Clustering for Customer Segmentation & Understanding

Architecture

[Pick the date]

Contents

Abstract

Not all customers are the same. To know which group is your customer and their Preferences are a big part of success in your business. Unsupervised machine learning can help marketers know their audience globally and engage them with their products accordingly. Here, we can classify millions of people's interests through their social media activity and also through other surveys, online and offline, and cluster them into a specific group of their interest.

Introduction

What is Low-Level design document?

The goal of LLD or a low-level design document (LLDD) is to give the internal logical design of the actual program code for Clustering-for-Customer-Segmentation-Understanding. LLD describes the class diagrams with the methods and relations between classes and program specs. It describes the modules so that the programmer can directly code the program from the document.

Scope

Low-level design (LLD) is a component-level design process that follows a step-by step refinement process. This process can be used for designing data structures, required software architecture, source code and ultimately, performance algorithms. Overall, the data organization may be defined during requirement analysis and then refined during data design work

Constraints

It requires specifying the number of clusters (k) in advance. It cannot handle noisy data and outliers. It is not suitable to identify clusters with non-convex shapes.

Architecture Description

Data Description:

The sample dataset summarizes the usage behaviour of about 9,000 active credit cards. Holders during the last 6 months. The file is at the customer level, with 18 behavioural variables.

- Variables of the Dataset
- Balance
- Balance Frequency
- Purchases
- One-off Purchases
- Instalment Purchases
- Cash Advance
- Purchases Frequency
- One-off Purchases Frequency
- Purchases Instalment Frequency
- Cash Advance Frequency
- Cash Advance
- TRX Purchases
- TRX Credit Limit
- Payments
- Minimum Payments
- PRC
- Full Payment
- Tenure Cluster

From this dataset, we need to calculate some patterns, as it is an unsupervised method, so we don't know what to calculate exactly.

Dataset overview:

The sample dataset summarizes the usage behaviour of about 9,000 active credit cards. Holders during the last 6 months. The file is at the customer level, with 18 behavioural variables.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	CUST_ID	BALANCE	BALANCE_	PURCHASE	ONEOFF_I	INSTALLM	CASH_ADV	PURCHASE	ONEOFF_I	PURCHASE	CASH_ADV	CASH_ADV	PURCHASE	CREDIT_LI	PAYMENT	MINIMUM	PRC_FULL_	TENURE
2	C10001	40.90075	0.818182	95.4	0	95.4	0	0.166667	0	0.083333	0	0	2	1000	201.8021	139.5098	0	12
3	C10002	3202.467	0.909091	0	0	0	6442.945	0	0	0	0.25	4	0	7000	4103.033	1072.34	0.222222	12
4	C10003	2495.149	1	773.17	773.17	0	0	1	1	0	0	0	12	7500	622.0667	627.2848	0	12
5	C10004	1666.671	0.636364	1499	1499	0	205.788	0.083333	0.083333	0	0.083333	1	1	7500	0		0	12
6	C10005	817.7143	1	16	16	0	0	0.083333	0.083333	0	0	0	1	1200	678.3348	244.7912	0	12
7	C10006	1809.829	1	1333.28	0	1333.28	0	0.666667	0	0.583333	0	0	8	1800	1400.058	2407.246	0	12
8	C10007	627.2608	1	7091.01	6402.63	688.38	0	1	1	1	0	0	64	13500	6354.314	198.0659	1	12
9	C10008	1823.653	1	436.2	0	436.2	0	1	0	1	0	0	12	2300	679.0651	532.034	0	12
10	C10009	1014.926	1	861.49	661.49	200	0	0.333333	0.083333	0.25	0	0	5	7000	688.2786	311.9634	0	12
11	C10010	152.226	0.545455	1281.6	1281.6	0	0	0.166667	0.166667	0	0	0	3	11000	1164.771	100.3023	0	12
12	C10011	1293.125	1	920.12	0	920.12	0	1	0	1	0	0	12	1200	1083.301	2172.698	0	12
13	C10012	630.7947	0.818182	1492.18	1492.18	0	0	0.25	0.25	0	0	0	6	2000	705.6186	155.5491	0	12
14	C10013	1516.929	1	3217.99	2500.23	717.76	0	1	0.25	0.916667	0	0	26	3000	608.2637	490.207	0.25	12
15	C10014	921.6934	1	2137.93	419.96	1717.97	0	0.75	0.166667	0.75	0	0	26	7500	1655.891	251.138	0.083333	12
16	C10015	2772.773	1	0	0	0	346.8114	0	0	0	0.083333	1	0	3000	805.648	989.9629	0	12
17	C10016	6886.213	1	1611.7	0	1611.7	2301.491	0.5	0	0.5	0.166667	4	11	8000	1993.439	2109.906	0	12
18	C10017	2072.074	0.875	0	0	0	2784.275	0	0	0	0.25	3	0	3000	391.9746	376.5796	0	8

Dataset overview:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8950 entries, 0 to 8949
Data columns (total 18 columns):
 #   Column                            Non-Null Count  Dtype  
---  ------                            --------------  -----  
 0   CUST_ID                           8950 non-null   object 
 1   BALANCE                           8950 non-null   float64
 2   BALANCE_FREQUENCY                 8950 non-null   float64
 3   PURCHASES                         8950 non-null   float64
 4   ONEOFF_PURCHASES                  8950 non-null   float64
 5   INSTALLMENTS_PURCHASES            8950 non-null   float64
 6   CASH_ADVANCE                      8950 non-null   float64
 7   PURCHASES_FREQUENCY               8950 non-null   float64
 8   ONEOFF_PURCHASES_FREQUENCY        8950 non-null   float64
 9   PURCHASES_INSTALLMENTS_FREQUENCY  8950 non-null   float64
 10  CASH_ADVANCE_FREQUENCY            8950 non-null   float64
 11  CASH_ADVANCE_TRX                  8950 non-null   int64  
 12  PURCHASES_TRX                     8950 non-null   int64  
 13  CREDIT_LIMIT                      8949 non-null   float64
 14  PAYMENTS                          8950 non-null   float64
 15  MINIMUM_PAYMENTS                  8637 non-null   float64
 16  PRC_FULL_PAYMENT                  8950 non-null   float64
 17  TENURE                            8950 non-null   int64  
dtypes: float64(14), int64(3), object(1)
```

Predicting:

The goal of cluster analysis in marketing is to accurately segment customers in order to achieve more effective customer marketing via personalization. A common cluster analysis method is a mathematical algorithm known as *k-means cluster analysis*, sometimes referred to as scientific segmentation. The clusters that result assist in better customer modelling and predictive analysis and are also used to target customers with offers and incentives personalized to their wants, needs, and preferences.

Deployment

Building the **Streamlit App**. Creating the Streamlit UI. Loading the Saved Model & Making Real-Time Predictions. Deploying Machine Learning Models with Python and Streamlit.

Streamlit

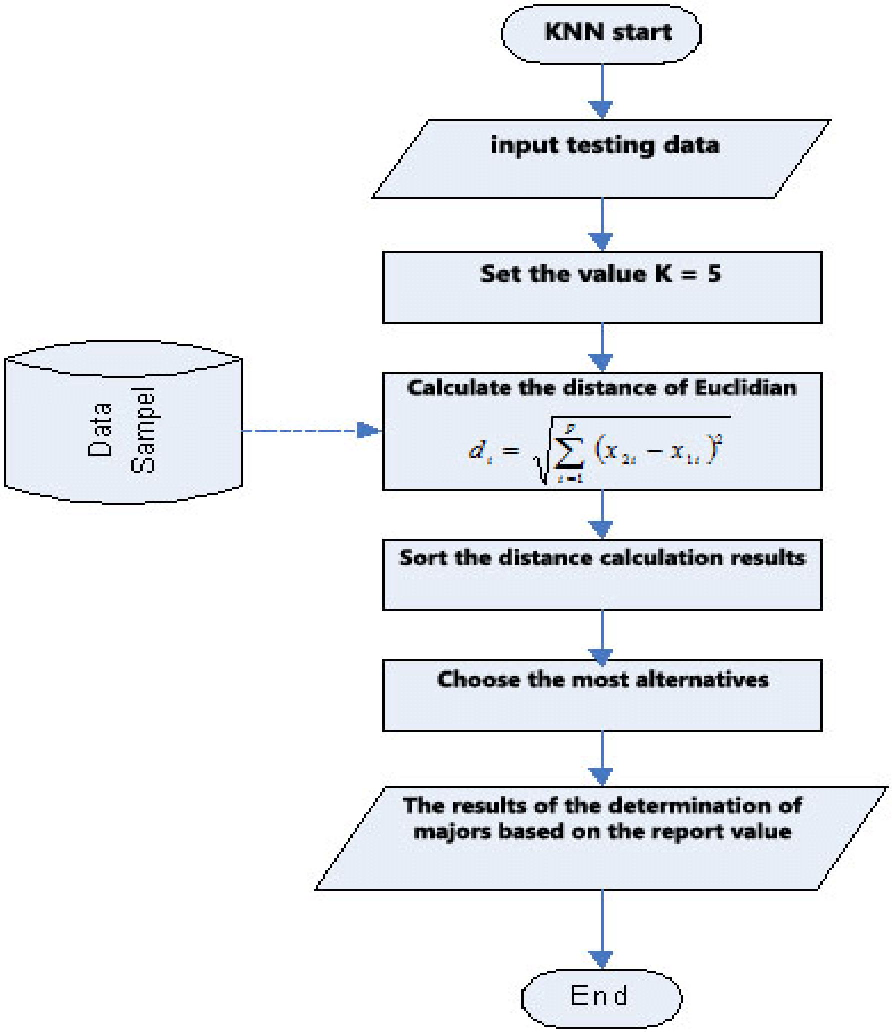
PROPOSED SOLUTION:

In the context of customer segmentation, customer clustering analysis is the use of a mathematical model to discover groups of similar customers based on finding the smallest variations among customers within each group. These homogeneous groups are known as "customer archetypes" or "personas".

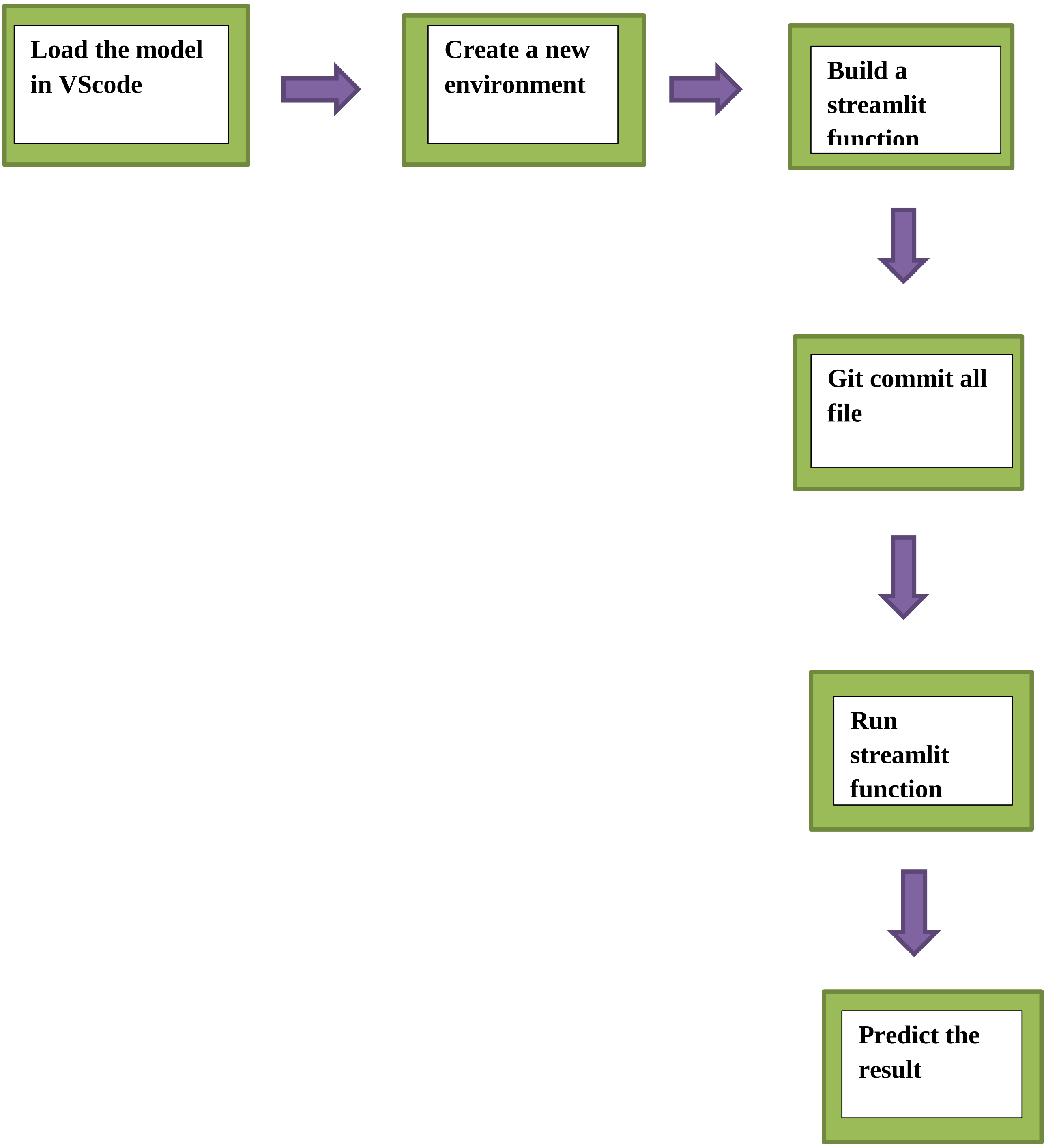
The goal of cluster analysis in marketing is to accurately segment customers in order to achieve more effective customer marketing via personalization. A common cluster analysis method is a mathematical algorithm known as *k-means cluster analysis*, sometimes referred to as scientific segmentation. The clusters that result assist in better customer modelling and predictive analysis and are also used to target customers with offers and incentives personalized to their wants, needs, and preferences.

Process Flow

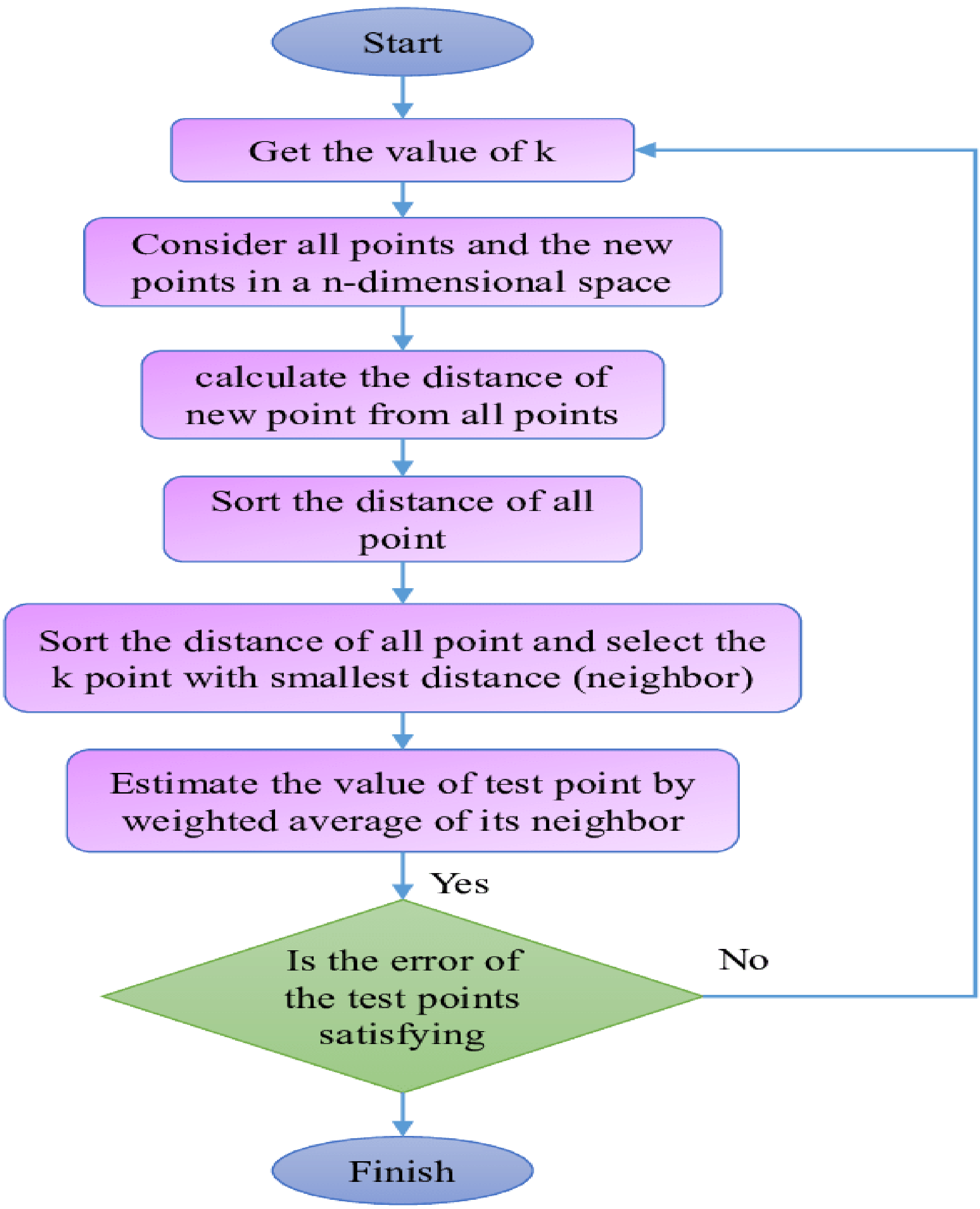
The flow chart for the KNN algorithm given below



Deployment Process



Architecture:



Key performance indicator

- Simple and easy to implement: The k-means algorithm is easy to understand and implement, making it a popular choice for clustering tasks.
- Fast and efficient: K-means is computationally efficient and can handle large datasets with high dimensionality.
- Scalability: K-means can handle large datasets with a large number of data points and can be easily scaled to handle even larger datasets.
- Flexibility: K-means can be easily adapted to different applications and can be used with different distance metrics and initialization methods.