IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 3, March 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303115 |

Customer Segmentation Using Ensemble Learning

Vamsi Desam[1], Ravi Sankar Beeram [2], Sethu Madhav Doredla[3], Bhargavi Lakshmi Chennupati[4],Yashwanth Kumar Addagiri[5]

Assistant Professor, Dept. of Computer Science and Engineering, Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India[1]

Undergraduate Student , Dept. of Computer Science and Engineering, Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India[2,3,4,5]

ABSTRACT: This project addresses the pivotal role of customer segmentation in refining marketing strategies and improving overall customer experience. Recognizing the limitations of traditional clustering methods in handling diverse and complex customer datasets, we propose a novel approach leveraging ensemble clustering. The study aims to bridge the existing research gap by evaluating the effectiveness of ensemble clustering in comparison to conventional techniques like K-means, Agglomerative, and Spectral. Our methodology involves the integration of multiple clustering algorithms and the assessment of segmentation quality using established metrics like silhouette score. Results indicate that the ensemble approach outperforms traditional methods, showcasing enhanced accuracy and robustness in capturing nuanced customer segments. This research underscores the significance of ensemble clustering for businesses seeking more accurate and actionable insights into customer behavior, thereby paving the way for refined marketing strategies and improved customer satisfaction. The projected results show that the silhouette score of ensemble clustering outperforms all other clustering algorithms with a silhouette score of 0.400891747

KEYWORDS: machine learning; ensemble clustering; k-means clustering; agglomerative clustering; silhouette score; spectral clustering

I. INTRODUCTION

Clustering is a fundamental technique in unsupervised machine learning used to group similar data points together based on certain features or characteristics. The goal of clustering is to partition a dataset into groups, or clusters, such that data points within the same cluster are more similar to each other than to those in other clusters. Clustering is widely used in various domains including data analysis, pattern recognition, image processing, and customer segmentation. Unlike labeled data, which has annotations or ground truth labels indicating the category or class to which each data point belongs, unlabeled data lacks such annotations. Unlabeled data is common in many real-world scenarios, and its analysis often falls under the domain of unsupervised learning. In unsupervised learning, the goal is to extract meaningful patterns, structures, or relationships from the data without the presence of explicit labels. Clustering, dimensionality reduction, and density estimation are some common tasks performed on unlabeled data.

Unsupervised machine learning can be achieved using two primary methods: dimensionality reduction and clustering. Clustering appears to aid in the surface analysis of unstructured data. A few factors that affect cluster formation are graphing, the shortest distance, and data point density. The method of clustering involves calculating an object's degree of relatedness using a metric known as the similarity measure. Smaller feature sets make it simpler to find similarity metrics. As the number of features rises, creating similarity metrics becomes more difficult. There are various kinds of clustering techniques to evaluate the unlabeled data.

- Centroid-based Clustering (Partitioning methods)
- Density-based Clustering (Model-based methods)
- Connectivity based Clustering (Hierarchical Clustering)
- Distribution-based clustering

There are many types of clustering algorithms, but K-means and hierarchical clustering are the most widely available in

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303115 |

data science tools. The spectral clustering technique which is used for handling non-convex structures, sensitivity to initializations, spectral technique does not require specifying the number of clusters at the initial stages since the spectral algorithm can detect the number of clusters based on the eigen structure of the similarity graph.

The goal of dimensionality reduction is to minimize the number of features in a dataset while preserving the majority of the pertinent data. Stated differently, it is the process of converting high-dimensional data into a lower-dimensional space while maintaining the essential characteristics of the source data. High-dimensional data in machine learning refers to data that has a lot of features or variables. A prevalent issue in machine learning is known as the "curse of dimensionality," which states that a model's performance declines with an increase in features. This is because as the number of characteristics in the model increases, so does its complexity, making it harder to come up with a workable solution. High-dimensional data can also result in overfitting, a phenomenon in which the model fits the training set too closely and performs poorly when applied to fresh data. Dimensionality reduction can lessen these issues by making the model simpler and enhancing its ability to generalize.

This paper contributes by applying the ensemble and individual clustering techniques on the online store data to analyze the unlabeled data. The proposed methodology is designed to apply clustering and dimensionality reduction to understand the different purchasing behaviors of the customers as well as enhance the marketing strategies based on the outcome. The rest of the paper is organized as follows: Section 2 introduces the related work. The data collection and data engineering are given in Section 3. Section 4 introduces the methodology of the proposed clustering technique's performances based on the silhouette score. The experiment results are given in Section 5. The conclusion and future works are given in Section 6. The references are given in Section 7.

II. RELATED WORK

Over the years, as there has been very strong competition in the business world, organizations have had to enhance their profits and business by satisfying the demands of their customers and attracting new customers according to their needs. The identification of customers and satisfying the demands of each customer is a very complex and tedious task. This is because customers may be different according to their demands, tastes, preferences, and so on. Instead of a “one-size-fits-all” approach, customer segmentation clusters the customers into groups sharing the same properties or behavioral characteristics [8]. The segmentation process can help businesses and companies understand their customer groups, target the right groups and develop effective marketing strategies for different targeted groups [10]. Customer segmentation allows companies to visualize what the customers are buying which will prompt the companies to better serve their customers resulting in customer satisfaction, it also allows the companies to find who their target customers are and improvise their marketing tactics to generate more revenues from them [9].

In their work documented in [2], hemashree and their colleagues used the k-means clustering technique to perform the clustering in which they used the k value as 5 which is an optimal value obtained when they performed the elbow method to find the optimal number of clusters. They evaluated the performance of the model using a silhouette score which measures how similar the data point is related to others. They also In the research outlined in [3], this research was performed by Yun Chen and his team and is based on chine telecom dataset, they used k-means clustering for the purpose of data mining with k value as 4 and nearly 196 attributes are used to find churn and performed analysis. as outlined in [3]. Patcharin Poniyam and his colleagues

performed a data mining research on customer behavior analysis and Experimental data were obtained from a store at Khon Kaen Thailand total of 22,450 instants. The components and attributes subject to the analysis included customer type, product type, and bill no. and product list. They used apriori and k-means algorithm to find the behavior of clients. The k value obtained after performing k-means is 5 by using elbow method produced good results when compared to apriori algorithm [4].Parameshwara Joga and colleagues, as detailed in [6], performed a comparative analysis of various algorithms. The paper analysis the efficiency of four machine learning algorithms, namely, K-Means, DBSCAN, Agglomerative Clustering, and PCA with KMeans, in solving the customer segmentation problem. The paper compares their performance in segmenting mall customer dataset. They performed PCA on 4 attributes and reduced it to 2 Attributes. They used to Davis Boulder and silhouette score to measure the performance , They used WCSS(within cluster sum of squares)method to find the optimal number of clusters

E.Y.L Nandapala and collaborators, in their research work[7] published the work by using various algorithms to find the optimal number of clusters. They are the Elbow method, gap static method, and finally silhouette

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303115 |

Method. After finding an optimal number of clusters they used k-means along with PCA to cluster the data and found different behaviors of customers by clustering the data.

III. PROPOSED METHODOLOGY

The project is divided into multiple important components, each with a distinct function. Data collection includes a collection of data from the Kaggle open data source for clustering. Data cleaning and feature engineering are essential steps in the machine learning pipeline that directly impact the quality, performance, interpretability, and efficiency of the resulting models. Data processing techniques allow for the transformation of raw data into a more suitable format for analysis. This may include converting categorical data into numerical format, scaling numerical features, or encoding text data into a format that machine learning algorithms can understand. Dimensionality reduction is useful for efficient and effective analysis of high-dimensional data while preserving essential information and improving the performance of machine learning models. Clustering facilitates pattern recognition, data summarization, and decision-making through data analysis and machine learning. After performing individual clustering and ensemble clustering or consensus clustering these models are evaluated individually by using silhouette analysis. Profiling plays a crucial role in clustering analysis by providing a deeper understanding of clusters, facilitating the interpretation and validation of results, and supporting decision-making processes in various applications.
Our proposed methodology depends on k-means, hierarchical, spectral clustering, and ensemble clustering methods for the clustering task, specifically aimed at identifying various hidden patterns and purchasing behavior of customers. The outline specifies the step-by-step approach of our project.

- dataset acquisition.
- Importing libraries.
- Data Cleaning and Feature Engineering.
- Data preprocessing and Feature scaling.
- Dimensionality reduction using Principal Component Analysis.
- Clustering the Data
- Evaluation of Model Based on Silhouette Score.

3.1. Dataset acquisition:

The dataset comprises 2240 entries with 29 different features. The images were obtained from the Kaggle and Git websites. All brain MRIs are split into training and testing folders. The Dataset used in this project contains four sub-parts: Customer's Information, Product's Information, Place, and Promotional Information. A representative sample of the dataset is depicted in Fig. 1. The size of the dataset is 220.19 KB.

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency
0	5524	1957	Graduation	Single	58138.0	0	0	04-09-2012	58
1	2174	1954	Graduation	Single	46344.0	1	1	08-03-2014	38
2	4141	1965	Graduation	Together	71613.0	0	0	21-08-2013	26
3	6182	1984	Graduation	Together	26646.0	1	0	10-02-2014	26
4	5324	1981	PhD	Married	58293.0	1	0	19-01-2014	94

5 rows × 29 columns

Fig.1: Online Store Sample Dataset

Dataset-Link:
https://www.kaggle.com/code/karnikakapoor/customer-segmentation-clustering

3.2. Importing libraries:

Various libraries are imported while working with this project and each libraries contains unique purpose. The libraries include

- numPy: for numerical computing and array operations.
- pandas: for data loading and manipulation.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303115 |

- matplotlib, seaborn, mpl_toolkits: for data visualization.
- sklearn: for clustering and data evaluation.
- scipy: for ensemble clustering.
- yellowbrick: for finding k-value.
- datetime: for feature engineering.
- warnings: to handle and control warnings.

3.3. Data Cleaning and Feature Engineering:

In the machine learning (ML) pipeline, data cleaning is an essential phase that entails finding and eliminating any duplicate, irrelevant, or missing data. Ensuring that the data is reliable, consistent, and error-free is the aim of data cleaning, since inconsistent or inaccurate data can have a detrimental effect on the performance of the machine learning model.

In data cleaning we have removed the outliers using dropna method as we have only 2% of outliers depicted in the Fig2. The method used to remove the entries with null values.

Fig.2: Removing Null values using dropna method

```
data = data.dropna()
print("The total number of data-points after removing the rows with missing values a
```

```
The total number of data-points after removing the rows with missing values are: 2
216
```

Feature engineering is creating new features from the existing features according to our requirements.

[Customer_For] from [Dt_Customer]: No of days the customer has done shopping relative to the newest. [Age] from [Year Birth]: Age of the customers.

[spent] from [various products]: Total spending of customers.

[living with] from [martial_status]: So we divide 4 categories into 2[Partner, alone].

[children] from [kidhome]+[teenhome]: redundancy removing.

[family_size]: Total size of family.

[Is_Parent]: Parent or not.

[Education]: Reducing the no of categories of education.

3.4. Data Preprocessing and Feature Scaling:

Label encoding the categorical features: Converting the categorical attributes into numerical values. The Fig.3 describes the conversion.

Fig .3:converting categorical into numerical values

```
s = (data.dtypes == 'object')
object_cols = list(s[s].index)

print("Categorical variables in the dataset:", object_cols)
```

```
Categorical variables in the dataset: ['Education', 'Living_With']
```

```
LE=LabelEncoder()
for i in object_cols:
    data[i]=data[[i]].apply(LE.fit_transform)

print("All features are now numerical")
```

```
All features are now numerical
```

Feature Scaling: Scaling all the features between 0 and 1[both included].The Fig :4 depicts the scaling of features

Fig :4 converting categorical into numerical values

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303115 |

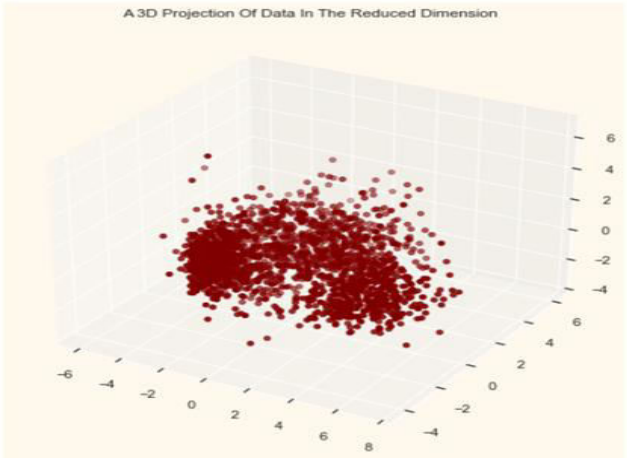
```
ds = data.copy()
# creating a subset of dataframe by dropping the features on deals accepted and promotions
cols_del = ['AcceptedCmp3', 'AcceptedCmp4', 'AcceptedCmp5', 'AcceptedCmp1','AcceptedCmp2', 'Complain', 'Response']
ds = ds.drop(cols_del, axis=1)
#Scaling
scaler = StandardScaler()
scaler.fit(ds)
scaled_ds = pd.DataFrame(scaler.transform(ds),columns= ds.columns )
print("All features are now scaled")
```

```
All features are now scaled
```

3.5. Dimensionality Reduction:

The amount of data needed to provide a statistically significant result grows exponentially with the number of features or dimensions in a dataset. This may result in problems including overfitting, longer computation times, and decreased machine learning model accuracy. When dealing with high-dimensional data, issues known as the "curse of dimensionality" can occur. The number of feature combinations that can be obtained grows exponentially with the number of dimensions, making it more computationally demanding to collect a representative sample of the data and more costly to carry out tasks like clustering or classification. For Reducing dimensions we are using Principal Component Analysis (PCA).In Fig.5 the dataset is converted into three dimensional graph

Fig.5: Converting Entire Dataset into Three dimensions



3.6. Clustering the Data:

Three clustering techniques and ensembling using majority voting of all these individual clusters are used to cluster the data and analyse the clusters based on their features and find the hidden patterns.

K-means:

K means clustering and assigns data points to one of the K clusters depending on their distance from the center of the clusters. The centroid of the cluster is first randomly assigned to the space. Based on its distance from the cluster centroid, each data point is then assigned to a cluster. Each point is assigned to a cluster after which fresh cluster centroids are assigned. Until it identifies a suitable cluster, this procedure iteratively runs. Assuming that the number of clusters is predetermined, we must assign points to each group during the analysis.
[7,10].Here we are using elbow method to find the optimal number of clusters, after clustering we found k=4 is optimal. Fig.6 depicts the elbow method to find the optimal number of clusters

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303115 |

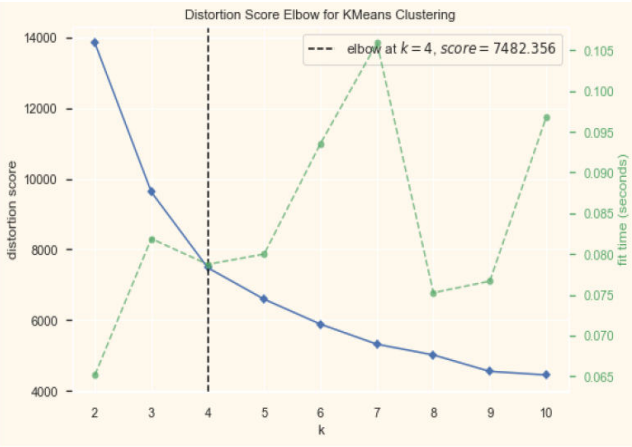


Fig.6 : Elbow method to find the optimal number of clusters

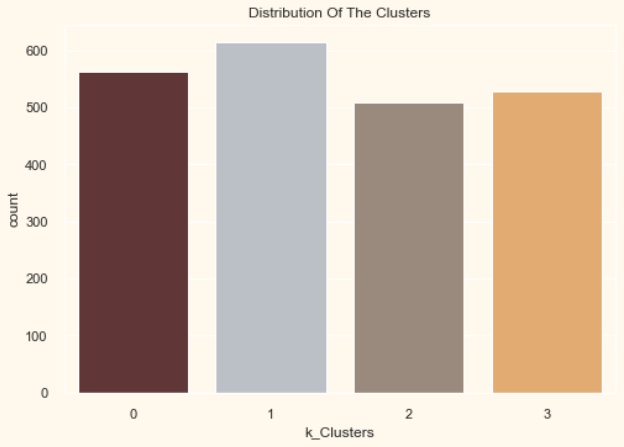
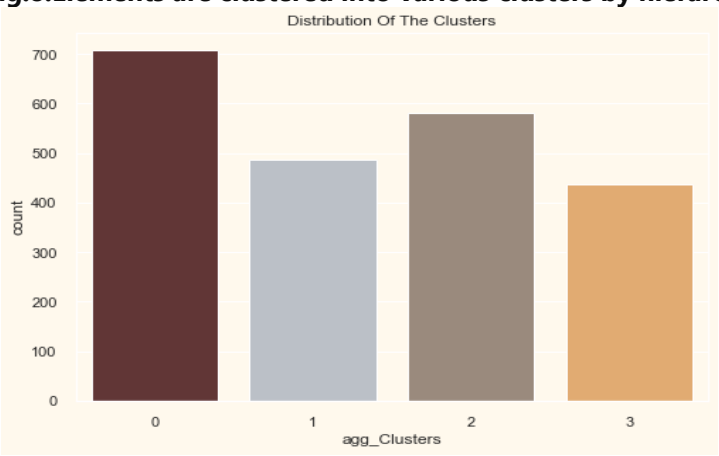


Fig.7 : Distribution of points in the clusters using kmeans

Agglomerative clustering:

Agglomerative clustering is a hierarchical clustering technique that builds clusters by iteratively merging smaller clusters into larger ones. It is a bottom-up approach where each data point initially starts as its own cluster, and then pairs of clusters are merged together based on a specified linkage criterion until a single cluster containing all data points is formed. The Fig.8 shows the points that are assigned in various clusters for n=4

Fig.8:Elements are clustered into various clusters by hierarchial clustering



Spectral Clustering:

Spectral clustering is a technique used for partitioning data into clusters based on the similarity between data points. Unlike traditional clustering methods like K-means, which operate in the original feature space, spectral clustering operates on a transformed representation of the data derived from the eigenvectors of a similarity matrix.Fig.9 shows how the points are clustered for n=4

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303115 |

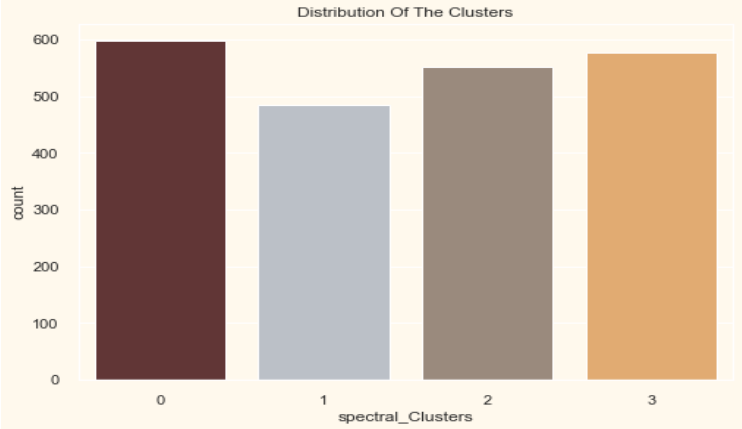


Fig.9: Distribution of elements in the spectral clustering

Ensemble Clustering using Consensus Technique:

Ensemble clustering techniques offer a powerful framework for improving the stability, robustness, generalization, and interpretability of clustering results, making them valuable tools in various machine learning and data analysis tasks. Fig.10 displays how the elements are distributed in various clusters.

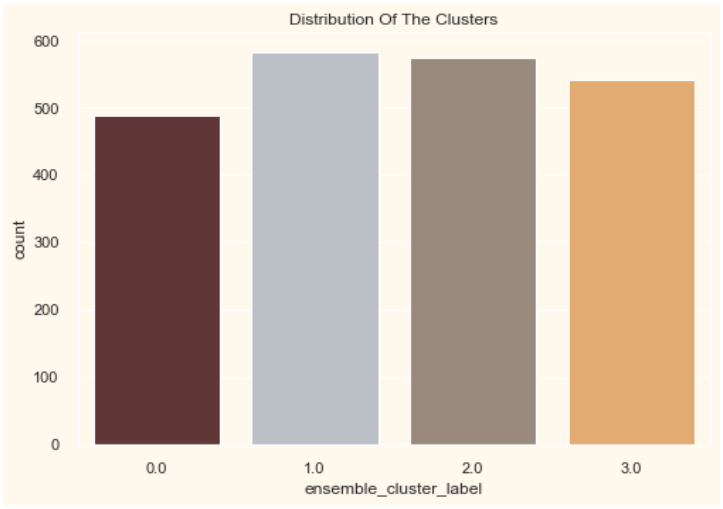


Fig.10 :The distribution of elements in various clusters

IV. EXPERIMENTAL RESULTS

The study involved the analysis of approximately 2200 entries with 29 features.The performance evaluation is done using silhouette score. Silhouette Score measures how similar an object is to its own cluster compared to other clusters. It ranges from -1 to 1, where a higher value indicates better clustering.The performance of various clustering techniques are depicted in table 1 and also graph is also used to display the results using picture shown in Fig.11

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

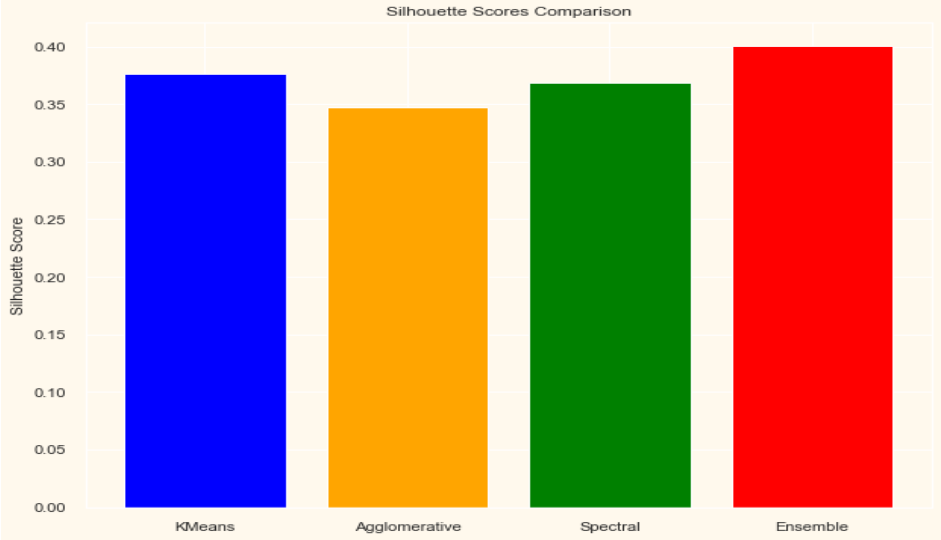
|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303115 |

Table 1. Experimental result Analysis

S.No	Algorithm	Silhouette Score
1	K-means	0.3759726431626535
2	Agglomerative	0.3472956525154227
3	Spectral	0.3687240653254949
4	Ensemble clustering	0.4008917470215087

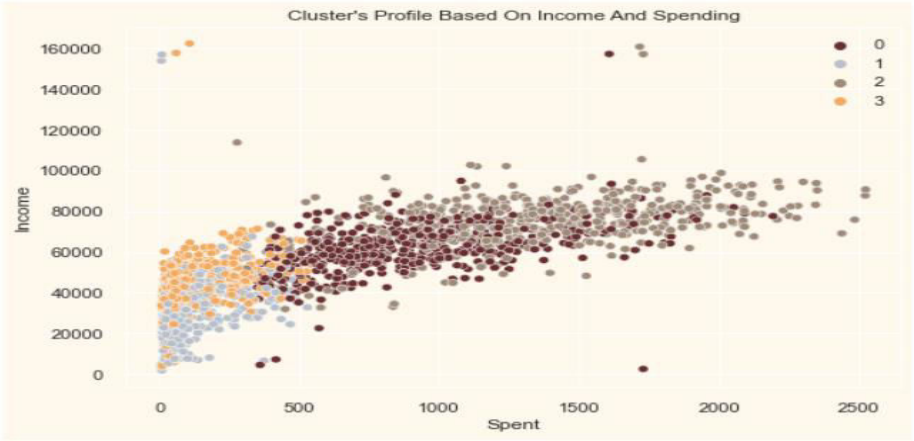
Fig.11. Graphical representation of Silhouette Scores



As evident from the data presented in Table 1 and the graphical representation in Figure 6, the ensemble learning using majority voting outperforms all other individual clusters with a silhouette score of 0.400891

After performing model evaluation we tried to understand the purchasing behaviours of our customers using scatter plot and the results are depicted in the Fig.12

Fig.12: Scatter Plot of Spent vs Income



After Understanding each cluster group with income vs spent these insights can be drawn

group 0: high spending & high income: The people with high income spends high amount for purchasing

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303115 |

group 1: low spending & low income: The people in this cluster has low income and spends low money
group 2: high spending & average income: The people in this cluster has average income but high spending
group 3: high spending & low income: The people in the cluster has low income but they spend more

V. CONCLUSION AND FUTURE WORK

Our project employs a comprehensive approach to enhance customer segmentation by leveraging advanced data processing techniques, ensuring data consistency, employing innovative clustering strategies, conducting in-depth analysis, and utilizing impactful visualization methods, leading to more precise and actionable insights for optimizing marketing strategies and enhancing customer satisfaction. It empowers businesses to tailor marketing and services, ultimately improving customer satisfaction and competitiveness in today's dynamic market

Customer segmentation future development could work on artificial intelligence for selecting the group of base algorithms, continuous Monitoring of customer behavior, automated analysis instead of manual analysis, dynamic algorithms selection, interactive visualizations, and dashboards.

.

REFERENCES

1. Zhou Yuping, Petra Jílková, Chen Guanyu, David Weisl. “New Methods of Customer Segmentation and Individual Credit Evaluation Based on Machine Learning”, Atlantis Press, New Silk Road: Business Cooperation and Prospective of Economic Development (NSRBCPED 2019).
2. Hemashree Kilari, Sailesh Edara, Guna Ratna Sai Yarra, Dileep Varma Gadhiraju, “Customer Segmentation using k-means Clustering”, International Journal of Engineering Research & Technology (IJERT).
3. Yun Chen, Guozheng Zhang, Dengfeng Hu, Chuan Fu, "Customer segmentation based on survival character", Published online: July 20017 © Springer Science+Business Media, LLC 20017.
4. Patcharin Ponyiam, Somjit Arch-int, Khon Kaen University, "Customer Behavior Analysis Using Data Mining Techniques" , 2018 International Seminar on Application for Technology of Information and Communication (iSemantic).
5. Vaidisha Mehta, Ritvik Mehra, Sourabh Singh Verma, “A Survey on Customer Segmentation Using Machine Learning Algorithms to Find Prospective Clients”, Conference Paper · September 2021.
6. Parmeshwara Joga, B. Harshini, and Rashmi Sahay, “Comparative Analysis of Machine Learning Models for Customer Segmentation”, Department of Computer Science and Engineering, Faculty of Science and Technology (IcfaiTech), The ICFAI Foundation for Higher Education, Hyderabad, India
7. E.Y.L Nandapala , K.P.N Jayasena,” The practical approach in Customers segmentation by Using the K-Means Algorithm”, 15th (IEEE) International Conference on Industrial and Information Systems (ICIIS) 2020
8. Eka Miranda, Evaristus Didik Madyatmadja, Mediana Aryuni,"Customer Segmentation in XYZ Bank using K-Means and K-Medoids Clustering", International Conference on Information Management and Technology (ICIMTech) 2020
9. Tushar Kansal, SurajBahuguna, Vishal Singh, Tanupriya Choudhury, "Customer Segmentation using K-means Clustering", International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS) 2018
10. Azad Abdulhafedh, "Incorporating K-means, Hierarchical Clustering and PCA in Customer Segmentation", Journal of City and Development, 2021, Vol. 3, No. 1, 12-30, Published by Science and Education Publishing, DOI:10.12691/jcd-3-1-3

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details