SECP2753 Data Mining

Group Task 3

Prepared by group 9

Ahmad Ziyaad	A23CS0206
Goe Jie Ying	A23CS0224
Teh Ru Qian	A23CS0191

Q1 - What is classification in data mining?

Definition 1:

Classification is a form of data analysis that extracts models describing important data classes. Such models, called classifiers, predict categorical (discrete, unordered) class labels. [1]

Definition 2:

Data classification is a two-step process, consisting of a learning step (where a classification model is constructed) and a classification step (where the model is used to predict class labels for given data). [2]

Q2 – Explain the importance and real-world applications of classification in data mining.

- **Enhances data protection and compliance**: Helps prioritize data protection efforts, thereby improving security and adherence to regulatory requirements [3].
- **Optimizes operations**: Reduces costs, boosts user productivity, and facilitates faster decision-making by filtering out irrelevant or redundant data [3].
- **Enables accurate predictions**: Allows for the construction of models that can reliably predict the class labels of new data instances based on input features [4].
- **Supports financial risk management**: Assists banks and financial institutions in identifying potential defaulters, aiding in decisions related to loan approvals, credit cards, and other financial products [4].

Q2 – Explain the importance and real-world applications of classification in data mining. [5]

Continue...

Area/Field	Application/Tool
Healthcare	Doctors use classification to group symptoms or indicators for better diagnostics. It also helps segment patients by age, medical condition, or treatment needs.
Weather Forecasting	Meteorologists classify data such as temperature, humidity, and wind speed to predict weather patterns and generate forecasts.
Finance	Banks and financial institutions classify customer data to assess creditworthiness, detect fraud, and approve loans or credit cards more effectively.
E-commerce	Online retailers use classification to analyze user behavior and recommend products based on past purchases or browsing history.

Q3 – Identify at least TEN (10) classification algorithms. Then, provide a concise explanation (including an example) for THREE (3) of them.

- k-Nearest-Neighbor classifier
- Bayesian Classifiers
- Rule-based classifiers
- Statistical based classifiers
- Neural networks
- Support Vector Machine
- Genetic algorithm
- Rough set
- Decision Tree Classifiers
- Support Vector Machines (SVM)

Q3(a) Neural Networks

Definition:

Neural Networks are inspired by the structure of the human brain. They consist of layers of interconnected nodes (neurons) that process input features through weighted connections and activation functions to make predictions.

Structure:

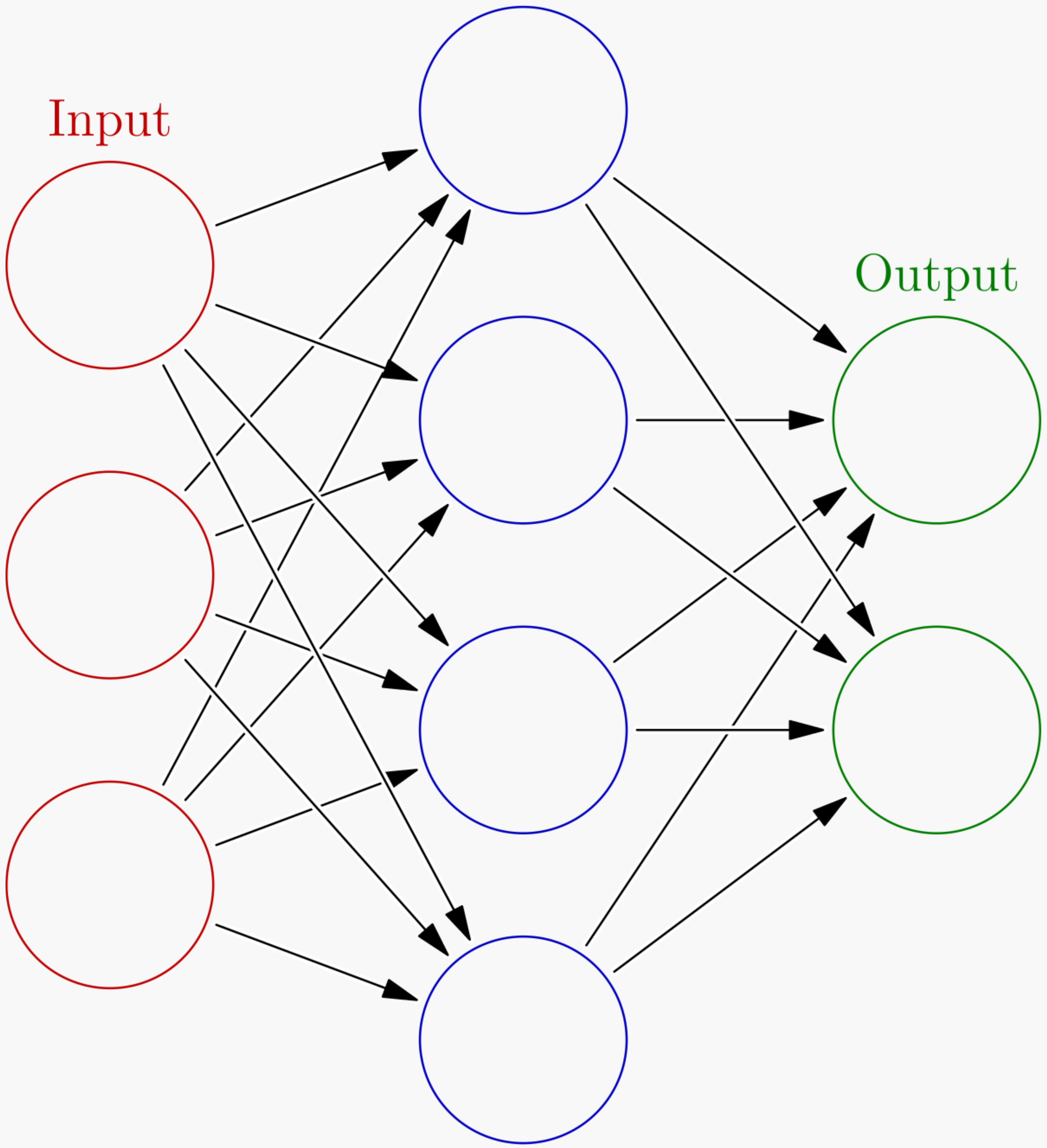
- Input Layer: Receives features (e.g., pixel values in an image).
- Hidden Layers: Perform complex computations (deep networks have many).
- Output Layer: Produces final class probabilities or labels.

Example of Usage:

- Great for large, complex datasets (e.g., images, audio).
- Can capture non-linear relationships.
- Scales well with data and compute power.

Example Use Case:

In a handwritten digit recognition task (like MNIST dataset), a neural network can classify an input image (say, a 28x28 pixel grayscale image) into one of the 10 digit classes (0–9).



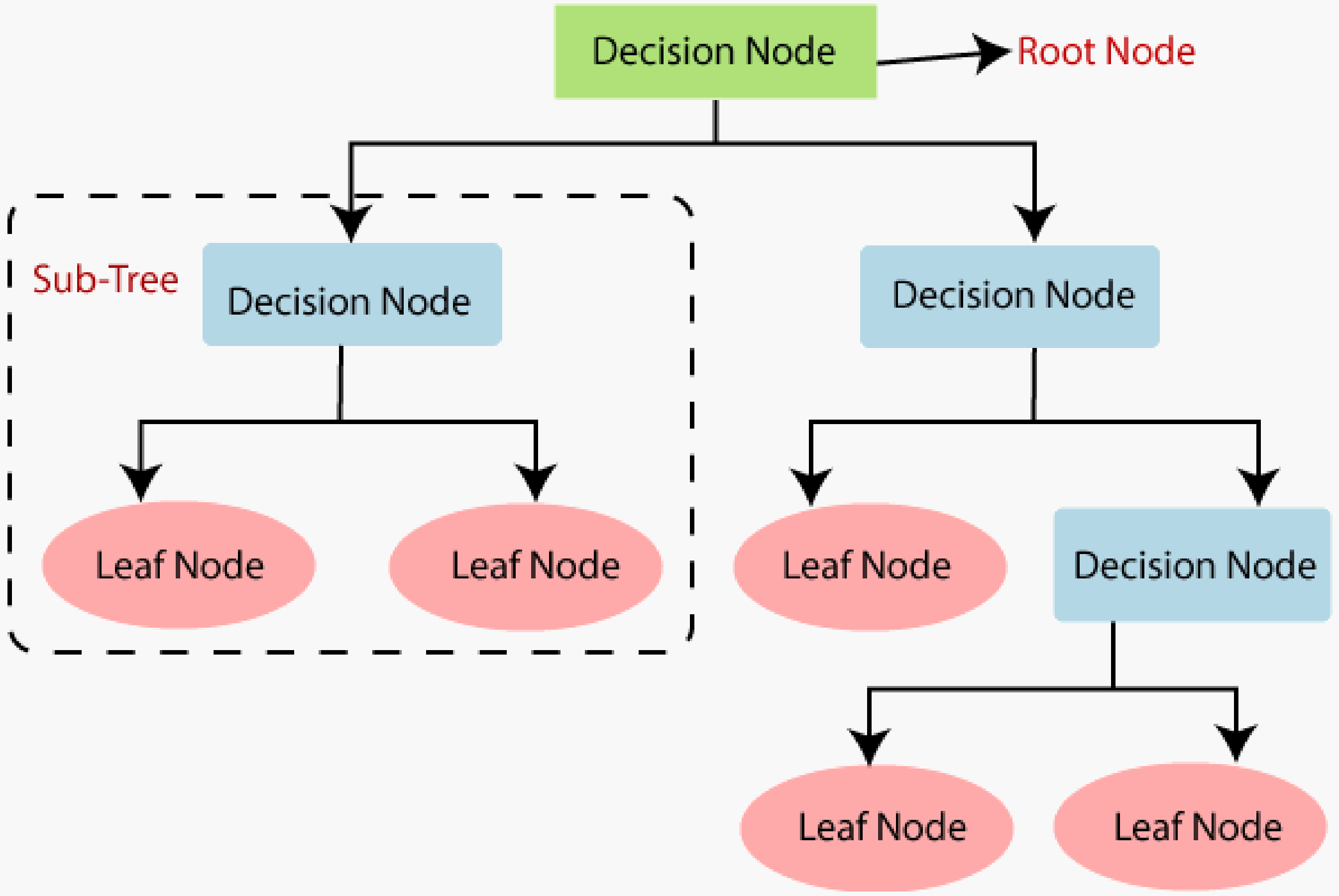
Q3(b) Decision Tree Classifiers

Definition:

Decision Tree Classifier is a supervised machine learning algorithm used for classification and regression tasks. It works by learning **simple decision rules** inferred from data features to make predictions. Each **internal node** represents a condition or decision on a feature, each **branch** represents an outcome of the decision, and each **leaf node** represents a class label (final decision).

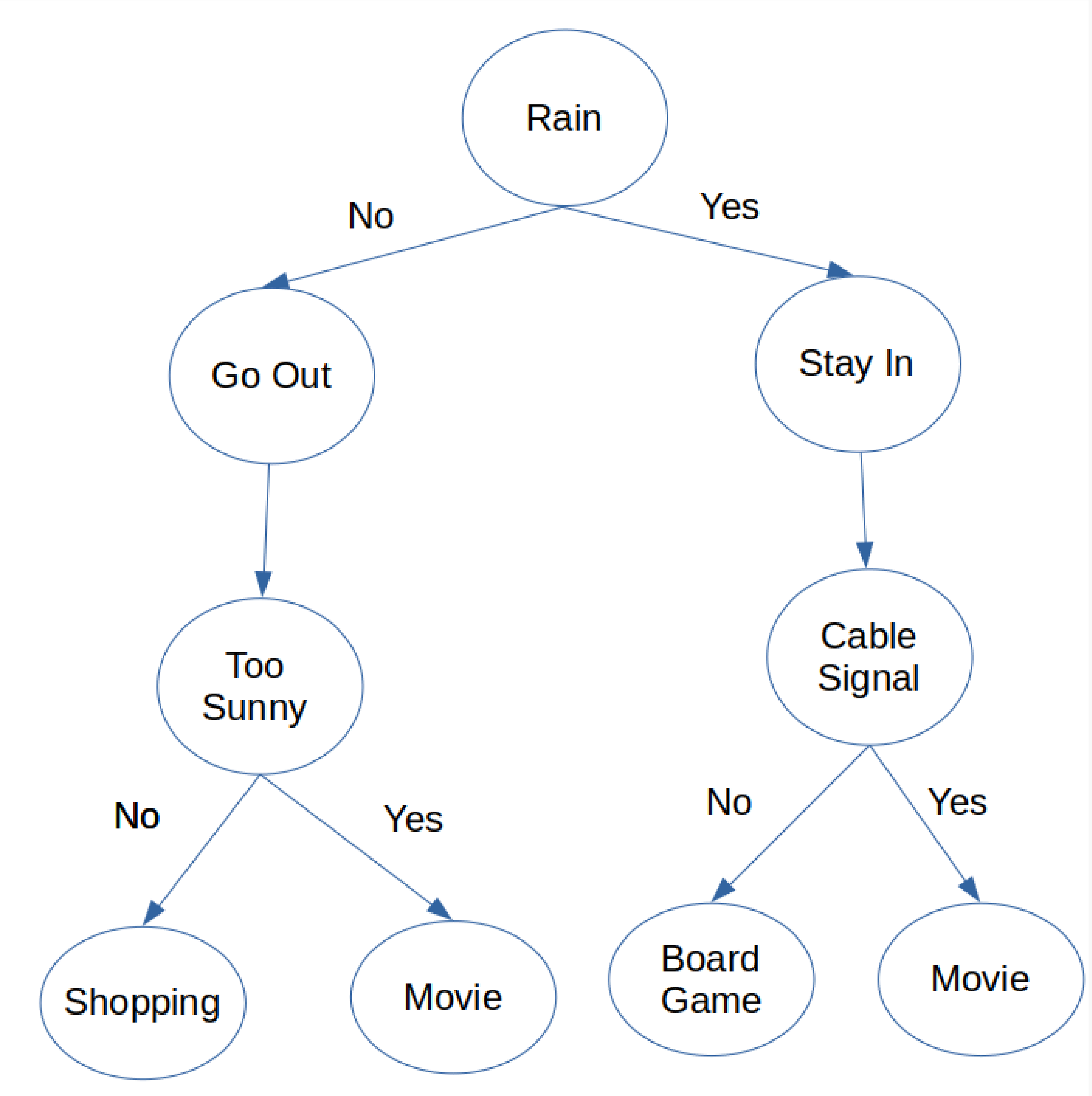
Structure:

- **Root Node:** The first feature the algorithm splits on.
- **Internal Node:** Test conditions (e.g., "Is age > 30?")
- **Branches**: Outcomes of those conditions (e.g., "Yes", "No")
- **Leaf Nodes**: Final decision or class (e.g., "Approve", "Reject")



Q3(b) Decision Tree Classifiers *Continue...*

Example of Usage:



Example Use Case:

- **Loan Approval Systems** – Like in banks or credit institutions.
- **Medical Diagnosis** – Classifying patients based on symptoms/test results.
- **Fraud Detection** – Identifying suspicious transactions.
- **Customer Churn Prediction** – Predicting if a customer will stop using a service.
- **Marketing** – Deciding whether a customer is likely to buy a product.

Q3(c) k-Nearest-Neighbor classifier

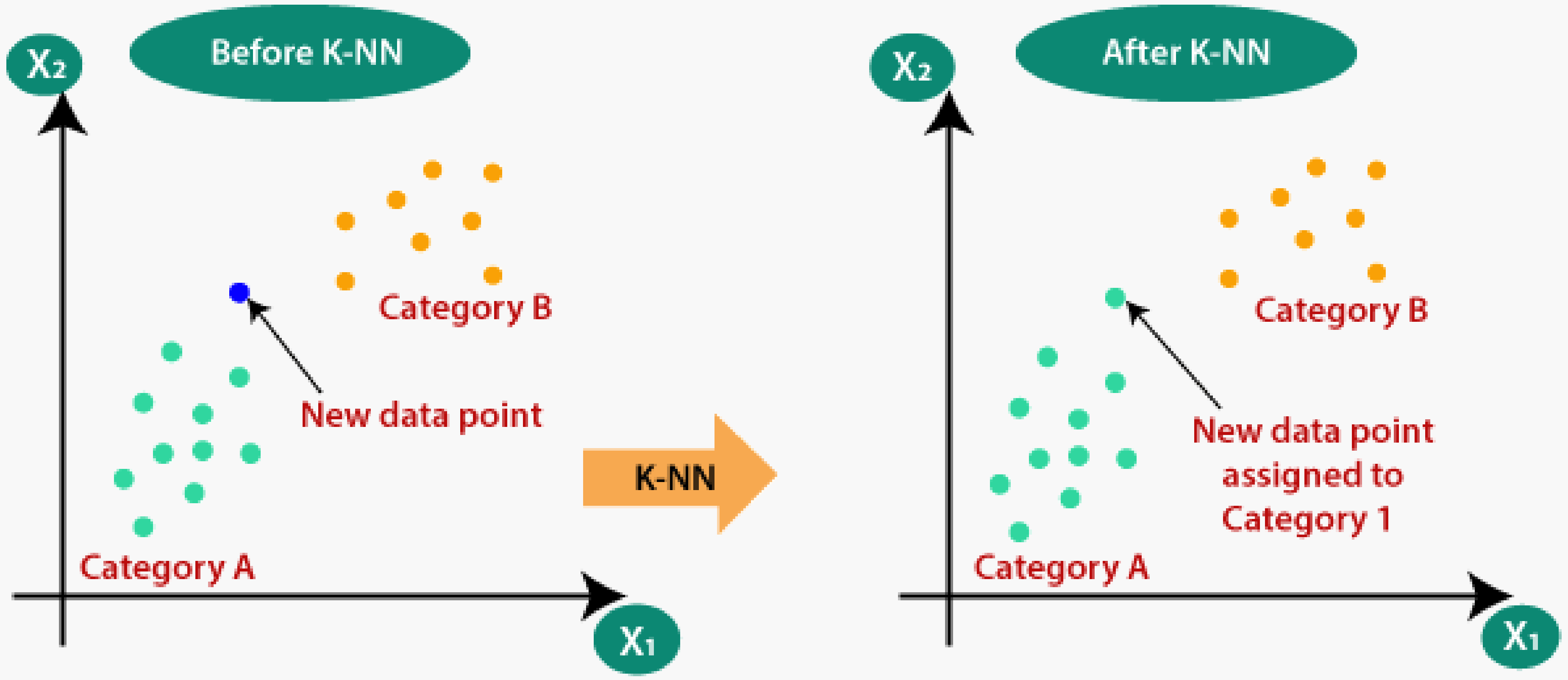
Definition:

The k-nearest neighbor (k-NN) classifier is a classification method that predicts the class of a given test tuple by comparing it to the k most similar training tuples, also called its nearest neighbors, in an n-dimensional space. It is a lazy-learners technique which memorizes the entire training data. [2]

Structure:

- Store all training data as points in an n-dimensional space.
- Given a test tuple, calculate its distance to all training tuples.
- Identify the k nearest neighbors (tuples with the smallest distances).
- Assign the most frequent class label among those neighbors to the test tuple.

Example of Usage [6]:



Example of Use Case [7]:

- Data Preprocessing: Fills in missing values through imputation.
- Recommendation Systems: Suggests content based on user behavior patterns.
- Finance: Assesses loan risk, predicts market trends, detects fraud.
- Healthcare: Predicts disease risks like heart attacks and cancer.

Q4 – How does classification works? Describe the general steps involved in building a classification model. [2]

1. **Learning Step (Training Phase):**
   - A classification model is constructed using a set of training data, where each data record (called a tuple) has a known class label.
   - Each tuple is represented by an attribute vector X = (x1,x2,...,xn), and the class label is a separate attribute.
   - A classification algorithm is used to learn a mapping function y=f(X) that can predict the class label y based on the input attributes X.
   - The output may be in the form of classification rules, decision trees, or mathematical formulas.
   - This process is called supervised learning, as the correct class labels are known during training

Q4 – How does classification works? Describe the general steps involved in building a classification model.

Continue...

2. Classification Step (Testing/Prediction Phase):

- The trained model is applied to new or unseen data to predict their class labels.
- To estimate the model's accuracy, it is tested on a separate test set, which also contains tuples with known class labels but was not used during training.
- The model's accuracy is the percentage of test tuples that are correctly classified.
- Once validated, the model can be used to classify future data where the class label is unknown.

Q5 – Identify and explain common evaluation metrics used in classification in data mining (e.g. accuracy, precision, recall, F1-score). Include relevant formulas.

Evaluation Metric	Explanation	Formula
Accuracy	Measures the proportion of total correct predictions	$Accuracy = \frac{TP+TN}{All}$
Precision	Measures how many of the predicted positives are actually correct	Positive Precision Value (PPV) = $\frac{TP}{TP+FP}$
Recall	Measures how many of the actual positives were correctly predicted	True Positive Recall (TPR) = $\frac{TP}{TP+FN}$
F1-score [8]	Harmonic mean of Precision and Recall; balances both in one metric	$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$

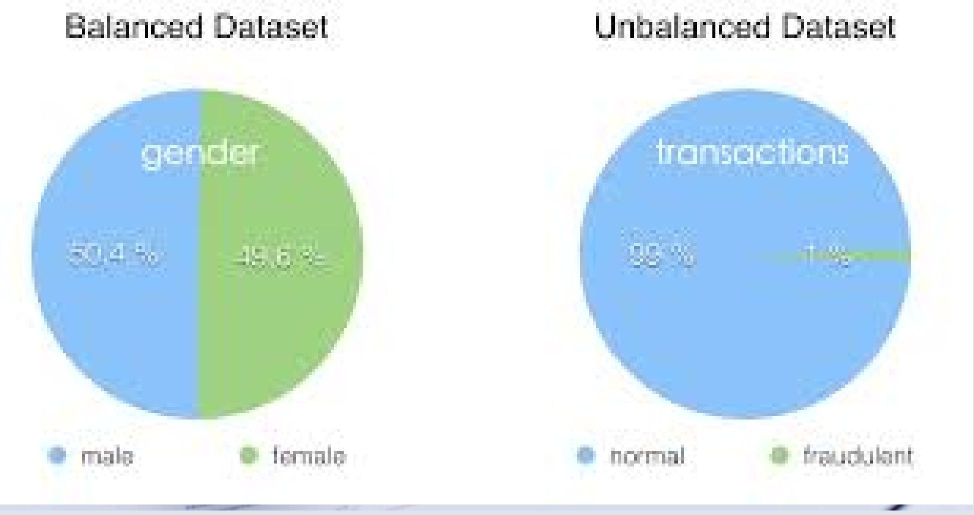
Q6 - Compare Supervised Learning and Unsupervised Learning with examples.

Aspect	Supervised Learning	Unsupervised Learning
Definition	Learning with labeled data	Learning with unlabeled data
Goal	Predict outcomes or classify data	Find hidden patterns or structure in data
Example	Email spam detection (spam or not spam), Predicting house prices based on features	Customer segmentation for marketing, Organizing news articles into categories without labels

Q7- Differentiate between binary classification and multi-class classification. Provide an example for each.

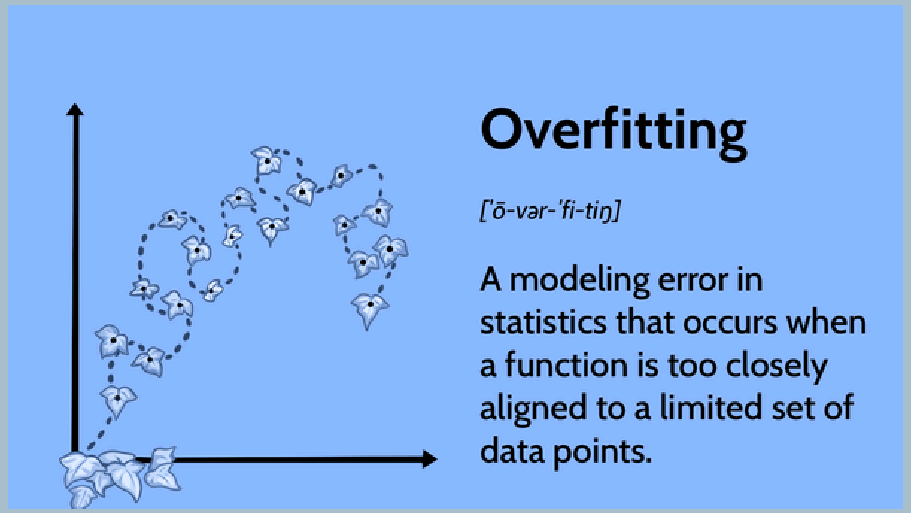
Aspect	Binary Classification	Multi-class Classification
Definition	Classification task with only two possible output classes	Classification task with more than two output classes
Number of Classes	2	3 or more
Example	Email spam detection (spam or not spam), Disease diagnosis: positive or negative	Handwritten digit recognition: digits 0–9, Classifying types of animals: cat, dog, rabbit, bird

Q8- Discuss common challenges encountered in classification tasks.



Imbalanced Data

- Issue: Model performs well on training data but poorly on unseen data.
- Effect: Fails to generalize.
- Solution: Use cross-validation, regularization, and simpler models.



Overfitting

[ˈō-vər-ˈfi-tiŋ]

A modeling error in statistics that occurs when a function is too closely aligned to a limited set of data points.

Overfitting

- Issue: Model performs well on training data but poorly on unseen data.
- Effect: Fails to generalize.
- Solution: Use cross-validation, regularization, and simpler models.

Underfitting

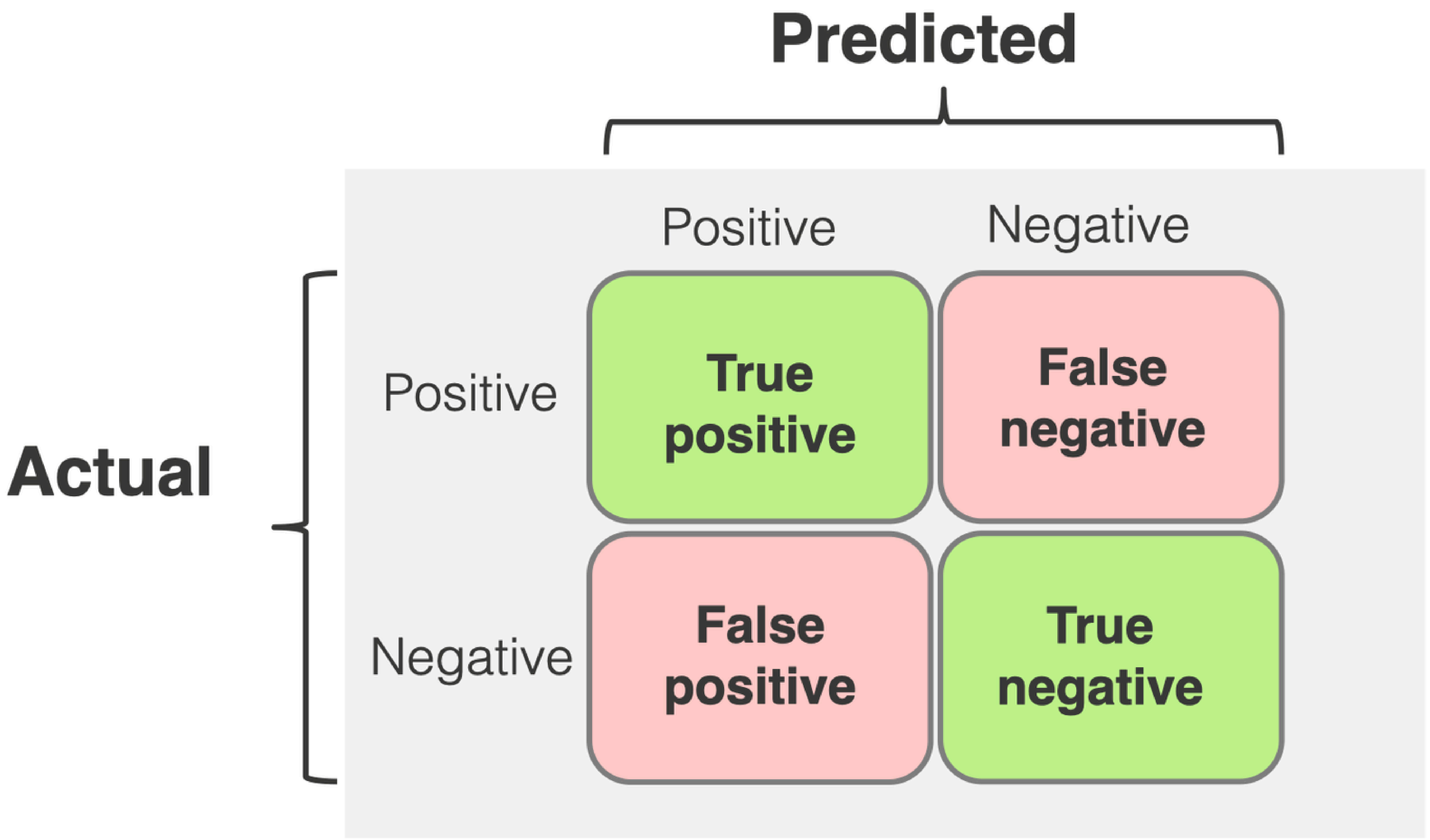
- Issue: Model is too simple to capture underlying patterns.
- Effect: Poor performance on both training and test data.
- Solution: Use more complex models or better feature engineering.

Q9- What is a confusion matrix and how is it used to evaluate classification models? Provide a labeled example.

Confusion Matrix:

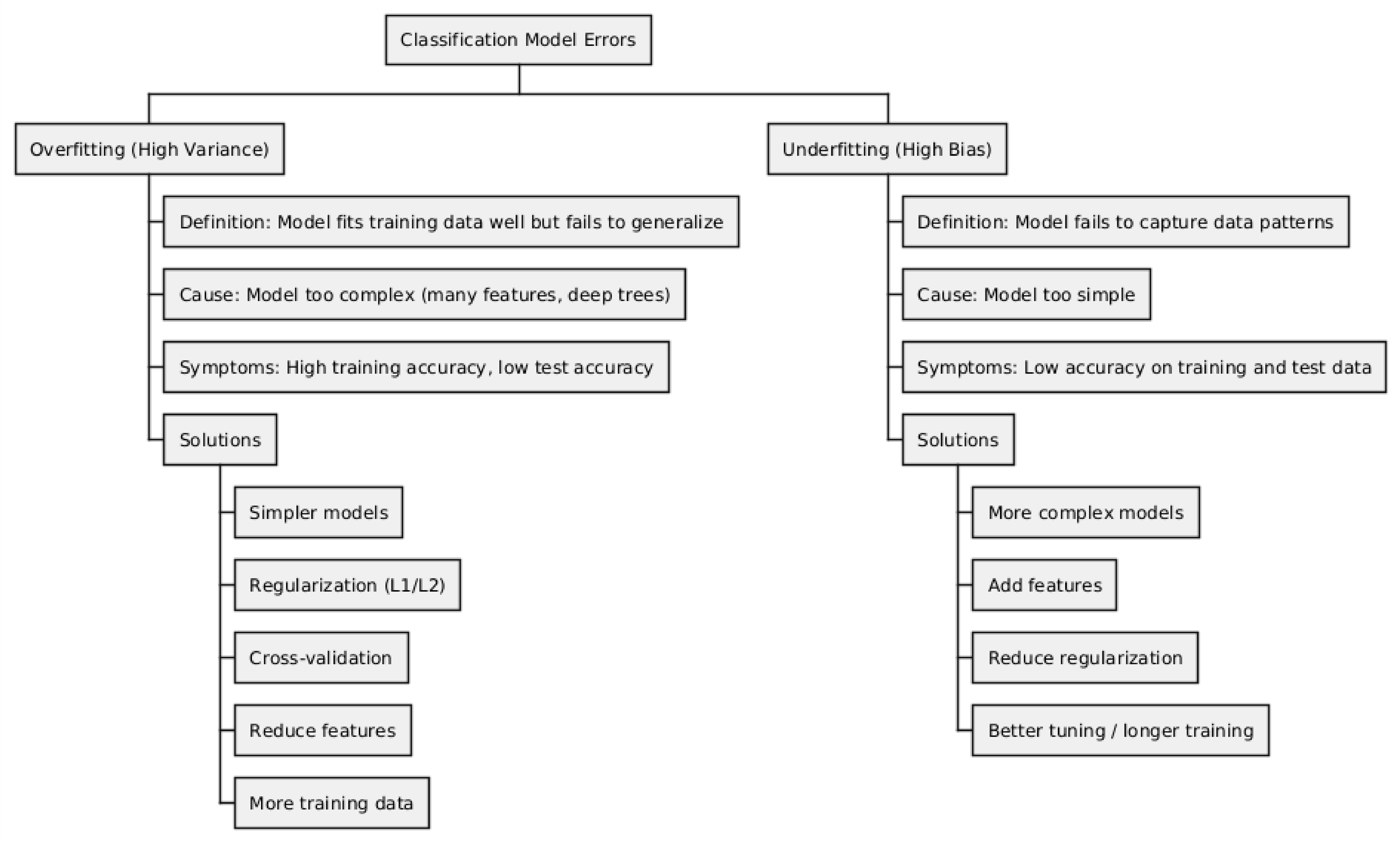
Axes:

- Rows (Actual): These represent the true labels from the dataset:
  - "Positive" means the actual class is positive.
  - "Negative" means the actual class is negative.
- Columns (Predicted): These represent the predictions made by the model:
  - "Positive" means the model predicted the class as positive.
  - "Negative" means the model predicted the class as negative.



		Predicted	
		Positive	Negative
Actual	Positive	True positive	False negative
	Negative	False positive	True negative

Q10- – Explain the concept of overfitting and underfitting in classification models. How can each issue be addressed?



- Classification Model Errors
  - Overfitting (High Variance)
    - Definition: Model fits training data well but fails to generalize
    - Cause: Model too complex (many features, deep trees)
    - Symptoms: High training accuracy, low test accuracy
    - Solutions
      - Simpler models
      - Regularization (L1/L2)
      - Cross-validation
      - Reduce features
      - More training data
  - Underfitting (High Bias)
    - Definition: Model fails to capture data patterns
    - Cause: Model too simple
    - Symptoms: Low accuracy on training and test data
    - Solutions
      - More complex models
      - Add features
      - Reduce regularization
      - Better tuning / longer training

Q11- Describe the role of cross-validation in evaluating the performance of a classification algorithm.

Cross-Validation

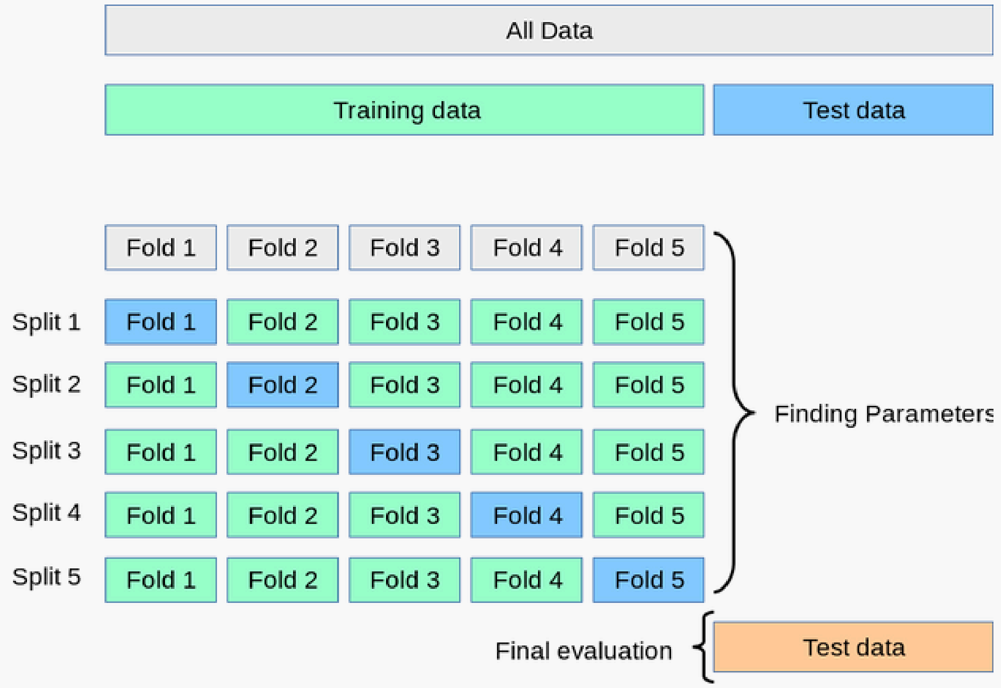
- A method to **check how well a classification model performs.**

How It Works (Using k-Fold Cross-Validation)

1. **Split** the dataset into **k parts** (called "folds").
2. **Train** the model on **k-1 folds** and **test** on the **remaining 1 fold**.
3. **Repeat** the process **k times**, each time with a different fold used for testing.
4. **Average** the results to get a final performance score.

Example (5-Fold Cross-Validation)

1. Data split into 5 folds.
2. Model trained 5 times, each time testing on a different fold.
3. Final score = average of the 5 test results.



Benefits

- Gives a **more accurate** estimate of model performance.
- Helps detect **overfitting or underfitting**.
- Uses the whole dataset for both training and testing (more efficient).

Q12- How can feature selection improve classification model performance?

Feature selection helps improve the performance of a classification model by:

- **Removing irrelevant or redundant features**, which can confuse the model.
- **Reducing overfitting**, by simplifying the model and making it generalize better.
- **Speeding up training and prediction**, because fewer features mean less computation.

Filter Method	**Wrapper Method**
Uses statistical techniques to rank and select features based on their relationship with the target variable.	Tries different combinations of features and evaluates model performance for each set.
Example techniques: • Correlation coefficient • Chi-square test • Mutual information	Example techniques: • Forward Selection (start with none, add one by one) • Backward Elimination (start with all, remove one by one)

Q13- Compare decision trees and logistic regression in terms of model interpretability, accuracy, and use cases.

Aspect	Decision Trees	Logistic Regression
Interpretability	• Easy to interpret (visual tree structure) • Clear decision paths	• Also interpretable (weights show influence) • Based on linear relationship
Accuracy	• Can overfit if not pruned • Good for complex relationships	• Often better for linearly separable data • May struggle with non-linear patterns
Use Cases	• When decisions can be broken down into rules (e.g., loan approval, medical diagnoses) • Handles both numerical and categorical data well	• When outcome depends on linear combination of inputs (e.g., spam detection, customer churn) • Best with numerical, binary/class labels

Q14 - Explain what ensemble methods are and how they improve classification performance.

Ensemble Methods

- combine **multiple models** to make a **stronger, more accurate model.**

Bagging (Bootstrap Aggregating)	Boosting	Random Forest
Trains many models on **different random samples** of the training data (with replacement), then **averages or votes** on the predictions.	Trains models **one after another**, where each new model focuses on the **errors** of the previous ones.	An ensemble of **decision trees** using **Bagging + feature randomness**.
Reduces variance.	**Reduces bias.**	Accurate, handles missing data and avoids overfitting better than a single tree.
Example: **Random Forest**	Examples: **AdaBoost, Gradient Boosting, XGBoost**	-

Reference

[1] SECP2753 Data Mining Lecture Slide Module 3a - Classification

[2] jiawei han, micheline kamber, and jian pei, “Data Mining Third Edition,” 2011. Available: https://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

[3] L. Bevin, “What Is Data Classification & Why Is It Important?,” ZenGRC, Dec. 10, 2023. https://www.zengrc.com/blog/what-is-data-classification-why-is-it-important/

[4] GeeksforGeeks, “Basic Concept of Classification (Data Mining) - GeeksforGeeks,” GeeksforGeeks, Apr. 16, 2019. https://www.geeksforgeeks.org/basic-concept-classification-data-mining/

[5] S. Suresh, “OPIT - Open Institute of Technology,” OPIT - Open Institute of Technology, Jul. 2023. https://www.opit.com/magazine/classification-in-data-mining/

[3b] A. K. Singh, “All You Need To Know About Decision Tree Algorithm - NashTech Blog,” NashTech Blog, Aug. 08, 2022. https://blog.nashtechglobal.com/all-you-need-to-know-about-decision-tree-algorithm/ (accessed Apr. 30, 2025).

Reference

[6] “Welcome To Zscaler Directory Authentication,” Medium.com, 2025.
https://medium.com/@abhishekjainindore24/everything-about-k-nearest-neighbour-and-the-mathematics-behind-it-5c774945d77c

[7] IBM, “What is the k-nearest neighbors (KNN) algorithm?,” Ibm.com, Oct. 04, 2021.
https://www.ibm.com/think/topics/knn

[8] P. Kashyap, “Understanding Precision, Recall, and F1 Score Metrics,” Medium, Dec. 02, 2024.
https://medium.com/@piyushkashyap045/understanding-precision-recall-and-f1-score-metrics-ea219b908093

[11.1]scikit learn, “3.1. Cross-validation: Evaluating Estimator Performance — scikit-learn 0.21.3 Documentation,” Scikit-learn.org, 2009. https://scikit-learn.org/stable/modules/cross_validation.html

Thank you