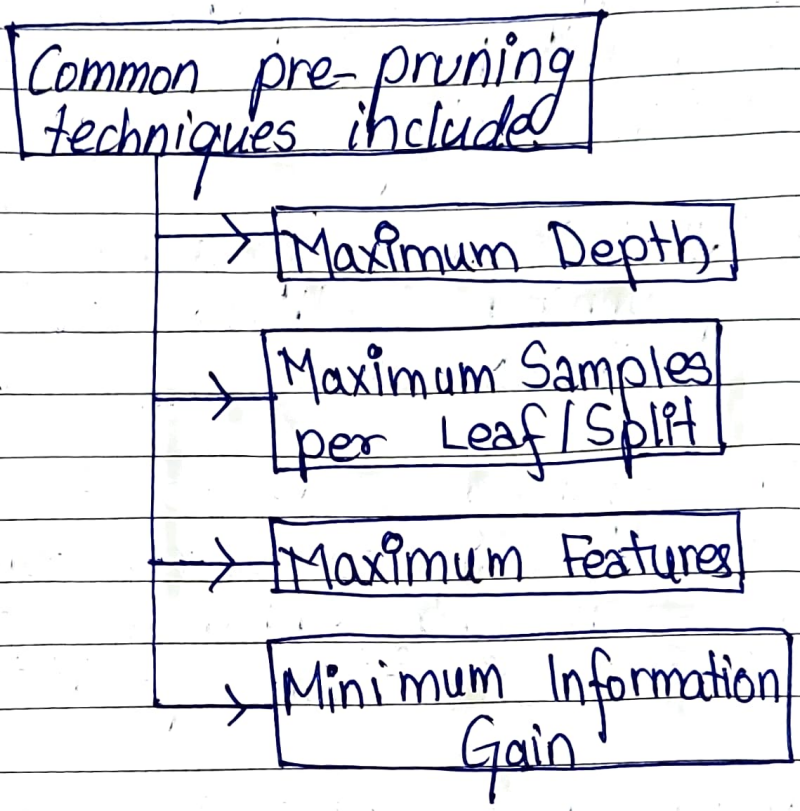
Pruning helps prevent overfitting, which occurs when a decision tree is so tailored to the training data that it performs poorly on unseen data.

By making tree smaller & less complicated, pruning improves accuracy, speed, & interpretability of the model.

(*)→ Pre-pruning

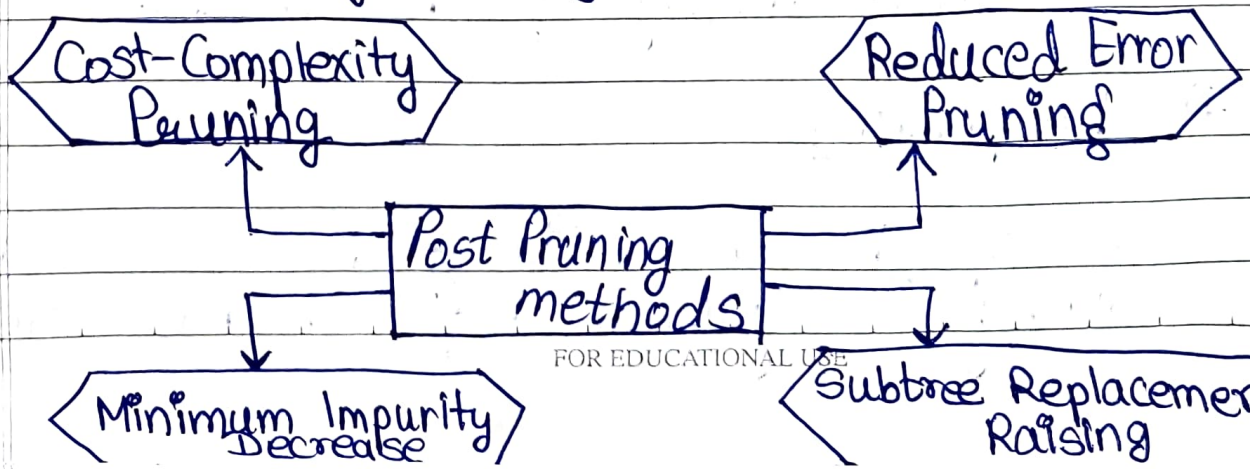
It stops the construction of a decision tree during its growth, before it has learned

all possible splits. The idea is to prevent the tree from becoming too detailed & capturing noise from the training data.



(*) **Post-Pruning**

It is done after tree has been fully constructed, often allowing it to overfit, & then removing branches or nodes that don't contribute to accuracy or generalization.



Q5. Define & Calculate Accuracy, Error Rate, Precision, Recall, F1-score, FPT, TPR, Sensitivity, Specificity

→ In confusion matrix, we define all metrics to understanding classification performance more structuredly.

↝ Accuracy: Measures overall correctness of the model or percentage of test set tuples that are correctly classified.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

TP: Correctly predicted +ve cases ; FP = Incorrectly predicted as +ve

TN: Correctly predicted -ve cases ; FN = Incorrectly predicted as -ve

↝ Error Rate: Measures the proportion of incorrect predictions, or errors made over the records used for testing.

$$\text{Error Rate} = \frac{FP + FN}{TP + TN + FP + FN} = 1 - \text{Accuracy}$$

↝ Precision shows how many predicted positives were actually positive:

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

↝ Recall also called 'Sensitivity' or 'True Positive Rate', measures how many actual positives were correctly identified:

$$\text{Recall} \,/\, \underset{TPR}{\text{Sensitivity}} = \frac{TP}{TP + FN}$$

↝ F1-Score combines precision & recall as their harmonic mean

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

↝ False Positive Rate (FPR) shows proportion of negatives incorrectly classified as positives:

$$FPR = \frac{FP}{FP + TN}$$

↝ Specificity is proportion of actual negatives correctly identified:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

DMBI Assignment - 1

Chapter 4 : Clustering & Outlier Detection

Name: Atharva Lotankar
Class: D15C ; RNum: 27
Batch B ; DMBT

Q1. Define 'Cluster Analysis'. What are its basic objectives in data mining ?

→ Clustering Analysis is a fundamental technique in data mining that involves grouping a collection of data objects into clusters, where objects within the same cluster are more similar to each other than to those in other clusters.

This unsupervised learning method helps uncover hidden patterns or structures in data without relying on predefined categories or labels.

The process relies on algorithms & mathematical models to assign data points into clusters based on measures of similarity, such as

- Euclidean distance = $\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}$
- Manhattan distance = $|x_2-x_1|+|y_2-y_1|$
- Minkowski distance = $(|x_1-x_2|^p+|y_1-y_2|^p)^{1/p}$

where $p \in \{1, \infty\}$ or Natural nos.

* Basics Objectives of Cluster Analysis in Data Mining

⇒ The primary objective of cluster analysis is to identify natural groupings or patterns within a dataset, making it easier to extract meaningful insights.

⇒ It seeks to maximize the similarity among data points within same cluster & minimize similarity between different clusters, thus achieving cohesive & well-separated groups.

⇒ It helps in exploratory data analysis by revealing potential trends or patterns that may require deeper investigation.

⇒ The technique enables effective resource allocation by identifying groups or areas requiring targeted intervention, as in healthcare or logistics

⇒ Used to organize large, unlabelled datasets into manageable segments, which assists in tasks such as market segmentation, customer profiling, fraud detection, & image recognition.

Q2. Differentiate between K-means & K-medoids clustering methods.

→

Point of Distinction	K-means	K-medoids
1. Cluster representation	Uses central mean (centroid) of points, may not be actual data point.	Uses medoid, the most centrally located actual data point.
2. Sensitivity to Outliers	Highly sensitive to outliers.	More robust & less affected by outliers.
3. Distance Metric	Typically uses Euclidean distance.	Can use any distance metric → Euclidean, Manhattan, Minkowski
4. Computationally Effiency.	Generally faster and more efficient.	Slower due to complex medioid selection process.
5. Cluster Shape Assumption	Assumes spherical clusters with similar size.	Does not assume specific cluster shape.
6. Data Type Compatibility	Suitable mainly for numerical data.	Can handle numerical and categorical data.

7. Convergence speed	Usually converges faster, may get stuck in local minima.	Slower convergence but can reach more global optimum.
8. Interpretability	Centroids may not be actual data points.	Medoids are actual data points.
9. Objective Function	Minimizes sum of squared distances (variance) btwn data points & cluster centroid.	Minimizes sum of pairwise dissimilarities b/w data points & medoid.
10. Use case stability	Best for large datasets where speed is essential & data is mostly numerical.	Suited for datasets requiring robustness against noise/outliers & mixed data types

Q3. Explain the terms agglomerative and divisive hierarchical clustering with examples.

→ 'Agglomerative' Hierarchical Clustering is a bottom-up approach, where each data point starts in its own individual clustering. At each step, the

two closest clusters are merged until all points are in a single cluster or a stopping criterion is met.

The algorithm is first calculated by estimating distance between every data points. Initially, each point forms a seperate cluster. The two clusters with the smallest distance are merged, then process repeats (updating distances as clusters merge) until only one cluster remains.

Example: Suppose we have four fruits with weights: apple (100g), banana (120g), cherry (50g), grape (30g).

1. Start

	Apple	Banana	Cherry	Grape
Apple	0	20	50	70
Banana	20	0	70	90
Cherry	50	70	0	20
Grape	70	90	20	0

2. merge closest clusters i.e. cherry & grape of dist = 20 units.

→ apple (100g)
banana (120g)

$C_1 = \{cherry, grape\}$ $C_1 = \{50g, 30g\}$

3. Using single linkage method of minimal distance

Dist. btw/n Apple-Banana = 20 units

Then, $d_{apple, C1} \rightarrow (|100-50|, |100-30|)_{min}$

$= 50$

Then, $d_{banana, C1} \rightarrow (|80-50|, |120-30|)_{min}$

$= 70$ units

merge second pair:

$C_2 = \{apple, banana\}$

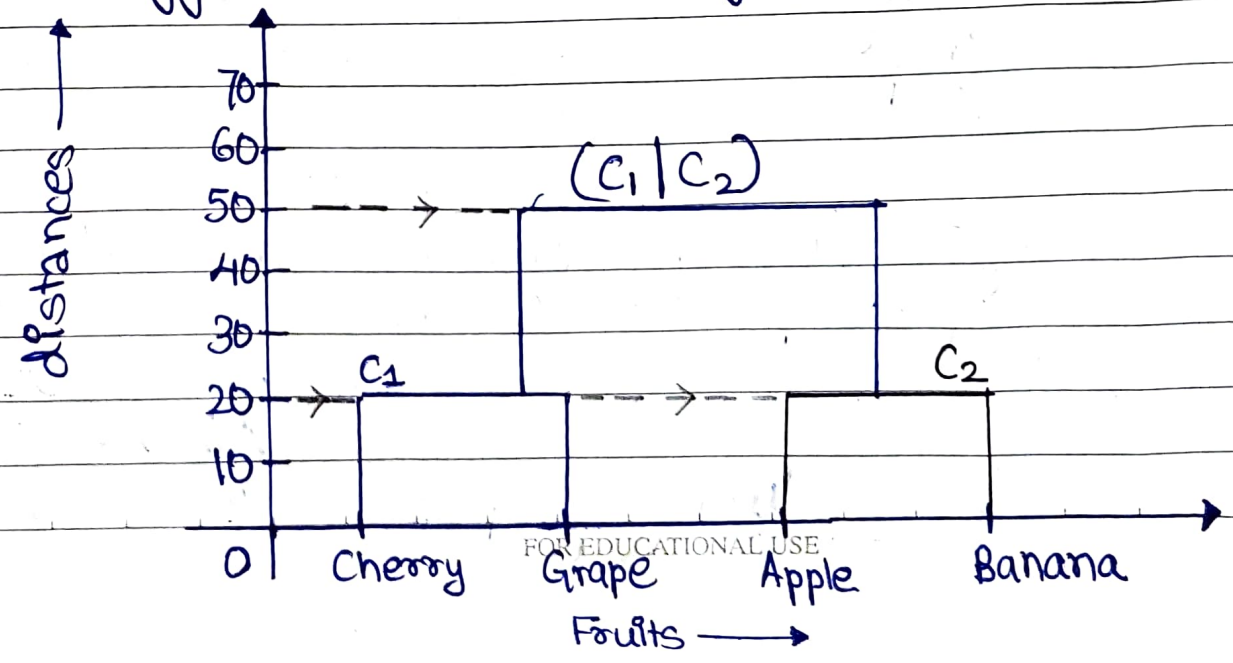
as minimal dist of 20. units

4. Final merge: $d_{C1, C2}$

$\rightarrow [|(50-100), (50-120), (30-100), (30-120)|]_{min}$

$\rightarrow 50$ units

So Agglomerative dendogram:



Divisive Hierarchical Clustering is a top-down approach, where all data points start in one single cluster. At each step, most dissimilar cluster is split into two until each point is alone in its own cluster.

Primarily start with all points in a single group. At each step, the algorithm identifies the cluster with the most heterogeneity (least similarity) within & splits it into two subclusters.

This continues until every data point is a seperate cluster.

Example: Apply divisive analysis for this dataset

	a	b	c	d	e
a	0	9	3	6	11
b	9	0	7	5	10
c	3	7	0	9	2
d	6	5	9	0	8
e	11	10	2	8	0

Step 1: Initially all dataset belong to one cluster

$$C_1 = \{a, b, c, d, e\}$$

Step 2: Find dissimilarities in all:

For a: $\frac{1}{4} \times (d_{ab} + d_{ac} + d_{ad} + d_{ae})$

$= \frac{1}{4}(9+3+6+11) = 7.25$

similarly

For b → 7.75 ; For c: 5.25

For d → 7.00 ; For e: 7.75

Step 3: Find highest avg. dist, here it's b & a, choose anyone like – b.

Now second cluster has 'b'

$C_1 = \{a, c, d, e\}$ & $C_2 = \{b\}$

Step 4: Now find dissim. in existing cluster

$D_a = \frac{1}{3}(d_{ac} + d_{ad} + d_{ae}) - \frac{0_1}{1}(d_{ab})$

$= -2.33$

$D_c = -2.33$; $D_d = 0.67$; $D_e = 0$

highest

so again;

$C_1 = \{a, c, e\}$ & $C_2 = \{b, d\}$

Again

$$D_a = \frac{1}{2}[d(a,c) + d(a,e)] - \frac{1}{2}[d(a,b) + d(a,d)]$$

$$= -0.5$$

$D_c = -13.5$; $D_e = -2.5$

(*) since all -ve we have to stop & form clusters.

Thus;
cluster 1 = (a, c, e) & Cluster 2 = (b, d)

Q4: What is BIRCH? Why is it suitable for large datasets?

→ 'BIRCH' or (Balanced Iterative Reducing & Clustering using Hierarchies) is a hierarchical clustering algorithm designed specifically to handle very large data sets efficiently.

It works by first creating a compact summary of dataset called a Clustering Feature (CF) Tree. This CF tree is a height-balanced tree where each leaf node represents a subcluster summarized by 3 values: the number of points (N),

the linear sum of points (LS), & the squared sum of points (SS).

BIRCH incrementally and dynamically clusters incoming data points to minimize memory usage & processing time, usually requiring only a single scan of the dataset.

-- It is suitable for Large Datasets

- Compact Summary: BIRCH reduces the large dataset into smaller, dense subclusters via CF Tree, enabling it to handle large datasets.

- Incremental Clustering: It incrementally clusters data as it reads it, so it can work with streaming data or datasets too large to fit in memory at once.

- Scalability: It's tree str. & use of cluster, allows it to scale well in both time + memory, making it fast & effient for very large dataset.

- Noise Handling: BIRCH can effectively handle noisy data points that don't

belong to any cluster due to its hierarchical summarization approach.

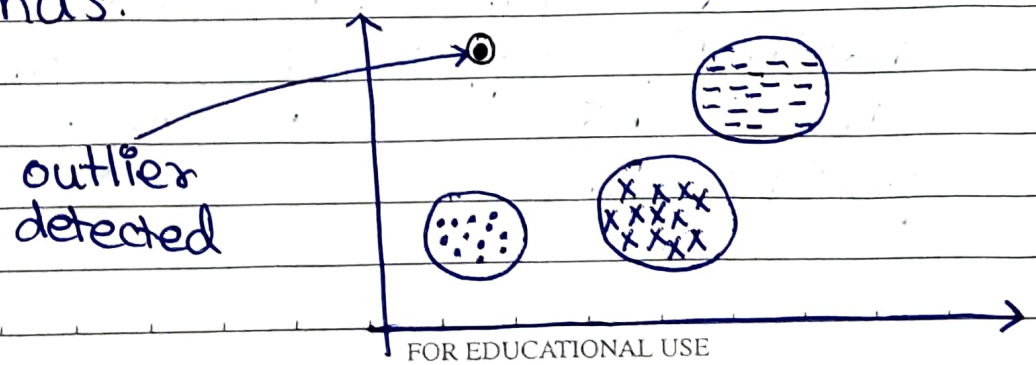
Q5. Define outliers. Give two real-world examples where outlier detection is important.

→ Outliers are data points that significantly differ from the majority of observations in a dataset.

They represent anomalies or rare events that do not conform to the expected pattern or distribution of the data.

Outliers can occur due to measurement errors, data entry mistakes, natural variability, or genuine but rare phenomena.

Detecting outliers is crucial because they can distort statistical analyzes affect the accuracy of machine learning models, and reveal important insights such as fraud, errors, or novel trends.



Two real-world examples where outlier detection is important.

(E1) → Fraud detection in Financial Transactions

In banking and credit card systems, outliers may represent fraudulent transactions that deviate from typical spending behaviours.

Detecting these anomalies early helps prevent financial loss & protects customers from fraud.

(E2) → Healthcare and Medical Diagnosis

In medical data, outliers might indicate abnormal patient conditions or errors in data collection.

Identifying these can help to diagnosed rare diseases, monitor treatment effectiveness, or flag errors that could affect patient care.