Text

Description automatically generated

**Activity based**

**Project Report on**

**DWDM**

**Submitted to Vishwakarma University, Pune**

**Under the Initiative of**

**By**

**Arun Adhikari**

**SRN No : 202100406**

**Roll No : 38**

**Div : G**

**Third Year Engineering**

**Department of Computer Engineering**

**Faculty of Science and Technology**

**Academic Year**

**2023-2024**

**Problem Statement**

**Implement a k-means Clustering Algorithm.**

**Business Requirements**

**Stakeholders: Executive leadership, customers,**

**Data Sources: Customer database, content database**

import pandas as pd

from sklearn.cluster import KMeans

from sklearn.preprocessing import StandardScaler

# Load the dataset (replace 'file\_path' with the actual file path of the Excel file)

file\_path = 'Online\_Retail.xlsx'

data = pd.read\_excel(file\_path, usecols=["InvoiceNo", "CustomerID", "Quantity", "UnitPrice", "InvoiceDate"])

# Drop rows with missing CustomerID values as clustering is based on customer ID

data = data.dropna(subset=['CustomerID'])

# Calculate the 'TotalSpent' for each transaction

data['TotalSpent'] = data['Quantity'] \* data['UnitPrice']

# Aggregate the data per CustomerID to create clustering features

customer\_data = data.groupby('CustomerID').agg({

'TotalSpent': 'sum', # Total spending per customer

'InvoiceNo': 'nunique' # Frequency of purchases (unique invoices per customer)

}).rename(columns={'InvoiceNo': 'PurchaseFrequency'})

# Standardize the data to ensure fair clustering

scaler = StandardScaler()

customer\_data\_scaled = scaler.fit\_transform(customer\_data)

# Apply K-means clustering with 3 clusters

kmeans = KMeans(n\_clusters=3, random\_state=42)

customer\_data['Cluster'] = kmeans.fit\_predict(customer\_data\_scaled)

# Display the cluster means to understand the segments

cluster\_summary = customer\_data.groupby('Cluster').mean()

print(cluster\_summary)

import matplotlib.pyplot as plt

# Scatter plot for visualizing clusters

plt.figure(figsize=(10, 6))

# Plot each cluster with a different color

for cluster in customer\_data['Cluster'].unique():

cluster\_data = customer\_data[customer\_data['Cluster'] == cluster]

plt.scatter(

cluster\_data['PurchaseFrequency'],

cluster\_data['TotalSpent'],

label=f'Cluster {cluster}'

)

# Labeling the plot

plt.xlabel('Purchase Frequency')

plt.ylabel('Total Spent')

plt.title('Customer Segments based on Purchase Frequency and Total Spent')

plt.legend()

plt.grid(True)

plt.show()

import matplotlib.pyplot as plt

from sklearn.tree import plot\_tree

# Plot the decision tree structure

plt.figure(figsize=(15, 10))

plot\_tree(

dt\_model,

feature\_names=['TotalSpent', 'PurchaseFrequency'],

class\_names=[f'Cluster {i}' for i in set(y)],

filled=True,

rounded=True

)

plt.title("Decision Tree Structure for Customer Segmentation")

plt.show()



