**✅ Batch 1: Core ML Concepts**

1. **What is Machine Learning?**  
   Machine Learning is a subset of AI that enables systems to learn patterns from data and make decisions or predictions without being explicitly programmed.
2. **Difference between Data Mining and Machine Learning?**  
   Data Mining focuses on discovering hidden patterns in large datasets; ML focuses on using data to train algorithms that can make predictions or decisions.
3. **What is Overfitting in ML?**  
   Overfitting occurs when a model learns noise or irrelevant details in the training data, performing well on training but poorly on new, unseen data.
4. **Why does Overfitting happen?**  
   Overfitting happens when the model is too complex, has too many parameters, or is trained for too long on limited or noisy data.
5. **How to avoid Overfitting?**  
   Use simpler models, cross-validation, regularization (L1/L2), pruning (for trees), early stopping, or increase training data.
6. **What is Inductive Machine Learning?**  
   It’s the process where a system infers general rules or patterns from specific training examples and uses them to predict future data.
7. **Five popular ML algorithms?**
   * Linear Regression
   * Decision Trees
   * Random Forest
   * K-Nearest Neighbors (KNN)
   * Support Vector Machines (SVM)
8. **Different Algorithm Techniques in ML?**
   * Supervised Learning
   * Unsupervised Learning
   * Semi-supervised Learning
   * Reinforcement Learning
9. **Three stages to build hypotheses or models?**
   * Model Building
   * Model Evaluation
   * Model Deployment
10. **Standard approach to supervised learning?**  
    Split data into training and testing sets, train a model on labeled data, evaluate using metrics (accuracy, precision, etc.), and tune parameters.

**✅ Batch 2: Training, Testing, Learning Approaches**

**11. What is a Training Set and a Test Set?**

* **Training set** is used to train the model by finding patterns in labeled data.
* **Test set** evaluates how well the model generalizes to unseen data.

**12. List down various approaches for Machine Learning:**

| **Approach** | **Description** |
| --- | --- |
| Supervised Learning | Learn from labeled data (input → output mapping) |
| pUnsupervised Learning | Discover hidden patterns in unlabeled data |
| Semi-Supervised | Mix of labeled and unlabeled data |
| Reinforcement Learning | Learn via feedback from actions in an environment |

**13. What is not Machine Learning?**

If a system is purely rule-based (hard-coded logic, not data-driven), it’s not ML. Traditional programming where logic is manually written falls outside ML.

**14. What is the function of Unsupervised Learning?**

To identify hidden patterns or groupings in unlabeled data, e.g., clustering or dimensionality reduction.

**15. What is the function of Supervised Learning?**

To train a model on input-output pairs so it can learn to predict outputs for new, unseen inputs.

**16. What is Algorithm-Independent Machine Learning?**

It's a learning strategy not tied to a specific algorithm. It emphasizes the process (like bias/variance trade-off) rather than the model type.

**17. Difference between Artificial Learning and Machine Learning?**

Artificial Learning isn't a standard term. If used, it may refer to rule-based AI systems. Machine Learning is data-driven and adaptive.

**18. What is a Classifier in ML?**

A classifier is a model that categorizes data into classes based on training data. E.g., spam vs non-spam email.

**19. Advantages of Naive Bayes:**

* Simple and fast to train
* Works well with high-dimensional data
* Performs well on text classification problems (e.g., spam detection)

**20. What is Inductive Logic Programming (ILP)?**

ILP combines logic programming with ML. It learns logical rules from observed examples and background knowledge, often used in bioinformatics and NLP.

**✅ Batch 3: Model Selection, SVM, Ensemble, Calibration**

**21. What is Model Selection in ML?**

Model selection is the process of choosing the best-performing model from a set of candidates based on evaluation metrics and cross-validation.

**22. Two methods used for calibration in Supervised Learning:**

| **Calibration Method** | **Purpose** |
| --- | --- |
| Platt Scaling | Logistic regression on model outputs |
| Isotonic Regression | Non-parametric calibration method |

**23. Which method is frequently used to prevent overfitting?**

**Cross-validation** (especially k-fold) is widely used. Others include **regularization** (L1/L2), **early stopping**, and **dropout** in neural nets.

**24. Why is instance-based learning called Lazy Learning?**

Because it delays generalization until a query is made. It stores training data and compares each test point to find predictions (e.g., KNN).

**25. Two classification methods SVM can handle:**

| **Type** | **Description** |
| --- | --- |
| Binary Classification | Separates two classes using a hyperplane |
| Multiclass Classification | Uses one-vs-rest or one-vs-one strategies |

**26. What is Ensemble Learning?**

It combines predictions from multiple models (weak learners) to produce a more accurate and robust final prediction.

**27. When to use Ensemble Learning?**

When single models underperform or have high variance/bias. It improves accuracy, generalization, and reduces overfitting.

**28. Two paradigms of Ensemble Methods:**

| **Paradigm** | **Description** |
| --- | --- |
| Bagging | Trains multiple models in parallel on bootstrapped samples |
| Boosting | Trains models sequentially, focusing on previous errors |

**29. General principle of Ensemble Method & Bagging vs Boosting:**

* **General principle:** Combine multiple models to improve predictive performance.
* **Bagging:** Reduces variance by averaging.
* **Boosting:** Reduces bias by focusing on hard-to-predict samples.

**30. Bias-Variance Decomposition in Ensemble Methods:**

* **Bias:** Error due to wrong assumptions (underfitting).
* **Variance:** Error from model sensitivity (overfitting).
* Ensemble methods aim to balance both to reduce total error.

**✅ Batch 4: Incremental Learning, PCA, SVM, Learning Types**

**31. What is an Incremental Learning Algorithm in Ensemble?**

It updates the model as new data arrives without retraining from scratch. Used in streaming data and online learning scenarios (e.g., SGDClassifier in scikit-learn).

**32. What is PCA, KPCA, and ICA used for?**

| **Method** | **Full Form** | **Use Case** |
| --- | --- | --- |
| PCA | Principal Component Analysis | Reduce dimensionality by finding new axes |
| KPCA | Kernel PCA | Non-linear version of PCA using kernels |
| ICA | Independent Component Analysis | Separate mixed signals into independent sources |

**33. What is Dimensionality Reduction in ML?**

It’s the process of reducing the number of input features while preserving important information. It improves model performance and interpretability.

**34. What are Support Vector Machines (SVM)?**

SVMs are supervised ML models that find the best hyperplane to separate classes by maximizing the margin between them.

**35. Differentiate Inductive vs Deductive Learning:**

|  |  |
| --- | --- |
| Type | Description |
| Inductive | Learns from specific data to general rules (data → model) |
| Deductive | Applies general rules to predict outcomes (rules → data) |

**36. Difference between Data Mining and Machine Learning:**

* **Data Mining:** Focuses on discovering patterns from large datasets.
* **Machine Learning:** Focuses on learning from data to make predictions or decisions.

**37. Supervised vs Unsupervised ML:**

|  |  |  |
| --- | --- | --- |
| Feature | Supervised Learning | Unsupervised Learning |
| Labeled Data | Required | Not required |
| Goal | Predict outputs | Find patterns/groups |
| Examples | Regression, Classification | Clustering, Association |

**38. Machine Learning vs Deep Learning:**

|  |  |  |
| --- | --- | --- |
| Feature | Machine Learning | Deep Learning |
| Feature Engg. | Often required manually | Automatic via neural nets |
| Data Needs | Small to medium | Large datasets |
| Algorithms | SVM, RF, KNN, etc. | CNN, RNN, Transformers |

**39. How is KNN different from K-means?**

|  |  |  |
| --- | --- | --- |
| Feature | KNN (K-Nearest Neighbors) | K-Means Clustering |
| Type | Supervised Learning | Unsupervised Learning |
| Use Case | Classification/Regression | Clustering |
| How it works | Based on nearest labeled points | Based on centroid distance |

**40. Different types of Algorithm methods in ML:**

* **Supervised:** Regression, Classification
* **Unsupervised:** Clustering, Dimensionality Reduction
* **Semi-supervised:** Mix of labeled/unlabeled
* **Reinforcement Learning:** Learn via feedback from environment

**✅ Batch 5: RL, Bias-Variance, Classification, Regression, Missing Data**

**41. What do you understand by Reinforcement Learning?**

It’s a type of ML where agents learn to take actions in an environment to maximize cumulative rewards. Think of it like training a dog—reward the good, discourage the bad. 🐶🎯

**42. What is the trade-off between Bias and Variance?**

|  |  |  |
| --- | --- | --- |
| Term | Meaning | Effect on Model |
| Bias | Error from simplifying assumptions | Underfitting |
| Variance | Error from sensitivity to training data | Overfitting |

➡️ **Trade-off:** We want low bias *and* low variance, but minimizing one often increases the other.

**43. How do classification and regression differ?**

|  |  |  |
| --- | --- | --- |
| Feature | Classification | Regression |
| Output | Discrete categories | Continuous values |
| Example | Spam detection | Predicting house prices |

**44. Three stages of building hypotheses/model in ML:**

1. **Model Selection** – Choose algorithm (e.g. SVM, RF)
2. **Training** – Fit model to training data
3. **Evaluation** – Validate model performance on test data

**45. Describe ‘Training Set’ and ‘Test Set’:**

|  |  |
| --- | --- |
| Set Type | Purpose |
| Training | Used to train the model |
| Test | Used to evaluate how well the model generalizes |

**46. Common ways to handle missing data:**

|  |  |
| --- | --- |
| Method | Use Case |
| Remove rows/columns | When missingness is high |
| Imputation (mean/median) | For numerical features |
| Mode imputation | For categorical features |
| Predictive modeling | Use ML to impute |

**47. Necessary steps in an ML project:**

1. Problem Definition
2. Data Collection
3. Data Cleaning & Preprocessing
4. Feature Engineering
5. Model Building
6. Model Evaluation
7. Deployment & Monitoring

**48. Describe Precision and Recall:**

|  |  |
| --- | --- |
| Metric | Meaning |
| Precision | True Positives / (True Positives + False Positives) |
| Recall | True Positives / (True Positives + False Negatives) |

➡️ Precision = “Of what I predicted positive, how many were correct?”  
➡️ Recall = “Of all actual positives, how many did I catch?”

**49. What do you understand by Decision Tree in ML?**

It’s a flowchart-like tree structure where internal nodes represent features, branches represent decisions, and leaves represent outcomes.

**50. What do you understand by Algorithm Independent ML?**

It refers to techniques like **feature selection, resampling, or data transformation** that can be applied across different algorithms without being tied to any specific one.

**✅ Batch 6: Classifiers, Confusion Matrix, Accuracy vs Performance, Ensemble Methods**

**51. Describe the classifier in machine learning.**

A **classifier** is an algorithm that maps input data to a **specific category or label**. Examples: Logistic Regression, SVM, Decision Trees, KNN.

**52. What is SVM in machine learning? What classification methods can it handle?**

* **SVM (Support Vector Machine):** A powerful supervised learning model that finds the optimal hyperplane to separate classes.
* **It can handle:**
  + **Binary Classification** – e.g., Spam vs Not Spam
  + **Multiclass Classification** – via One-vs-Rest or One-vs-One strategies

**53. What do you understand by the Confusion Matrix?**

|  |  |  |
| --- | --- | --- |
| Actual \ Predicted | Positive (P) | Negative (N) |
| Positive | TP | FN |
| Negative | FP | TN |

It’s a table used to evaluate classification models by comparing actual vs predicted values.

**54. Explain TP, TN, FP, FN with example:**

|  |  |  |
| --- | --- | --- |
| Term | Meaning | Example – Email Spam |
| TP | Predicted Spam, Actually Spam | Correctly caught spam ✉️🚫 |
| TN | Predicted Not Spam, Actually Not Spam | Correctly allowed a legit email |
| FP | Predicted Spam, Actually Not Spam | Good email wrongly filtered |
| FN | Predicted Not Spam, Actually Spam | Spam email sneaked through 😤 |

**55. What’s more important: Model Accuracy or Model Performance?**

* **It depends.** Accuracy alone is misleading, especially in **imbalanced datasets**.
* You should look at:
  + **Precision/Recall**
  + **F1 Score**
  + **AUC-ROC**

📌 *Performance > Accuracy when the cost of errors matters (medical diagnosis, fraud detection, etc.)*

**56. What is Bagging and Boosting?**

|  |  |  |
| --- | --- | --- |
| Feature | Bagging | Boosting |
| Goal | Reduce variance | Reduce bias and variance |
| Example | Random Forest | AdaBoost, Gradient Boosting |
| Approach | Train models independently | Train models sequentially |

**57. Similarities and Differences between Bagging and Boosting:**

* ✅ **Similarities:**
  + Both are ensemble techniques
  + Use multiple models to improve performance
* ❌ **Differences:**
  + Bagging reduces **variance**, Boosting reduces **bias**
  + Bagging uses **parallel** models, Boosting is **sequential**

**58. What do you understand by Cluster Sampling?**

A **probability sampling** method where the population is divided into groups (clusters), and some clusters are randomly selected for analysis.

📦 Example: Pick random schools (clusters), then survey all students in selected schools.

**59. What is the F1 Score?**

It’s the **harmonic mean** of Precision and Recall.

F1=2⋅Precision⋅RecallPrecision+RecallF1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}F1=2⋅Precision+RecallPrecision⋅Recall​

📌 Use F1 Score when you need a balance between Precision and Recall, especially with **imbalanced datasets**.

**60. How is a Decision Tree pruned?**

|  |  |
| --- | --- |
| Method | Description |
| Pre-Pruning | Stop tree growth early (e.g., max depth) |
| Post-Pruning | Grow full tree first, then remove weak branches |

🎯 *Pruning prevents overfitting by simplifying the tree.*

**✅ Batch 7: Rec Systems, Regularization, Encoding, ML in Real Life**

**61. What are the Recommended Systems?**

Systems designed to predict users' preferences and suggest items, e.g., Netflix recommending movies or Amazon suggesting products. Types:

|  |  |  |
| --- | --- | --- |
| Type | How It Works | Example |
| Collaborative | Based on similar users' behavior | “People like you bought…” |
| Content-Based | Based on item features user likes | “Because you watched X…” |
| Hybrid | Combination of both | Netflix’s hybrid approach |

**62. When does regularization become necessary in Machine Learning?**

When the model **overfits** the training data—regularization adds a penalty to model complexity to keep it simple and generalizable.

**63. What is Regularization? What problems does it solve?**

Adds a penalty term (like L1 or L2 norm) to the loss function, discouraging overly complex models. It solves **overfitting** by:

* Shrinking coefficients (L2: Ridge)
* Performing feature selection (L1: Lasso)

**64. Why convert categorical variables into factors? Which functions do this?**

ML algorithms expect numeric inputs, so categorical variables are encoded as factors to enable modeling.

* In **Python (pandas)**: pd.get\_dummies(), LabelEncoder
* In **R**: as.factor()

**65. Would treating categorical variables as continuous improve model?**

Nope, it misleads the model since categories don’t have a natural order or numeric distance. Encoding preserves meaning without implying false relationships.

**66. How is Machine Learning used in day-to-day life?**

From voice assistants like Siri, spam filters, personalized ads, fraud detection, to smart recommendations — ML powers tons of seamless experiences we barely notice.

**67. How do you handle missing or corrupted data?**

* Remove or drop missing data (if minor)
* Impute (mean, median, mode, or model-based)
* Use algorithms that handle missingness (like XGBoost)

**68. How to choose a classifier based on training set size?**

|  |  |
| --- | --- |
| Training Size | Recommended Classifiers |
| Small (<1000 samples) | Simpler models (Logistic Regression, Naive Bayes) |
| Medium (1000–100k) | SVM, Random Forest |
| Large (>100k) | Deep Learning, Gradient Boosting |

**69. Applications of Supervised Machine Learning in Business:**

* Fraud detection in banking
* Customer churn prediction
* Demand forecasting
* Email spam filtering
* Credit scoring

**70. What is Semi-supervised Machine Learning?**

A hybrid approach where the model learns from a small amount of labeled data combined with a large amount of unlabeled data — useful when labeling is costly.

**✅ Batch 8: KNN, Naive Bayes, Recommendation Engines, Amazon**

**71. Compare K-means and KNN Algorithms**

|  |  |  |
| --- | --- | --- |
| Aspect | K-means | KNN (K-Nearest Neighbors) |
| Type | Unsupervised clustering | Supervised classification |
| Purpose | Group data into k clusters | Classify new points based on neighbors |
| Output | Cluster centroids and membership | Predicted class label |
| Distance metric | Euclidean (usually) | Distance-based (Euclidean, Manhattan, etc.) |
| Training | Yes (iterative centroid update) | No explicit training, just stores data |

**72. What is ‘naive’ in Naive Bayes Classifier?**

The "naive" assumption is **feature independence** — it assumes all features independently contribute to the outcome, which is rarely true, but surprisingly effective.

**73. How will you know which ML algorithm to choose for classification?**

* Understand data size, dimensionality, and type
* Check interpretability needs
* Benchmark models with cross-validation
* Use domain knowledge and computational budget

**74. How is Amazon able to recommend other things to buy? How does the recommendation engine work?**

* **Collaborative Filtering:** Uses purchase/browsing history of similar users
* **Content-Based Filtering:** Uses product attributes
* **Hybrid Models:** Combine above with deep learning & session-based data
* They leverage massive data and real-time user behavior!

**75. When will you use classification over regression?**

Use **classification** when the target variable is categorical (e.g., spam or not), and **regression** when the target is continuous (e.g., house price prediction).

**76. How do you design an email spam filter?**

* Collect labeled emails (spam/not spam)
* Extract features (words, sender, links)
* Train a classifier (Naive Bayes, SVM)
* Test and optimize for recall/precision
* Deploy to filter incoming mail in real-time

**77. What is a Random Forest?**

An ensemble of decision trees built on different random subsets of data and features; it aggregates predictions by majority vote (classification) or averaging (regression).

**78. What is Pruning in Decision Trees, and how is it done?**

Pruning cuts back overgrown branches to prevent overfitting, usually by setting constraints (max depth) or removing branches that don’t improve validation error.

**79. Briefly explain Logistic Regression.**

A supervised classification algorithm that models the probability of a binary outcome using the logistic (sigmoid) function, mapping any real-valued number to a value between 0 and 1.

**80. Explain the K Nearest Neighbor Algorithm.**

Classifies a data point based on the majority class among its *k* closest neighbors using a distance metric (e.g., Euclidean). It’s simple, lazy learning (no training phase).

**✅ Batch 9: Kernel SVM, Dimensionality Reduction, Errors, Correlation**

**81. What is Kernel SVM?**

Kernel SVM uses a kernel function to transform data into a higher-dimensional space where it becomes linearly separable—think of it as magic carpet ride for non-linear data (popular kernels: RBF, polynomial).

**82. What are some methods of reducing dimensionality?**

* **PCA (Principal Component Analysis)**
* **KPCA (Kernel PCA)**
* **ICA (Independent Component Analysis)**
* **t-SNE, UMAP (non-linear methods)**

Purpose: reduce features while preserving structure, combatting the “curse of dimensionality.”

**83. What is Principal Component Analysis (PCA)?**

A linear dimensionality reduction method that projects data onto directions (principal components) that maximize variance, simplifying data while retaining most info.

**84. What do you understand by Type I vs Type II error?**

|  |  |  |
| --- | --- | --- |
| Error Type | Description | Consequence |
| Type I (False Positive) | Rejecting true null hypothesis | Detecting an effect that isn’t there |
| Type II (False Negative) | Failing to reject false null hypothesis | Missing a true effect |

**85. Explain Correlation and Covariance?**

|  |  |  |
| --- | --- | --- |
| Measure | Meaning | Scale |
| Covariance | Directional relationship between two variables | Unbounded (depends on units) |
| Correlation | Strength and direction of linear relationship (standardized covariance) | Between -1 and 1 |

**86. What are Support Vectors in SVM?**

Data points closest to the decision boundary that “support” or define the margin. They’re crucial because the model depends only on these points, not the entire dataset.

**87. What is Cross-Validation?**

A method to evaluate model performance by splitting data into training/testing sets multiple times (e.g., k-fold), ensuring robustness and reducing overfitting risk.

**88. What are the different methods to split a tree in a decision tree algorithm?**

* **Gini Impurity** (measures impurity)
* **Entropy / Information Gain** (measures uncertainty reduction)
* **Variance Reduction** (for regression trees)

**89. How does the Support Vector Machine algorithm handle self-learning?**

SVM is not inherently self-learning; it’s a supervised learner that optimizes a margin given labeled data. Self-learning would need integration with active learning or semi-supervised approaches.

**90. What are the assumptions you need to take before starting with linear regression?**

* Linear relationship between predictors and outcome
* Homoscedasticity (constant variance of errors)
* Independence of errors
* Normal distribution of errors
* No or little multicollinearity among predictors

**✅ Batch 10: Lasso & Ridge, Entropy, Epoch, Classification vs Regression**

**91. What is the difference between Lasso and Ridge regression?**

|  |  |  |
| --- | --- | --- |
| Feature | Lasso Regression (L1) | Ridge Regression (L2) |
| Penalty | Sum of absolute coefficients | Sum of squared coefficients |
| Effect on coefficients | Can shrink some coefficients exactly to zero (feature selection) | Shrinks coefficients but none to zero |
| Use case | When you want feature selection | When you want to keep all features with smaller weights |

**92. What is Entropy in Machine Learning?**

Entropy measures the **uncertainty or impurity** in a dataset; in decision trees, lower entropy means a cleaner split. Calculated as:

Entropy=−∑pilog⁡2piEntropy = -\sum p\_i \log\_2 p\_iEntropy=−∑pi​log2​pi​

where pip\_ipi​ is the probability of class iii.

**93. What is Epoch in Machine Learning?**

One complete pass through the entire training dataset during model training. Multiple epochs usually improve learning but too many can lead to overfitting.

**94. Differentiate between Classification and Regression in Machine Learning**

|  |  |  |
| --- | --- | --- |
| Aspect | Classification | Regression |
| Output type | Categorical labels | Continuous values |
| Goal | Assign input to classes | Predict a numerical value |
| Algorithms | Logistic Regression, SVM, KNN | Linear Regression, SVR, Decision Tree Regression |

**95. How is the suitability of a Machine Learning Algorithm determined for a problem?**

* Nature of problem (classification vs regression)
* Dataset size and dimensionality
* Interpretability needs
* Computation time and resources
* Accuracy and performance on validation data

**96. What is ROC Curve and what does it represent?**

Receiver Operating Characteristic curve plots True Positive Rate vs False Positive Rate at various thresholds. It visualizes model’s classification performance and trade-offs.

**97. Both being tree-based algorithms, how is Random Forest different from Gradient Boosting Machine (GBM)?**

|  |  |  |
| --- | --- | --- |
| Aspect | Random Forest | Gradient Boosting Machine (GBM) |
| Training | Parallel trees trained independently | Sequential trees, each correcting previous errors |
| Performance | Robust, less prone to overfitting | Higher accuracy but more prone to overfitting |
| Speed | Faster due to parallelism | Slower, sequential process |

**98. What do you understand about the P-value?**

Probability of observing data as extreme as you did, assuming the null hypothesis is true. A low p-value (< 0.05) suggests evidence against the null hypothesis.

**99. Suppose your model has high variance. Which algorithm can handle this and why?**

**Random Forest** can reduce variance by averaging many uncorrelated trees, improving generalization and reducing overfitting.

**100. What is Rescaling of Data and how is it done?**

Rescaling (normalization or standardization) transforms features to a common scale (like 0-1 or mean=0, std=1) to help algorithms converge faster and perform better.

**✅ Batch 11: Data Handling, Classifiers, Semi-supervised, K-means vs KNN**

**101. How Do You Handle Missing or Corrupted Data in a Dataset?**

* **Remove rows/columns** with missing values (if small portion)
* **Imputation**: fill with mean/median/mode or use algorithms (KNN imputer, regression)
* **Predictive modeling** to estimate missing values
* **Flag missingness** as a feature

**102. How Can You Choose a Classifier Based on a Training Set Data Size?**

|  |  |  |
| --- | --- | --- |
| Data Size | Preferred Classifiers | Reason |
| Small | Naive Bayes, Logistic Regression | Simple models with less overfitting risk |
| Medium | SVM, Random Forest | Balance complexity & generalization |
| Large | Deep Learning, Gradient Boosting | Handle high dimensionality & complexity |

**103. What Are the Applications of Supervised Machine Learning in Modern Businesses?**

* Fraud detection
* Customer churn prediction
* Email spam filtering
* Sales forecasting
* Credit scoring

**104. What is Semi-supervised Machine Learning?**

A hybrid approach that learns from a small amount of labeled data plus a large amount of unlabeled data — the best of both worlds when labels are scarce or costly.

**105. Compare K-means and KNN Algorithms.**

|  |  |  |
| --- | --- | --- |
| Aspect | K-means | KNN (K-Nearest Neighbors) |
| Type | Unsupervised learning (clustering) | Supervised learning (classification/regression) |
| Goal | Group data into k clusters | Classify or predict based on neighbors |
| Training | No explicit training, iterative cluster assignment | Stores entire training dataset (lazy learner) |
| Output | Cluster centers | Class label or regression value |

**✅ Batch 12: Naive Bayes, Algorithm Choice, Recommendations, Spam Filters**

**106. What Is ‘naive’ in the Naive Bayes Classifier?**

The “naive” assumption is that all features are **conditionally independent** given the class label — a simplification that makes computation easy, even if it’s not always true.

**107. How Will You Know Which Machine Learning Algorithm to Choose for Your Classification Problem?**

* Data size & dimensionality
* Feature types & distributions
* Model interpretability needs
* Computational resources
* Accuracy and speed trade-offs
* Cross-validation performance

**108. How is Amazon Able to Recommend Other Things to Buy? How Does the Recommendation Engine Work?**

By leveraging **collaborative filtering** (users with similar tastes) and **content-based filtering** (similar item features), plus hybrid models and deep learning for personalization.

**109. When Will You Use Classification over Regression?**

* When the output is **categorical** (e.g., spam/not spam, disease/no disease)
* When you want to assign discrete labels instead of predicting continuous values

**110. How Do You Design an Email Spam Filter?**

* Collect labeled spam/non-spam emails
* Extract features (keywords, sender info, metadata)
* Train classifiers like Naive Bayes or SVM
* Evaluate with metrics (precision, recall, F1-score)
* Deploy with ongoing retraining for evolving spam tactics

**✅ Batch 13: Random Forest, Pruning, Logistic Regression, KNN, Kernel SVM**

**111. What is a Random Forest?**

An ensemble of decision trees trained on random subsets of data and features, whose predictions are aggregated (voted or averaged) to improve accuracy and reduce overfitting.

**112. What is Pruning in Decision Trees, and How Is It Done?**

Pruning removes branches that add little predictive power, preventing overfitting. Done by:

* **Pre-pruning**: stop tree growth early based on criteria (max depth, min samples)
* **Post-pruning**: grow full tree, then remove branches based on validation performance

**113. Briefly Explain Logistic Regression.**

A classification algorithm that models the probability of a binary outcome using the logistic (sigmoid) function applied to a linear combination of input features.

**114. Explain the K Nearest Neighbor Algorithm.**

A lazy learner that classifies a new sample by majority voting among the k closest samples in the training data based on distance metrics (e.g., Euclidean).

**115. What is Kernel SVM?**

An SVM that uses a kernel function (e.g., RBF, polynomial) to transform data into higher-dimensional space, enabling separation of data that is not linearly separable.