Assignment 8 – Unsupervised Learning

Name: Rabia Abdul Sattar

Roll No: 2225165022

Course: Applied Data Science with AI

Week #: 8

Project Title: Customer Churn Prediction

---

1. Reading Summary

Reading Material:

- Intro to ML with Python – GitHub
- Scikit-Learn Clustering

Key Learnings:

- **K-Means is a powerful tool** for finding natural patterns and segments in unlabeled data, especially in business datasets like telecom churn.
- **PCA helps simplify complex data** into two or three components, making it easy to visualize clusters and relationships between features.
- **Customer segmentation before prediction** improves business understanding and helps build more effective churn prediction models in future steps.

```
kmeans = KMeans(n_clusters=3, random_state=42)
df['Cluster'] = kmeans.fit_predict(X_scaled)
```

Reflection:

- The official *Scikit-Learn* documentation was studied to understand clustering implementations. It explained how to perform **K-Means clustering**, determine the optimal number of clusters using methods such as the **Elbow Method**, and interpret the results

through metrics like inertia and silhouette scores.

2. Classroom Task Documentation

Task Performed:

1. **K-Means Clustering:**
    - Learned how to divide data into *k* clusters based on feature similarity.
2. **Principal Component Analysis (PCA):**
    - Learned how PCA reduces data dimensions while retaining most of the variance.

3. Weekly Assignment Submission

Assignment Title: Apply Clustering on Telco Customer Churn Dataset

Step 1: Data Preparation

The dataset was first loaded into Python and prepared for clustering:

- **Data Cleaning:**
  Missing and inconsistent values (especially in *TotalCharges*) were handled by converting data types and filling nulls.

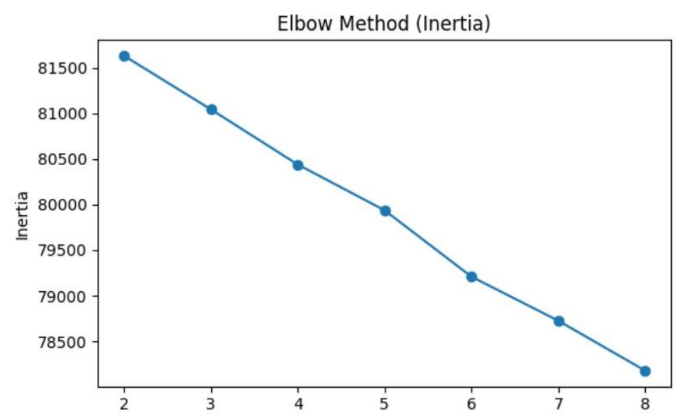
- **Feature Selection:**
  Only relevant numeric and encoded categorical features such as tenure, MonthlyCharges, TotalCharges, and Contract type were selected.

- **Encoding Categorical Variables:**
  Non-numeric columns like *gender*, *Contract*, and *PaymentMethod* were converted to numeric using **Label Encoding** and **One-Hot Encoding**.

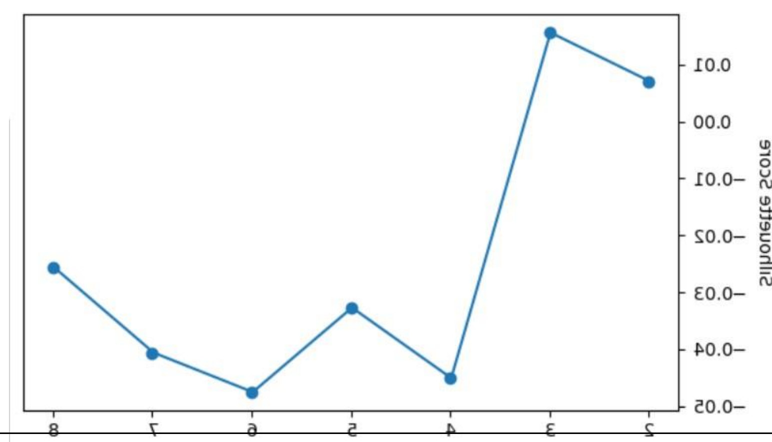
```python
from sklearn.preprocessing import LabelEncoder, StandardScaler

df['TotalCharges'] = pd.to_numeric(df['TotalCharges'], errors='coerce')
df.dropna(inplace=True)
df_encoded = pd.get_dummies(df, drop_first=True)
scaler = StandardScaler()
X_scaled = scaler.fit_transform(df_encoded)
```

Step 2: K-Means Clustering

- **Model Implementation:**
  The **KMeans** algorithm from *scikit-learn* was applied to the preprocessed dataset.
- **Choosing Number of Clusters (k):**
  The **Elbow Method** was used to find the optimal value of *k* by plotting the inertia (sum of squared distances) for different cluster values.

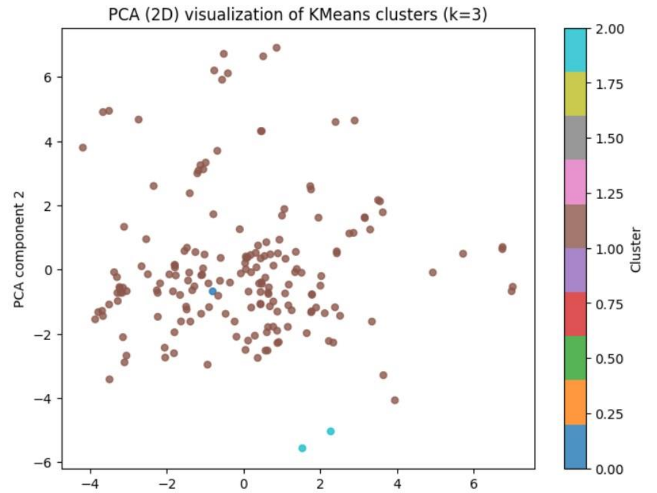


- The elbow curve suggested that **k = 3** was the most suitable choice.
- **Cluster Formation:**
  After selecting *k = 3*, the algorithm was retrained to assign each customer to one of the three clusters.



Step 3: PCA Visualization

Since the dataset had many dimensions, **PCA (Principal Component Analysis)** was used to reduce the data to **two principal components** for visualization.



Step 4: Insights and Interpretation

After analyzing the clustered groups, the following patterns were observed:

- **Cluster 0:**
  Customers with **low monthly charges and long tenure**, likely **loyal and satisfied** users who are less likely to churn.
- **Cluster 1:**
  Customers with **medium tenure and average charges**, showing **neutral churn risk**.
- **Cluster 2:**
  Customers with **high monthly charges and short tenure**, possibly **new users or dissatisfied customers**, indicating a **higher likelihood of churn**.

  These insights can be very helpful for:
- Developing targeted retention strategies.
- Identifying groups needing promotional offers or service improvements.

Conclusion

Through **K-Means clustering** and **PCA visualization**, the dataset was successfully segmented into meaningful customer groups.

Even without using churn labels, patterns of **loyal vs. at-risk customers** emerged clearly.

This demonstrates the power of unsupervised learning for **data exploration and customer segmentation**, which can guide marketing and retention strategies before applying supervised models.

Challenges Faced:

• During this week, the main challenges included handling missing and non-numeric values in the *TotalCharges* column, choosing the optimal number of clusters for K-Means, and ensuring accurate scaling of mixed data types. Visualizing high-dimensional data meaningfully using PCA also required careful preprocessing and experimentation.

GitHub Link:

https://github.com/Rabia-Abdul-Sattar/Customer-Churn-Prediction

4. Project Progress Milestone

- By completing Week 8, the project successfully implemented unsupervised learning through K-Means clustering and PCA visualization.
- Customer segments were identified, revealing patterns useful for churn risk analysis.

5. Self-Evaluation

☑ **Completed:** dataset loading, preprocessing, encoding, feature scaling, clustering using K-Means, and dimensionality reduction using PCA. Visualized clusters in 2D and analyzed customer group patterns for churn insights.