Clustering Students Based on Study Methods Using KMeans

A PROJECT REPORT

Submitted by

MOHAMMAD DILSHAD

in partial fulfillment for the award of the degree
of

Bachelor of Technology

KIET GROUP OF INSTITUTIONS

Introduction

This project focuses on applying the KMeans Clustering algorithm to classify students based on their study methods. Clustering is an unsupervised learning technique used to group similar data points together. The dataset includes various attributes of students' learning habits, and this project aims to group them into meaningful clusters for analysis.

Methodology

1.Data Import: Used pandas to load the CSV file student_methods.csv.

2. Data Cleaning: Selected only numeric columns for clustering since KMeans works on numerical data.

3. Elbow Method: Used to determine the optimal number of clusters by plotting WCSS.

4. KMeans Clustering: Applied KMeans with the optimal number of clusters (assumed 3).

5. Visualization: Plotted the clusters using the first two numeric attributes.

Code

```python
# Step 1: Install required libraries (if not already installed)
!pip install pandas scikit-learn matplotlib seaborn

# Step 2: Import the necessary libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans

# Step 3: Load your dataset
file_path = '/content/student_methods.csv'
df = pd.read_csv(file_path)

# Step 4: Preview the data
print(df.head())

# Step 5: Select only numeric columns
numeric_df = df.select_dtypes(include=['int64', 'float64'])

# Step 6: Elbow Method
wcss = []
for i in range(1, 10):
    kmeans = KMeans(n_clusters=i, init='k-means++', random_state=42)
    kmeans.fit(numeric_df)
    wcss.append(kmeans.inertia_)

plt.plot(range(1, 10), wcss, marker='o')
plt.title('Elbow Method')
plt.xlabel('Number of Clusters')
plt.ylabel('WCSS')
plt.show()

# Step 7: Apply KMeans
kmeans = KMeans(n_clusters=3, init='k-means++', random_state=42)
df['Cluster'] = kmeans.fit_predict(numeric_df)

# Step 8: Display the clustered data
print(df)

# Step 9: Visualize the clustering result
sns.scatterplot(data=df, x=numeric_df.columns[0], y=numeric_df.columns[1], hue='Cluster')
plt.title('Students Classification')
plt.show()
```
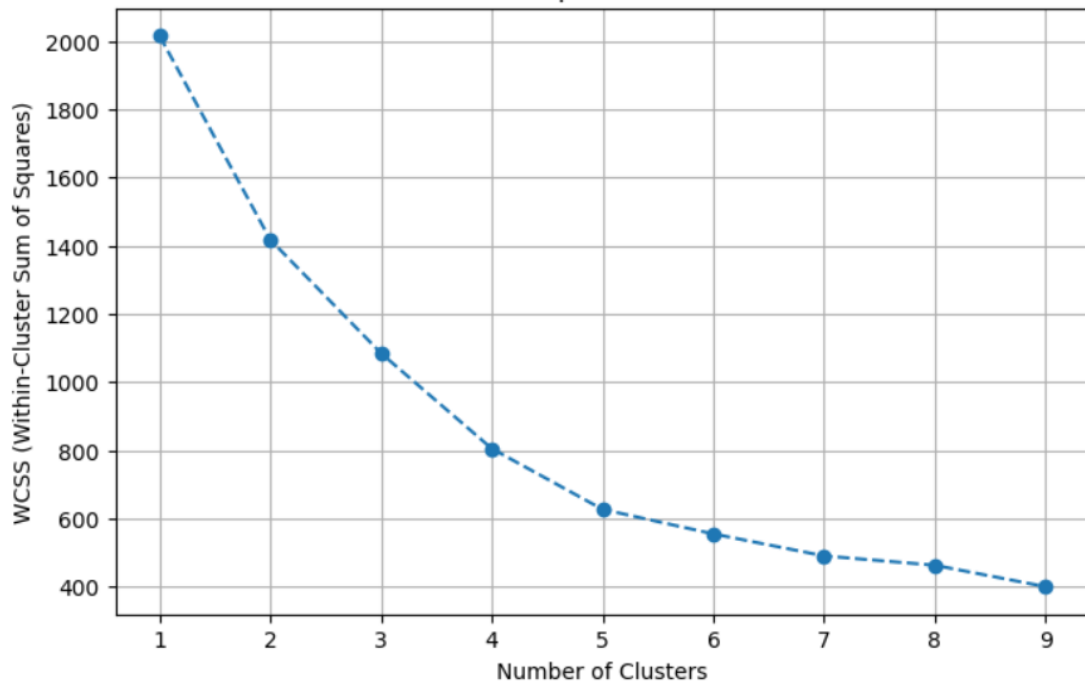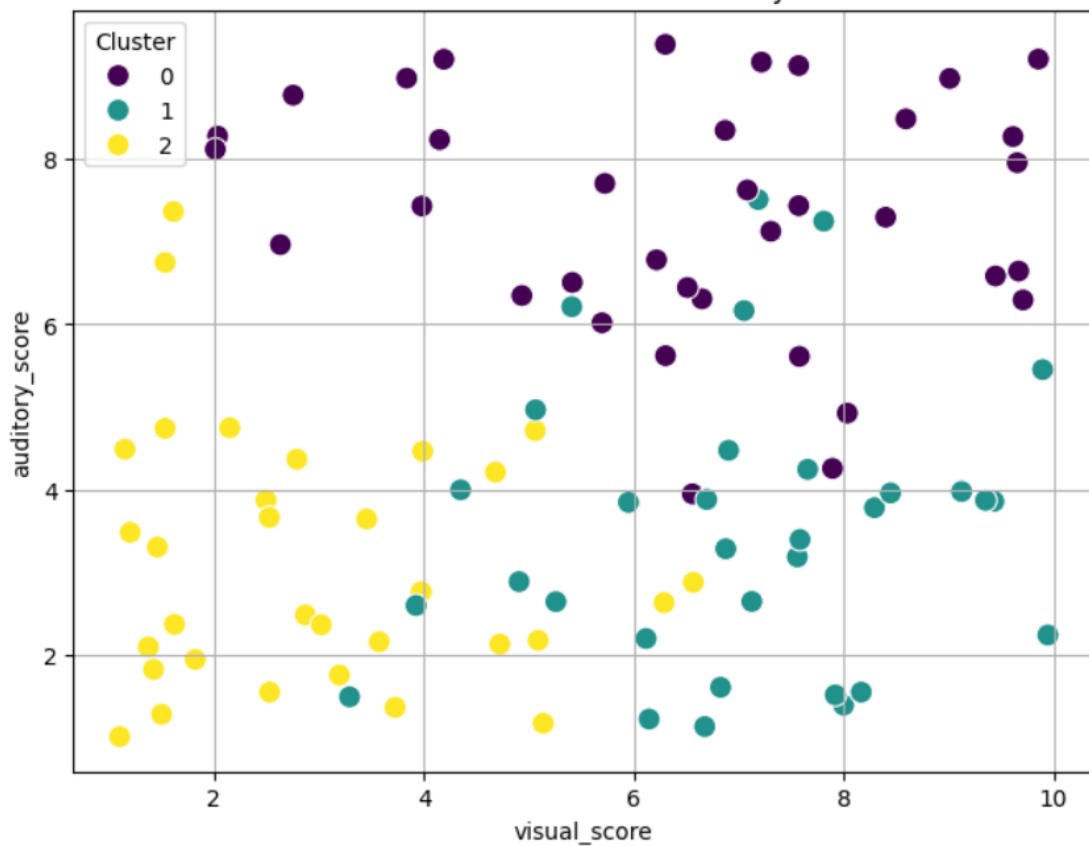
Output/Result

Elbow Method for Optimal Number of Clusters

Students Classification Based on Study Methods

References/Credits

- scikit-learn documentation: https://scikit-learn.org
- Matplotlib documentation: https://matplotlib.org
- Dataset: student_methods.csv (provided by instructor)