A PROJECT REPORT ON

Customer Segmentation using Unsupervised Machine Learning

Submitted in partial fulfillment of requirements
for the award of the degree of

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING

Submitted by:

D. V. Sathya Vardhan Reddy	**(20091A05D9)**
S. Afifa	**(20091A0504)**
M. Radhakrishna	**(20091A0505)**
R. Aishwarya	**(20091A0508)**

Under the Guidance of

Dr. R. Kaviarasan., M.Tech., Ph.D.,

Associate Professor, Dept. of CSE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

RAJEEV GANDHI MEMORIAL COLLEGE OF ENGINEERING & TECHNOLOGY
(AUTONOMOUS)
Approved by AICTE, New Delhi; Affiliated to JNTUA-Ananthapuramu,
Accredited by NBA (6-Times); Accredited by NAAC with 'A+' Grade (Cycle-3), New Delhi;
World Bank Funded Institution; Nandyal (Dist)-518501, A.P
(Estd-1995)

YEAR: 2023-2024

Rajeev Gandhi Memorial College of Engineering &Technology

(AUTONOMOUS)

Approved by AICTE, New Delhi; Affiliated to JNTUA-Ananthapuramu,
Accredited by NBA (6-Times); Accredited by NAAC with 'A+' Grade (Cycle-3), New Delhi;
World Bank Funded Institution; Nandyal (Dist)-518501, A.P

(ESTD – 1995)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that **D. SATHYA VARDHAN REDDY** (*20091A05D9*), **S. AFIFA** (*20091A0504*), **M. RADHAKRISHNA** (*21095A0508*) and **R. AISHWARYA** (*20091A0505*) of IV- B. Tech II- semester, have carried out the major project work entitled "**CUSTOMER SEGMENTATION USING UNSUPERVISED MACHINE LEARNING**" under the supervision and guidance of **Dr. R. KAVIARASAN,** Associate Professor, CSE Department, in partial fulfillment of the requirements for the award of Degree of **Bachelor of Technology** in **Computer Science and Engineering** from **Rajeev Gandhi Memorial College of Engineering & Technology (Autonomous)**, Nandyal is a bonafied record of the work done by them during 2023-2024.

Project Guide

Dr. R. Kaviarasan M.Tech., Ph.D.

Associate Professor, Dept. of CSE

Head of the Department

Dr. K. Subba Reddy M.Tech., Ph.D.

Professor, Dept. of CSE

Place: Nandyal
Date:

External Examiner

Candidate's Declaration

We hereby declare that that the work done in this project entitled **"CUSTOMER SEGMENTATION USING UNSUPERVISED MACHINE LEARNING"** submitted towards completion of major project in *IV Year II Semester of B. Tech (CSE)* at the **Rajeev Gandhi Memorial College of Engineering & Technology**, Nandyal. It is an authentic record our original work done under the esteemed guidance of **Dr. R. Kaviarasan,** Associate Professor, Department of **Computer Science and Engineering**, RGMCET, Nandyal.

We have not submitted the matter embodied in this report for the award of any other Degree in any other institutions for the academic year 2023-2024.

By

(D. Sathya vardhan redddy)

(S. Afifa)

(M. Radhakrishna)

(R. Aishwarya)

Dept. of CSE,
RGMCET.

Place: Nandyal
Date:

ACKNOWLEDGEMENT

We manifest our heartier thankfulness pertaining to your contentment over our project guide **Dr. R. Kaviarasan,** Associate Professor of Computer Science and Engineering Department, with whose adroit concomitance the excellence has been exemplified in bringing out this project to work with artistry.

We express our gratitude to **Dr. K. Subba Reddy,** Head of the Department of Computer Science Engineering department, all the **Teaching Staff Members** of the Computer Science and Engineering department of Rajeev Gandhi memorial College of Engineering and Technology for providing continuous encouragement and cooperation at various steps of our project successful.

Involuntarily, we are perspicuous to divulge our sincere gratefulness to our Principal, **Dr. T. Jaya Chandra Prasad**, who has been observed posing valiance in abundance towards our individuality to acknowledge our project work tangentially.

At the outset we thank our honourable **Chairman Dr. M. Santhi Ramudu,** for providing us with exceptional faculty and moral support throughout the course.

Finally we extend our sincere thanks to all the non- teaching **Staff Members** of CSE Department who have co-operated and encouraged us in making our project successful.

Whatever one does, whatever one achieves, the first credit goes to the **Parents** be it not for their love and affection, nothing would have been responsible. We see in every good that happens to us their love and blessings.

BY

D.V. Sathya Vardhan Reddy	(20091A05D9)
S. Afifa	(20091A0504)
M.Radhakrishna	(21095A0508)
R.Aishwarya	(20091A0505)

ABSTRACT

In today's world, companies view data as the new money.Customer data is crucial for measuring their worth to a business, particularly in marketing. Customer segmentation categorizes clients based on comparable behaviors and purchase patterns. Customer segmentation helps marketing firms identify important consumers and target them to maximize revenue. K-means clustering is the most often used technique for consumer segmentation. The disadvantage of employing k-means clustering is its inefficiency in grouping data points and its high processing cost. Additionally, the computing time of K-means grows as the number of data points increases K-means is computationally costly and becomes more time-consuming with larger datasets. In this study, micro batch K-means clustering is employed to segment clients. The concept is to employ random batches of little data with a fixed size.

Keywords: Machine Learning, Classification, RFM analysis, Clustering, Mini batch K-means algorithm.

CONTENTS

LIST OF FIGURES

1. INTRODUCTION

In today's data-driven business environment, companies aim to leverage information to strengthen their advantage over competitors and drive growth. Data has become the new currency, driving innovation and changing marketing tactics in a variety of sectors. Businesses hoping to prosper in this changing climate must have a thorough understanding of client behavior and preferences.

The primary strategy used by businesses to divide consumers into discrete groups according to characteristics and behaviors they have in common is customer segmentation. Through efficient client segmentation, organizations can tailor their marketing efforts, provide customized goods and services, and maximize the use of their resources.

The goal of this research is to create a machie learning model for customer segmentation by combining the Mini-batch K-Means clustering method with RFM (Recency, Frequency, Monetary) analysis. RFM analysis measures consumer involvement by assessing a customer's recent purchase history (Recency), frequency of purchases (Frequency), and amount of expenditure (Monetary).

1.1 Project Structure

This project encompasses distinct modules, including data collection and preprocessing, RFM analysis, Mini-batch K-Means clustering, evaluation metrics, and visualization. By systematically integrating these modules, we aim to deliver a comprehensive customer segmentation system that empowers businesses to make data-driven decisions and foster strongercustomerrelationships.

In the subsequent sections, we will delve into the implementation details of each module, showcasing how RFM analysis is performed, Mini-batch K-Means clustering is applied, and how the effectiveness of our segmentation model is evaluated. Through this project, we endeavor to demonstrate the transformative potential of machine learning in refining customer segmentation strategies and driving business growth.

1.2 Objectives

- Design and develop a machine learning model using RFM and Mini-batch K-Means for customer segmentation.
- Enable marketers to engage with customers more effectively based on their segmentation.
- Improve product promotion and customer relationship management.
- Empower marketers with actionable insights derived from RFM analysis.
- Inform product development and marketing tactics based on segmented customer behaviors.

2. LITERATURE REVIEW

2.1 Introduction to Survey

A literature survey is a thorough evaluation of published research in a certain subject or study. Critically assessing and synthesizing previous research might highlight gaps, limitations, and potential for future study. Literature surveys are crucial for various reasons:

- **Identify gaps and opportunities for research:** A literature survey helps identify thegaps and limitations in the existing research, which can inform the development of new research questions and hypotheses. Additionally, a literature survey can identify emerging research trends and opportunities for future research.
- **Avoid duplication of research:** Conducting a literature survey can help researchers avoid duplicating existing research by identifying the research that has already been conducted in the area. This can help save time and resources and prevent unnecessary repetition of research.
- **Evaluate and critique existing research:** A literature survey involves critically evaluating and synthesizing the existing research, which can help identify the strengths and weaknesses of the research. This can help researchers identify areas where further research is needed to address the limitations of the existing research.
- **Inform research methodology:** Literature surveys can help researchers identify the most appropriate research methods and techniques for their research questions. By evaluating the existing research methods and techniques used in the area, researchers can identify the most appropriate methods and techniques for their research.
- **Build on existing research:** By synthesizing the existing research, researchers can identify the most promising areas for further research and build on the existing research to advance the field.

2.2 Related Work

Related work refers to the research that has been conducted in a particular field or area of study that is related to the research being conducted. Related work is typically reviewed and analyzed in a literature survey or literature review.

1. **Customer segmentation model using K-means clustering of E-commerce**

   Author Name: E.Y.L Nandapala

   Summary: The using the K-mean clustering algorithm to slove the segmentation process to segment the customers K defines the number of pre-defined clusters that need to be created in the process.In tis algorithm use the three methods are Elbow method,Silhouette method and Gap statistic method.The elbow method is used to finding the optimal number of clusters it requires the equation for the with in cluster sum of squares or the wcss.Silhouette method is used to measure of how similer a data points is with in cluster compared to other cluster.Gap statistic method is used to selecting the optimal number cluster(k) in a clustering algorithm it is identify a k values that results in a larger gap between the actual clustering quality and the expectedclustering quality in a randam data.

   Advantage:

   - Using customer segmentation the proposed system get to know about the profitable customer so accordingly there will be giving a special discount to customer depending upon their purchase and expenditure
   - Easy to understand and implement.
   - Using Silhouette method and Elbow method give the most accurate.

   Disadvantages:

   - K-means clustering only considers a limited number of variables, which may not captures all the factors that influence customer behaviour and preferences
   - Requires pairwise distance computations for each data point

2. **A case study on customer segmentation by using ML methods**

Author Name: Sokru Ozan, K.P.N. Jayasena

Summary: Here is use the three methods are NEM method,LiRM method and LoRM method.The NEM method is an analytical apporach to linear regression with a least square cost function we can use the normal equation to directly compute the parameters of a model that minimizes the sum of the squared differences between the actual term and the predicted term.LiRM is a supervise machine learning algorithm that measures the degree of linear relationships between multiple independent and variables.LoRM is also a supervised machine learning algorithm mainly used for classification tasks where the goal is to predict the probabilitly that an instance belongs to a given class or not it analyzes the relationship between a set of independent and dependent binary variables.The advamtage is very fast at classifying unknown records and flexible and adaptable disadvantage is constructs linear boundaries the drawback is these methods give the 89.43% efficiency percentage to customer segmentation

Advantages

- Analytical Solution: Provides a direct analytical solution to linear regression using the normal equation, avoiding iterative optimization algorithms.
- Efficiency for Small Datasets: Well-suited for small to moderate-sized datasets where computing the normal equation is feasible.
- Efficient for Linear Relationships: Effective for capturing linear relationships between multiple independent variables and a continuous dependent variable.

Disadvantages

- Computational Cost for Large Datasets: Computing the normal equation can become computationally expensive for large datasets due to matrix operations.
- Assumption of Linearity: Limited to capturing linear relationships, which may not be suitable for complex, nonlinear data patterns.
- Linear Boundaries: Constructs linear decision boundaries, which may not capture complex decision boundaries in the data.

3. **Customer segmentation based on activity monitoring applications for the recommendation system**

Author Name: Vaidisha Mehta, Iiona Pawelogzek,

Summary: A survey on Customer Segmentation using Machine Learning Algorithms to find prospective clients. Stages in Customer segmentation: Foundation , Analysis , Data Collection, Synthesis. Market Segmentation types include Demographic, Psychographic, Geographic, Behavioral Segmentation . Major technologies used for customer segmentation are K-means, Agglomerative, Divisive, Mean-Shift, GMM, DBSCAN Algorithms. For measurement of accuracy they uses Silhouette Score and Davies Bouldin Score. Advantages K-means algorithm is better than other clustering algorithms for customer segmentation in terms of performance and accuracy both which can also be seen from the obtained Davies and Silhouette scores of various clustering algorithms. K-Means takes O(n) time as compared to others like Hierarchical taking O(n^2) time.

Adavantages

- Efficiency: K-means is computationally efficient and scales well with large datasets, as it has a time complexity of $O(n)$ O(n).
- Simple and Easy to Implement: Straightforward to understand and apply, making it suitable for a wide range of clustering tasks.
- Robust to Noise and Outliers: Identifies clusters based on dense regions in the data space, ignoring sparse areas.
- Automatic Cluster Detection: Automatically determines the number of clusters and can handle arbitrary cluster shapes.

Disadvantages

- Parameter Sensitivity: Performance can be sensitive to the choice of epsilon (neighborhood distance) and minimum points parameters.
- Memory Intensive for Large Datasets: Requires maintaining a neighborhood list for each data point, which can be memory-intensive for large datasets**.**

2.3 Challenges

- **Data Quality and Preparation:** One of the major challenges in customer segmentation is the quality and preparation of data. The effectiveness of clustering algorithms heavily relies on the quality and relevance of the input data. Noisy or incomplete data can lead to inaccurate segmentation results.
- **Algorithm Selection and Parameters:** Choosing the right clustering algorithm and tuning its parameters appropriately can be challenging. Different algorithms have different assumptions and are sensitive to parameters like the number of clusters (k), distance metrics, or other settings. It's essential to understand the characteristics of the data and the behavior of each algorithm to make informed choices.
- **Handling Large Data:** Customer datasets in e-commerce or other domains can be large and high-dimensional, which poses challenges for traditional clustering algorithms. Memory constraints and computational efficiency become critical factors when dealing with big data. Some algorithms may struggle to scale effectively with increasing dataset size.
- **Interpretability of Clusters:** While clustering algorithms can effectively group customers based on similarities in behavior or attributes, interpreting these clusters and translating them into actionable insights can be challenging. Understanding what each cluster represents and how it relates to business goals requires domain knowledge and post-analysis.
- **Dynamic and Evolving Customer Behavior:** Customer behavior and preferences can change over time, leading to evolving segments. Building customer segmentation models that can adapt to dynamic changes and updates in customer behavior is a non-trivial task. Continuously updating and re-evaluating segmentation strategies is essential for maintaining relevance.
- **Evaluation Metrics and Validation:** Assessing the quality of segmentation results and choosing appropriate evaluation metrics can be challenging. Metrics like silhouette score, Davies-Bouldin index, or others mentioned in the summaries are used to evaluate the quality of clusters, but interpreting these metrics and drawing meaningful conclusions requires expertise.

2.4 Problem Statement

Using K-Means clustering is time consuming presents challenges related to computational complexity, memory usage, initial centroid selection, convergence and interpretability. The distance calculations between data points and centroids become time-consuming, and memory limitations may hinder storage of the entire dataset. Selecting suitable initial centroids becomes more complex, affecting convergence rates. Managing a large number of clusters and data points complicates result interpretation. Addressing these issues often requires variations like Mini-Batch K-Means to ensure efficient and effective clustering in large-scale datasets.

To address this challenge, the code creates a consumer segmentation model that combines RFM (Recency, Frequency, and Monetary) analysis with clustering algorithms. The algorithm creates RFM ratings for each client using past transaction data, capturing important features of their purchasing behavior such as how recently they made a purchase, how frequently they purchase, and the overall monetary value spent. These ratings offer a quantifiable measure of consumer involvement and value.

The following phase uses the MiniBatch K-means clustering technique to divide clients into various categories depending on their RFM ratings. Clustering identifies homogenous groups of clients who have similar purchasing habits. This segmentation enables firms to better identify client groupings, such as high-value or loyal consumers vs occasional buys, resulting in more targeted marketing tactics and tailored experiences.

3.SYSTEM ANALYSIS

3.1 Existing system

The existing system for customer segmentation utilizes several machine learning algorithms, including K-means clustering, Hierarchical clustering, and DBSCAN. The mainly K-means clustering is used to group customers based on similarity, but its limitations include the need to specify the number of clusters in advance and the potential for producing incoherent segments.

Specific issues with the existing K-means approach include its time-consuming nature, particularly with large datasets, and the need to predefine the number of clusters, leading to potential improper clustering with varied data types. These drawbacks hinder scalability and efficiency, especially when dealing with complex datasets.

3.1.1 Disadvantages of Existing System

1. k-means algorithm is used for customer segmentation where it requires more time.
2. K-means algorithm it does not work for huge data where it gives accurate solution for only small data .
3. Where previously we need to classify K value before the start of clustering and if there are various types of data it leads to improper clustering.
4. It takes more time if we specify large k value.
5. K-Means can be slower on large datasets since it considers the entire dataset in each iteration.

3.2 Proposed system:

The proposed approach uses RFM (Recency, Frequency, Monetary) analysis and Mini-Batch K-Means clustering to improve client segmentation. Unlike standard K-Means, this technique overcomes constraints while providing significant benefits.

RFM analysis evaluates customer behavior based on recency, frequency, and monetary value, allowing for adaptive clustering without predefined cluster numbers (K). This strategy provides useful information on customer segmentation based on unique purchasingbehaviors.

Mini-Batch K-Means clustering increases efficiency by updating centroids using randomly picked data subsets, hence lowering memory usage and improving scalability for large datasets. Mini-Batch K-Means may sacrifice some accuracy for speed, but it excelsincomputingefficiency.

3.2.1 Advantages of Introduced System:

- Reduces the computational cost for finding the cluster.
- Using only the objective function of k-means does not give a complete idea The proposal is to use also validity measures used to compare the behavior of cluster algorithms.
- The main advantage of the proposed system is usage of Mini batch K-Means algorithm over k-means to segmentation.
- Mini-batch K-means algorithm that it reduces the computational cost of finding a cluster
- Enhances accuracy and efficiency in customer segmentation.

4.FEASIBILITY STUDY

In the evolving landscape of commerce, data-driven insights are becoming as vital as electricity. This feasibility study delves into the application of unsupervised machine learning algorithms for customer segmentation, aiming to enhance marketing strategies and customer relationship management. The study compares five clustering algorithms to identify the most effective approach for grouping customers based on their behavioral patterns. By leveraging customer data and innovative methodologies, businesses can optimize interactions with distinct customer segments to maximize customer value and competitiveness in the market.

Feasibility study should be performed on the basis of various criteria and parameters. The various feasibility studies are:

- Technical Feasibility
- Operation Feasibility
- Economic Feasibility

4.1 Technical Feasibility:

1. Software Requirements: Python 3.7 is used for the project, along with widely-available and well-documented libraries like scikit-learn, NumPy, and Pandas.

2. Algorithm Implementation: Using well-known machine learning frameworks, RFM analysis and Mini-batch K-Means clustering are easily implemented.

3. Data Processing: Python modules can be used to efficiently carry out pre-processing tasks, such as data collecting, cleaning, and feature engineering.

4. Computational Resources: The project needs access to widely available computer resources that can handle the volume of data and train the machine learning model.

5. Scalability: The system can handle bigger datasets and grow in the future because to the scalability provided by the technologies and algorithms selected.

4.2 Operational Feasibility:

1. Data Accessibility: The success of the project depends on the availability and quality of customer data, which should be accessible to the project team.

2. Stakeholder Collaboration: Collaboration and cooperation among stakeholders, including marketing managers and data analysts, are essential for project success.

3. User Training: Adequate training and support will be provided to stakeholders involved in implementing and using the segmentation model.

4. Integration: The model should seamlessly integrate into existing systems and workflows to ensure smooth operational adoption.

5. Feedback Mechanism: A feedback mechanism will be established to incorporate user feedback and continuously improve the segmentation model based on operational needs.

4.3 Economic Feasibility:

1. Initial Investment: The project requires investment in software tools and possibly computing resources, which are justifiable considering the potential benefits.

2. Cost-Benefit Analysis: The expected benefits, including improved marketing strategies and customer relations, outweigh the initial investment and ongoing operational costs.

3. Return on Investment (ROI): The project aims to deliver tangible ROI through increased revenue, cost savings, and enhanced customer satisfaction.

4. Resource Optimization: By efficiently utilizing machine learning techniques, the project reduces the need for manual segmentation efforts, optimizing resource allocation.

5. Long-Term Viability: The project's economic feasibility is supported by its long-term viability, as it aligns with the company's strategic goals and contributes to sustainable growth.

5.SYSTEM REQUIREMENT SPECIFICATION

5.1 Requirement analysis:

To improve marketing strategies and gain a deeper understanding of customer behavior, the project's primary goal is to segment customers using mini-batch K-means clustering. In order to drive targeted interaction and customized product offerings, the project will distill actionable insights from preprocessing diverse customer data to showing clustering outcomes. The technology aims to identify subtle trends and preferences by breaking down consumer datasets into observable clusters. This method helps decision-makers make well-informed decisions and increases customer happiness and loyalty.

In this sense, the project includes every step of the analytical process, from preparing data to explaining clustering results. To obtain thorough insights into consumer profiles, this requires competently managing a variety of client data types, including as demographic information, transaction histories, and behavioral measurements. The system must effectively preprocess the data, run the mini-batch K-means algorithm for clustering, assess the effectiveness of the clustering process using pertinent metrics, and then present the segmented customer data in an understandable and eye-catching way. A thorough approach like this guarantees a sophisticated comprehension of client segments, enabling companies to efficiently customize their product offers and marketing tactics to appeal to a wide range of customers.

5.2 Software Requirements Specification (SRS):

The Software Requirements Specification (SRS) delineates the blueprint for developing a robust customer segmentation system through mini-batch K-means clustering. The primary objective is to enhance marketing strategies and deepen insights into customer behavior. This entails a comprehensive approach encompassing data preprocessing, application of the mini-batch K-means algorithm, evaluation of clustering quality, and visualization of segmented customer data. By distilling actionable insights from the clustering analysis, the system aims to empower businesses to tailor their strategies and offerings to better resonate with the diverse needs and preferences of their customer base.

Within this scope, the project sets forth functional requirements geared towards ensuring the efficiency and efficacy of the system. This includes seamless data preprocessing to handle diverse customer datasets, implementation of the mini-batch K-means algorithm to generate meaningful clusters, and robust evaluation of clustering quality using appropriate metrics. Additionally, the system is tasked with providing clear and intuitive data visualization capabilities, facilitating the interpretation and communication of clustering results. Non-functional requirements encompass critical aspects such as system performance, scalability, usability, and security, ensuring smooth operation and safeguarding sensitive customer data throughout the process. By addressing these requirements, the project endeavors to deliver a comprehensive customer segmentation solution that empowers businesses to make data-driven decisions and foster stronger relationships with their customers.

5.3 Functional requirements:

The system's functional requirements encompass several key aspects essential for the successful implementation of mini-batch K-means clustering for customer segmentation. Firstly, it should seamlessly preprocess the customer data, ensuring its cleanliness and readiness for clustering analysis. This includes handling missing values, outliers, and appropriately scaling features to ensure the accuracy and reliability of the clustering results. Moreover, the system must effectively implement the mini-batch K-means clustering algorithm, allowing for customization of parameters such as the number of clusters and batch size. This functionality is crucial for accurately segmenting customers into distinct clusters based on their shared characteristics and behaviors. Furthermore, the system should provide robust clustering evaluation capabilities, employing metrics like silhouette score and Davies-Bouldin index to assess the quality of clustering results. These evaluations will enable stakeholders to gauge the effectiveness of the segmentation process and make informed decisions based on the identified customer clusters.

5.4 Non-Functional Requirements:

Performance: The system should efficiently handle large datasets and maintain optimal processing speed.

Scalability: It should scale seamlessly to accommodate growing data volumes and evolving business needs.

Usability: The interface should be intuitive, supported by clear documentation, and customizable for user preferences.

Security: Robust measures must ensure data privacy, access control, and encryption to protect sensitive customer information.

5.5 Software Requirements

Programming Language : Python (version 3.x)

Development Environment: Google Colab

Libraries: Numpy, Pandas, matplotlib

Data Visualization Tools: Matplotlib, Seaborn

Data: Customer dataset

Operating System: Compatible with Windows

5.6 Hardware requirements:

Hard Disk Drive : 40 GB

Processor : Intel core i5

Main Memory : 8 GB RAM

6. SYSTEM DESIGN

6.1 MODULES

The Six modules are:

- Data Preprocessing
- Calculating RFM value
- Finding the RFM-Score
- Mini-batch Formation
- Cluster Evaluation
- Output and Reporting Module

6.1.1 Data Preprocessing:

This module is in charge of efficiently reading and processing client data for analysis. It includes activities like loading CSV data, changing date formats for consistency, and handling missing or irrelevant data. By handling these preconditioning phases, the module makes sure that the dataset is of high quality and ready for analysis.

6.1.2 Calculating RFM value:

RFM values are critical metrics for customer analytics that help segment and understand consumer behavior. Each component—recency, frequency, and monetary—provides useful information about client engagement and spending trends.

The term "recency" refers to a customer's most recent purchase since their last transaction. It tracks the time since the customer's most recent purchase. Customers have made recent purchases have a higher probability to be active and responsive.

Frequency refers to how frequently a customer makes a purchase during a given time period. It counts the number of transactions a client has completed, indicating their level of involvement with the firm. Customers that make frequent purchases are more likely to be loyal and valuable.

Monetary refers to a customer's purchasing amount. It calculates the total worth of the customer's transactions during a specified time period. Customers with higher monetary values generate more revenue and are frequently sought for premium services or specialized offers.

To calculate RFM values, each customer's transaction history is analyzed. For Recency, the time difference between the customer's most recent purchase date and a reference date (e.g., today's date) is calculated. Frequency is determined by counting the number of transactions made by the customer within a defined timeframe. Monetary value is calculated by summing the total amount spent by the customer on transactions.

6.1.3 Finding the RFM-Score:

RFM (Recency, Frequency, Monetary) scoring is a strategic approach used by firms to examine customer transactional patterns and behavior. The procedure begins with data collection from transaction databases, which includes crucial parameters such as purchase dates and times, the amount of the transaction, and client identities. This data serves as the foundation for determining RFM scores, beginning with Recency, which assesses the time since a customer's previous purchase.

To assign RFM scores, clients are ranked according to percentage range for each metric—Recency, Frequency, and Monetary. Customers are assigned a score from 1 to 5 depending on their transactional activity, with higher scores indicating more recent purchases, higher transaction frequency, and greater monetary worth.

RFM scores are used to segment customers into distinct groups based on similar behaviors and characteristics. This segmentation enables targeted marketing strategies, such as personalized promotions or retention campaigns, tailored to each customer segment's needs and preferences.

6.1.4 Mini-batch Formation:

In the context of customer segmentation using RFM (Recency, Frequency, Monetary) scores, the mini-batch formation process plays a critical role in efficiently clustering customers based on their transactional behavior. Once the RFM scores are calculated for each customer, these scores are utilized as input for the mini-batch K-means clustering algorithm.

The mini-batch K-means algorithm is a variant of the traditional K-means clustering method, optimized for handling large datasets by processing data in smaller, randomly sampled batches. This approach ensures computational efficiency while maintaining the effectiveness of the clustering process. The algorithm works by iteratively assigning data points (in this case, customers represented by their RFM scores) to clusters based on the similarity of their features.

Using the RFM scores as input, the mini-batch K-means algorithm segments customers into clusters according to their transactional characteristics. Customers within the same cluster exhibit similar RFM profiles, reflecting shared patterns in their purchasing behavior and engagement levels. This segmentation enables businesses to identify distinct customer segments with unique preferences and needs.

6.1.5 Cluster Evaluation:

Cluster evaluation is crucial in assessing the quality of customer segmentation achieved through machine learning models. The Silhouette Score measures how well data points within clusters are grouped, with higher scores indicating more distinct clusters. The Davies-Bouldin Index assesses cluster separation, with lower values indicating better-defined clusters and minimal overlap. These evaluation metrics provide quantitative insights into the effectiveness of clustering algorithms, guiding decisions on segmenting customers for targeted marketing strategies and personalized customer engagement.

6.1.6 Output and Reporting Module:

Finally, the client Segmentation System returns the segmented client groups, together with their RFM values. The Marketing Manager uses this output to obtain insights into consumer segments and their unique characteristics, allowing for customized marketing campaigns.

6.2 System architecture

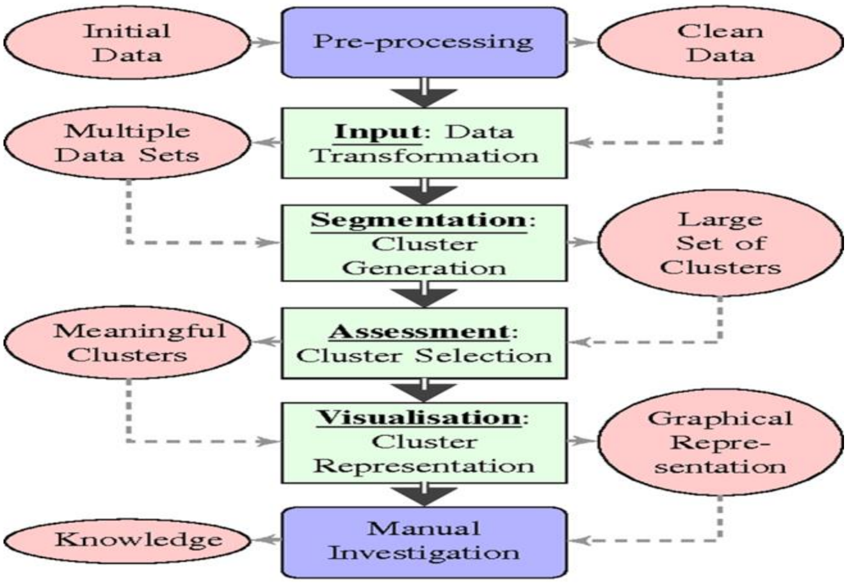

Initial Data
Pre-processing
Clean Data
Multiple Data Sets
Input: Data Transformation
Segmentation: Cluster Generation
Large Set of Clusters
Meaningful Clusters
Assessment: Cluster Selection
Visualisation: Cluster Representation
Graphical Repre-sentation
Knowledge
Manual Investigation

Fig 1:system architecture

Initial Data: This is the data you'll use to segment your consumers. It might originate from a number of sources, including your customer relationship management (CRM) system.

Pre-processing and Cleaning: Before segmenting your customers, ensure that your data is clean and ready for analysis. This can include duties like deleting missing values, fixing formatting problems, and eliminating outliers.

Segmentation: it is the process of splitting your clients into groups (or segments) with comparable characteristics. There are several clustering algorithms that can be utilized for this task.

Assessment: Once you've classified your consumers, you must evaluate the quality of the clusters. This can include assessing how well-separated the clusters are from one another.

Finally, you can visualize the clusters to better understand the structure of your data. This can be done using a scatter plot and other graphical tools.

6.3 Introduction to UML

The Unified Modeling Language (UML) is a widely used modeling language in object-oriented software engineering. It was developed in the mid-1990s by Grady Booch, Ivar Jacobson, and James Rumbaugh at Rational Software Corporation. UML combines their different methodologies for designing and visualizing system structures, behaviors, and business processes. It provides a standardized way to represent the design of a system, making it easier for software developers and other stakeholders to communicate and understand system requirements.

UML became an industry standard in 1997 when it was adopted by the Object Management Group (OMG). Since then, OMG has been responsible for managing its further development. In 2000, the International Organization for Standardization (ISO) also recognized UML as an official ISO standard. This recognition established UML as a universally accepted modeling language, not just for software engineering but also for modeling various aspects of systems and processes in different domains beyond software development.

6.3.1 Class Diagram

Class diagrams are essential UML diagrams used to illustrate the structure of object-oriented systems, displaying classes, their attributes, and operations. In class diagrams, each class is typically represented with its name at the top, followed by attributes (properties) in the middle, and operations (methods) at the bottom.

6.3.2 Component Diagram

A component diagram in UML illustrates the structural relationships between components within a software system. Components, such as classes or modules, communicate through interfaces defined by connectors. This diagram is valuable for visualizing and designing complex systems with multiple interacting components.

6.3.3 Deployment Diagram

Deployment diagrams model the physical architecture of a system. Deployment diagrams show the relationships between the software and hardware components in the system and the physical distribution of the processing.

6.3.4 Object Diagram

Object Diagram sometimes referred to as instance diagrams, resemble class diagrams but use real-world examples to show relationships between objects. They depict how a system appears at a specific moment in time.

6.3.5 Use Case Diagram

Use Case Diagram One of the most recognized types of behavioral UML diagrams, use case diagrams provide a visual overview of system actors, the functions they require, and how these functions interact. They are ideal for initiating project discussions by identifying key actors and system processes.

6.3.6 Activity Diagram

Activity diagrams represent workflows graphically, describing business or component operational workflows. They are often used as alternatives to state machine diagrams.

6.3.7 State Machine Diagram

State machine diagrams, also known as state diagrams or state chart diagrams, are similar to activity diagrams but focus on object behavior according to their current state.

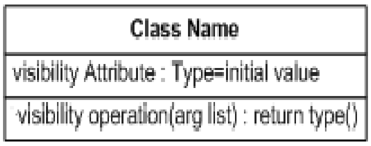
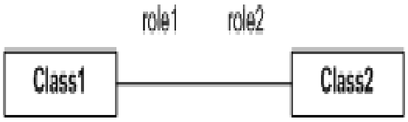
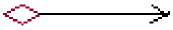
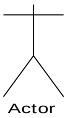
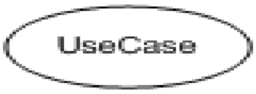
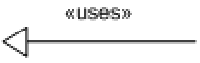
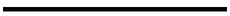
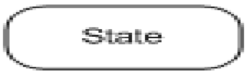
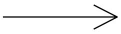
6.3.8 Sequence Diagram

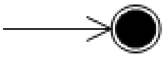
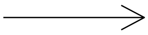
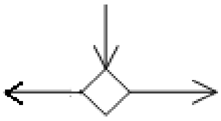
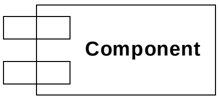
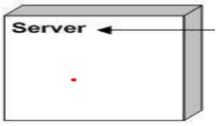
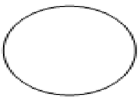
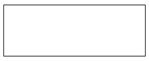
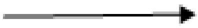
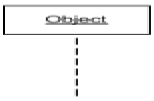
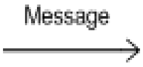
Sequence diagrams illustrate object interactions and the order in which these interactions occur for a specific scenario. Processes are depicted vertically with interactions shown as arrows.

6.3.9 Collaboration Diagram

Collaboration diagrams, similar to sequence diagrams, emphasize messages exchanged between objects. The same information can be represented using a sequence diagram with different object perspectives.

6.3.10 List of UML Notations

S.NO	SYMBOL NAME	SYMBOL	DESCRIPTION
1	Class	 Class Name visibility Attribute : Type=initial value visibility operation(arg list) : return type()	Classes represent a collection of similar entities grouped together.
2	Association		Association represents a static relationship between classes.
3	Aggregation		One type of relationship is aggregation. Itaggregates several classes into single class.
4	Actor		Actors are the users of the system and other external entity that react with the system.
5	Use Case		A use case is an interaction between the system and the external environment.
6	Relation (Uses)		It is used for additional process communication.
7	Communication		It is the communication between various use cases.
8	State		It represents the state of a process. Each state goes through various flows.
9	Initial State		It represents the initial state of the object.

10	Final State		It represents the final state of the object.
11	Control Flow		It represents the various control flow between the states.
12	Decision Box		It represents the decision making process from a constraint.
13	Component		Components represent the physical components used in the system.
14	Node		Deployment diagrams use the nodes for representing physical modules, which is a collection of components.
15	Data Process/State		In DFD, a state or process that has been initiated as a result of an occasion or action is represented by a circle.
16	External Entity		It represents any external device, including a keyboard and sensors.
17	Transition		It represents any communication that occurs between the processes.
18	Object Lifeline		Object lifelines represents the vertical dimension that objects communicates.
19	Message		It represents the messages exchanged.

6.4 UML diagrams

6.4.1 Usecase Diagram:

- Marketing Manager inputs the customer data
- Customer Segmentation System performs RFM analysis on the customer data
- Customer Segmentation System applies Mini-batch K-Means to group customers based on RFM values
- Customer Segmentation System outputs the customer segments
- Marketing Manager views the customer segments and their respective RFM values.

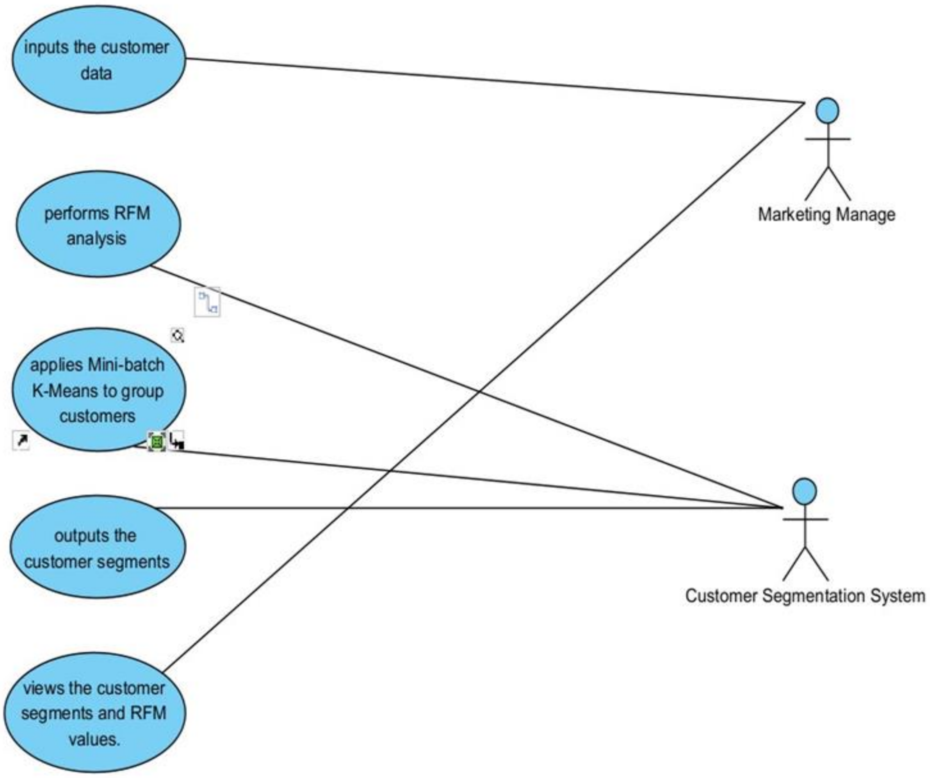

inputs the customer data
performs RFM analysis
applies Mini-batch K-Means to group customers
outputs the customer segments
views the customer segments and RFM values.
Marketing Manage
Customer Segmentation System

Fig 2: Usecase diagram of system

6.4.2 Class Diagram

This class diagram represents the objects and their relationships in the customer segmentation system. The Customer class represents the customer data with attributes of Recency, Frequency, Monetary, and Segment. The Customer Segmentation System class performs the RFM analysis and applies Mini-batch K-Means to group the customers into segments. The Marketing Manager class inputs the customer data and views the customer segments.

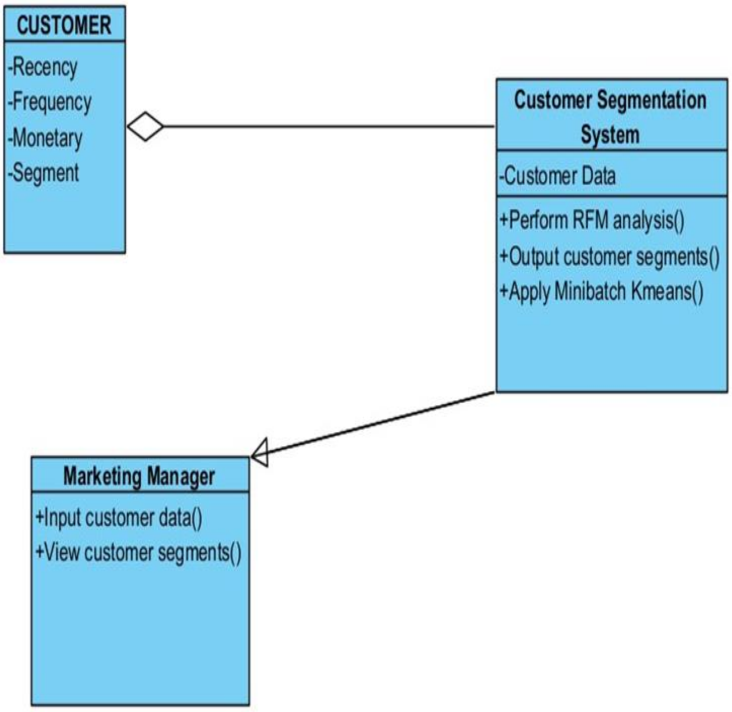


Fig 3: Class diagram of system

6.4.3 Sequence Diagram

This activity diagram represents the flow of activities in the customer segmentation system. The Marketing Manager inputs the customer data, which is then received by the Customer Segmentation System. The Customer Segmentation System performs RFM analysis and applies Mini-batch K-Means to group the customers into segments. The customer segments and their respective RFM values are outputted and viewed by the Marketing Manager. The goal of the system is to group customers based on their RFM values, so that targeted marketing campaigns can be created for each segment

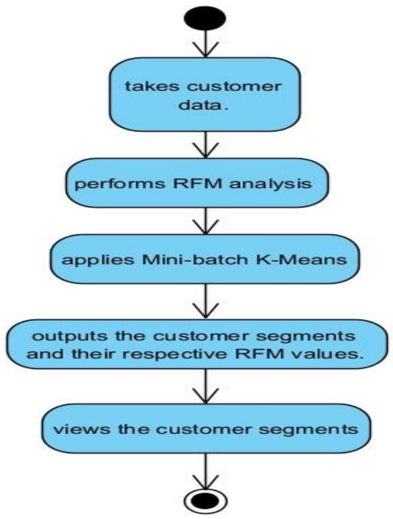

takes customer data.
performs RFM analysis
applies Mini-batch K-Means
outputs the customer segments and their respective RFM values.
views the customer segments

Fig 4: Sequence diagram of system

6.4.4 Activity Diagram

The activity diagram depicts the flow of actions within the customer segmentation system, from data entry to the creation of segmented customer groups. Each phase, including RFM analysis, Mini-batch K-Means clustering, and segment characterization, helps to achieve the overarching goal of optimizing consumer interaction and marketing strategies using data-driven insights. This methodical technique improves understanding of consumer behavior and facilitates decision-making for corporate success and satisfied customers.

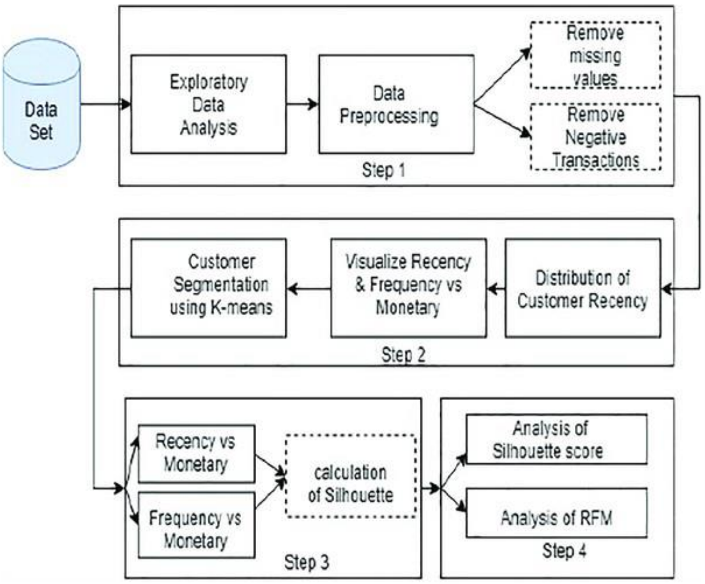

Data Set
Exploratory Data Analysis
Data Preprocessing
Remove missing values
Remove Negative Transactions
Step 1
Customer Segmentation using K-means
Visualize Recency & Frequency vs Monetary
Distribution of Customer Recency
Step 2
Recency vs Monetary
Frequency vs Monetary
calculation of Silhouette
Step 3
Analysis of Silhouette score
Analysis of RFM
Step 4

Fig 5: Activity diagram of syatem

7. SYSTEM IMPLEMENTATION

7.1 Technologies

7.1.1 Python

Python is a popular high-level interpreted programming language that is easy to learn and has a lot of adaptability. Web development, data analysis, automation, scientific computing, artificial intelligence, and other fields all make extensive use of it. Python's syntax is meant to be easy to understand and comprehend, with a style that is similar to pseudo-code. This makes it suitable for novices and speeds up the process of developing and prototyping. Python is an interpreted language, meaning that code is run by an interpreter line by line, facilitating interactive debugging and development. Python is dynamically typed, which means that variables' types are decided upon at runtime. This allows for flexibility, but type safety must be carefully considered. Python has an extensive standard library that offers pre-built functionality for common tasks, making development simpler. The library contains modules and packages for a wide range of operations. Because Python is platform-independent, programming written in it can run without changes on a variety of operating systems. Python is a popular choice for many programming tasks since it is extensively utilized in many different industries and has many libraries and frameworks that are particular to certain domains.

7.1.1.1 History

Guido van Rossum developed the programming language Python in the late 1980s in response to his need for a language that could be used for both small and large-scale projects and was understandable and simple to learn. It made its official release in February 1991 with the publication of Python 0.9.0, which embodied the ideas of simplicity, explicitness, and readable code. In homage to the British comedy television series "Monty Python's Flying Circus," the name "Python" was selected to convey the language's approachable and amiable qualities. During its development, Python has achieved several noteworthy achievements. The first stable release of Python was version 1.0, which was released in January 1994 and included functions like lambda, map, filter, and reduce. The Python 2.x series, which ran from 2000 to 2008, introduced features including garbage collection, list comprehensions, and support for Unicode. Beginning in 2008, Python 3.x improves upon the flaws and inconsistencies of previous versions with improvements to syntax and better support for Unicode. With Python 2.x reaching its end of support in 2020, Python 3.x has emerged as the standard version. Python's open-source nature encourages cooperation and creativity within its vibrant community, guaranteeing its lasting significance and influence in the programming industry.

7.1.2 Features of Python

- **Simple to Learn and Use:** The syntax of Python is succinct and straightforward, akin to plain English. In comparison to other languages, this makes it simpler to learn and comprehend.
- **Interpreted Language:** An interpreter runs Python code line by line, doing away with the need for further compilation stages. Faster development cycles and simpler debugging are made possible by using python.
- **Dynamically Typed**: Python does not need you to define variable types in advance, in contrast to statically typed languages. Although this results in a more compact code, it may also generate runtime errors.
- **Large-scale Standard Library**: A vast array of pre-built modules and functions for a variety of activities, including file handling, networking, and data processing, are included with Python. This lessens the need to write code from scratch, which saves time.
- **Improved Productivity:**Python is an extremely useful language. Python's simplicity allows developers to concentrate on finding the solution. They don't have to invest a lot of time learning the syntax or syntax of the programming language**.**
- **Free and Open-Source:** Python is licensed under an open-source agreement authorized by OSI. It is therefore free to use and share. You are able to download the source code, edit it, and even share your customized Python version. Organizations who wish to alter a certain behavior and utilize their version for development will find this handy.

7.1.3 Python Modules

7.1.3.1 Pandas:

Pandas is similar to your reliable Python data handling helper. Series and DataFrame are the two primary data structures it provides. A series is similar to a single column of data; it can be thought of as an integrated list or labeled array. Working with one-dimensional data, like time series or sensor measurements, is a breeze with it. Conversely, a DataFrame, with its rows and columns, resembles a table or spreadsheet. With each column denoting a distinct variable or feature, it's ideal for organizing and analyzing structured data. Pandas makes it easy to load data, manipulate it with user-friendly techniques, and carry out sophisticated operations like grouping, aggregating, and filtering. Additionally, Pandas makes statistical analysis, visualization, and computations on your data easy, making difficult tasks seem simple. Pandas has you covered whether you're working with database tables, CSV files, or spreadsheets. For data scientists, analysts, and anyone who deals with data on a regular basis. It's also simple to learn and use thanks to its clear documentation and simple syntax. You may effectively and confidently handle data-related problems using Pandas, gaining insights and using your data to make wise judgments.

7.1.3.2 Numpy

NumPy is a fundamental package in Python that forms the basis of activities involving numerical computation. The fundamental component of it is the ndarray, a flexible data structure that can work with multidimensional arrays and matrices. In contrast to conventional Python lists, ndarrays provide a contiguous memory layout and homogeneous data type, which guarantees numerical operations are efficient. The library includes a large number of mathematical functions designed specifically for working with arrays. NumPy offers a wide range of functions, from simple arithmetic operations to intricate mathematical calculations. Among the many functions it provides are logarithms, exponentials, trigonometric functions, and statistical procedures.

NumPy's user-friendly interface streamlines operations and manipulation of arrays. Efficient access to items and subarrays within arrays is made possible by its strong indexing and slicing features. Furthermore, users can easily execute element-wise operations across arrays of varying forms thanks to NumPy's broadcasting functionality. NumPy's capabilities are particularly useful for jobs involving linear algebra. It provides a full range of functions for linear algebra, such as matrix multiplication, inversion, and decomposition. These operations serve as the foundation for applications in science and engineering, making complex calculations simple to perform.

NumPy's array manipulation skills are complemented by its ability to streamline random number generation, which is essential for statistical analysis and simulations. The versatility of the library is increased by the functions it offers for creating random numbers and arrays with different probability distributions. Furthermore, NumPy forms a strong ecosystem for scientific computing, data analysis, and visualization through its smooth integration with other Python libraries such as SciPy, Matplotlib, and pandas. NumPy is now even more useful and solidified as a mainstay of the Python scientific computing environment thanks to this inclusion.

7.1.3.3 Matplotlib

For Python data visualization, Matplotlib is your creative collaborator. It enables you to produce eye-catching plots, charts, and graphs to bring your data to life, much like an artistic toolset. Complex datasets are simpler to comprehend and analyze when you use Matplotlib to visualize your data in 2D and 3D. Plot types are abundant, ranging from basic scatter plots and line charts to histograms, bar charts, and pie charts. Whether you're presenting research, comparing values, or investigating trends, There is a plot suitable for every situation in Matplotlib. Furthermore, it offers an extensive range of customization choices that let you personalize the look of your plots to fit your preferences and style. You may use Matplotlib to turn unprocessed data into visually appealing graphics that captivate your audience and convey a compelling tale. For analysts, data scientists, and anybody else who wishes to effectively convey insights through graphics, it is an indispensable tool.

7.1.3.4 Datetime

Python's datetime library facilitates working with timings, dates, and periods of time in Python programs. Datetime facilitates the manipulation of dates and times, as well as the performance of computations such as day addition and subtraction, and date formatting for display. It is ideal for time-based data analysis, work scheduling, and event tracking. Datetime can be used to parse dates from strings, calculate time differences, and manage time zones. Datetime is a flexible tool for a range of applications because it connects with other Python modules with ease. You can make sure that your projects are completed on time and that time-sensitive tasks are completed accurately and precisely by using datetime.

7.1.3.5 Sklearn

The Python machine learning library scikit-learn, sometimes known as sklearn, is an extensive and flexible library. You will discover a wide range of algorithms created for various uses. For instance, let's say you are working on a categorization task. when classifications or classes need to be predicted. With methods like logistic regression, decision trees, random forests, k-nearest neighbors (KNN), support vector machines (SVM), and naive Bayes classifiers, sklearn has you covered. For regression algorithms. bwhere continuous numerical values are being predicted. Regression techniques from sklearn include decision tree regression, random forest regression, support vector regression (SVR), ridge regression, Lasso regression, and linear regression. When performing activities such as clustering, the goal is to combine related data points into groups. Algorithms including K-means, mini-batch K-means clustering, DBSCAN, hierarchical clustering, and Gaussian mixture models (GMM) are all implemented by sklearn. It's not just about algorithms with sklearn. In addition, it provides tools for hyperparameter tuning, model evaluation, cross-validation, and model selection. These tools assist you in selecting the most appropriate model for your data, optimizing its parameters, and precisely evaluating its performance.

7.1.3.6 Slihouette Score

The more comprehensible silhouette score helps in assessing how well your clusters are constructed. By comparing each data point's distance from its own cluster to that of other clusters, it evaluates the quality of the clusters. The silhouette score computes two things for every data point:

- The typical separation between a data point and other points within the same cluster.
- The average separation between a data point and its closest surrounding cluster's points.
- It is desirable for a data point to be well within its own cluster and remote from other clusters, as indicated by a score near 1.
- The data point is on the border between two clusters if its score is near zero.
- A negative score suggests that the data point may have been incorrectly classified into a different cluster.

An intelligible statistic for evaluating the cohesion and separation of clusters is the silhouette score. While a lower score can point to overlapping or poorly separated clusters, a higher silhouette score indicates that the clusters are clearly defined and distinct.

7.1.4 Google colab

Google Colab is an online collaboration environment similar to a virtual playground where you may build and run Python code. You don't need to install anything on your computer because it is hosted on Google's servers. It's quite convenient that you may access it immediately through your computer browser. You can open and create new notebooks once you're in Colab. With these notebooks, which function similarly to interactive documents, you may write Python code, include comments and explanations, and even produce visualizations. It's a really useful tool for code organization and work documentation.

The fact that Colab gives users access to strong computing resources like CPUs, GPUs, and TPUs is one of its most amazing aspects. This implies that you can use Google's servers, which are far more powerful than the typical laptop or desktop, to run your Python code. Everything is free.

Google Colab is excellent framework for running Python code. It's robust, user-friendly, and totally free. Colab includes all you need to write and run Python code with ease, regardless of your level of experience whether you're just beginning learning the language or an advanced programmer working on challenging projects.

7.2 Sample code

```
import pandas as pd
import numpy as np
import datetime as dt
df = pd.read_csv('/content/Dataset1.csv')
df.head()
df.shape
df.isna().sum()
```

```
df['invoice_date'] =
pd.to_datetime(df['invoice_date'],infer_datetime_format=True,form
at='mixed',utc=True, errors='ignore')
```

```
df_recency = df.groupby(by='customer_id',
     as_index=False)['invoice_date'].max()
df_recency.columns = ['customer_id', 'invoice_date']
recent_date = df_recency['invoice_date'].max()
df_recency['Recency'] = df_recency['invoice_date'].apply(
    lambda x: (recent_date - x).days)
df_recency.head()
```

```
frequency_df = df.drop_duplicates().groupby(
  by=['customer_id'], as_index=False)['invoice_date'].count()
frequency_df.columns = ['customer_id', 'Frequency']
frequency_df.head()
```

```
monetary_df =
df.groupby('customer_id')['price'].sum().reset_index()
monetary_scores = pd.qcut(monetary_df['price'], q=4,
labels=range(1, 5))
monetary_df.columns = ['customer_id', 'Monetary']
monetary_df.head()
```

```
rf_df = df_recency.merge(frequency_df, on='customer_id')
rfm_df = rf_df.merge(monetary_df, on='customer_id').drop(
  columns='invoice_date')
rfm_df.head()
```

```
rfm_df['R_rank'] = rfm_df['Recency'].rank(ascending=False)
rfm_df['F_rank'] = rfm_df['Frequency'].rank(ascending=True)
rfm_df['M_rank'] = rfm_df['Monetary'].rank(ascending=True)

# normalizing the rank of the customers
rfm_df['R_rank_norm'] =
(rfm_df['R_rank']/rfm_df['R_rank'].max())*100
rfm_df['F_rank_norm'] =
(rfm_df['F_rank']/rfm_df['F_rank'].max())*100
rfm_df['M_rank_norm'] =
(rfm_df['F_rank']/rfm_df['M_rank'].max())*100
rfm_df.drop(columns=['R_rank', 'F_rank', 'M_rank'], inplace=True)
rfm_df.head()
```

```
rfm_df['RFM_Score'] = 0.15*rfm_df['R_rank_norm']+0.28 * \
  rfm_df['F_rank_norm']+0.57*rfm_df['M_rank_norm']
rfm_df['RFM_Score'] *= 0.05
rfm_df = rfm_df.round(9)
rfm_df[['customer_id', 'RFM_Score']].head(7)
```

```
from sklearn.cluster import MiniBatchKMeans
from sklearn.preprocessing import StandardScaler
batch_size = 100
scaler = StandardScaler()
rfm_scaled = scaler.fit_transform(rfm_df)
```

```
kmeans = MiniBatchKMeans(n_clusters = 10,random_state=42,
batch_size=100)
kmeans.fit(rfm_scaled)
cluster_labels = kmeans.predict(rfm_scaled)
```

```
kmeans.fit(rfm_scaled)
print(rfm_df['Cluster'].value_counts())
```

```
import matplotlib.pyplot as plt
%matplotlib inline
plt.scatter(rfm_df['Frequency'], rfm_df['Monetary'],
c=rfm_df['Cluster'])
plt.xlabel('Frequency')
plt.ylabel('Monetary')
plt.show()
```

7.3 Results&ScreenShorts

7.3.1 Recency

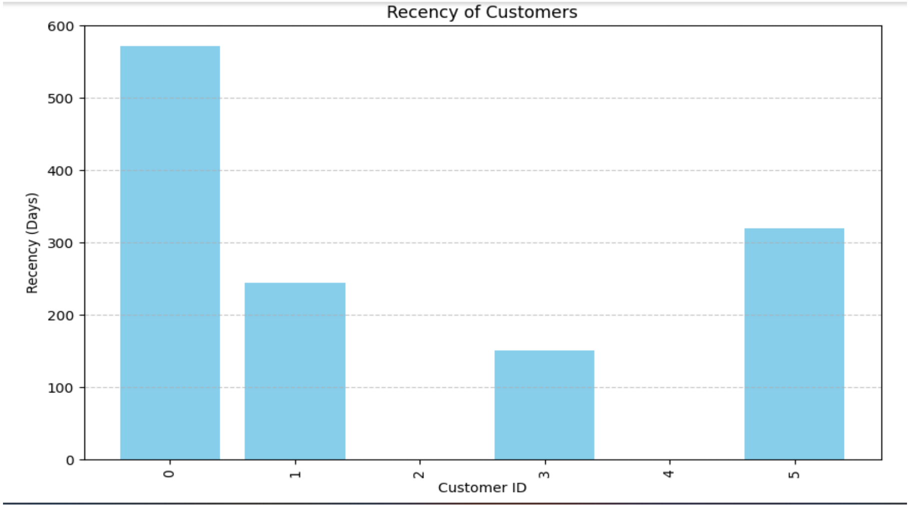
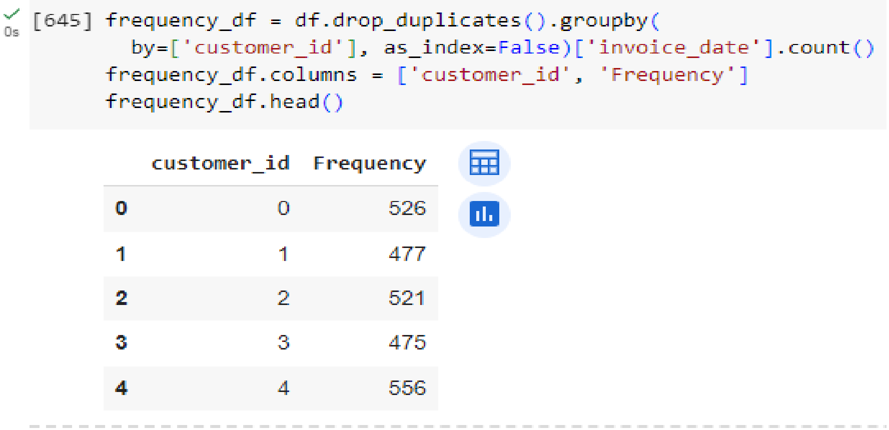


Fig 6 : Recency

The Fig represents the recency of each customer. It is calculated as by subtracting the purchase date of the customer with the maximum purchase date of the dataset. The output shows the number of days that the customer has recently purchased the product. Recency, in customer relationship management (CRM), is typically a measurement of time elapsed since a customer's last purchase. Businesses use recency to target their marketing campaigns and segment their consumer base. Consumers who have made recent purchases are often valued higher than those who have not made any purchases in a long period. In order to entice recent buyers to make another purchase, they may target them with exclusive deals or discounts.

7.3.2 Frequency



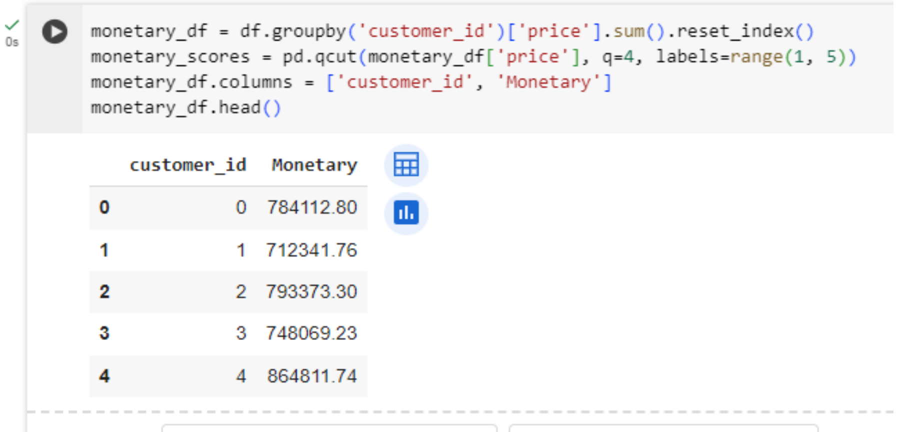
```
[645] frequency_df = df.drop_duplicates().groupby(
        by=['customer_id'], as_index=False)['invoice_date'].count()
      frequency_df.columns = ['customer_id', 'Frequency']
      frequency_df.head()
```

```
   customer_id  Frequency
0            0        526
1            1        477
2            2        521
3            3        475
4            4        556
```

Fig.7 : Frequency

The Fig shows the frequency of each customer. which represent the number of times the customer had done the transaction or purchased the item. In marketing, frequency describes the number of times a customer makes a purchase in a certain amount of time. Marketing managers need to understand frequency since it provides important information about the behavior and loyalty of their customers. Marketing managers may determine which clients are the most devoted to them, learn about their purchase habits, and adjust their marketing tactics to successfully target various customer segments by examining frequency data. High frequency consumers, or those who buy from the company frequently, are probably devoted supporters of the brand. Marketing managers have the opportunity to further cultivate brand loyalty and stimulate repeat purchases by providing special offers, discounts, or loyalty programs to these devoted customers.

7.3.3 Monetary



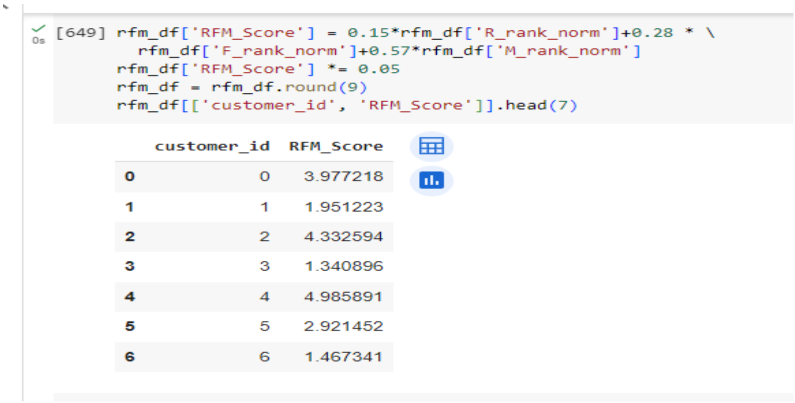
```
monetary_df = df.groupby('customer_id')['price'].sum().reset_index()
monetary_scores = pd.qcut(monetary_df['price'], q=4, labels=range(1, 5))
monetary_df.columns = ['customer_id', 'Monetary']
monetary_df.head()
```

	customer_id	Monetary
0	0	784112.80
1	1	712341.76
2	2	793373.30
3	3	748069.23
4	4	864811.74

Fig.8:Monetary

In the context of marketing, monetary value is the total amount of money that each consumer spends during a given period of time on purchases. It is an indicator of the financial support a consumer provides to the company. Marketing managers need to understand monetary value in order to identify high-value customers, prioritize marketing initiatives, and optimize return on investment. Consumers that make large purchases are regarded as high-value clients and are frequently the most lucrative for the company. Marketing managers can concentrate their efforts on providing individualized incentives and special offerings in order to keep and grow these valuable consumers.

7.3.4 Spending score



```
[649] rfm_df['RFM_Score'] = 0.15*rfm_df['R_rank_norm']+0.28 * \
        rfm_df['F_rank_norm']+0.57*rfm_df['M_rank_norm']
      rfm_df['RFM_Score'] *= 0.05
      rfm_df = rfm_df.round(9)
      rfm_df[['customer_id', 'RFM_Score']].head(7)
```

```
   customer_id  RFM_Score
0            0   3.977218
1            1   1.951223
2            2   4.332594
3            3   1.340896
4            4   4.985891
5            5   2.921452
6            6   1.467341
```

Fig.9: Spending score

The fig shows the spending-score for each customer. Spending score is a metric that measures each customer's value to the company throughout the course of their relationship. It is sometimes referred to as customer profitability score or customer lifetime value (CLV). To determine a score that reflects the total profitability or value of the consumer, it considers variables including the frequency, monetary value, and recentness of purchases. They may discover and prioritize high-value consumers who are most likely to bring in the most money and profit for the company by using the spending score.

7.3.5 cluster visualization

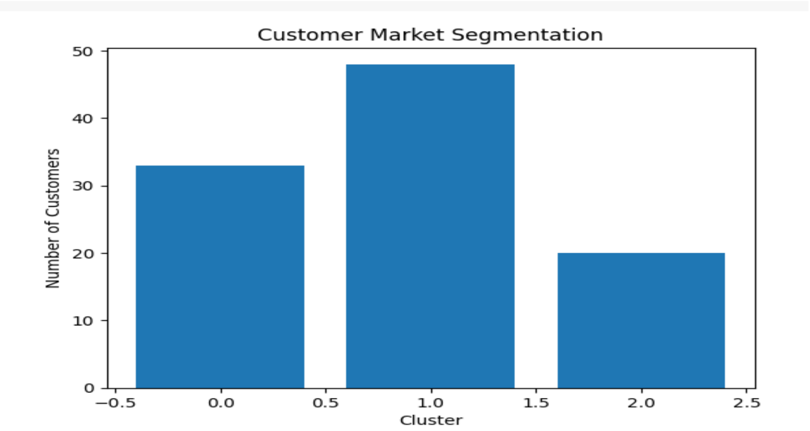


Fig.10:clustering

The Fig 10 represents the number of customers are assigned to each cluster. Based on the visualization the marketing manager comes to know who are the most valuable and trusted customers for their company. And what are the strategies should be used to retain more customers.

8. SYSTEM TESTING

8.1 Testcase Description

Testing is the process of detecting errors. Testing performs a very critical role for quality assurance and for ensuring the reliability of software. The results of testing are used later on during maintenance also.

Psychology of Testing

The aim of testing is often to demonstrate that a program works by showing that it has no errors. The basic purpose of testing phase is to detect the errors that may be present in the program. Hence one should not start testing with the intent of showing that a program works, but the intent should be to show that a program doesn't work. Testing is the process of executing a program with the intent of finding errors.

Testing Objectives

The main objective of testing is to uncover a host of errors, systematically and with minimum effort and time. Stating formally, we can say,

- Testing is a process of executing a program with the intent of finding an error.
- A successful test is one that uncovers an as yet undiscovered error.
- A good test case is one that has a high probability of finding error, if it exists.
- The tests are inadequate to detect possibly present errors.
- The software more or less confirms to the quality and reliable standards.

Levels of Testing:

In order to uncover the errors present in different phases we have the concept of levels of testing.

System Testing:

The philosophy behind testing is to find errors. Test cases are devised with this in mind. A strategy employed for system testing is code testing.

Code Testing:

This strategy examines the logic of the program. To follow this method we developed some test data that resulted in executing every instruction in the program and module i.e. every path is tested. Systems are not designed as entire nor are they tested as single systems. To ensure that the coding is perfect two types of testing is performed or for that matter is performed or that matter is performed or for that matter is performed on all systems.

8.2 Types of Testing:

Unit Testing:

Unit testing focuses verification effort on the smallest unit of software i.e., the module. Using the detailed design and the process specifications testing is done to uncover errors within the boundary of the module. All modules must be successful in the unit test before the start of the integration testing begins. Unit testing is first done on modules, independent of one another to locate errors.

Link Testing:

Link testing does not test software but rather the integration of each module in system. The primary concern is the compatibility of each module. The Programmer tests where modules are designed with different parameters, length, type etc.

Integration Testing:

After the unit testing, we must perform integration testing. The goal here is tosee if modules can be integrated properly, the emphasis being on testing interfaces between modules. This testing activity can be considered as testing the design and hence the emphasis on testing module interactions.

In this project integrating all the modules forms the main system. When integrating all the modules I have checked whether the integration effects working of any of the services by giving different combinations of inputs with which the two services run perfectly before Integration.

System Testing:

Here the entire software system is tested. The reference document for this process is the requirements document, and the goals to see if software meets its requirements.

Acceptance Testing:

Acceptance Test is performed with realistic data of the client to demonstrate that the software is working satisfactorily. Testing here is focused on external behavior of the system; the internal logic of program is not emphasized.

Test cases should be selected so that the largest number of attributes of an equivalence class is exercised at once. The testing phase is an important part of software development. It is the process of finding errors and missing operations and also a complete verification to determine whether the objectives are met and the user requirements are satisfied.

White Box Testing:

This is a unit testing method where a unit will be taken at a time and tested thoroughly at a statement level to find the maximum possible errors. I tested step wise every piece of code, taking care that every statement in the code is executed at least once. The white box testing is also called Glass Box Testing.

Black Box Testing:

This testing method considers a module as a single unit and checks the unit at interface and communication with other modules rather getting into details at statement level. Here the module will be treated as a block box that will take some input and generate output. Output for a given set of input combinations are forwarded to other modules.

9 CONCLUSION

Mini Batch K-means clustering was used to categorize customers based on their shopping habits. Data preparation and algorithm execution provided insights into client recency, frequency, monetary value, and spending score. The quality of the clustering findings was assessed using evaluation measures such as the silhouette score and the Davies-Bouldin Index. Visualizations helped to present the data and identify high-value client groupings. This demonstrated the need of data-driven approaches for understanding client behavior and optimizing marketing efforts. Finally, the analysis yielded practical recommendations for targeting certain client segments and increasing business growth. Mini Batch K-means clustering is excellent for segmenting clients and making informed decisions in marketing.

10 FUTURE ENHANCEMENT

Exploring creative methods and approaches for improving customer segmentation and behavior analysis across sectors. This includes exploring advanced machine learning methods and data processing frameworks to efficiently handle vast and complicated datasets.

Interdisciplinary research, including sentiment analysis and natural language processing, has the potential to enhance segmentation strategies. Extending client segmentation beyond retail to e-commerce, banking, and healthcare can reveal common concepts and best practices.

REFERENCES

[1] J. Zhu, Z. Jiang, G. D. Evangelidis, C. Zhang, S. Pang, and Z. Li, ``Ef_cient registration of multi-view point sets by K means clustering," Inf. Sci.,vol. 488, pp. 205_218, Jul. 2019.

[2] E. Y. L. Nandapala and K. P. N. Jayasena, “The practical approach in Customers segmentation byusing the K-Means Algorithm,” 2020 IEEE 15th Int.Conf. Ind. Inf. Syst. ICIIS2020Proc.,no.978,pp.344349,2020,doi:10.1109/ICIIS511 40.2020.9342639.

[3] Z. Lv, T. Liu, C. Shi, J. A. Benediktsson, and H. Du, ``Novel land cover change detection method based on k Means clustering and adaptive majority voting using bitemporal remote sensing images," IEEE Access, vol. 7,pp. 34425_34437, 2019.

[4] "UCI Machine Learning Repository: Online Retail Data Set", Archive.ics.uci.edu, 2020. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Online Retail.

[5] K. Torizuka, H. Oi, F. Saitoh, and S. Ishizu, “Benefit Segmentation of Online Cus tomer Reviews Using Random Forest,” IEEE Int. Conf. Ind. Eng. Eng. Manag., vol. 2019 Decem, pp. 487–491, 2019, doi:10.1109/IEEM.2018.8607697.

[6] B. K. Shah, A. K. Jaiswal, A. Shroff, A. K. Dixit, O.N. Kushwaha, and N. K. Shah, “Sentiments Detection for Amazon Product Review,” 2021 Int. Conf. Comput. Commun. Informatics,ICCCI2021,2021,doi:10.1109/ICCCI50826.2021. 9402414.

[7] E. Umuhoza, D. Ntirushwamaboko, J. Awuah, andB. Birir, “Using unsupervised machine learning techniques for behavioral-based credit card users segmentation in Africa,” SAIEE Africa Res. J., vol.111, no. 3, pp. 95–101, 2020, doi: 10.23919/saiee.2020.9142602.

[8] P. D. Hung, N. T. Thuy Lien, and N. D. Ngoc, “Customer segmentation using hierarchical agglomerative clustering,” ACM Int. Conf .Proceeding Ser., vol. Part F1483, pp. 33–37, 2019, doi: 10.1145/3322645.3322677.

[9] L. Abidar, D. Zaidouni, and A. Ennouaary, “Customer segmentation with machine learning: New strategy for targeted actions,” ACM Int. Conf. Proceeding Ser., pp. 221 226, 2020, doi: 10.1145/3419604.3419794.

[10] P. Monil, “Customer Segmentation using Machine Learnin,” Int. J. Res. Appl. Sci. Eng. Technol., vol.8, no. 6, pp. 2104–2108, 2020,doi:10.22214/ijraset.2020.6344.

[11] Yifei Wang Research on the analysis of commercial economic data based on hierarchical clustering algorithm”, IEEE InternationalConference on Power, Intelligent Computing and Systems (ICPICS),Year: 2020.

[12] Liang Li, Jia Wang, And Xuetao Li “Efficiency Analysis of MachineLearning Intelligent Investment Based on K Means Algorithm”, IEEE,Year: 2020.

[13] Sarker, Iqbal H., et al. “Behavdt: a behavioral decision tree learning tobuild a user-centric context-aware predictive model.” Mobile Networks and Applications 25.3 (2020): 1151-1161.

[14] Zhang, Pin, et al. “A novel hybrid surrogate intelligent model for creep index prediction based on particle swarm optimization and random forest.” Engineering Geology 265 (2020): 105328.

[15] Role of Big Data Analysis and Machine Learning in eCommerce - Customer Segmentation, Alphy Mathew, T J Jobin, National Conference on Emerging Computer Applications (NCECA)-2021, Vol.3, Issue.1, ISBN:978-93 5426-386-6@2021.

[16] A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA Maha Alkhayrat, Mohamad Aljnidi and Kadan Aljoumaa, Springer Open, Journal of Big Data, (2020) 7:9 Alkhayrat et al. J Big Data.

[17] H. Chen, L. Zhang, X. Chu, and B. Yan, “Smartphone customer segmentation based on the usage pattern,” Advanced Engineering Informatics, vol. 42, no. October, 2019. [Online].Available:https://www.sciencedirect.com/science/article/abs/pii/S1474034619305737.

[18] A. Maulina, Isti, "Data Mining Approach for Customer Segmentation in B2B Settings using Centroid-Based Clustering," 2019 16th International Conference on Service Systems and Service Management (ICSSSM), no. March, pp. 1–6, 2019.

[19] A. A. Aktas¸, O. Tunalı, and A. T. Bayrak, "Comparative unsupervised clustering approaches for customer segmentation," in 2021 2nd International Conference on Computing and Data Science (CDS). IEEE, 2021, pp. 530 535.