Clustering and Naive Bayes

Assignment 1: Basic - Engineer Skill Clustering

Context: One Hour AI Solution needs to group their AI engineers based on their skills to better understand their team's composition. Apply K-means clustering to identify natural groupings.

Dataset: engineer_skills.csv

Tasks:

1. Load and explore the dataset
2. Implement K-means clustering to identify natural groupings of engineers
3. Use the elbow method to determine the optimal number of clusters
4. Visualize the clusters using at least 2 features
5. Describe the characteristics of each identified cluster

Requirements:

- Implement K-means clustering (without using scikit-learn's implementation)
- Create at least one visualization of the clusters
- Include comments explaining your approach

Assignment 2: Intermediate - Session Request Classification

Context: One Hour AI Solution needs to automatically categorize client session requests into different topics. Implement a Naive Bayes classifier to categorize these requests.

Dataset: session_requests.csv

Tasks:

1. Preprocess the text data (lowercase, remove punctuation, tokenize)
2. Split the dataset into training (70%) and testing (30%) sets
3. Implement Multinomial Naive Bayes from scratch
4. Train the model on the training data
5. Evaluate the model on the test data using accuracy, precision, and recall

Requirements:

- Implement your own Naive Bayes algorithm (don't use scikit-learn's implementation)
- Create a confusion matrix visualization
- Include detailed comments explaining the Naive Bayes probability calculations

Assignment 3: Advanced - Combining Clustering and Classification

Context: One Hour AI Solution wants to both understand their engineers' expertise clusters AND classify incoming requests. Implement both techniques and show how they can work together.

Datasets: Use both datasets from the previous assignments:

1. engineer_skills.csv - Engineer skill ratings
2. session_requests.csv - Client session requests

Tasks:

1. Engineer Clustering Component:
    - Implement DBSCAN clustering to identify groups of engineers with similar skills
    - Compare the results with K-means clustering from Assignment 1
    - Visualize the differences between the clustering approaches
2. Request Classification Component:
    - Implement Naive Bayes to classify session requests
    - Create a visualization showing the distribution of topics
    - Analyze which words are most predictive for each topic
3. Integration Analysis:
    - For each cluster of engineers, identify which request topics they would be best suited to handle
    - Create a visualization showing the relationship between engineer clusters and request topics
    - Explain how One Hour AI Solution could use this integrated approach

Requirements:

- Implement both DBSCAN and Naive Bayes algorithms from scratch
- Create visualizations for both clustering and classification results
- Include detailed comments explaining both algorithms