Day 15 Training Report

11 July 2025

Clustering Mini Project + Presentation

On **Day 15**, students consolidated their learning by completing a **full unsupervised learning mini project**. This session emphasized **practical application, critical thinking, and presentation skills**.

---

1. Project Objectives

- Conduct an **end-to-end clustering analysis** on a chosen dataset (e.g., customer segmentation, e-commerce data, or sales data).
- Implement **data cleaning, preprocessing, clustering, evaluation, and visualization**.
- Present **findings and insights** in a structured manner.

---

2. End-to-End Workflow

Step 1: Data Loading & Exploration

- Loaded the dataset and inspected the columns.
- Performed **summary statistics** and identified key features for clustering.
- Detected missing values, duplicates, or outliers.

Step 2: Data Preprocessing

- Scaled features using **StandardScaler** or **MinMaxScaler** for better clustering performance.
- Converted categorical variables (if any) to numeric using **encoding techniques**.

Step 3: Choosing the Number of Clusters

- Used **Elbow Method** to visualize WCSS vs. number of clusters.
- Calculated **Silhouette Scores** to assess cluster quality.
- Selected the optimal k based on both metrics.

Step 4: Applying k-Means Clustering

- Applied k-Means with the chosen k.
- Assigned **cluster labels** to all data points.
- Visualized clusters using **scatter plots, pair plots, or 3D plots** if multiple features.

Step 5: Analyzing Clusters

- Examined **average values** of features within each cluster.

- Identified **patterns and behavioral segments**.
- Formulated **actionable insights** (e.g., targeted marketing, customer engagement strategies).

Step 6: Presentation Preparation

- Created **visualizations and charts** to summarize clusters.
- Highlighted key insights and recommended strategies.
- Prepared a **short presentation** explaining the workflow and findings.

---

3. Hands-on Example

```python
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt

# Load dataset
data = pd.read_csv('customer_data.csv')

# Feature scaling
scaler = StandardScaler()
X_scaled = scaler.fit_transform(data[['Annual Income', 'Spending Score']])

# Apply k-Means
kmeans = KMeans(n_clusters=5, random_state=42)
data['Cluster'] = kmeans.fit_predict(X_scaled)

# Visualize clusters
plt.scatter(data['Annual Income'], data['Spending Score'], c=data['Cluster'], cmap='viridis')
plt.xlabel('Annual Income')
plt.ylabel('Spending Score')
plt.title('Customer Segmentation')
plt.show()
```

---

4. Learning Outcomes

By the end of Day 15, we :

- Completed a **full unsupervised learning project from start to finish**.
- Learned to **preprocess, scale, and cluster real-world datasets**.
- Evaluated cluster quality using **Elbow Method and Silhouette Score**.
- Derived **actionable insights** and **interpreted patterns** for business decisions.
- Developed **presentation skills**, summarizing technical findings in a **clear, professional format**.