PROJECT:CUSTOMER SEGMENTATION USING DATA SCIENCE-PHASE 2

INNOVATION

Consider incorporating dimensionality reduction techniques like PCA or t-SNE to visualize high-dimensional customer data and discover underlying patterns.

1)Short explanation for the project

Customer segmentation using data science is a process of dividing a company's customer base into distinct groups or segments based on common characteristics and behaviors. This approach leverages data analysis, machine learning, and statistical techniques to identify meaningful patterns and insights within customer data. The goal is to better understand customers, tailor marketing efforts, and optimize business strategies. By segmenting customers, businesses can personalize their interactions, products, and services, ultimately improving customer satisfaction and driving growth. Data science helps identify hidden trends and correlations in large datasets, making it a valuable tool for businesses seeking to enhance their understanding of their diverse customer base.

2)Dataset and its detail

https://www.kaggle.com/datasets/kandij/mall-customers

The above dataset we used for this project in phase 2

A dataset for customer segmentation is a collection of data about customers that can be used to divide them into smaller groups based on their shared characteristics or behaviors. This data can include demographic information (e.g., age, gender, location, income), purchase history, website behavior, and other factors.

Here we used this dataset

- Mall Customer Segmentation Data (Kaggle): This dataset contains information about 20,000 customers of a shopping mall, including their age, gender, annual income, spending score, and purchase history.

3)Details about columns in the dataset:

We have used 5 columns in the dataset

1)Customer_id:integer datatype,quantitative data
2)Gender:varchar datatype,qualitative data
3)Age:integer datatype,quantitative data
4)Annual income:integer datatype,quantitative data
5)Spending score:integer datatype,quantitative data

4)Details of libraries

1)Numpy-NumPy is a Python library for numerical and mathematical operations.Go to command prompt and give “pip install numpy”

2)Pandas-Pandas is a Python library used for working with data sets.Go to command prompt and give “pip install pandas”

3)Matplotlib-Matplotlib is a Python library for creating static, animated, and interactive visualizations in various formats, such as line charts, scatter plots, bar charts, histograms, and more.Go to command prompt and give “pip install matplotlib”

4)Sklearn- It provides a wide range of tools and algorithms for tasks such as classification, regression, clustering, dimensionality reduction, and more.Go to command prompt and give “pip install sklearn”

5)Training and testing:

1. Split the dataset into training and testing sets. This will help you to evaluate the performance of your PCA,t-SNE model on unseen data. A common split is 80% training and 20% testing.
2. Preprocess the data. This may involve scaling the data, handling missing values, and removing outliers.
3. Train the PCA,t-SNE model. This can be done using a Python library such as scikit-learn.
4. Apply the PCA,t-SNE model to the training and testing sets. This will produce two lower-dimensional representations of the data, one for the training set and one for the testing set.
5. Evaluate the PCA,t-SNE model. You can do this by comparing the lower-dimensional representations of the training and testing sets. If the model has preserved the pairwise similarities between the data points, then the lower-dimensional representations should be similar.

6)Algorithm explanation:

1.PCA

Principal Component Analysis (PCA) is a dimensionality reduction technique commonly used in data analysis and machine learning. It helps simplify the complexity in high-dimensional data while retaining trends and patterns. PCA transforms the original variables into a new set of orthogonal variables called principal components, which are linear combinations of the original variables.

2.t-SNE

T-distributed Stochastic Neighbor Embedding (t-SNE) is a dimensionality reduction technique used in machine learning and data visualization. It is particularly useful for visualizing high-dimensional data in a lower-dimensional space while preserving the pairwise similarities between data points. t-SNE is commonly employed for exploratory data analysis, clustering, and visualization tasks.

PROGRAM:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.decomposition import PCA
from sklearn.manifold import TSNE
data = pd.read_csv('D:\project\Mall_Customers.csv')
```

```
numerical_features = [ 'Age', 'Annual Income (k$)','Spending Score (1-100)']
categorical_features = ['Genre']
numerical_transformer = Pipeline(steps=[('scaler', StandardScaler())])
categorical_transformer = Pipeline(steps=[('onehot', OneHotEncoder(drop='first')) ])
preprocessor = ColumnTransformer(transformers=[('num', numerical_transformer,
numerical_features),('cat', categorical_transformer, categorical_features)   ])
X = data[numerical_features + categorical_features]
X_processed = preprocessor.fit_transform(X)

pca = PCA(n_components=2, random_state=42)
X_pca = pca.fit_transform(X_processed)
tsne = TSNE(n_components=2, perplexity=30, random_state=42)
X_tsne = tsne.fit_transform(X_processed)

plt.figure(figsize=(12, 6))
plt.subplot(1, 2, 1)
plt.scatter(X_pca[:, 0], X_pca[:, 1], alpha=0.5)
plt.title('PCA Visualization')
plt.xlabel('PCA Dimension 1')
plt.ylabel('PCA Dimension 2')

plt.subplot(1, 2, 2)
plt.scatter(X_tsne[:, 0], X_tsne[:, 1], alpha=0.5)
plt.title('t-SNE Visualization')
plt.xlabel('t-SNE Dimension 1')
plt.ylabel('t-SNE Dimension 2')

plt.tight_layout()
plt.show()
```

OUTPUT: