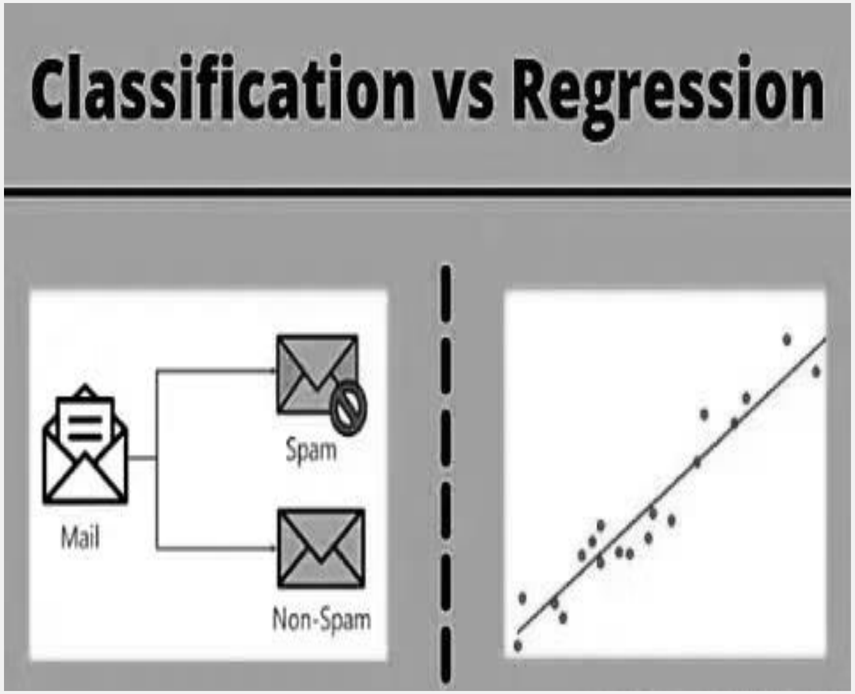
CLASSIFICATION IN MACHINE LEARNING

Definition, Process, and Algorithms

WHAT IS CLASSIFICATION?

- Classification is a supervised learning method where input data is mapped to discrete labels (e.g., spam vs. not spam). Unlike regression, it predicts categories, not continuous values.



CLASSIFICATION PROCESS

1. Define the problem
2. Collect data
3. Preprocess data (cleaning, encoding, scaling)
4. Train-test split
5. Train model on training data
6. Evaluate using accuracy, precision, recall, F1-score, ROC-AUC
7. Deploy for predictions

LINEAR MODELS

1. Logistic Regression

- Uses sigmoid to output probabilities
- Linear decision boundary
- Pros: Simple, good for binary problems
- Cons: Assumes linearity

2. Support Vector Machine

- Finds linear hyperplane
- Pros: Works in high dimensions
- Cons: Cannot capture non-linear data

NON-LINEAR MODELS (PART 1)

1. K-Nearest Neighbors (KNN)
- Classifies based on neighbors
- Pros: Simple, no assumptions
- Cons: Slow with large datasets

2. Naive Bayes
- Based on Bayes' theorem, assumes independence
- Pros: Works well for text classification
- Cons: Independence assumption unrealistic

NON-LINEAR MODELS (PART 2)

3. Decision Tree

- Splits data into branches using feature thresholds

- Pros: Easy to interpret

- Cons: Prone to overfitting

4. Random Forest

- Ensemble of decision trees

- Pros: Reduces overfitting, higher accuracy

- Cons: Less interpretable

SUMMARY TABLE

Algorithm	Type	Pros	Cons
Logistic Regression	Linear	Simple, interpretable	Assumes linearity
SVM	Linear	Works in high dimensions	Fails on non-linear data
KNN	Non-linear	Simple, no training needed	Slow, sensitive to scaling
Naive Bayes	Non-linear	Fast, good for text	Independence assumption
Decision Tree	Non-linear	Easy to interpret	Overfits easily
Random Forest	Non-linear	Accurate, reduces overfitting	Less interpretable