E-COMMERCE USER SEGMENTATION

A PROJECT REPORT

Submitted by

SHARAN SUNIL (18113045)

T.P MOHANA MAHENDIRA (18113032)

Under the guidance of

Dr. S. SATHYALAKSHMI

PROFESSOR

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY
in
COMPUTER SCIENCE AND ENGINEERING

HINDUSTAN INSTITUTE OF TECHNOLOGY AND SCIENCE
CHENNAI - 603 103
MAY 2022

HINDUSTAN
INSTITUTE OF TECHNOLOGY & SCIENCE
(DEEMED TO BE UNIVERSITY)
CHENNAI

BONAFIDE CERTIFICATE

Certified that this project report **E-COMMERCE USER SEGMENTATION** is the bonafide work of **SHARAN SUNIL (18113045), T.P MOHANA MAHENDIRA(18113032)** who carried out the project work under my supervision during the academic year **2021-2022**.

Dr. J. THANGAKUMAR,
HoD
Department of CSE.

Dr. S. SATHYALAKSHMI
SUPERVISOR
Department of CSE.

INTERNAL EXAMINER

Name:

Designation:

EXTERNAL EXAMINER

Name:

Designation:

Project Viva-voce conducted on:

TABLE OF CONTENTS

ACKNOWLEDGEMENT

First and foremost we would like to thank **ALMIGHTY** who has provided us the strength to do justice to our work and contribute our best to it.

We wish to express our deep sense of gratitude from the bottom of our hearts to our guide **Dr.S.Sathyalakshmi,(Professor), Computer Science and Engineering,** for her motivating discussions, overwhelming suggestions, ingenious encouragement, invaluable supervision, and exemplary guidance throughout this project work.

We would like to extend our heartfelt gratitude to **Dr.J.Thangakumar, Ph.D,Head of the Department, Department of Computer Science and Engineering** for his valuable suggestions and support in successfully completing the project.

We thank the management of **HINDUSTAN INSTITUTE OF TECHNOLOGY AND SCIENCE** for providing us with the necessary facilities and support required for the successful completion of the project.

As a final word, we would like to thank each and every individual who has been a source of support and encouragement and helped us to achieve our goal and complete our project work successfully.

ABSTRACT

In this project, we aim to successfully perform customer segmentation on an E-Commerce website data set by successfully clustering it using the K-Prototype Clustering Algorithm method which is a hybrid algorithm – a combination of K-Means and K-Modes algorithm. It has the ability to handle both Categorical and Numerical variables. Customer Segmentation plays an important role in businesses, especially E-commerce businesses these days. Key factors for any E-Commerce business are to contain the customer needs and then identify the right customers for the particular products. This process should be a checklist every now and then by any business out there. In this paper, we deal with two datasets of an E-Commerce website that contains the list of orders and order details of different customers and their purchases along with clustering done using the K-Prototypes clustering method. We aim to find out the segments in which a customer can be divided and the different attributes of the customer and also compare K-Prototype with other clustering methods which are often used, and find the comparisons among them. We use the same data set to perform clustering using other methods – K-Modes Clustering Algorithm, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), and Hierarchical Agglomerative Clustering (HAC). Finally, after performing the segmentation, we obtain the required results as clusters and we can observe certain comparisons among the algorithms. While applying the K-Prototype algorithm we are successfully able to get the required number of clusters using the dataset having mixed variables which is a major advantage.

LIST OF TABLES

LIST OF FIGURES

CHAPTER – 1

INTRODUCTION

1.1 Overview

The world is getting smarter day by day. The role of E-Commerce in our life is becoming unavoidable. Nowadays right from socks to shoes, dresses to apparel, electronics to appliances, everything we totally depend on E-Commerce websites for the best products. It is a modern-day business that works extensively with the help of technology. In simple words, the process of buying products online is E-commerce shopping. Nowadays slowly the trend of moving from traditional window shopping or going physically to stores is declining. People tend to think more of shopping online than going to shops. Customer is the most important aspect of any business, a business cannot achieve success without this key factor. They need to satisfy customer needs always. In order to do that customer data need to be analyzed from their websites. This process of analyzing and grouping people according to certain attributes is known as customer segmentation.

1.2 Motivation of the project

Customer Segmentation is mainly on the basis of Customer Relationship Management. It helps in the retention of customers and maintains the marketing strategy and competitiveness in the field. It is the process of dividing a group of customers into subgroups according to various attributes of the customer. One of the most common methods of segmentation is the clustering method. Clustering is the process of using Machine Learning and algorithm to identify how data are similar and group them on the basis of certain attributes. Among the clustering methods, K-Means clustering is one of the most commonly used methods. But sometimes when we have a lot of categorical data, it is impossible for K-means to find out the required number of clusters from the given data. So there are other methods used to solve that issue such as the K-Modes Clustering algorithm which is used to handle categorical data, Kprototypes Clustering Algorithm used to handle both – Categorical and Object data types. So in this project, we use the Kprototype Clustering technique to find out the ways in which customers can be segmented.

1.3 Problem Statement

In most cases, clustering is mostly done using the K-Means algorithm. When it comes to handling data in the e-commerce industry or any other business line, taking care of the customer interests, understanding their involvement, average money spent, states/cities from where customer involvements are more; these all factors are huge in terms of profit as well as better future of a

business. Every time analyzing numerical data like quantities, the amount spent, etc cannot alone bring the different segments of customers. It is more difficult to analyze the clusters. A solution for this issue is to use a hybrid algorithm, the K-Prototype algorithm for clustering categorical as well as to object data types.

1.4 Organization of report

Chapter 1 discusses the introduction and problem statement of this project. Chapter 2 showcases the literature review of different papers referred for this project. Chapter 3dealss with the project details, and comparisons between the existing and proposed systems. Chapter 4 discusses about the system design, architecture diagram, use case diagram, activity diagram, and sequence diagram of the project. Chapter 5 discusses the project's software requirements and Chapter 6 deals with the different modules into which the project has been divided. Chapters 7 and 8 deal with implementation parts and the respective results. Chapters 9 and 10 discuss the future enhancement works and info about the team members and their individual objectives in this project.

1.5 Summary

This chapter dealt with the introduction about the project , motivaton to do this project and mention aot the prolem statement of this project. Organization of reort discusses about the different chapters that will be present in this report.

CHAPTER – 2

LITERATURE REVIEW

2.1 Introduction

The first idea of customer segmentation dates back to 1956 where under the article ”Product Differentiation and Market Segmentation as Alternative Marketing Strategies”.

We deal with an Indian E-Commerce data set, with which we perform customer segmentation through clustering using the K-Prototype Clustering Algorithm which handles mixed variables. In order to compare its performance, we perform clustering using three other commonly used algorithms – K-Modes, DBSCAN, and Hierarchical Agglomerative Clustering Algorithms.

2.2 K-modes Clustering (2001, Anil Chaturvedi., Paul E.Green., J.Douglas Carroll)

The authors present a simple procedure for clustering data. The procedure used is the K-Modes clustering algorithm. They propose a method of bilinear clustering model, using loss function L0 norm and that results in k-modes clustering procedure. They explain the drawbacks and advantages of k-modes in market/customer segmentation.

2.3 Concept-based document clustering using K-Prototype Algorithm (2018, Sneha Pasarate, Rajashree Shedge)

The authors propose a method of document clustering using the k-prototype algorithm. The proposed system starts with document extraction(input) and proceeds further for doing the data pre-processing, generating named entities using feature selection. K-prototype clustering model is performed and clusters are obtained successfully. They use Java(JDK 1.8) along with Netbeans for the implementation of the project.

2.4 A K-Prototypes Algorithm Based on Adaptive Determination of the Initial Centroids (2018, Dongwei Guo., Yingjie Chen., Jingwen Chen)

The authors propose the idea of an improved k-prototypes algorithm model. They provide an improved solution for finding adaptive centroids and the required mathematical calculations. They propose a newly improved algorithm model where a dataset is loaded, the average distance between samples is determined and the radius along with threshold distance is found too. Dissimilarities between the dataset and k centroids are calculated. The algorithm reduces input parameters and brings the algorithm closer to the concept of Machine Learning.

2.5 Customer Segmentation Using Hierarchical Agglomerative Clustering International Conference on Information Science and Systems. (2019, Phan Duy Hung., Nguyen Thi Thuy Lien.,Nguyen Duc Ngoc)

The authors propose a method of performing Hierarchical Agglomerative Clustering on a Credit card dataset. The aim is to perform customer segmentation on 18 features of the Credit card dataset. They use Ubuntu, R Language, Python, and Java for the implementation of the project. They experiment with various methods to find the k value, by using the gap static method and silhouette method.

2.6 Customer Segmentation Techniques on E-Commerce (2021, Sumit Koul, Trissa Merrin Philip)

The authors profess the knowledge about the different types of segmentation in general types, mainly: psychographic segmentation, geographic segmentation, value-based segmentation demographic segmentation, and propensity-based segmentation. Then they find out the best clustering methods. maisegmentationthe types of segmentation and customer segmentation techniques which is clustering. They describe each clustering technique in detail with proofs derived from other papers written and found by other authors.

2.7 Application of Clustering Algorithm for Effective Customer Segmentation in E-Commerce (2021, Rita Punhani, V.P.S Arora, Sai Sabitha, Vinod Kumar Shukla)

The authors propose a method of segmentation using open source software Rapid Miner and predict the segments and customer behavior attributes. They take a dataset from Kaggle and then normalize the data using Rapid Miner, they take the number of clusters as 4 since it had the least DBL. And as they cluster and compare to find out the average order id and payment used by customers.

2.8 Summary

This chapter deals with the different approaches and techniques used in the customer segmentation process. From each of the papers, we can infer the different methods and points to be taken further to make our project much more efficient. Most of the papers dealt with the K-Prototype algorithm and K-Modes algorithm, which are used to handle mixed and categorical variables respectively.

CHAPTER – 3

PROJECT DESCRIPTION

3.1 Objective

Through this project, we aim to:

(i) Perform Customer Segmentation on an E-Commerce website Data set through K-Prototype Clustering Algorithm, and derive the different clusters in which customers are divided and the decisive parameters.

(ii) Comparison of clustering among other clustering methods. In order to compare K-Prototype Algorithm's performance and its advantages over others, we do clustering of the same data set with three other widely used algorithms for clustering and analyze the differences, distance metrics, evaluation metrics, and flaws of each one:

(a) K-Modes Clustering Algorithm

(b) DBSCAN Algorithm

(c) Hierarchical Agglomerative Clustering.

3.2 Existing System

The most common way of implementing customer segmentation is by using K-Means clustering. K-Means clustering works only with numerical data and cannot handle categorical data. This factor is a major disadvantage. For every instance, it is not possible to only cluster on the basis of numerical values. There are other important factors that could be categorical that needs to be dealt with. In order to solve that issue, there is the K-Modes algorithm which deals with only categorical data. But again this is a problem. In order to solve both the issues, Huang came up with a hybrid algorithm called the K-Prototype algorithm for clustering using both numerical and categorical data and providing the clusters accordingly.

3.3 Shortcomings of Existing System

- Using the K-Means algorithm, it is difficult to predict the K-Value.
- When initial clusters are different, it can result in a different cluster finally.
- It does not cluster large data well having different densities and sizes.
- Can handle only numerical variables of the same density and sizes.

3.4 Proposed System

We propose a method of clustering the e-commerce dataset containing 1500 rows and the following columns – Order id, Order Date, Customer Name, State, City, Amount, Profit, Quantity, Category, and Sub-Category. After loading the dataset, we search for missing values. Then we find any duplicate values and remove them. After that, we check the data distribution to find the various trends in the dataset and know the customer characteristics.

Next thing is to proceed further with clustering using the K-Prototypes algorithm. The first thing to do is convert the data frame to a matrix and also fetch the categorical column positions respectively. Next, we use the elbow method to find the optimal number ok 'K'. Using KneeLocator we can confirm the point at which the elbow is formed and the optimal K value. Next, we fit the K Prototype model for predicting the clustering results. Next, we add cluster labels to the data frame and do the cluster interpretation in which we get the desired result. We obtain the clusters along with the different columns correlated. We can observe the clusters and their trends.

The next objective is to do clustering using DBSCAN, Hierarchical Agglomerative Clustering, and K-Modes Clustering algorithm. At last, we find out how the clustering is done using different models on the dataset.

3.5 Advantages of Proposed System

- Works with mixed data types, unlike the K-Means Clustering algorithm.
- Measures distance using Euclidean distance (like K-Means), but measures distance between categorical features using the number of matching categories.
- Categorical data doesn't need any pre-processing to apply K-Prototype Clustering.
- K-Prototype Clustering Algorithm produces clusters more superior and clearer than the ones produced by K-Means.

CHAPTER – 4

SYSTEM DESIGN

4.1 Flow Diagram

Firstly we take the dataset, in our case, it is an E-Commerce dataset, load them using pandas. Next, we search for any missing or duplicate values in any of the rows, as the K-Prototype Algorithm cannot handle any missing or duplicate values. Next, we seek the categorical column positions from the dataset, and we convert the data frame to a matrix. Next, we find the optimal number of 'K' using the elbow method and then we build the k-prototype model and use the fit_predict for it. Next, we add the cluster labels and then do interpretation to get the different clusters in which we can identify the customer spending traits.

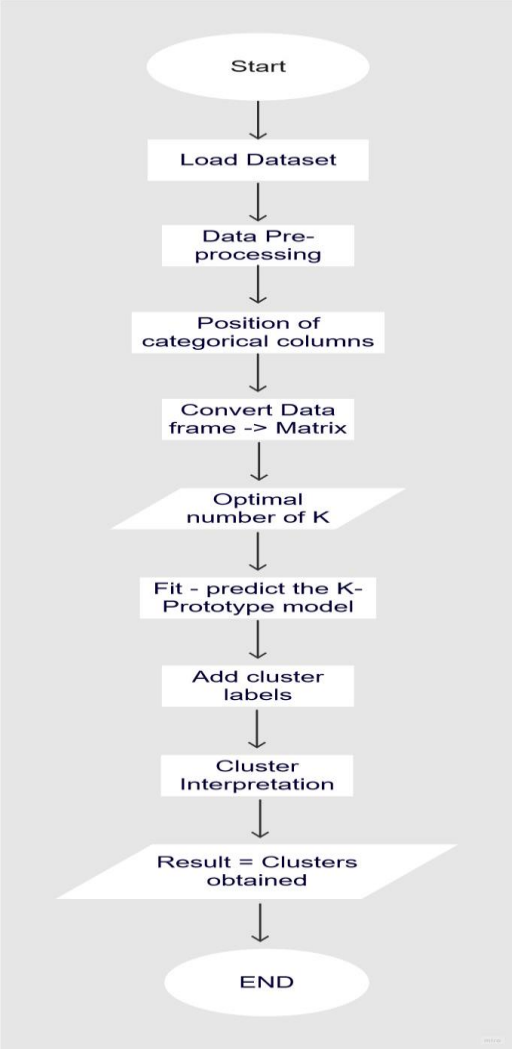

Fig 4.1 Design Flow Diagram

4.2 UML Diagrams

UML diagram is a UML (Integrated Model Language) diagram for the purpose of visually representing a program and its main characters, roles, actions, art objects, or categories, in order to better understand, modify, store, or document information about the system.

ML drawings can be used as a preview of a project before it happens or as a project document later. But the overall purpose of UML drawings is to allow teams to visualize how a project works or will work, and can be used in any field, not just software engineering.

4.3 Use case Diagram

A user case diagram is a clear picture of possible user interaction with a system. The application case diagram shows the different operating conditions and different types of users the system has and will often be compatible with other types of drawings. Terms of use are represented by circles or ellipses. Characterized diagrams show and explain the context and requirements of the whole system or key components of the system. You can model a complex system with a single user case diagram, or create character charts to model system components. You can usually create diagrams of the characters used in the early stages of a project and refer to them throughout the development process.

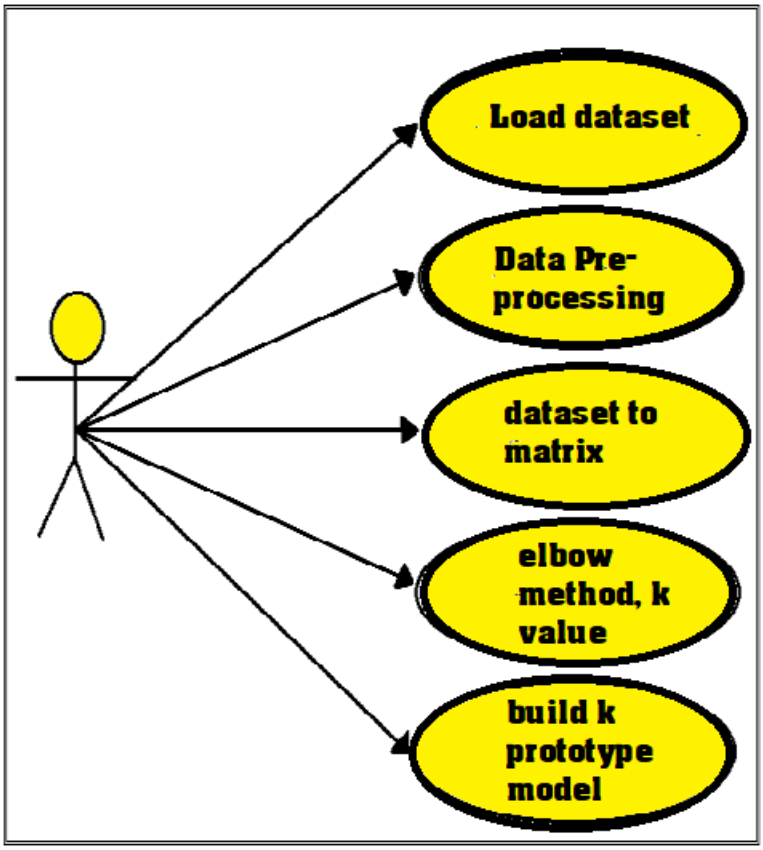

Fig 4.2 Use case Diagram

4.3 Activity Diagram

The activity diagram is another important diagram of the UML diagram to describe the changing aspects of a program. An activity diagram is an advanced version of the flow chart that simulates the flow from one task to another function. Activity diagrams describe how tasks are integrated to provide a service that can be at different levels of output. Often, an event needs to be accomplished through a particular task, especially when the task is intended to accomplish many different things that require collaboration, or how events in a single-use context are related, in particular, to the use of contexts in which activities. it may overlap and require cooperation.

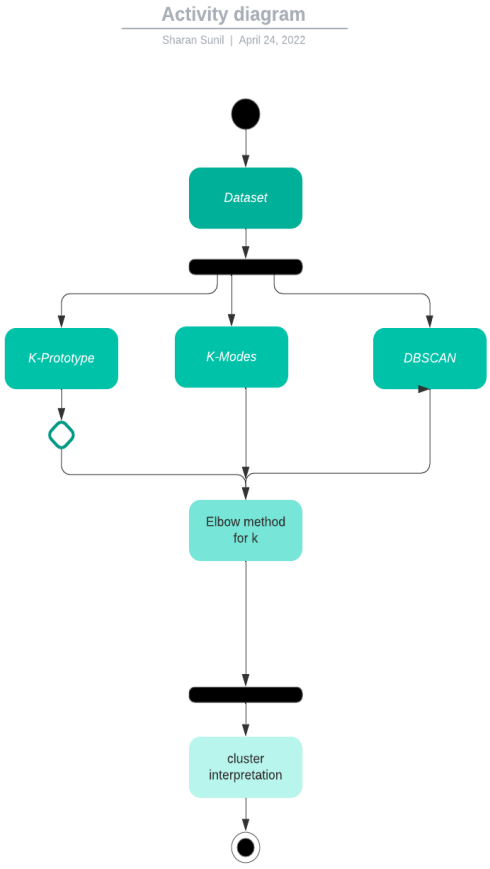

Fig 4.3 Activity Diagram

4.4 Sequence Diagram

Sequential diagram simply shows the interaction between objects in sequential sequence i.e. how these interactions occur. We can also use the words event drawings or event scenes to refer to a sequence drawing. Sequential diagrams describe how things in a system work and in what order.

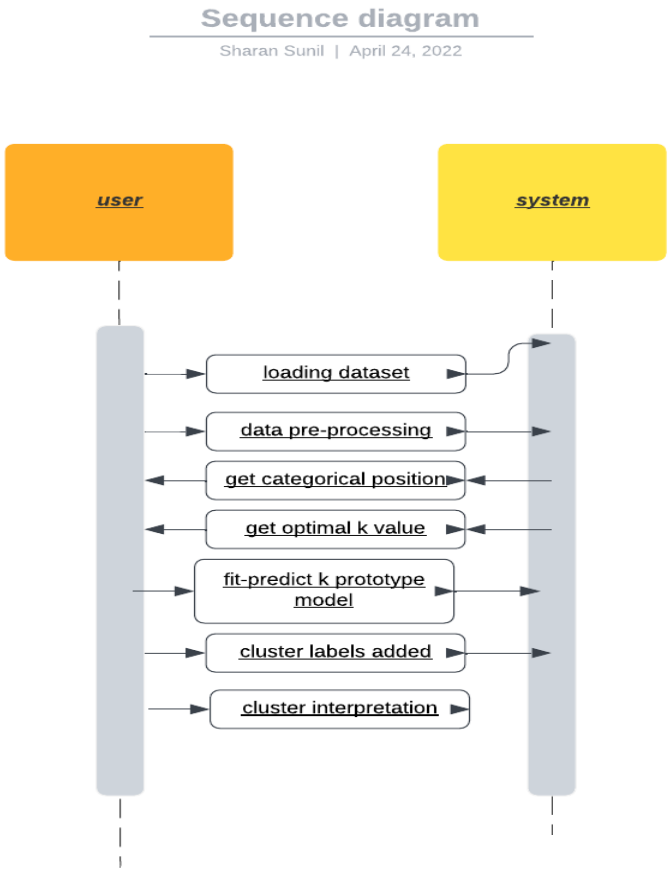

Fig 4.4 Sequence Diagram

4.5 Summary

This chapter deals with the system design before we head into the implementation parts. The use case, activity, and sequence UML diagrams give us an idea about the system's structure and how we need to move forward in implementing the project

CHAPTER – 5

SOFTWARE REQUIREMENTS

5.1 Introduction

This chapter gives an insight into the system requirements of this project. We list the software and hardware requirements of this project below.

5.2 Software Requirements

- Operating System: Windows 10
- Programming Language: Python
- IDE: Jupyter Notebook
- Database: Indian E-Commerce Website

5.3 Hardware Requirements

- System: Intel i5
- Hard Disk: 500 GB
- Monitor: 14’ Colour Monitor
- Input Devices: Mouse, Keyboard
- RAM: 8 GB

5.4 Summary

This chapter discusses the system requirements needed to implement this project successfully.

CHAPTER – 6

MODULE DESCRIPTION

6.1 Introduction

This chapter deals with the different modules into which this project is divided. The project is divided into modules based on the system design and flow diagram. With this, it will be possible to successfully perform customer segmentation using the k –prototypes algorithm efficiently and gain the required clusters.

6.2 Modules

After analyzing the flow diagram and objectives of the proposed system, we have divided the project into 2 modules mainly:

1. K-Prototypes Clustering algorithm.
2. Comparisons with K-Modes, DBSCAN, and HAC algorithms.

6.3 Module Description

6.3.1 K-Prototypes Clustering algorithm

This module mainly contains the steps such as data pre-processing which involves the loading of the e-commerce dataset and importing the necessary libraries, checking for missing data and duplicate rows as the k-prototype algorithm cannot handle any missing data or duplicate values.

We take two E-Commerce datasets as discussed earlier, and merge them into a single pandas data frame to include a mix of both categorical and numerical variables. K-Prototype has an advantage because it’s not too complex and is able to handle large data and is better than hierarchical-based algorithms.

Firstly before heading over to the K-Prototype algorithm, we look to find the categorical column positions. And then we convert the data frame to a matrix. We then proceed to do the elbow method to find the optimal ‘K’ value. And with the ‘K’ value we go on to build the K-Prototype model for successful clustering.

6.3.2 K-Modes, DBSCAN, and HAC algorithms

In the earlier module, we discussed how to go with the K-Prototype clustering. In order to compare the overall efficiency of the K-Prototype clustering algorithm along with other Clustering algorithms, we perform clustering on the same dataset with other often used algorithms – K-Modes Clustering algorithm, DBSCAN(Density-Based Spatial Clustering with Added Noise), and (HAC) Hierarchical Agglomerative Clustering. K-Modes Clustering is a widely used clustering algorithm that is used to handle only categorical variables mainly. It is a great choice over K-Means which cannot handle categorical data. DBSCAN Algorithm is a density-based algorithm that divides the denser clusters from the least dense ones. Hierarchical Agglomerative Clustering is one of the most common types of hierarchical clustering which clusters objects based on similarities.

6.4 Summary

This chapter dealt with the different modules in which the project is divided such as the K-Prototypes clustering module and a module comprising comparisons between K-Prototypes to that of K-Modes, DBSCAN, and HAC.

CHAPTER – 7

IMPLEMENTATION

7.1 Dataset

The dataset we are using is an e-commerce sales dataset which is from an Indian e-commerce website. We have obtained it from Kaggle. We are considering two important datasets and using them mainly: (a) List of orders and (b) order details.

(a)List of Orders-This dataset contains purchase information. The information includes ID, Date of Purchase, and customer details

(b) Order Details- This dataset contains order ID, with the order price, quantity, profit, category, and subcategory of the product

Since we are doing clustering using the K-Prototypes algorithm, we use a mix of datatypes containing objects as well as categorical data. We form a new data frame by combining the columns of the two datasets, dropping a few columns, and naming it 'df_merged'. It contains 1500 rows and 10 columns. Further in the upcoming parts of implementation we remove a couple of unnecessary columns and take the necessary ones.

```
1 df_merged
```

	Order ID	Order Date	CustomerName	State	City	Amount	Profit	Quantity	Category	Sub-Category
0	B-25601	01-04-2018	Bharat	Gujarat	Ahmedabad	1275.0	-1148.0	7	Furniture	Bookcases
1	B-25601	01-04-2018	Bharat	Gujarat	Ahmedabad	66.0	-12.0	5	Clothing	Stole
2	B-25601	01-04-2018	Bharat	Gujarat	Ahmedabad	8.0	-2.0	3	Clothing	Hankerchief
3	B-25601	01-04-2018	Bharat	Gujarat	Ahmedabad	80.0	-56.0	4	Electronics	Electronic Games
4	B-25602	01-04-2018	Pearl	Maharashtra	Pune	168.0	-111.0	2	Electronics	Phones
...	...	...	...	...	...	...	...	...	...	...
1495	B-26099	30-03-2019	Bhishm	Maharashtra	Mumbai	835.0	267.0	5	Electronics	Phones
1496	B-26099	30-03-2019	Bhishm	Maharashtra	Mumbai	2366.0	552.0	5	Clothing	Trousers
1497	B-26100	31-03-2019	Hitika	Madhya Pradesh	Indore	828.0	230.0	2	Furniture	Chairs
1498	B-26100	31-03-2019	Hitika	Madhya Pradesh	Indore	34.0	10.0	2	Clothing	T-shirt
1499	B-26100	31-03-2019	Hitika	Madhya Pradesh	Indore	72.0	16.0	2	Clothing	Shirt

1500 rows × 10 columns

Fig 7.1 The merged dataset - 'df_merged'

7.2 Data Exploration

It involves importing the data set, importing the necessary libraries, and checking for missing values and duplicate rows, since K-Prototype cannot handle missing values or duplicate rows.

After that, we check for the different data distributions in the dataset to get a deep insight before we do the clustering with different algorithms. When we explored the e-commerce dataset we could analyze different trends in the dataset of customers. The explorations are visualized below in the form of count plots and hist plots.

Widely sought-after categories are **FURNITURE, CLOTHING CATEGORIES, and ELECTRONICS**.

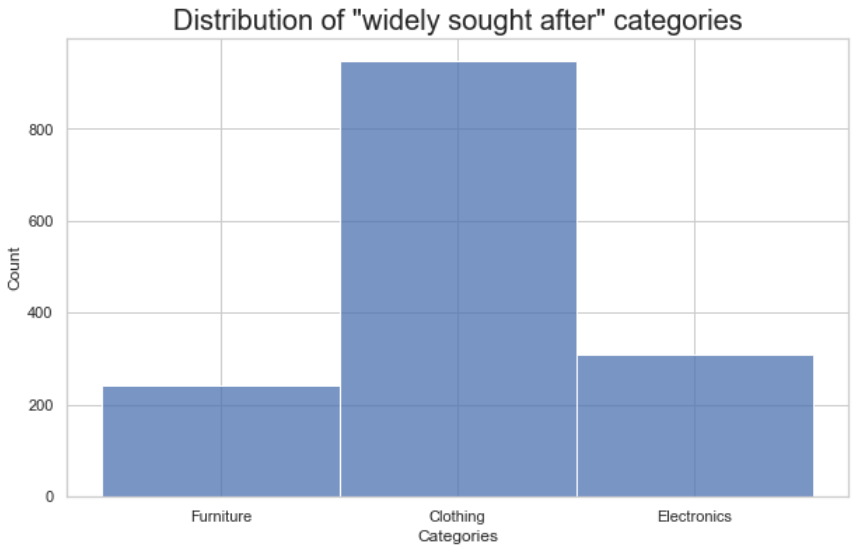

Fig 7.2 Sought after categories

Popular Sub-categories are: **STOLES, HANDKERCHIEFS, SAREES, PHONES**

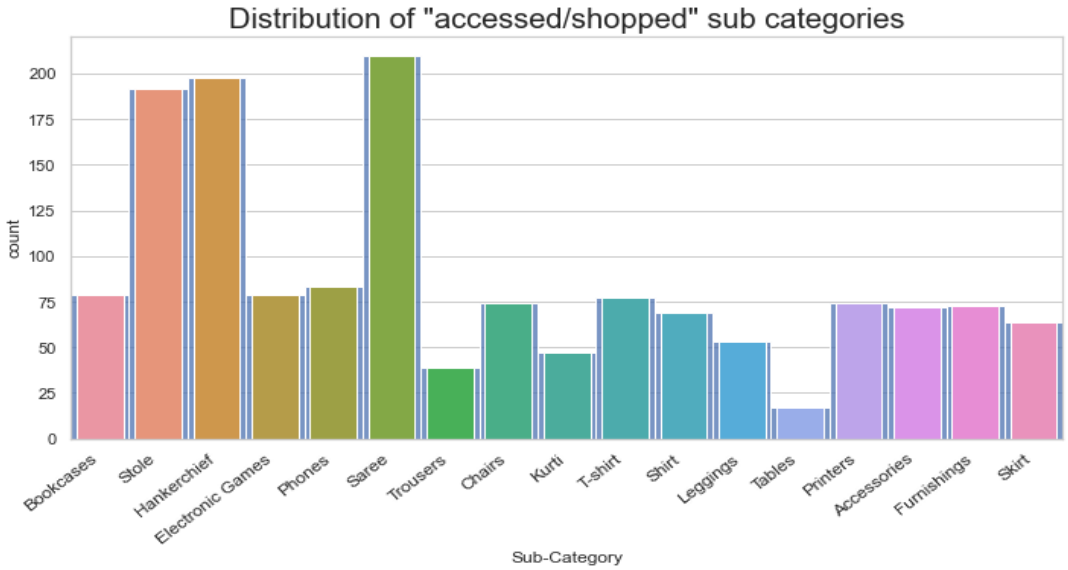

Fig 7.3 Popular Sub-Categories

States with the most number of orders: **MAHARASHTRA, M.P, RAJASTHAN**

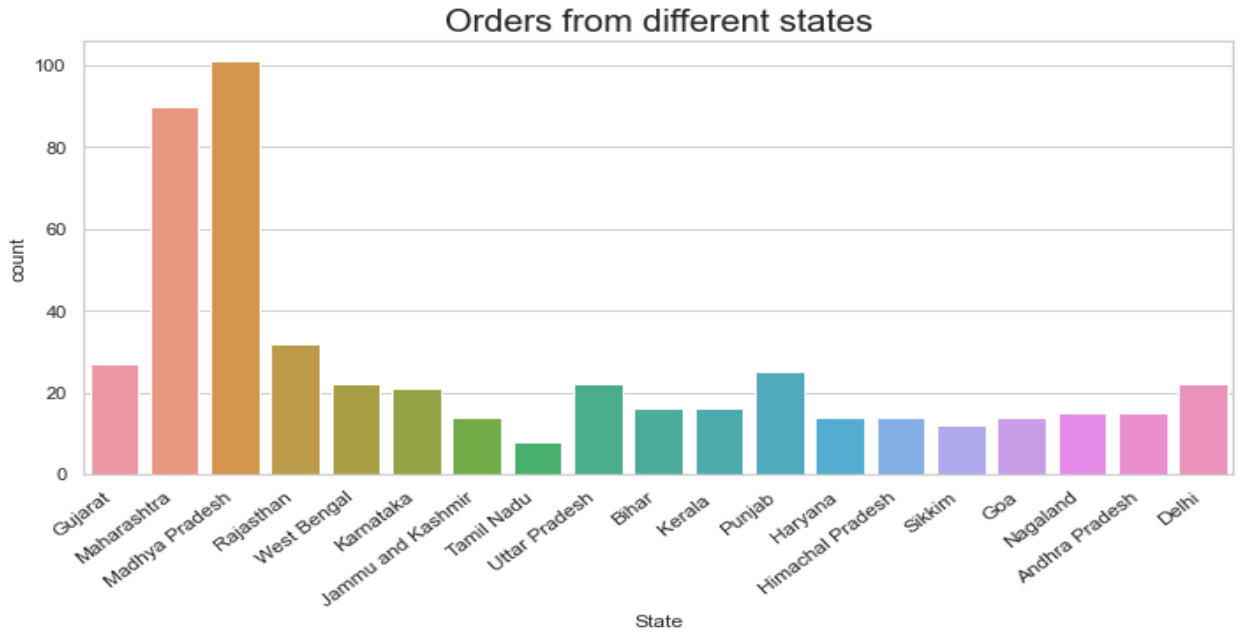

Fig 7.4 States with the most number of orders

Cities with the most number of orders: **Indore(M.P), Mumbai(Maharashtra), Delhi, Chandigarh(Punjab)**

7.3 K-Prototype Clustering

Firstly we try to extract the categorical column positions of the data. The k-Prototype algorithm requires the position of columns. We store the positions in a variable ‘categColumnPos’. This would be helpful for cluster analysis.

```
Categorical columns          : ['Order ID', 'Order Date', 'CustomerName', 'State', 'City', 'Category', 'Sub-Category
']
Categorical columns position  : [0, 1, 2, 3, 4, 8, 9]
```

Fig 7.5 Categorical column positions

Next, we convert the data from the data frame to a matrix. This helps the kmodes module to run the K-Prototype clustering algorithm properly. We save the data matrix to a variable called ‘dfMatrix’.

```
3 dfMatrix
```

```
array([['B-25601', '01-04-2018', 'Bharat', ..., 7, 'Furniture',
        'Bookcases'],
       ['B-25601', '01-04-2018', 'Bharat', ..., 5, 'Clothing', 'Stole'],
       ['B-25601', '01-04-2018', 'Bharat', ..., 3, 'Clothing',
        'Hankerchief'],
       ...,
       ['B-26100', '31-03-2019', 'Hitika', ..., 2, 'Furniture', 'Chairs'],
       ['B-26100', '31-03-2019', 'Hitika', ..., 2, 'Clothing', 'T-shirt'],
       ['B-26100', '31-03-2019', 'Hitika', ..., 2, 'Clothing', 'Shirt']],
      dtype=object)
```

Fig 7.6 Data to matrix

We are using the Elbow method to find out the optimal number of 'K' where K denotes the number of clusters. Instead of calculating WSSE(Within Sum-of Squares Errors) using Euclidean distance like in K-Means and some other clustering algorithms, the K-Prototype Algorithm provides a 'Cost' function that combines the calculation of mixed variables. We use the elbow method and visualize its result. Through the graph, it seems that the line bends near 3, so we assume K=3.

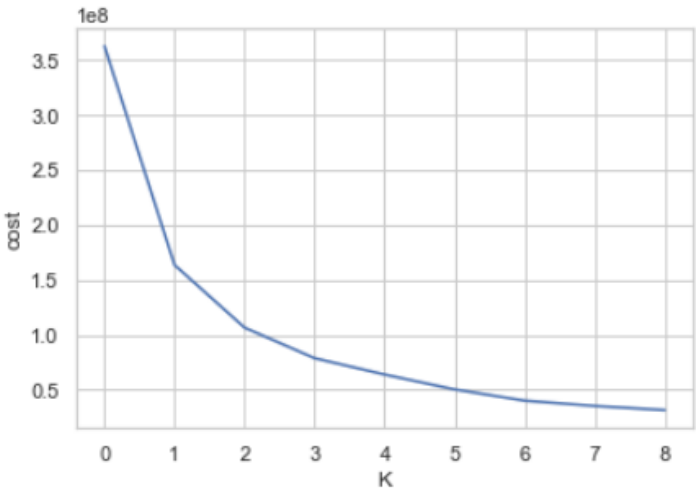

Fig 7.7 Elbow Method

In order to confirm that K=3, we use another method called KneeLocator which can find out k using the convex or concave length of bends. It is is imported using a module named Kneed.

```
3  from kneed import KneeLocator
4  cost_knee_c3 = KneeLocator(
5          x=range(1,10),
6          y=cost,
7          S=0.1, curve="convex", direction="decreasing", online=True)
8
9  K_cost_c3 = cost_knee_c3.elbow
10 print("elbow at k =", f'{K_cost_c3:.0f} cluster')
```

```
elbow at k = 3 cluster
```

Fig 7.8 KneeLocator

Next, we fit and predict the K-Prototype model and add cluster labels to the data frame, and finally, we interpret the clusters. To interpret the cluster, for the numerical variables, it will be using the average while the categorical use the mode. But there are other methods that can be implemented such as using median, percentile, or value composition for categorical variables. Through this, we get the required clusters.

```
# Cluster interpretation
df_merged.rename(columns = {'Cluster Labels':'Total'}, inplace = True)
df_merged.groupby('Segment').agg(
    {
        'State': lambda x: x.value_counts().index[0],
        'City': lambda x: x.value_counts().index[0],
        'Amount': 'mean',
        'Quantity': 'mean',
        'Category': lambda x: x.value_counts().index[0],
        'Sub-Category': lambda x: x.value_counts().index[0],
    }
).reset_index()
```

Fig 7.8 Cluster Interpretation

7.4 K-Modes, DBSCAN, HAC

K- Modes algorithm is one of the unsupervised Machine Learning algorithms that is used to cluster categorical variables. K-Modes can be imported through a module named kmodes. We use the same dataset here, just rename it as ‘df_kc’, and drop some unnecessary columns. We use the elbow method to get the optimal ’K’ value. After initializing centroids and clusters, it plots the graph and gives K=2. Next, we build the model with 2 clusters and then finally we obtain the required clusters.

DBSCAN requires only two parameters: epsilon and minPoints. Epsilon is the radius of the circle to be created around each data point to check the density and minPoints is the minimum number of data points required inside that circle for that data point to be classified as a Core point. We visualize the clusters through a scatterplot and then calculate the silhouette score also.

Hierarchical Agglomerative Clustering is one of the most commonly used hierarchical clustering methods which groups objects in clusters based on their similarity. Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects. The result is a tree-based representation of the objects, named dendrogram. We verify the cluster tree and then we cut the dendrogram via a threshold value. We can cut the hierarchical tree at a given height in order to partition the data into clusters.

7.5 Summary

This chapter describes the various parts of implementation, the necessary steps involved in each phase is given in a detailed manner along with necessary visualizations and outputs in order to have a clear understanding of the implementations.

CHAPTER – 8

RESULTS

8.1 Introduction

In this chapter, we display the results obtained through implementations as discussed in the previous chapter. Firstly we show the results obtained in the first module where we applied the K-Prototypes algorithm for performing customer segmentation of the E-Commerce Dataset. Secondly, we perform other clustering algorithms with the same data set, just renaming it with different names in different scenarios. Lastly, we show a table comparison of all algorithms used, their calculation metrics, and remarks about them.

8.2 Clustering using K-Prototype

Segment	State	City	Category	Sub-Category
First	Madhya Pradesh	Indore	Electronics	Trousers
Second	Madhya Pradesh	Indore	Electronics	Printers
Third	Madhya Pradesh	Indore	Clothing	Hankerchief

Fig 8.1 Clusters obained using K-Prototype algorithm

8.3 Clustering using K-Modes

```
2 df_kc
```

	Cluster	Order ID	Order Date	CustomerName	State	City	Amount	Profit	Quantity	Category	Sub-Category	Total
0	0	B-25601	01-04-2018	Bharat	Gujarat	Ahmedabad	1275.0	-1148.0	7	Furniture	Bookcases	1
1	0	B-25601	01-04-2018	Bharat	Gujarat	Ahmedabad	66.0	-12.0	5	Clothing	Stole	2
2	0	B-25601	01-04-2018	Bharat	Gujarat	Ahmedabad	8.0	-2.0	3	Clothing	Hankerchief	2
3	0	B-25601	01-04-2018	Bharat	Gujarat	Ahmedabad	80.0	-56.0	4	Electronics	Electronic Games	2
4	0	B-25602	01-04-2018	Pearl	Maharashtra	Pune	168.0	-111.0	2	Electronics	Phones	2
...	...	...	...	...	...	...	...	...	...	...	...	...
1495	1	B-26099	30-03-2019	Bhishm	Maharashtra	Mumbai	835.0	267.0	5	Electronics	Phones	1
1496	1	B-26099	30-03-2019	Bhishm	Maharashtra	Mumbai	2366.0	552.0	5	Clothing	Trousers	0
1497	0	B-26100	31-03-2019	Hitika	Madhya Pradesh	Indore	828.0	230.0	2	Furniture	Chairs	1
1498	0	B-26100	31-03-2019	Hitika	Madhya Pradesh	Indore	34.0	10.0	2	Clothing	T-shirt	2
1499	0	B-26100	31-03-2019	Hitika	Madhya Pradesh	Indore	72.0	16.0	2	Clothing	Shirt	2

1500 rows × 12 columns

Fig 8.2 Clusters obained using K-Modes algorithm

8.4 Clustering using DBSCAN

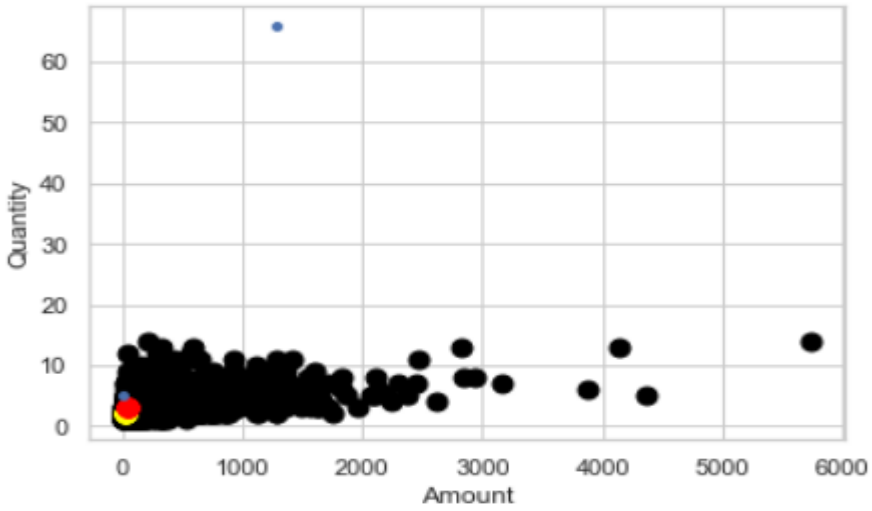

Fig 8.3 Scatterplot of Clusters obtained using DBSCAN

Here we get the required graph containing the clusters formed using density based clustering algorithm. The black ones are the more denser ones long with the yellow and red clusters.

8.5 Clustering using HAC

- Dendrogram

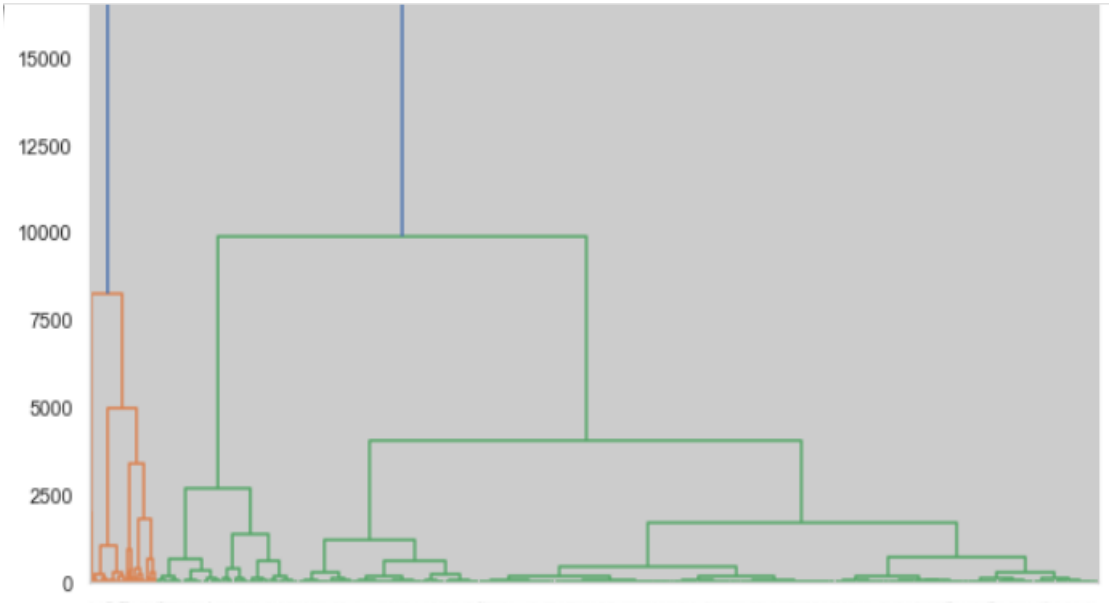

Fig 8.4 Dendrogram of clusters

Here we plot dendrograms for Hierarchical clustering, here we can see many clusters formed and the tall lines represent the distance between certain clusters. Here we take the x-axis as amount and y-axis as quantity, and the clusters formed are the results between these two factors.

- Dendrogram cut into different groups

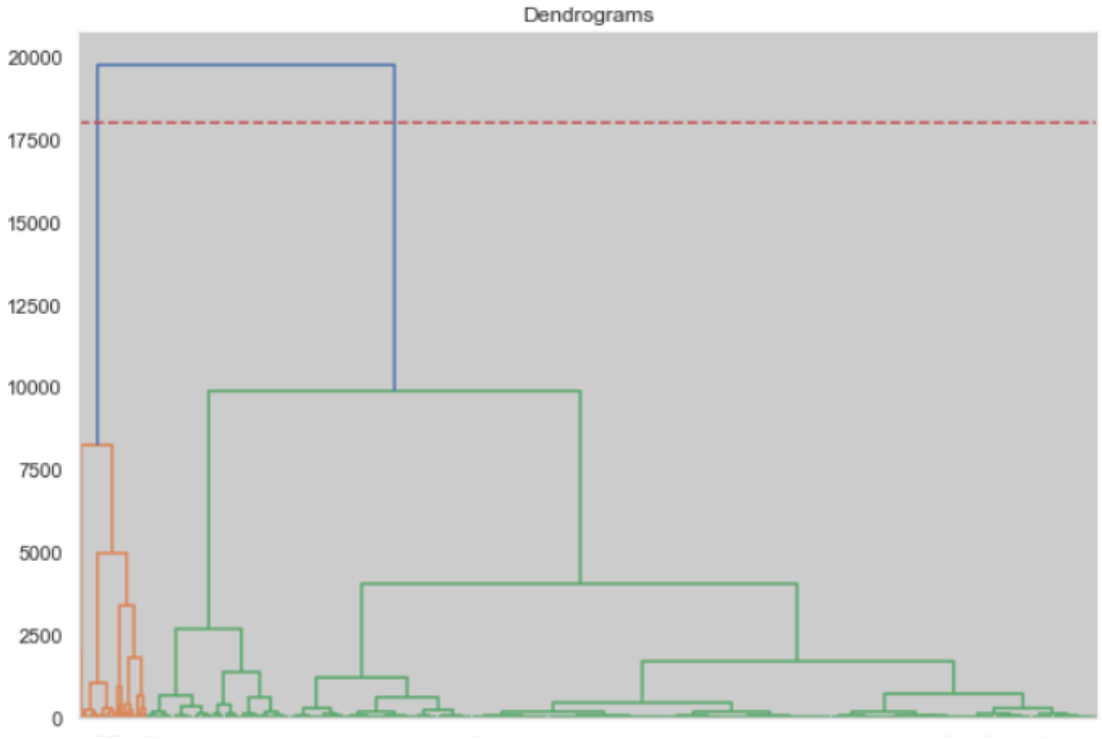

Fig 8.5 Dendrogram cut into different groups

Here we cut the dendrogram with an ideal threshold value, x axis is the quantity and y axis is the amouunt spent.

- Cluster plot

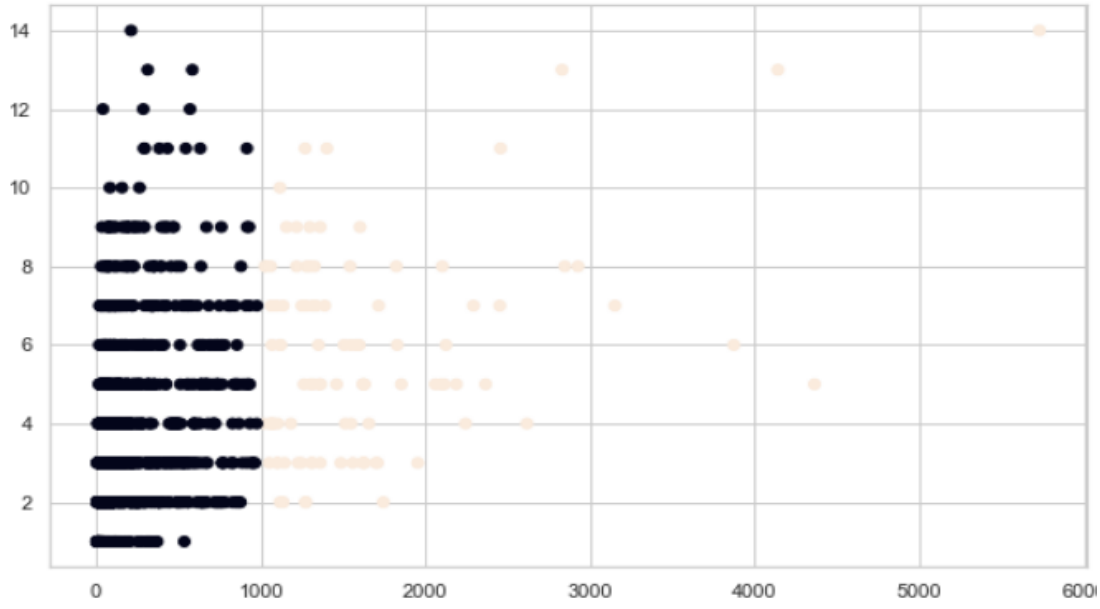

Fig 8.6 Cluster Plot

8.6 Comparisons

CLUSTERING ALGORITHMS	EVALUATION METRICS	DATA TYPES	DISTANCE METRICS	REMARKS
K-PROTOTYPE ALGORITHM	Gowers Distance	Mixed	Euclidean dissimilarity	• Easy to use. • handles mixed data. • Cannot find silhouette score because of categorical variables. • Clear representation of clusters with large data.
K- MODES ALGORITHM	Gowers Distance	Categorical	Euclidean dissimilarity	• Handles only categorical variables • Cannot perform silhouette score, instead used Gowers distance metrics as in K-Prototype. • Clusters formed according to categorical data only.
DBSCAN	Silhouette Score	Numerical	Euclidean	• Silhouette score close to -1 indicating poor clustering. • Not necessary to give labels or k-value • Clustering was possible only with the numerical variables. • Densely clustered scatter plot
AGGLOMERATIVE HIERARCHICAL CLUSTERING	Silhouette Score	Numerical	Euclidean	• Easy method to plot clusters and visualize them using dendrograms. • Silhouette score close to 1, indicating good clustering of results.

Table 1.1 Comparison of clustering algorithms

8.7 Summary

This chapter dealt with the results obtained through various stages of the implementation and they are displayed in a well-organized way in order to clearly understand the outputs obtained.

CHAPTER – 9

CONCLUSION AND FUTURE ENHANCEMENTS

9.1 Conclusion

Through this project, we could work on the customer segmentation of an E-Commerce website dataset using the K-Prototype Clustering algorithm and also find out the data distributions in it. We could analyze the different clusters in which customers were divided and based on different parameters. We also used the same data frame to perform clustering with the help of other clustering algorithms mainly – K-Modes, DBSCAN, and Agglomerative Hierarchical Clustering. In the end, we could find a lot of comparisons between the algorithms, the advantages, and the disadvantages. In the end, we feel that the K-Prototype Clustering algorithm being a hybrid can handle mixed data and therefore can handle large data very easily and obtain clusters from it efficiently.

9.2 Future Enhancements

adding more characteristics into it like concentrating on the different dynamic clusters which can help the company in finding out the interests and predict the necessary profits they could make from various sections/categories of products or services that they provide through their e-platform. The current dataset is smaller in size, a larger dataset with huge records could be tested out using the same and find out how it performs.

CHAPTER – 10

TEAM REPORT

10.1 Individual Objective

NAME: SHARAN SUNIL

ROLL NUMBER: 18113045

SECTION & CLASS: CSE 7 A

The objective is to identify the problem statement and divide the project into the required number of modules and identification of relevant papers related to the project.

NAME: T.P MOHANA MAHENDIRA

ROLL NUMBER: 18113032

SECTION & CLASS: CSE 7 A

The objective is to work on identifying the datasets and collecting the information on the algorithms used in this project.

10.2 ROLE

NAME: SHARAN SUNIL

ROLL NUMBER: 18113045

ROLE: Worked on the first module involving clustering using the K-Prototype clustering algorithm, data pre-processing, and also offering helping hands on the other modules as well.

NAME: T.P MOHANA MAHENDIRA

ROLL NUMBER: 18113032

ROLE: Worked on the second module involving clustering using DBSCAN and Hierarchical Agglomerative clustering algorithm, and also involved in identifying the right datasets for this project and related information.

10.3 Contributions

The contributions were equally divided on the basis of modules present in the project with which we could analyze deep into certain domains and take up parts in contributing towards the betterment of the project. Mohana Mahendira handled the implementations containing clustering using DBSCAN and HAC. Responsible for the day-to-day analysis of papers relevant to the project. Find out the different methods used to solve different scenarios of a problem. Sharan handled implementation containing K-Prototype and K-Modes algorithm. During the first few weeks, he handled the dataset distribution parts and analyzed various parts of how to go forward with the project.

REFERENCES

[1]Sneha Pasarate.,Rajashree Shedge(2018).Concept-based document clustering using K prototype Algorithm. International Conference on Control, Power, Communication and Computing Technologies (ICCPCCT)

[2] Dongwei Guo.,Yingjie Chen.,Jingwen Chen(2018). A K-Prototypes Algorithm Based on Adaptive Determination of the Initial Centroids. International Conference on Machine Learning and Computing.

[3] Anil Chaturvedi., Paul E.Green., J.Douglas Carroll(2001). K-modes Clustering. Journal of Classification 18:35-55.

[4]Sumit Koul, Trissa Merrin Philip(2021)Customer Segmentation Techniques on E-Commerce. International Conference on Advanced Computing and Innovative Technologies in Engineering(ICACITE)

[5] Rita Punhani ,V.P.S Arora ,Sai Sabitha ,Vinod Kumar Shukla,(2021)

Application of Clustering Algorithm for Effective Customer Segmentation in E-Commerce, IEEE.

[6] Phan Duy Hung., Nguyen Thi Thuy Lien.,Nguyen Duc Ngoc(2019). Customer Segmentation Using Hierarchical Agglomerative Clustering International Conference on Information Science and Systems.

[7] E-Commerce Data | KaggleSales details from the Indian e-commerce website.https://www.kaggle.com/datasets/benroshan/ecommerce-data. Accessed: 18 Feb 2022.

[8]KModes Clustering Algorithm for Categorical data. https://www.analyticsvidhya.com/blog/2021/06/kmodes-clustering-algorithm-for-categorical-data/.Accessed: 8 Mar 2022.

[9] Customer Segmentation using the k-prototypes algorithm. https://medium.com/analytics-vidhya/customer-segmentation-using-k-prototypes-algorithm-in-python-aad4acbaaedeAccessed: 21 Mar 2022

[10] Precomputed distance matrix in DBSCAN https://stackoverflow.com/questions/62695842/precomputed-distance-matrix-in-dbscanAccessed: 04 Apr 2022

[11]Clustering Distance Measures https://www.datanovia.com/en/lessons/clustering-distance-measures/Accessed: 10 Apr 2022

[12] The k-prototype as Clustering Algorithm for Mixed Data Type (Categorical and Numerical)

https://towardsdatascience.com/the-k-prototype-as-clustering-algorithm-for-mixed-data-type-categorical-and-numerical-fe7c50538ebbAccessed: 10 Apr 2022

[13] K-Prototypes - Customer Clustering with Mixed Data Types https://antonsruberts.github.io/kproto-audience/ Accessed: 15 Mar 2022

APPENDIX A

SAMPLE SCREEN

```
1 df_merged[df_merged['Segment']=='First'].head(20)
```

	Order ID	Order Date	CustomerName	State	City	Amount	Profit	Quantity	Category	Sub-Category	Total	Segment
6	B-25602	01-04-2018	Pearl	Maharashtra	Pune	2617.0	1151.0	4	Electronics	Phones	0	First
36	B-25613	12-04-2018	Mukesh	Haryana	Chandigarh	1603.0	0.0	9	Clothing	Saree	0	First
67	B-25629	24-04-2018	Hemant	Kerala	Thiruvananthapuram	1560.0	421.0	3	Clothing	Trousers	0	First
89	B-25639	27-04-2018	Lisha	Madhya Pradesh	Bhopal	1629.0	-153.0	3	Electronics	Phones	0	First
235	B-25681	04-06-2018	Bhawna	Madhya Pradesh	Indore	1625.0	-77.0	3	Electronics	Phones	0	First
250	B-25686	11-06-2018	Pooja	Himachal Pradesh	Simla	1829.0	-56.0	6	Furniture	Tables	0	First
462	B-25755	19-08-2018	Shourya	Kerala	Thiruvananthapuram	1709.0	564.0	3	Clothing	Trousers	0	First
472	B-25757	21-08-2018	Mohit	Madhya Pradesh	Indore	3151.0	-35.0	7	Clothing	Trousers	0	First
483	B-25761	25-08-2018	Surabhi	Maharashtra	Mumbai	2188.0	1050.0	5	Furniture	Bookcases	0	First
504	B-25768	01-09-2018	Shreyoshe	Karnataka	Bangalore	1582.0	-443.0	6	Clothing	Trousers	0	First
548	B-25786	19-09-2018	Abhishek	Karnataka	Bangalore	1854.0	433.0	5	Furniture	Bookcases	0	First
585	B-25797	30-09-2018	Sauptik	Madhya Pradesh	Indore	1630.0	-802.0	5	Furniture	Tables	0	First
589	B-25798	01-10-2018	Shishu	Andhra Pradesh	Hyderabad	2830.0	-1981.0	13	Furniture	Bookcases	0	First
653	B-25823	18-10-2018	Rohan	Maharashtra	Mumbai	2103.0	322.0	8	Electronics	Electronic Games	0	First
673	B-25830	26-10-2018	Aastha	Himachal Pradesh	Simla	1954.0	782.0	3	Electronics	Phones	0	First
694	B-25842	02-11-2018	Sheetal	Madhya Pradesh	Indore	1543.0	370.0	8	Electronics	Printers	0	First
743	B-25853	08-11-2018	Gaurav	Gujarat	Ahmedabad	2093.0	721.0	5	Furniture	Chairs	0	First
780	B-25858	13-11-2018	Swapnil	Maharashtra	Mumbai	2457.0	665.0	11	Electronics	Electronic Games	0	First
789	B-25862	15-11-2018	Amol	Bihar	Patna	2061.0	701.0	5	Furniture	Bookcases	0	First
834	B-25881	25-11-2018	Pooja	Uttar Pradesh	Allahabad	2244.0	247.0	4	Clothing	Trousers	0	First

:

	Order ID	Order Date	CustomerName	State	City	Amount	Profit	Quantity	Category	Sub-Category	Total	Segment
0	B-25601	01-04-2018	Bharat	Gujarat	Ahmedabad	1275.0	-1148.0	7	Furniture	Bookcases	1	Second
7	B-25602	01-04-2018	Pearl	Maharashtra	Pune	561.0	212.0	3	Clothing	Saree	1	Second
9	B-25603	03-04-2018	Jahan	Madhya Pradesh	Bhopal	1355.0	-60.0	5	Clothing	Trousers	1	Second
22	B-25608	08-04-2018	Aarushi	Tamil Nadu	Chennai	1364.0	-1864.0	5	Furniture	Tables	1	Second
23	B-25608	08-04-2018	Aarushi	Tamil Nadu	Chennai	476.0	0.0	3	Furniture	Chairs	1	Second
25	B-25608	08-04-2018	Aarushi	Tamil Nadu	Chennai	856.0	385.0	6	Electronics	Printers	1	Second
26	B-25609	09-04-2018	Jitesh	Uttar Pradesh	Lucknow	485.0	29.0	4	Electronics	Electronic Games	1	Second
28	B-25610	09-04-2018	Yogesh	Bihar	Patna	1076.0	-38.0	4	Electronics	Printers	1	Second
31	B-25610	09-04-2018	Yogesh	Bihar	Patna	781.0	-594.0	6	Electronics	Printers	1	Second
37	B-25614	13-04-2018	Vandana	Himachal Pradesh	Simla	494.0	54.0	4	Furniture	Bookcases	1	Second
52	B-25622	22-04-2018	Monisha	Rajasthan	Jaipur	534.0	0.0	3	Clothing	Saree	1	Second

APPENDIX B

SAMPLE CODE

```
import import_ipynb
#importing Numpy and pandas
import numpy as np
import pandas as pd
# For visualizations
import matplotlib.pyplot as plt
%matplotlib inline
# For regular expressions
import re
# For handling string
import string
# For performing mathematical operations
import math
#data visualization library based on matplotlib.
import seaborn as sns
import sklearn
from sklearn import metrics
from sklearn.datasets import make_blobs
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.preprocessing import MinMaxScaler
from sklearn.cluster import Kmeans
```

```
list_of_orders= pd.read_csv('list of orders.csv')
list_of_orders
list_of_orders.head()
order_details=pd.read_csv('order details.csv')
order_details
print(f"Missing values in each variable: \n{list_of_orders.isnull().sum()}")
print(f"Duplicated rows: {list_of_orders.duplicated().sum()}")
list_of_orders.drop_duplicates()
list_of_orders.drop_duplicates(inplace=True)
print(f"Duplicated rows: {order_details.duplicated().sum()}")
df_merged = pd.merge(list_of_orders,order_details,how='inner')
df_merged
plt.figure(figsize=(10, 6))
sns.set(style = 'whitegrid')
sns.histplot(order_details['Category'])
plt.title('Distribution of "widely sought after" categories', fontsize = 20)
plt.xlabel('Categories')
plt.ylabel('Count')
plt.figure(figsize=(10, 6))
sns.set(style = 'whitegrid')
sns.histplot(order_details['Sub-Category'])
plt.title('Distribution of "accessed/shopped" sub categories', fontsize = 20)
plt.xlabel('Sub-Categories')
plt.ylabel('Count')
ax = sns.countplot(x="Sub-Category", data=order_details)
ax.set_xticklabels(ax.get_xticklabels(), fontsize=12,rotation=40, ha="right")
plt.show()
```

APPENDIX C

PLAGIARISM REPORT

Curiginal

Document Information

Analyzed document	REPORT18.pdf (D134530034)
Submitted	2022-04-25T05:08:00.0000000
Submitted by	
Submitter email	slakshmi@hindustanuniv.ac.in
Similarity	3%
Analysis address	slakshmi.hits@analysis.urkund.com

Sources included in the report

W	URL: https://www.hindawi.com/journals/mpe/2020/5143797/ Fetched: 2020-11-25T08:44:12.0000000	2
W	URL: https://antonsruberts.github.io/kproto-audience/ Fetched: 2022-04-25T05:08:00.0000000	1
W	URL: https://towardsdatascience.com/the-k-prototype-as-clustering-algorithm-for-mixed-data-type-categorical-and-numerical-fe7c50538ebb Fetched: 2022-04-25T05:08:00.0000000	6
W	URL: https://www.analyticsvidhya.com/blog/2021/06/kmodes-clustering-algorithm-for-categorical-data/ Fetched: 2022-04-25T05:08:00.0000000	1

1/11

APPENDIX D

PUBLICATIONS DETAILS

ICIRET

IFERP
connecting engineers... developing research

Ref No : 2429

Date : 25/04/2022

Conference Secretariat - Vietnam

Letter of Acceptance

2nd International Conference on Innovative Research in Engineering and Technology (ICIRET-22)

15th & 16th July 2022 | Vietnam

Abstract ID : ICIRET-2022_VIE_0078

Paper Title : E-COMMERCE USER SEGMENTATION

Author Name : SHARAN SUNIL

Co-Author Name : T.P MOHANA MAHENDIRA, S.SATHYALAKSHMI

Institution : HINDUSTAN INSTITUTE OF TECHNOLOGY AND SCIENCE, CHENNAI

Hello SHARAN SUNIL

Congratulations!!!

The scientific reviewing committee is pleased to inform your article "E-COMMERCE USER SEGMENTATION" is accepted for Oral/Poster Presentation at "ICIRET-2022" On 15th – 16th July 2022 at Vietnam. The Paper has been accepted after our double-blind peer review process and plagiarism check.

Your paper will be published in relevant Scopus Indexed Journals as well as in the Conference Proceedings.

Authors are recommended to proceed for registration to confirm their slots in relevant scientific sessions by following the link given.

https://iciret.net/conference-registration.php

For further more details and other affiliated journals feel free to contact us to: info@iciret.net

Registration Guidelines : https://iciret.net/conference-registration-guidelines.php

APPENDIX E

TEAM DETAILS

NAME	CONTACT NO.	MAIL ID	ROLE
Dr. S.SATHYALAKSHMI	9884745667	slakshmi@hindustanuniv.ac.in	Supervisor

NAME	ROLL NO.	CONTACT NO.	MAIL ID	ROLE
SHARAN SUNIL	18113045	8921374602	18113045@student.hindustanuniv.ac.in	Team Member
T.P MOHANA MAHENDIRA	18113032	9715005766	18113032@student.hindustanuniv.ac.in	Team Member