IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, June 2024

Applying K Means Clustering Techniques on Retail Shop

Chandra Mohini C.P[1] and V. Raghavendran[2]
Research Scholar, Department of Computer Science[1]
Assistant Professor Department of Information Technology[2]
Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, Tamil Nadu, India
mohinisuryan@gmail.com, raganand78in@gmail.com

Abstract: *Implement a customer segmentation system using k-mean clustering for a Retail business. Collect Data and pre-processing to clear missing data and inconsistence data and then visualization to analysis(EDA), apply K-Means to divide customer into distinct groups based on purchasing behaviour. Analyse and create customer profiles for each segment to optimize marketing strategies and enhance customer Engagement. Main goal of in this project on price optimization, Increase revenue, brand awareness and increase relationship between customer and enhance marketing Strategies and increasing sales.*

Keywords: EDA

I. INTRODUCTION

In the highly competitive world of retail, understanding customer behaviour and preferences is crucial for business success. Customer segmentation is a data-driven technique that allows retailers to group customers with similar characteristics and behaviours. These segments can then be targeted with tailored marketing strategies, leading to improved customer satisfaction and increased sales.

One of the powerful methods for customer segmentation is K-Means clustering. K-Means is an unsupervised machine learning algorithm that identifies natural groupings within large datasets. When applied to retail shopping details, it can reveal distinct customer segments based on purchasing patterns, demographic information, and other factors.This research aims to leverage K-Means clustering to uncover valuable insights from retail shopping data. By dividing customers into meaningful segments, retailers can:

- Optimize Inventory: Understand which products are popular among specific segments, helping in stock management and product recommendations
- Enhance Customer Experience: Deliver a more personalized shopping experience, such as customized product recommendations
- Improve Customer Retention: Recognize high-value customers and implement strategies to retain them.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, June 2024

Flow Diagram on Customer Segmentation Using K-Means Techniques

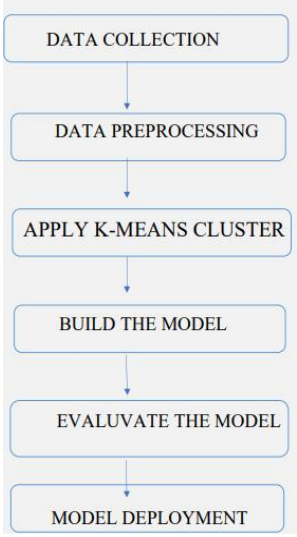


Figure 1 – Flow diagram

Customer segmentation:-

Customer segmentation is the process of dividing a company's customer base into distinct and homogeneous groups or segments based on shared characteristics or behaviours. The goal of customer segmentation is to gain a better understanding of a company's customers and to tailor marketing strategies, products, and services to the specific needs and preferences of each segment. By treating different customer segments differently, businesses can improve customer satisfaction, increase sales, and enhance overall customer experiences.

Types of Customer Segmentation:

• Demographic Segmentation:

- Age: Grouping customers by age ranges (e.g., teens, young adults, seniors). Gender: Segmenting customers based on their gender.
- Income: Dividing customers into income brackets (e.g., lowincome, middle-income, high-income).
- Education: Categorizing customers by their educational attainment.

• Geographic Segmentation:

- Location: Segmenting customers by geographic regions, such as countries, states, cities, or neighborhoods.
- Urban vsRural: Distinguishing between customers in urban and rural areas. Climate: Segmenting based on climate zones or weather patterns.

• Psychographic Segmentation:

- Lifestyle: Grouping customers by their lifestyle choices, interests, and values.
- Personality: Segmenting based on personality traits and behaviors.
- Opinions and Attitudes: Categorizing customers by their opinions and attitudes towards specific issues or products.

• Behavioral Segmentation:

- Purchase Behavior: Segmenting based on how frequently and how much customers buy.
- Usage Rate: Categorizing customers by how often they use a product or service.
- Brand Loyalty: Distinguishing between loyal, occasional, and disloyal customers.
- Occasion-based: Segmenting based on specific occasions or events when customers make purchases.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, June 2024

Demographic customer segmentation

• Demographic customer segmentation is a method of categorizing and dividing a customer base into distinct groups based on specific demographic characteristics or attributes. Demographic factors are quantifiable and straightforward, making them one of the most common and easily accessible ways to segment customers. These attributes help businesses understand the basic characteristics of their customers and tailor marketing strategies accordingly.

SCHEMATIC DIAGRAM TOCUSTOMER SEGEMENTATION USING K-MEANS CLUSTER

TRAIN AND TEST THE MODEL:-

The scikit-learn library (commonly referred to as sklearn), you can follow these general steps. Scikit-learn is a popular Python library for machine learning tasks.

Assuming you have your data loaded into arrays or DataFrames (X for features and y for labels), here's how you can create a train-test split and train a machine learning model

Import Necessary Libraries

You need to import the required libraries for your machine learning project. Common libraries include scikit-learn for machine learning tools, NumPy for numerical operations, and pandas for data handling

Load and Prepare Your Data:

Load your dataset and prepare it by separating the features (X) an the target variable (y).

Ensure that your data is in a format suitable for machine learning

Split Data into Training and Testing model:

Split your data into training set and a testing set. This is typically done using the train_test_split function. In this example, 20% of the data is reserved for testing, and random_state ensures reproducibility.

Define and Train Your Model:

Choose a machine learning model and initialize it. Then, train the model on the training data

Make Predictions on the Test Data:

Use the trained model to make predictions on the test data

Evaluate Model Performance:

Use appropriate metrics to evaluate your model's performance. Classification tasks, you can use accuracy, precision, recall,F1- score, etc.

DECISION TREE CLASSIFIER TO FIND ACCURACY SCORE:-

A Decision Tree Classifier is a popular machine learning algorithm used for both classification and regression tasks. It's a type of supervised learning algorithm that creates a tree-like model of decisions and their possible consequences. In the case of classification, it assigns a class label to an input based on a series of decisions, and in the case of regression, it predicts a continuous target variable. Here's a more detailed definition:

Decision Tree Classifier:

Tree Structure: A decision tree is a tree-like structure where each internal node represents a decision or test on a feature (attribute), each branch represents an outcome of the decision, and each leaf node represents a class label (in classification) or a numerical value (in regression).

Decision-Making Process:

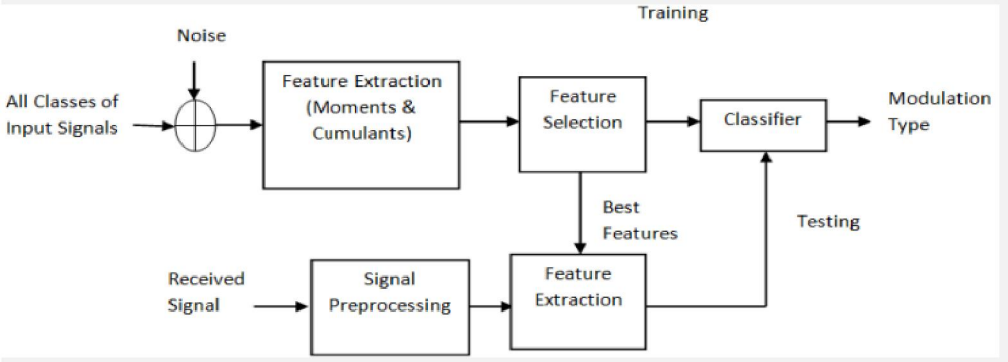
To make a prediction, the algorithm starts at the root node and moves down the tree based on the outcome of each decision until it reaches a leaf node. The path taken through the tree represents a series of decisions that lead to a final prediction.

Prediction:

Once the tree is constructed, it can be used to make predictions on new, unseen data by traversing the tree from the root to a leaf node.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, June 2024

DECISION TREE CLASIFIER MODEL UML DAIGRAM



KEY ADVANTAGES:

- Easy to interpret: Decision trees can be visualized, making it easy to understand the decision-making process.
- Can handle both categorical and numerical features.
- Non-linear relationships between features and the target variable can be captured.
- Can handle missing data and outliers.

KEY DISADVANTAGES:

- Prone to overfitting, especially if the tree is deep.
- Can be unstable, as small changes in the data can lead to different tree structures.
- Limited in expressing complex relationships between features.
- Not suitable for tasks where the decision boundary is highly irregular.
- In Decision Tree Classifiers are often used as part of ensemble methods, like Random Forests or Gradient Boosting, to improve their predictive performance and reduce overfitting. Decision trees can be a valuable tool in machine learning, but they should be used in combination with other techniques and with careful consideration of hyperparameters to achieve a specific problem.

DECISION TREE CLASSIFIER RESULTS:-

Accuracy :

$$\frac{TP + TN}{TP + FP + TN + FN}$$

The most commonly used metric to judge a model is actually not a clear indicator of the performance. The worst happens when classes are imbalanced.

Precision:

$$\frac{TP}{TP + FP}$$

Percentage of positive instances out of the total predicted positive instances.
ACCURACY SCORE :0.8601950532877539 % Accuracy

DECISION TREE CLASSIFIER RESULTS:

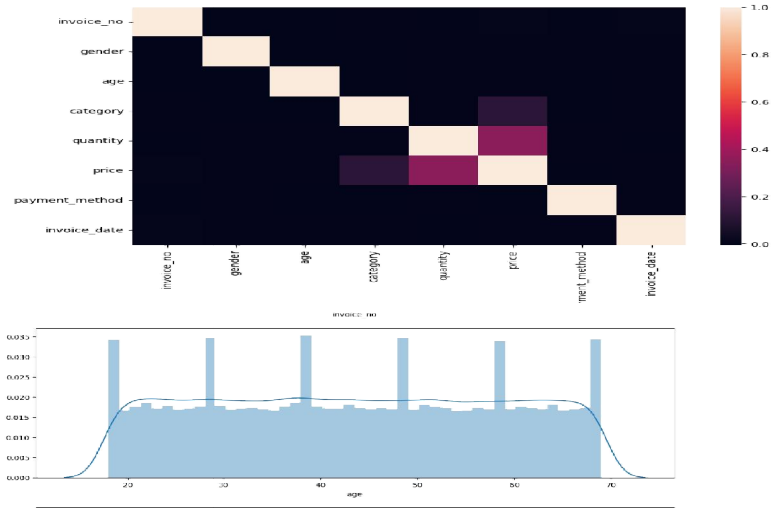
FINALLY SAVE THE MODEL USING JOBLIB AND PICKLE:-

Joblib is a Python library, often used in the context of machine learning and scientific computing, that provides tools for efficiently saving and loading Python objects, especially NumPy arrays and scikit-learn machine learning models. It is similar in purpose to the built-in Python pickle module but has some key differences and advantages, particularly when working with large data arrays or complex machine learning models.

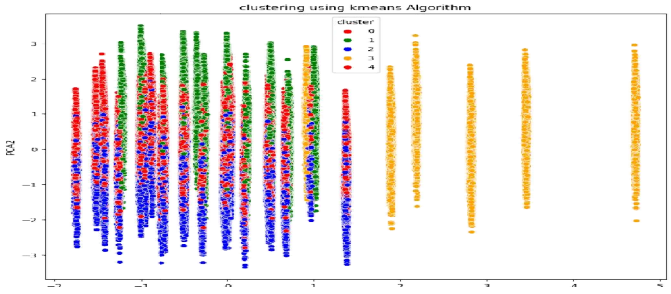
IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, June 2024

Key characteristics and functions of Joblib

- Efficiency: Joblib is optimized for efficiently storing large data, particularly NumPy arrays. It uses memory-mapping to reduce memory consumption and is designed to work well with numerical data.
- Parallel Processing: Joblib can leverage parallel processing, which can be helpful for certain operations, such as parallelizing model training or data processing.
- Scikit-Learn Integration: Joblib is often used in combination with the scikit-learn machine learning library. When you save a scikit-learn model using Joblib, it can save additional metadata and structures, making it easier to load and use the model in other Python environments.
- Pickle is a Python module that is used for serializing (pickling) and deserializing (unpickling) Python objects. Serialization is the process of converting a Python object into a byte stream, and deserialization is the process of reconstructing a Python object from a byte stream



K-Means_model=K-Means(5) K-Means_model.fit_predict(scaled_df) pca_data_K-Means=pd.concat([pca_df,pd.DataFrame({'cluster':K-Means_model.labels_})],axis=1) plt.figure(figsize=(10,8)) ax=sns.scatterplot(x="PCA1",y="PCA2",hue="cluster",data=pca_data_K-Means,palette=['red','green','blue','orange']) plt.title("clustering using K-Means Algorithm") plt.show()



cluster_centers = pd.DataFrame(data=K-Means_model.cluster_centers_,columns=[df.columns]) cluster_centers = scalar.inverse_transform(cluster_centers) cluster_centers = pd.DataFrame(data=cluster_centers,columns=[df.columns]) cluster_centers

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, June 2024

	invoice_no	gender	age	category	quantity	price	payment_method	invoice_date
49701.552559	-3.091971e-14	43.251298	1.816891	3.448077	599.374636	1.509060	397.549221	
1	49838.316103	4.094786e-01	43.332521	6.356955	0.714487	396.428846	2.855769	145.967560
2	49578.649641	1.000000e+00	43.435082	1.764747	0.752060	398.722995	2.926679	539.417295
3	49768.326249	4.031817e-01	43.481357	4.997017	0.749441	398.872359	4.000621	3185.767588
4	49840.291563	-3.258505e-14	43.581735	1.787728	0.248665	399.782502	2.560380	487.286774

```
cluster_df=pd.concat([df,pd.DataFrame({'cluster':K-
Means_model.labels_})],axis=1) cluster_df
```

	invoice_no	gender	age	category	quantity	price	payment_method	invoice_date	cluster
0	12165.0	0.0	28.0	1.0	5.0	1500.40	1.0	687.0	0
67813.0	1.0	21.0	4.0	3.0	1800.51	2.0	86.0	2	
8722.0	1.0	20.0	1.0	1.0	300.08	0.0	776.0	2	
23002.0	0.0	66.0	4.0	5.0	3000.85	1.0	193.0	3	
74054.0	0.0	53.0	0.0	4.0	60.60	0.0	438.0	4	
...	...	...	...	...	...	...	...	...	...
37202.0	0.0	45.0	5.0	5.0	58.65	1.0	359.0	1	
70336.0	1.0	27.0	3.0	2.0	10.46	0.0	384.0	2	
93050.0	1.0	63.0	3.0	2.0	10.46	2.0	528.0	2	
88554.0	1.0	56.0	6.0	4.0	4200.00	0.0	189.0	3	
41383.0	0.0	36.0	5.0	3.0	35.19	1.0	178.0	1	

rows × 9 columns

DEPLOY THE MODEL USING STREAMLIT:-

Streamlit is an open-source Python library that is used for creating web applications with minimal effort. It is designed to be simple, efficient, and user-friendly, making it a popular choice for building data-driven web applications and interactive dashboards without the need for extensive web development knowledge.

Rapid Prototyping:

Streamlit enables you to create web apps and interactive dashboards quickly by turning Python scripts into shareable web applications. You can focus on the functionality and data visualization, and Streamlit handles the web interface.

Simplicity:

Streamlit is known for its simplicity and ease of use. It has a clean and straightforward Python API, and you can get started with just a few lines of code.

Data Visualization:

You can easily integrate data visualizations, plots, and charts into your web applications using popular Python libraries such as Matplotlib, Plotly, and Altair.

Widgets:

Streamlit provides a wide range of widgets (input elements) like sliders, text input, buttons, and select boxes, allowing users to interact with your application.

Customization:

While Streamlit is beginner-friendly, it also allows for customization and more advanced features for those who need them.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, June 2024

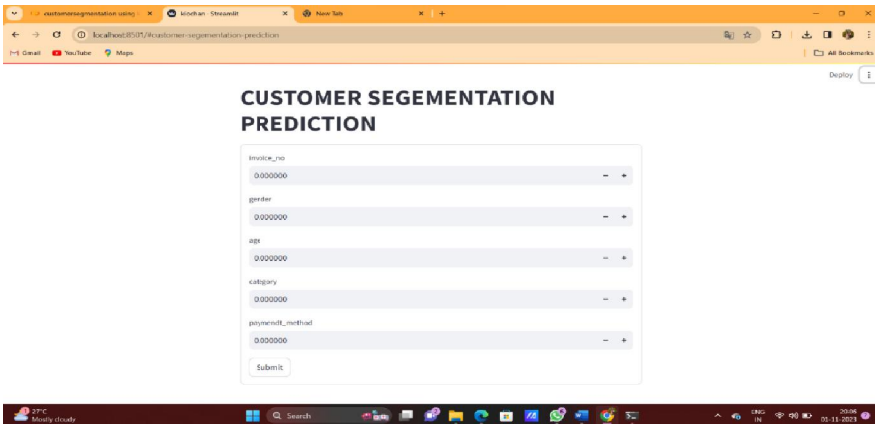
Deployment:

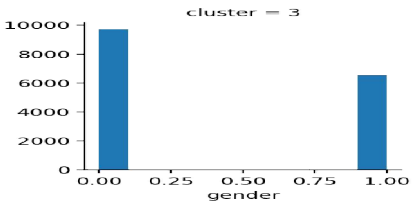
Streamlit apps can be deployed on various platforms, including Streamlit sharing, Heroku, AWS, and more. This makes it easy to share your apps with others.

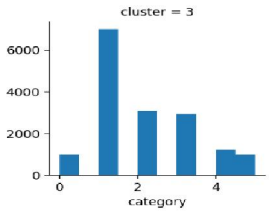
Integration:

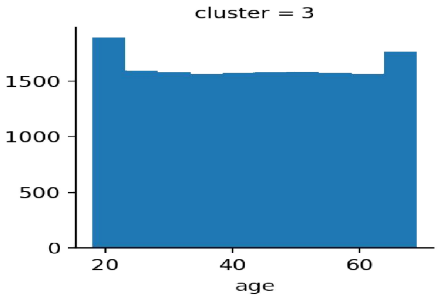
Streamlit can be integrated with data science and machine learning workflows, enabling you to create interactive data exploration tools or deploy machine learning models as web applications.

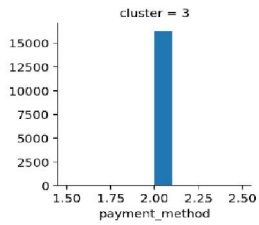
OUTPUT RESULT











II. CONCLUSION

Customer segmentation using K-Means clustering is a valuable technique in marketing and business analysis, and it provides insights that can guide various aspects of a company's strategy. Here's a conclusion to customer segmentation using K-Means clustering

Identifying Distinct Customer Groups

K-Means clustering helps identify natural groupings within a customer dataset. These groups may not be obvious from raw data but are revealed through the clustering process.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, June 2024

Improved Targeted Marketing:

Segmenting customers allows for tailored marketing strategies. Companies can create marketing campaigns that are specific to each segment's preferences and behaviors, leading to higher engagement and conversion rates.

Product and Service Customization:

Understanding customer segments enables businesses to customize their products or services to better meet the needs and desires of each group. This personalization can enhance customer satisfaction and loyalty.

Pricing Strategies:

Different customer segments may be willing to pay different prices for products or services. Segment-specific pricing strategies can maximize revenue without alienating customers.

Customer Retention :

Companies can use customer segmentation to identify at-risk customers and take proactive steps to prevent them from leaving. For example, they can offer special promotions or incentives to retain valuable customers.

The customer shopping previous data to analysis the customer likelihood products and grouping the customer on k means technique on customer age and purchase date, gender to analysis to grouping the customer.

REFERENCES

[1]. "Market Segmentation: A Review" by Paul Smith. (Journal of Marketing Management, 1956) - This seminar paper lays the foundation for modern market segmentation and provides a historical perspective on theconcept.

[2]. "Market Segmentation: Conceptual and Methodological Foundations" by Wagner A. Kamakura and Michel Wedel. (International Journal of Research in Marketing, 1997) - This paper delves into the conceptual and methodological aspects of market segmentation, providing an in-depth understanding of the process.

[3]. "Customer Segmentation and Clustering Using SAS Enterprise Miner" by Randall S. Collica. (Book, 2014) - This book covers practical approaches to customer segmentation using SAS Enterprise Miner, offering insights into the application of data analytics in segmentation.

[4]. "Customer Segmentation and Clustering Using Unsupervised Learning" by Deepak Arora. (Book, 2018) - This book focuses on customer segmentation using unsupervised learning techniques and provides a comprehensive guide to data-driven segmentation methods.

[5]. "A Review of Customer Relationship Management: Success Factors & Benefits" by Hussain Al-Hawari, T. Y. Hartley, and Moheeb Abu Saud. (Journal of Database Marketing & Customer Strategy Management, 2005) -This paper explores the relationship between customer relationship management (CRM) and customer segmentation, highlighting success factors and benefits.

[6]. "Customer Segmentation in Digital Marketing: Challenges and Future Research Directions" by ShamaNazSadaat, KhairilShazminKamaruddin, and NorkhairunnisaMazlan. (Telematics and Informatics, 2020) - This article discusses the challenges and future research directions in customer segmentation within the context of digital marketing.

[7]. "A Framework for Customer Relationship Management" by Richard J. Lusch and Patricia A. Vargo. (California Management Review, 2006) - This paper introduces a framework for understanding customer relationships and segmentation as part of relationship marketing.

[8]. "Customer Segmentation Based on Purchase Sequences" by Peter S. Fader and Bruce G.S. Hardie. (Marketing Science, 2007) - This research paper discusses a unique approach to customer segmentation using purchase sequences, a valuable method for certain industries.

[9]. "Customer Segmentation by Service Firms: Is It Worth It?" by Arthur J. Hughes and Amy L. Smith. (Journal of Marketing Research, 2003) - This paper explores the effectiveness of customer segmentation for service firms and discusses the potential benefits.

[10]. "Market Segmentation" by Philip Kotler, Gary Armstrong, and Veronica Wong. (Principles of Marketing, 2010) - This chapter from a widely-used marketing textbook provides a comprehensive overview of market segmentation as a core marketing concept.