# Fundamentals of Machine Learning (4341603) - Winter 2023 Solution

Milav Dabgar

February 2, 2024

## Question 1(a) [3 marks]

**Define human learning and explain how machine learning is different from human learning?**

**Human Learning** is the process of acquiring knowledge through experience, observation, and reasoning.

**Table 1.** Human Learning vs Machine Learning

| Aspect | Human Learning | Machine Learning |
|---|---|---|
| **Method** | Experience, trial and error | Data and algorithms |
| **Speed** | Slow, gradual | Fast processing |
| **Data Requirement** | Limited examples needed | Large datasets required |

**Machine Learning** is the automated learning from data using algorithms to identify patterns without explicit programming.

**Mnemonic**

"Humans Experience, Machines Analyze Data (HEMAD)"

## Question 1(b) [4 marks]

**Describe the use of machine learning in finance and banking.**

**Solution**

**Table 2.** Applications in Finance and Banking

| Application | Purpose | Benefit |
|---|---|---|
| **Fraud Detection** | Identify suspicious transactions | Reduce financial losses |
| **Credit Scoring** | Assess loan default risk | Better lending decisions |
| **Algorithmic Trading** | Automated trading decisions | Faster market responses |

**Risk Assessment**: ML analyzes customer data to predict creditworthiness. **Customer Service**: Chatbots provide 24/7 support using NLP. **Regulatory Compliance**: Automated monitoring for suspicious activities.

**Mnemonic**

"Finance Needs Smart Analysis (FNSA)"

# Question 1(c) [7 marks]

**Give difference between Supervised Learning, Unsupervised Learning and Reinforcement Learning.**

---

**Solution**

**Table 3.** Comparison

| Feature | Supervised | Unsupervised | Reinforcement |
|---------|-----------|--------------|---------------|
| **Data Type** | Labeled data | Unlabeled data | Environment interaction |
| **Goal** | Predict output | Find patterns | Maximize rewards |
| **Examples** | Classification | Clustering | Game playing |
| **Feedback** | Immediate | None | Delayed rewards |

**Supervised Learning**: Teacher-guided learning with correct answers provided.
**Unsupervised Learning**: Self-discovery of hidden patterns in data.
**Reinforcement Learning**: Learning through trial and error with rewards/penalties.

---

**Mnemonic**

"Supervised Teachers, Unsupervised Explores, Reinforcement Rewards (STUER)"

---

# Question 1(c) OR [7 marks]

**Explain different tools and technology used in machine learning.**

---

**Solution**

**Table 4.** ML Tools and Technologies

| Category | Tools | Purpose |
|----------|-------|---------|
| **Programming** | Python, R, Java | Algorithm implementation |
| **Libraries** | Scikit-learn, TensorFlow | Ready-made algorithms |
| **Visualization** | Matplotlib, Seaborn | Data visualization |
| **Data Processing** | Pandas, NumPy | Data manipulation |

**Key Technologies:**
- **Cloud Platforms**: AWS, Google Cloud for scalable computing
- **Development Environments**: Jupyter Notebook, Google Colab
- **Big Data Tools**: Spark, Hadoop for large datasets

---

**Mnemonic**

"Python Libraries Visualize Data Effectively (PLVDE)"

---

# Question 2(a) [3 marks]

**Define outliers with one example.**

---

---

> **Solution**
>
> **Definition**: Outliers are data points that significantly differ from other observations in a dataset.
>
> **Table 5.** Example: Student Heights
>
> | Student Heights (cm) | Classification |
> |---|---|
> | 165, 170, 168, 172 | Normal values |
> | 195 | Outlier (too tall) |
> | 140 | Outlier (too short) |
>
> **Detection**: Values beyond $1.5 \times$ IQR from quartiles. **Impact**: Can skew statistical analysis and model performance.

> **Mnemonic**
>
> "Outliers Stand Apart (OSA)"

## Question 2(b) [4 marks]

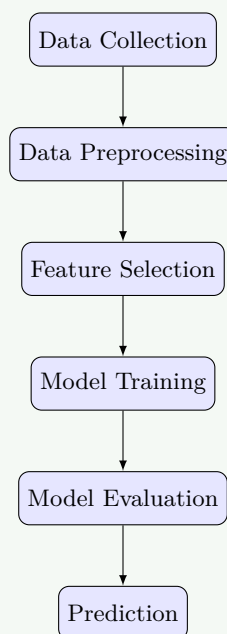**Explain regression steps in detail.**

> **Solution**
>
> 
>
> **Figure 1.** Regression Process Steps
>
> **Detailed Steps:**
> - **Data Collection**: Gather relevant dataset with input-output pairs
> - **Preprocessing**: Clean data, handle missing values, normalize features
> - **Feature Selection**: Choose relevant variables that affect target
> - **Model Training**: Fit regression line to minimize prediction errors

> **Mnemonic**
>
> "Data Preprocessing Features Train Evaluation Predicts (DPFTEP)"

---

# Question 2(c) [7 marks]

**Define Accuracy and for the following binary classifier's confusion matrix, find the various measurement parameters like 1. Accuracy 2. Precision.**

### Solution

**Confusion Matrix Analysis:**

**Table 6.** Given Confusion Matrix

|  | **Predicted No** | **Predicted Yes** |
|---|---|---|
| **Actual No** | 10 (TN) | 3 (FP) |
| **Actual Yes** | 2 (FN) | 15 (TP) |

**Calculations:**
- **Accuracy** formula: $(TP + TN)/(TP + TN + FP + FN)$
- Calculation: $(15 + 10)/(15 + 10 + 3 + 2) = 25/30 = 0.8333$
- **Accuracy Result: 83.33%**
- **Precision** formula: $TP/(TP + FP)$
- Calculation: $15/(15 + 3) = 15/18 = 0.8333$
- **Precision Result: 83.33%**

**Definitions:**
- **Accuracy**: Proportion of correct predictions out of total predictions.
- **Precision**: Proportion of true positive predictions out of all positive predictions.

### Mnemonic

"Accuracy Counts All, Precision Picks Positives (ACAPP)"

# Question 2(a) OR [3 marks]

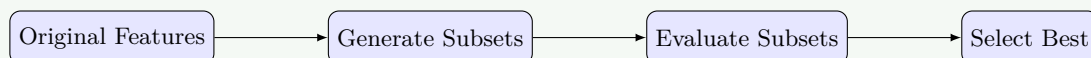**Identify basic steps of feature subset selection.**

### Solution



**Figure 2.** Feature Subset Selection Process

**Basic Steps:**
1. **Generation**: Create different combinations of features
2. **Evaluation**: Test each subset using performance metrics
3. **Selection**: Choose optimal subset based on criteria

### Mnemonic

"Generate, Evaluate, Select (GES)"

# Question 2(b) OR [4 marks]

**Discuss the strength and weakness of the KNN algorithm.**

> **Solution**
>
> **Table 7.** KNN Algorithm Analysis
>
> | Strengths | Weaknesses |
> |---|---|
> | Simple to understand | Computationally expensive |
> | No training required (Lazy) | Sensitive to irrelevant features |
> | Works with non-linear data | Performance degrades with high dimensions |
> | Effective for small datasets | Requires optimal K value selection |
>
> **Key Points:**
> - **Lazy Learning**: No explicit training phase required.
> - **Distance-Based**: Classification based on neighbor proximity.

> **Mnemonic**
>
> "Simple but Slow, Effective but Expensive (SBSEBE)"

## Question 2(c) OR [7 marks]

**Define Error-rate and for the following binary classifier's confusion matrix, find the various measurement parameters like 1. Error value 2. Recall.**

> **Solution**
>
> **Confusion Matrix Analysis:**
>
> **Table 8.** Given Confusion Matrix
>
> | | Predicted No | Predicted Yes |
> |---|---|---|
> | **Actual No** | 20 (TN) | 3 (FP) |
> | **Actual Yes** | 2 (FN) | 15 (TP) |
>
> **Calculations:**
> - **Error Rate** formula: $(FP + FN)/(Total)$
> - Calculation: $(3 + 2)/(15 + 20 + 3 + 2) = 5/40 = 0.125$
> - **Error Rate Result: 12.5%**
> - **Recall** formula: $TP/(TP + FN)$
> - Calculation: $15/(15 + 2) = 15/17 = 0.8824$
> - **Recall Result: 88.24%**
>
> **Definitions:**
> - **Error Rate**: Proportion of incorrect predictions out of total predictions.
> - **Recall**: Proportion of actual positives correctly identified.

> **Mnemonic**
>
> "Error Excludes, Recall Retrieves (EERR)"

## Question 3(a) [3 marks]

**Give any three examples of unsupervised learning.**

---

> **Solution**
>
> **Table 9.** Unsupervised Learning Examples
>
> | Example | Description | Application |
> |---|---|---|
> | **Customer Segmentation** | Group customers by behavior | Marketing strategies |
> | **Document Classification** | Organize documents by topics | Information retrieval |
> | **Gene Sequencing** | Group similar DNA patterns | Medical research |
>
> **Additional Examples:**
> - **Market Basket Analysis**: Finding product purchase patterns
> - **Anomaly Detection**: Detecting unusual patterns in data

> **Mnemonic**
>
> "Customers, Documents, Genes Group Automatically (CDGGA)"

# Question 3(b) [4 marks]

**Find Mean and Median for the following data: 4,6,7,8,9,12,14,15,20**

> **Solution**
>
> **Data**: 4, 6, 7, 8, 9, 12, 14, 15, 20 (Already sorted)
> **Mean Calculation:**
> - Sum $= 4 + 6 + 7 + 8 + 9 + 12 + 14 + 15 + 20 = 95$
> - Count $= 9$
> - **Mean** $= 95/9 = 10.56$
>
> **Median Calculation:**
> - N $= 9$ (Odd number)
> - Position $= (N+1)/2 = $ 5th value
> - 5th value in sorted list is 9
> - **Median** $= 9$

> **Mnemonic**
>
> "Mean Averages All, Median Middle Value (MAAMV)"

# Question 3(c) [7 marks]

**Describe k-fold cross validation method in detail.**
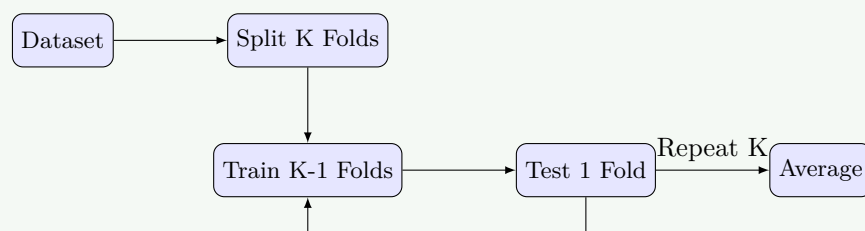
> **Solution**
>
> 
>
> **Figure 3.** K-Fold Cross Validation
>
> **Process Steps:**

---

1. **Data Division**: Split data into K equal parts (folds).
2. **Iterative Training**: Use K-1 folds for training the model.
3. **Validation**: Test the model on the remaining fold.
4. **Averaging**: Repeat K times and average the performance metrics.

**Advantages:**
- **Unbiased Estimation**: Each data point used for both training and testing.
- **Reduced Overfitting**: Multiple validation rounds increase reliability.

**Mnemonic**

"K-fold Keeps Keen Knowledge (KKKK)"

# Question 3(a) OR [3 marks]

**Give any three applications of multiple linear regression.**

**Solution**

**Table 10.** Multiple Linear Regression Applications

| Application | Variables | Purpose |
|---|---|---|
| **House Price Prediction** | Size, location, age | Real estate valuation |
| **Sales Forecasting** | Marketing spend, season | Business planning |
| **Medical Diagnosis** | Symptoms, age, history | Disease prediction |

**Mnemonic**

"Houses, Sales, Medicine Predict Multiple Variables (HSMPV)"

# Question 3(b) OR [4 marks]

**Find Standard Deviation for the following data: 4,15,20,28,35,45**

**Solution**

**Data**: 4, 15, 20, 28, 35, 45 (N=6)
**Step 1: Calculate Mean**
- Sum $= 4 + 15 + 20 + 28 + 35 + 45 = 147$
- Mean $(\bar{x}) = 147/6 = 24.5$

**Step 2: Calculate Squared Deviations**
- $(4 - 24.5)^2 = (-20.5)^2 = 420.25$
- $(15 - 24.5)^2 = (-9.5)^2 = 90.25$
- $(20 - 24.5)^2 = (-4.5)^2 = 20.25$
- $(28 - 24.5)^2 = (3.5)^2 = 12.25$
- $(35 - 24.5)^2 = (10.5)^2 = 110.25$
- $(45 - 24.5)^2 = (20.5)^2 = 420.25$

**Step 3: Calculate Variance and Std Dev**
- Sum of squared deviations $= 1073.5$
- Variance $(\sigma^2) = 1073.5/6 = 178.92$
- **Standard Deviation** $(\sigma) = \sqrt{178.92} = 13.376$

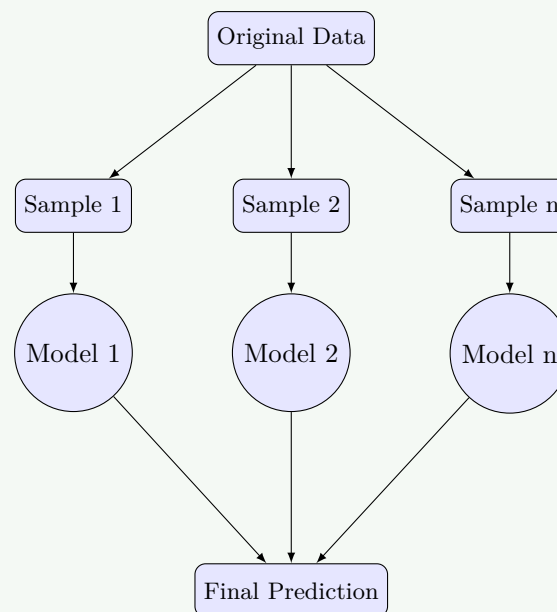# Question 3(c) OR [7 marks]

**Explain Bagging, Boosting in detail.**

**Solution**

**Table 11.** Bagging vs Boosting

| Aspect | Bagging | Boosting |
|---|---|---|
| **Strategy** | Parallel training | Sequential training |
| **Data Sampling** | Random with replacement | Weighted sampling |
| **Goal** | Reduces variance | Reduces bias |

**Bagging (Bootstrap Aggregating):** Trains multiple independent models in parallel using random subsets of data and averages their predictions.



**Figure 4.** Bagging Process

**Boosting:** Trains models sequentially, where each new model focuses on the errors made by previous models.

**Mnemonic**

"Bagging Builds Parallel, Boosting Builds Sequential (BBPBS)"

# Question 4(a) [3 marks]

**Define: Support, Confidence.**

---

**Solution**

**Table 12.** Association Rule Metrics

| Metric | Definition and Formula |
|--------|------------------------|
| **Support** | Frequency of itemset in transactions. |
| | $Support(A) = Count(A)/Total$ |
| **Confidence** | Conditional probability of rule. |
| | $Confidence(A \rightarrow B) = Support(A \cup B)/Support(A)$ |

**Example:**
- **Support**: 60% of transactions have Bread.
- **Confidence**: 80% of people buying Bread also buy Butter.

---

**Mnemonic**

"Support Shows Frequency, Confidence Shows Connection (SSFC)"

---

# Question 4(b) [4 marks]

**Illustrate any two applications of logistic regression.**

---

**Solution**

**Table 13.** Logistic Regression Applications

| Application | Description | Outcome |
|-------------|-------------|---------|
| **Email Spam** | Detect spam based on words | Spam/Not Spam |
| **Medical Diagnosis** | Predict disease from symptoms | Disease/Healthy |
| **Credit Approval** | Assess loan risk | Approve/Reject |

**Key Features:**
- **Binary Classification**: Predicts probability (0 to 1).
- **Sigmoid Function**: Maps output to S-shaped curve.

---

**Mnemonic**

"Logistic Limits Linear Logic (LLLL)"

---

# Question 4(c) [7 marks]

**Discuss the main purpose of Numpy and Pandas in machine learning.**

---

**Solution**

**NumPy** provides support for large, multi-dimensional arrays and matrices, along with mathematical functions.
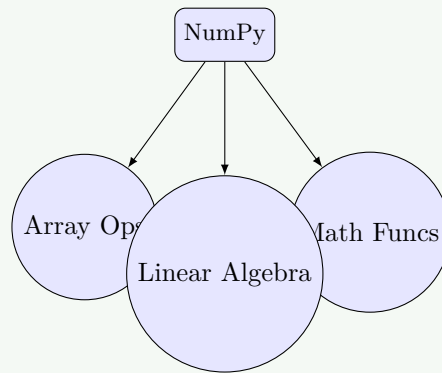**Pandas** offers data structures and operations for manipulating numerical tables and time series.

---

**Figure 5.** NumPy Features

**Table 14.** Comparison

| Library | Primary Purpose | Key Features |
|---------|-----------------|--------------|
| **NumPy** | Numerical Computing | N-dim arrays, broadcasting |
| **Pandas** | Data Manipulation | DataFrames, cleaning, analysis |

**Mnemonic**

"NumPy Numbers, Pandas Processes Data (NNPD)"

# Question 4(a) OR [3 marks]

**Give any three examples of Supervised Learning.**

**Solution**

**Table 15.** Supervised Learning Examples

| Example | Type | Input → Output |
|---------|------|----------------|
| **Email Classification** | Classification | Email features → Spam/Not |
| **House Price Prediction** | Regression | House features → Price |
| **Image Recognition** | Classification | Pixels → Object Class |

**Mnemonic**

"Emails, Houses, Images Learn Supervised (EHILS)"

# Question 4(b) OR [4 marks]

**Explain any two applications of the apriori algorithm.**

**Solution**

**Table 16.** Apriori Applications

| Application | Description |
|---|---|
| **Market Basket Analysis** | Finding products bought together (e.g., Bread & Butter) to optimize store layout. |
| **Web Usage Mining** | Discovering reliable navigation patterns to improve website UX. |

**Process**:
1. Generate frequent itemsets.
2. Prune infrequent items based on support.
3. Generate association rules based on confidence.

**Mnemonic**

"Apriori Analyzes Associations Automatically (AAAA)"

## Question 4(c) OR [7 marks]

**Explain the features and applications of Matplotlib.**

**Solution**

**Matplotlib** is a comprehensive library for creating static, animated, and interactive visualizations in Python.
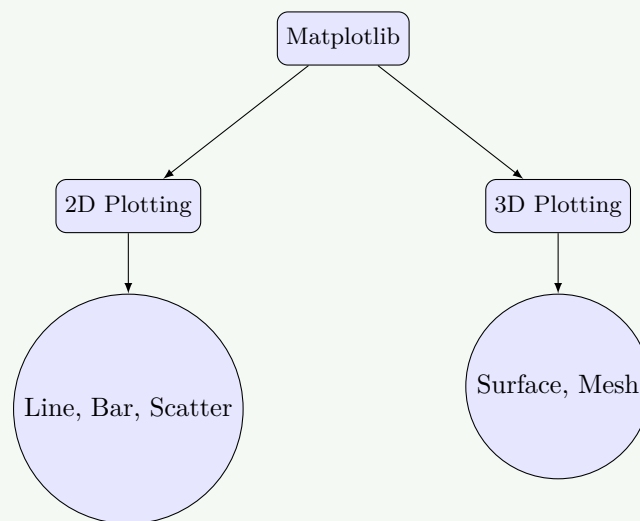


**Figure 6.** Matplotlib Capabilities

**Applications:**
- **Data Exploration**: Histograms, scatter plots to understand data.
- **Model Performance**: Plotting loss curves and accuracy.
- **Result Presentation**: Publication-quality figures.

**Mnemonic**

"Matplotlib Makes Meaningful Visual Displays (MMVD)"

## Question 5(a) [3 marks]

**List out the major features of Numpy.**

---

> **Solution**
>
> **Features of NumPy:**
> - **N-dimensional Arrays**: Fast and efficient multidimensional array object (ndarray).
> - **Broadcasting**: Functions to perform operations on arrays of different shapes.
> - **Linear Algebra**: Built-in support for matrix operations and Fourier transforms.
> - **C/C++ Integration**: Tools for integrating C/C++ and Fortran code.

> **Mnemonic**
>
> "NumPy Numbers Need Neat Operations (NNNNO)"

## Question 5(b) [4 marks]

**How to load an iris dataset csv file in a Pandas Dataframe program? Explain with example.**

> **Solution**
>
> ```python
> import pandas as pd
>
> # Method 1: Load from local CSV file
> df = pd.read_csv('iris.csv')
>
> # Method 2: From sklearn (common in ML)
> from sklearn.datasets import load_iris
> iris = load_iris()
> df_iris = pd.DataFrame(iris.data, columns=iris.feature_names)
>
> # Display first 5 rows
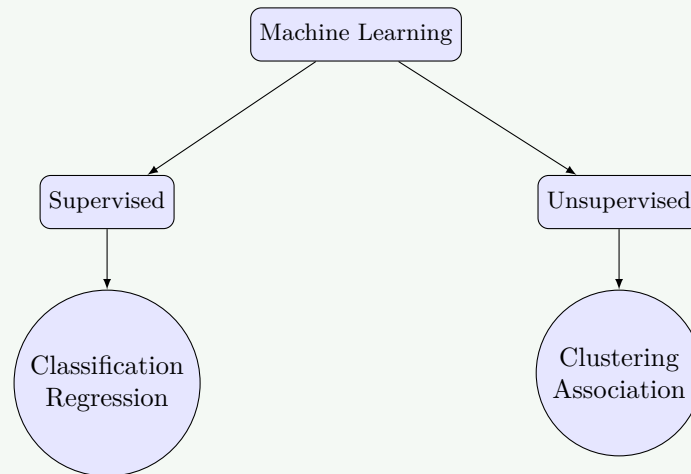> print(df.head())
> ```
>
> **Explanation:**
> - `pd.read_csv()`: Function to read comma-separated values.
> - `df.head()`: Returns the first n rows (default 5).

> **Mnemonic**
>
> "Pandas Reads CSV Files Easily (PRCFE)"

## Question 5(c) [7 marks]

**Compare and Contrast Supervised Learning and Unsupervised Learning.**

---

**Solution**



**Figure 7.** ML Learning Types

**Table 17.** Comparison

| Aspect | Supervised | Unsupervised |
|---|---|---|
| **Data** | Labeled | Unlabeled |
| **Goal** | Predict output | Discover patterns |
| **Feedback** | Direct feedback | No feedback |
| **Complexity** | Validation is easier | Validation is harder |

**Mnemonic**

"Supervised Seeks Specific Solutions, Unsupervised Uncovers Unknown (SSSUU)"

# Question 5(a) OR [3 marks]

**List out the applications of Pandas.**

**Solution**

**Table 18.** Pandas Applications

| Application | Description | Field |
|---|---|---|
| **Data Cleaning** | Handling missing data | General ML |
| **Financial Analysis** | Stock market trends | Finance |
| **Recommendation** | Analyzing user behavior | E-commerce |

**Mnemonic**

"Pandas Processes Data Perfectly (PPDP)"

# Question 5(b) OR [4 marks]

**How to plot a vertical line and horizontal line in matplotlib? Explain with examples.**

---

**Solution**

```python
import matplotlib.pyplot as plt

# Plot a simple line
plt.plot([1, 2, 3], [1, 4, 9])

# Vertical line at x = 2 (Red dashed)
plt.axvline(x=2, color='red', linestyle='--')

# Horizontal line at y = 4 (Green solid)
plt.axhline(y=4, color='green', linestyle='-')

plt.show()
```

**Functions:**

- `axvline(x)`: Adds a vertical line across the axes.
- `axhline(y)`: Adds a horizontal line across the axes.

# Question 5(c) OR [7 marks]

**Describe the concept of clustering using appropriate real-world examples.**

**Solution**

**Clustering** is an unsupervised learning technique that groups similar data points such that points in the same group are more similar to each other than to those in other groups.

**Table 19.** Clustering Applications

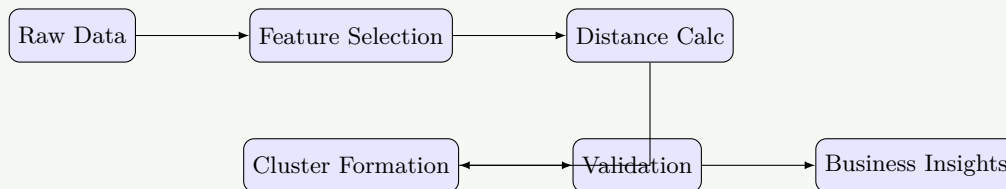| Type | Example | Impact |
|------|---------|--------|
| **Customer Seg.** | Group by purchase behavior | Targeted marketing |
| **Image Seg.** | Tumor detection in MRI | Improved diagnosis |
| **Gene Analysis** | Group genes by expression | Drug discovery |



**Figure 8.** Clustering Process

**Real-World Examples:**

1. **Customer Segmentation**: Identifying high-value customers vs. seasonal shoppers.
2. **Social Media Analysis**: Grouping users by interests (e.g., sports, tech).
3. **Market Research**: Finding segments with similar product needs.

**Mnemonic**

"Clustering Creates Clear Categories (CCCC)"