

Customer Segmentation Using Machine Learning to Improve Marketing

Ms. B Gnaneshwari Devi¹, K. Dinesh Santhosh Kumar², A. Rajesh³, Ch. Revanth⁴, G. Kalyan⁵

¹ Assistant Professor, Department of Computer Science

^[2-5] B.Tech Student, Department of Computer Science

^[1-5] Raghu Engineering College, Visakhapatnam

Abstract - Customer segmentation is the process of classifying customers according to characteristics they have in common. This enables organizations to effectively and accurately market to each group. In business-to-business marketing, a firm may divide its clientele based on a variety of criteria, such as location, industry, number of employees, and previous purchases made from the business. Marketers can more effectively target different audience subsets with their marketing efforts by using segmentation. These initiatives may pertain to product development as well as communications. In particular, segmentation aids in a business's ability to craft and deliver marketing communications that are specifically intended to appeal to some client segments while excluding others. Search and select the most effective communication channel. Based on the segmented clusters, this could be radio advertising, social media (Instagram, Telegram, Twitter, Youtube, etc.) email, or another strategy. We have used Gradio (an API for UI), the KNN Classification Algorithm (to classify new consumers), and the K-Means algorithm (to segment the customers) to accomplish this goal. Determine how to make goods or new opportunities for products or services better. It starts with collecting and evaluating data and concludes with taking suitable and useful action based on the information obtained.

Key Words: Segmentation, targeted marketing, K-Means, KNN, Gradio.

1. INTRODUCTION

In the modern business world, success depends on taking into account a wide range of technology due to the intense rivalry. The most important element in any business is data. We can carry out several operations to determine customer interests with the aid of unlabelled data. By grouping the vast amounts of data from the dataset according to factors like age, gender, and wealth, customer segmentation can be helpful. An alternative name for these groups is clusters. This enables us to determine, amongst other things, which goods are likely to sell a lot and which particular age group is making the purchases.

Customer segmentation is essential for every customer category, supporting business decision-making, managing and identifying products for each customer category, handling the availability of the item and consumer demand solution, displaying customer defection, and recognizing and concentrating on a prospective customer base. This is because it enables the modification of marketing plans to increase their

effectiveness. It gives firms greater leeway to engage in productive rivalry and to be more creative in their customer acquisition and retention strategies.

Customer segmentation is one of the most important initial stages toward improved customization. Here's where customization starts, and smart segmentation will assist you in selecting new products, services, prices, offers, and even recommended in-app purchases. On the other hand, manually segmenting then could take a long time.

A company's ability to build strong, individualized relationships with its clients is essential to its success. Customer segmentation assists a business in dividing its clientele into groups according to their requirements, purchasing patterns, financial status, and other pertinent variables. This aids a business in assessing the needs of its clientele as well as their priorities, interests, and spending patterns, all of which may be utilized to inform various marketing tactics and customer acquisition and retention programs.

The exploratory data analysis on additional customer data sets will be covered in this project. Thus, it assists us in optimizing the understanding of a dataset and its fundamental structure while offering all the particular elements that an analyst would seek to extract from a dataset. In this project, I'm going to apply clustering analysis to tackle this problem. To segment customers, this project makes use of KNN categorization. Following customer segmentation, analysts and decision-makers may begin focusing on specific consumers with whom to launch their marketing campaigns. By using segmentation, marketers may more effectively target certain audience segments with their campaigns.

2. LITERATURE SURVEY

“Peker, Serhat, Altan Kocyigit, and P. Erhan Eren-LRFMP model for customer segmentation in the grocery retail industry: a case study.”: [1]

P. Erhan Eren, Altan Kocyigit, and Serhat Peker's study presents the LRFMP model for customer segmentation in the grocery retail industry. This approach divides consumers into groups based on attributes including product preferences, monetary value, frequency, and recency. In this way, merchants may better customize their product offerings and marketing methods to match the unique requirements of each market group. A case study showcasing the model's practical application is probably included in the article.

“Hamka, Fadly, et al- Mobile customer segmentation based on smartphone measurement.”: [2]

This 2014 paper from Telematics and Informatics focuses on smartphone measurement data for mobile client segmentation. It probably looks at how segmenting mobile clients may be done with the help of smartphone data—like location, app choices, and usage habits. The project hopes to discover discrete groups of mobile users with comparable behavior or attributes by utilizing this data. The results could offer mobile service providers useful information that they can use to better target specific mobile client categories with their services and marketing campaigns.

“Aryuni, Mediana; Evaristus, Didik Madyatmadja; Miranda, Eka (2018)- [IEEE 2018 ICIMTech]”: [3]

Presenting at the IEEE 2018 ICIMTech, this study examines how Kmeans clustering methods are used within XYZ Bank to segment customers. Customer segmentation is an essential part of marketing and CRM that helps businesses target and understand their clientele more effectively. You may categorize your customers into groups based on similarities in preferences, behavior, or demographics by using K-Means and K-Medoids clustering.

“Wu, Jing; Lin, Zheng (2005) [The Seventh International Conference, ACM Press, Xi'an, China, August 15,–August 17, 2005]”: [4]

The paper by Wu Jing and Lin Zheng presents a model for customer segmentation using clustering techniques. It focuses on dividing customers into groups based on similarities in their behavior or characteristics. By employing clustering algorithms, businesses can better tailor their marketing strategies and product offerings to different customer segments. This approach aids in enhancing customer satisfaction and marketing effectiveness in electronic commerce.

3. PROPOSED SYSTEM

Below is a proposed machine learning-based customer segmentation:

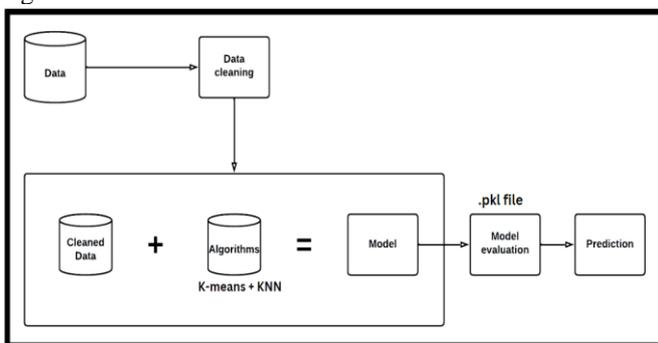


Fig 1: System Block Diagram

Data Collection: To construct a machine learning project for consumer segmentation, we must gather information from a variety of sources, including online stores, retail centers, user profiles, past purchases, website usage, etc. The dataset has 13766 samples once all the data have been combined. The variables in each row (sample) of the data are ID, Name, Email, Age, Gender, Spending Score, and Monthly Income.

Data cleaning: Once the data is collected, it needs to be pre-processed by cleaning, transforming, and normalizing it to prepare it for machine learning algorithms. By being aware of the data and also the quantity of varying variables present in the input file. The kinds of data, feature variables, null value checks, and other aspects of it may all be analyzed.

Feature Selection: The next step is to select relevant features that can help identify customer segments. For example, age, gender, spending score, and monthly income are essential features that are opted for in this project for customer segmentation.

Algorithms: After cleaning the data and performing feature selection, the cleaned data is used for applying machine learning algorithms. The ideal number of clusters to use for consumer segmentation is ascertained using the elbow approach. Subsequently, the k-means clustering technique is utilized to generate clusters and allocate every sample to the relevant cluster. Any fresh data samples are classified using KNN classification. KNN is a well-liked consumer segmentation algorithm due to its ease of implementation and capacity for handling large data sets. Outliers from fresh data samples can be handled by it.

Model: The model is the outcome of applying the machine learning algorithms to cleaned data. This model can now be used for classifying large datasets, new data samples, etc.

Model Evaluation: The trained model needs to be evaluated to determine its accuracy and performance. The evaluation can be done using techniques such as cross-validation, hyperparameter tuning, confusion matrix, accuracy score, AUC and ROC, etc. After evaluation, we will generate a .pkl file which can be portable and is used for prediction.

Prediction: To keep the model accurate, it must be regularly retrained and its performance monitored. There is a chance that newly collected data will exhibit distinct patterns from the training set of data due to data drift. Consequently, ongoing observation is essential to guarantee that the model stays accurate and relevant.

4. IMPLEMENTATION MODULES

Module 1: Cleaning the Dataset

Import Libraries: 1. re: Python libraries have a built-in module called 're', which can be used for evaluating and solving regular expressions in Python. By using a regex pattern along with the necessary methods, it can find the presence or absence of a string.

2. random: Python libraries have a built-in module called 'Random' which is used for generating random numbers. These are not actually random but the numbers are generated using a very complex process that looks like random. It can

also generate random elements from a sequence like lists, strings, tuple, etc.

Cleaning names: Using the findall() method of the random module along with the search pattern to remove unnecessary symbols, numbers, and extra spaces in the names.

Cleaning gender: The machine learning algorithm uses numerical values for performing analysis, and for data visualization also needs numerical values for plotting graphs, etc. We convert “male” or “m” in any case to 1 and “female”, or “f” in any case to 0.

Module 2: K-Means Clustering

Import Libraries: kmeans: By importing the kmeans module from the ‘sklearn.cluster’ package of the python library which is used for implementing the K-means clustering algorithm based on similarity.

Elbow Method: The elbow method is used for determining the ‘k’ no of clusters in the K-means algorithm. This is done by iteratively increasing the number of centroids and assigning data points having the closest distance to each centroid as a cluster.

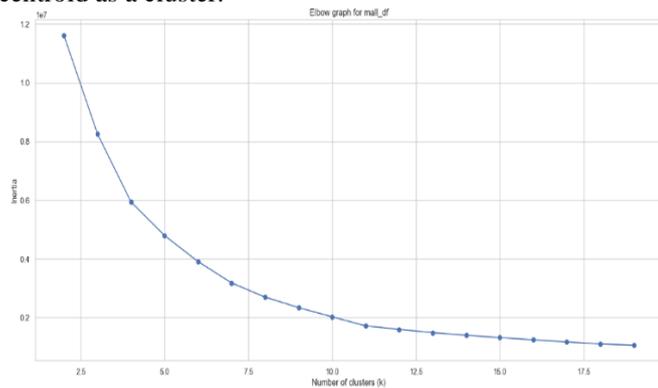


Fig 2: Elbow graph

Then find the inertia value using the WCSS distance method and plot that value on the graph. The value at which the graph drops slowly is considered as the value of ‘k’. The elbow graph for the given dataset is shown in above Fig 2.

From the above Fig 2, it is visible that the curve starts decreasing sharply from 1.2 inertia onwards and then starts to decrease slowly from 0.45 inertia onwards at 5 clusters. So we consider the ‘k’ value to be 5, but we can also choose ‘k’>5 also but for this project, creating that many clusters may be confusing for its main purpose i.e., marketing.

Kmeans Clustering: The first step in K-means clustering is choosing the number of clusters, K, which is done using the elbow method. The method then chooses K data points at random as the dataset’s first centroids. The clusters’ initial positions are represented by these centroids. Each data point is then matched with the closest centroid using a selected distance metric, most often the Euclidean distance. The first clusters are formed by this assignment process.

The algorithm then recalculates the centroid of each cluster by taking the mean of all the data points assigned to that cluster. Up until convergence, the procedure repeats, reassigning data samples to the nearest centroid and recalculating centroids. When there is no more redistribution of data points, convergence takes place, signifying stable cluster centroids. The procedure ends at this stage and the last clusters form.

The below fig. 3 shows the distribution of data samples around the centroids of each cluster.

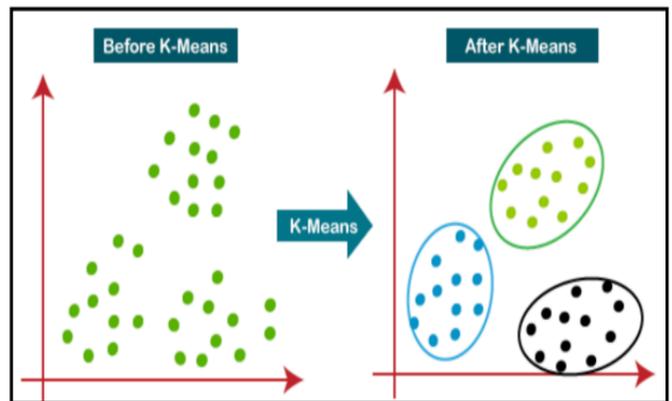


Fig 3: Clusters formed using Kmeans Clustering

Module 3: KNN Classification Algorithm

Import Libraries: KNeighborsClassifier: This is a part of the sklearn.neighbors module in Python, which is part of the scikit-learn library. It’s a type of supervised learning algorithm used for classification tasks.

KNN Classification Algorithm: The KNN classification algorithm is frequently applied to continuous numeric data regression as well as data classification. The fact that the machine learning technique is non-parametric suggests that it doesn’t make any assumptions regarding the data. Additionally, it makes no presumptions, instead, it classifies unknown or uncategorized data items by merely examining their neighboring data points.

Here, rather than the regression component of KNN, we will focus on its classification feature. The KNN machine learning technique is predicated on the idea that comparable data points are located in close proximity to one another and operate depending on the distance between the data points. If your neighbors are categorized in a certain way, you will most likely be placed in the same category as them due to your close proximity to them. Conversely, those who live further away yet are not in close proximity to you will likely fall into a separate group.

The gray color spots in Figure 4 above are expected to be red and blue. A decision boundary or contour (high-dimension space) is created by a predictive algorithm, and a prediction is generated based on a point’s placement on either side of the boundary or contour. Which spots are expected to be red or

blue are indicated by the decision border. Furthermore, as variable k changes, so does this decision boundary.

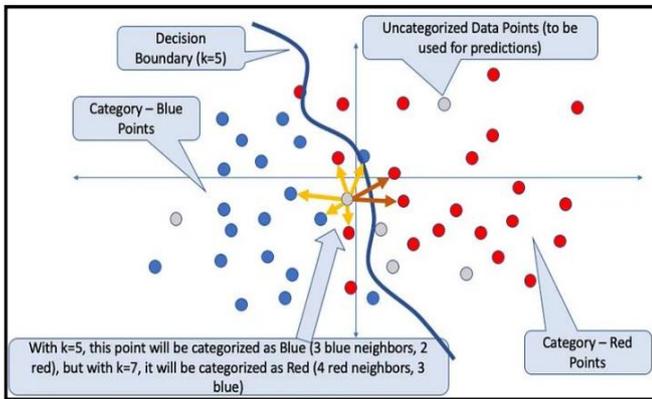


Fig 4: KNN Classification

By altering the value of k , we may derive a new boundary that might be more predictive of the outcomes. Choosing a value for k that results in a higher prediction accuracy rate is known as parameter tweaking. The KNN model contains many different parameters, much like k . To identify a collection of variables that perform well with the customer churn data set, we will experiment with these values.

5. RESULT AND ANALYSIS

A. Loading Dataset:

This particular dataset contains 13,766 records (samples) of data. This dataset consists of an ID, names, gender, email, age, spending_score, and income as seen in Fig. 5.

ID	names	gender	emails	age	spending_score	income	
0	1	barjraj	1	barjrrjraj9331@gmail.com	24	15	38
1	2	ramdin verma	1	ramrama44001@gmail.com	20	14	42
2	3	amit	1	amiammit5387@gmail.com	24	11	36
3	4	kushal	1	kushashushal79321@yahoo.com	23	16	44
4	5	kasid	1	kassid9735@gmail.com	21	12	46
5	6	shiv prakash	1	shivhikash763@gmail.com	19	12	41
6	7	vikram singh	1	vikvingh764@yahoo.com	19	14	47
7	8	sanjay	1	sanjannjay553@gmail.com	20	16	59
8	9	abhi	1	abhibhhabhi860@gmail.com	22	13	30
9	10	ram dutt gupta	1	ramrapta4123@gmail.com	24	16	50

Fig 5: Dataset

B. Data Distribution:

Age Distribution: The dataset contains the age variable whose distribution is shown below in Fig. 6. From Fig. we can observe that most people purchasing are between the ages of 20 to 27. And then just gradually fluctuating till the end.

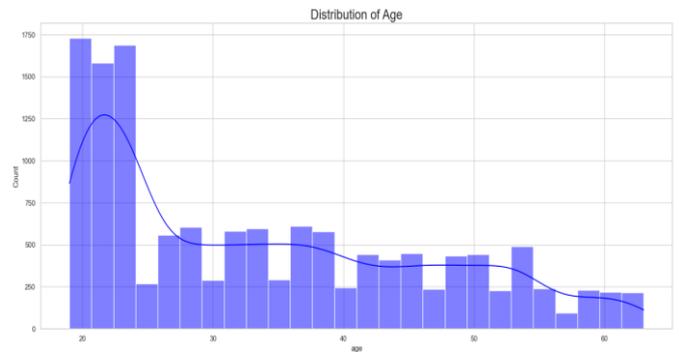


Fig 6: Age Distribution

Gender Analysis: We can observe from the gender analysis plot which is a pie chart, from Fig. 7, the dataset shows that there are more female clients than male clients. Here, females are 51.17% and males are 48.83% as shown from the below pie chart - Fig.7.

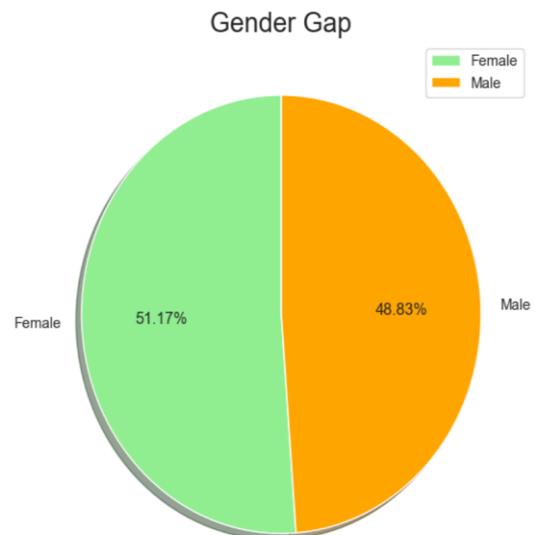


Fig 7: Gender Analysis

Spending Score Distribution: The spending_score distribution from the below Fig. 8 shows a histogram plot. We can observe that there are not many deep fluctuations in the plot with most people spending in the range of 40-80.

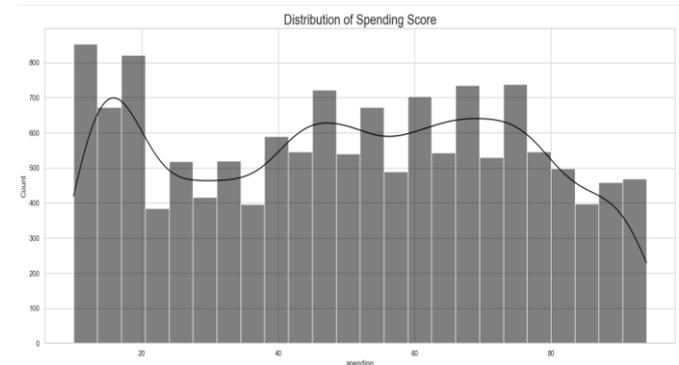


Fig 8: Spending Score Distribution

Income Distribution: Figure 9 shows that the majority of the customer's monthly incomes fall between Rs.55,000 and Rs.90,000.



Fig 9: Monthly Income Distribution

Gender vs Spending Score: By making a violin plot distribution of gender vs spending score, from Fig. 10, we can observe that more females (x-axis = 0) have spending scores between 60 – 80, whereas in the case of males (x-axis =1) mostly in between 35 – 60 as shown in the below fig.10.

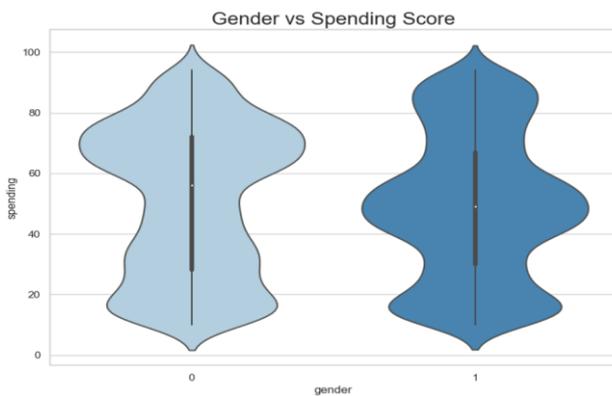


Fig 10: Gender vs Spending Score

Gender vs Income: Again, by making a violin plot distribution of gender vs income, we can observe that most females are earning between 10k-25k, 50k-100k only, very few females have an income between 25k-50k which can be seen as a dip in the plot. Whereas in the case of males, most of them have income in the ranges between 25k-60k, 80k-140k as shown in the below fig.11.



Fig 11: Gender vs Monthly Income

Elbow Graph: In an Elbow Graph, from Fig. 12, we can observe that inertia greatly decreases from 'k' 1 to 4, but from 'k' = 5 it decreases slowly indicating the ideal quantity of clusters to be taken for k-means clustering should be from $k \geq 5$.

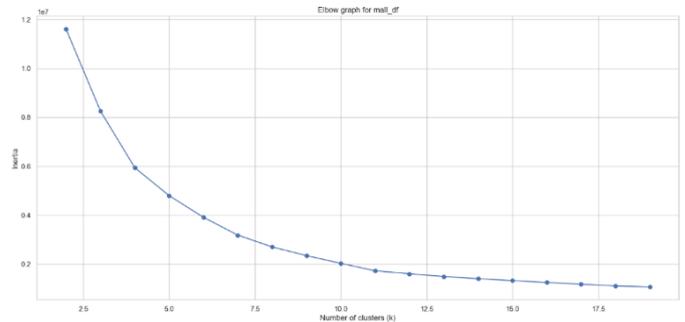


Fig 12: Elbow Graph

KMeans Clustering Algorithm (with 2 features): For this project, the kmeans clustering is done on 2 categorical features i.e., income, and spending_score. Below Fig. 13 shows the kmeans clustering algorithm for the 2 features.

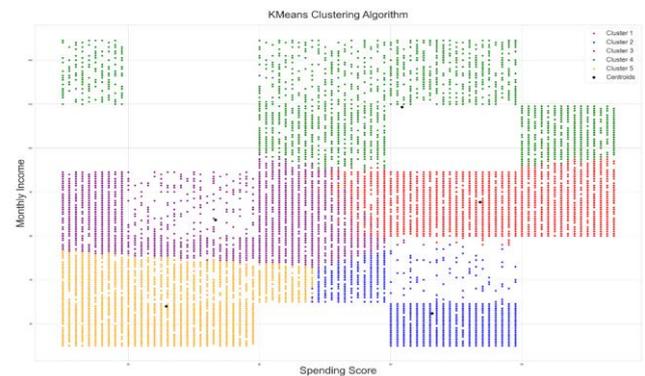


Fig 13: Kmeans Clustering (2 features)

KNN Classification: As shown in Figure 14 below, we were able to obtain a Test Accuracy of 99.31% and a Train Accuracy of 99.45% by utilizing KNN classification with "9" neighbors.

Train accuracy: 99.45%
 Test accuracy: 99.31%
 1 4 2 2 4 4 3 3 0 0 2 2 3 4 3 2 3 1 2 3 3 1 2 1 3 3 3 1 0 3 0 3 4 4 4 0 1 1 0

Fig 14: KNN Classifier Test & Train Accuracy

Project Result: The below fig.15 shows the outcome of the project, the interface is implemented using the Gradio module of Python. The input parameters are name, gender, age, spending score, and income along with their validations. In the below fig.15, Name= 'Santhosh Kumar', Gender= 'Male', Age=22, Spending Score= 30, Income= 45.

Customer Segmentation Using Machine Learning

Name
Santhosh Kumar

Gender
Male

Age
22

Spending Score
30

Income
45

Clear Submit

output
Santhosh Kumar belongs to cluster 4.

Flag

Use via API  Built with Gradio 

Fig 15: Customer Segmentation outcome

For these inputs, the trained model of the KNN classification algorithm is applied which classifies the given customer input data and gives the result as ‘Santhosh Kumar belongs to cluster 4.’

6. CONCLUSION

All things considered, a machine learning-based consumer segmentation project may offer insightful analysis and helpful suggestions to companies trying to raise revenue, increase customer happiness, and interact with customers. It is important to guarantee that the data employed is precise, pertinent, and inclusive of the intended audience. Additionally, the models and algorithms employed must undergo validation and accuracy testing. You may find underserved or unexplored consumer niches that offer your company new growth prospects by using customer segmentation. You may boost the likelihood of conversion and upselling by focusing on particular consumer categories with tailored offers, promotions, and suggestions. You may learn more about the unique requirements, problems, and driving forces of various client segments by using customer segmentation.

In this project, we have taken 2 categorical features like spending score, and income. But we can even use age, race, nationality, profession, work experience, and locality and perform segmentation based on 2 or more features. In the future, we can use other clustering algorithms like DBSCAN, Gaussian Mixture Model, and BIRCH algorithms for better utilization or resource optimization purposes. In our project we have achieved 99.31% accuracy by using the KNN Classification algorithm, we can also use other algorithms like Naïve Bayes, and Random Forest. We can also use the XGboost classifier to achieve even better accuracy.

REFERENCES

1. Peker, Serhat, Altan Kocyigit, and P. Erhan Eren- "LRFMP model for customer segmentation in the grocery retail industry: a case study." [1].
2. Hamka, Fadly, et al- "Mobile customer segmentation based on smartphone measurement." [2].
3. Aryuni, Mediana; Didik Madyatmadja, Evaristus; Miranda, Eka (2018)- *[IEEE 2018 International Conference on Information Management and Technology (ICIMTech)* [3].

4. Wu, Jing; Lin, Zheng (2005) [ACM Press the 7th international conference - Xi'an, China (2005.08.15-2005.08.17)] [4].
5. Bhade, Kalyani, et al- "A Systematic Approach to Customer Segmentation and Buyer Targeting for Profit Maximization." [5].
6. Hwang, Jinsoo, et al- "Customer segmentation based on dining preferences in full-service restaurants." [6].