Journal of Science, Computing and Engineering Research (JSCER)
Volume-8, Issue-3, March 2025.
DOI: *https://doi.org/10.46379/jscer.2025.080310*

Customer Classification Based on The Historical Purchase Data

R P. Swarajya Lakshmi[1], B. Akash Rao[2], N. Sidhartha[3], K. Kalyan Chary[4]

[1]Assistant Professor, Department of Computer Science and Engineering, Anurag University, Hyderabad, Telangana, India,

[2,3,4] Department of Computer Science and Engineering, Anurag University, Hyderabad, Telangana, India.

Article Information

Received : 10 Feb 2025
Revised : 15 Feb 2025
Accepted : 28 Feb 2025
Published : 06 Mar 2025

Corresponding Author:
P.Swarajya Lakshmi

Abstract— Customer segmentation plays a vital role in modern business strategies, especially in the competitive e-commerce landscape. Understanding customer needs and identifying potential buyers at the right time is crucial for businesses aiming to stay ahead. By categorizing customers into distinct segments, businesses can create targeted marketing strategies that enhance customer satisfaction and increase sales. This paper explores the application of clustering techniques, with particular emphasis on the K-Means algorithm. Known for its efficiency, simplicity, and proven effectiveness, K-Means offers businesses a powerful tool for achieving accurate and actionable customer segmentation results. The paper further investigates how K-Means can be applied to large-scale datasets, ensuring scalability and adaptability to various business needs. By examining case studies and real-world applications, it highlights how this algorithm contributes to informed decision-making and the development of personalized marketing campaigns. Ultimately, this approach not only improves customer engagement but also drives business growth through optimized resource allocation and targeted communication strategies.

Keywords: ***Dynamic latch cBig Data, Business Growth, Clustering, Consumer Analysis, Customer Behavior, Customer Profiling, Customer Segmentation, Data Mining, Data Visualization, DBSCAN Clustering, Decision Making, Demographic Data, E-commerce, Elbow Method, Hierarchical Clustering, K-Means Algorithm, Machine Learning, Marketing Strategies, Market Basket Analysis, Mean Shift Clustering, Purchase History, Scalability, Spending Score, Targeted Marketing, Unsupervised Learning***

Copyright © 2025: P. Swarajya Lakshmi, B. Akash Rao, N. Sidhartha, K. Kalyan Chary, This is an open access distribution, and reproduction in any medium, provided Access article distributed under the Creative Commons Attribution License the original work is properly cited License, which permits unrestricted use.

Citation: P. Swarajya Lakshmi, B. Akash Rao, N. Sidhartha, K. Kalyan Chary, "Customer Classification Based on The Historical Purchase Data", Journal of Science, Computing and Engineering Research, 8(01), March 2025.

I. INTRODUCTION

As more and more business being coming up every day, it has become significantly important for the old businesses to apply marketing strategies to stay in the market as the competition has been cut to throat. Change or die have become the simple rule of marketing in today's world. As the customer base is increasing day by day it has become challenging for the companies to cater to the needs of each and every customer, this is where Data mining serves a very important role to unravel hidden patterns stored in the company's database. Customer segmentation is one of the applications of data

mining which helps to segment the customers with similar patterns into similar clusters hence, making easier for the business to handle the large customer base. This segmentation can directly or indirectly influence the marketing strategy as it opens many new paths to discover like for which segment the product will be good, customizing the marketing plans according to the each segment, providing discounts for a specific segment, and decipher the customer and object relationship which has been previously unknown to the company. Customer segmentation allows companies to visualize what actually 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. Clustering has been proven effective to implement customer segmentation. Clustering comes under unsupervised learning, having ability to find clusters over unlabelled dataset. There are a number of clustering algorithm over which like k-means, hierarchical clustering, DBSCAN clustering etc. In this paper, three different clustering algorithms have been implemented over a dataset with two features with 200 records

1.1 Overview:

Customer segmentation plays a key role in making business decisions. In the competitive field of e-commerce, it is very important to satisfy the customer needs and to identify the potential customer and these things should be

done in the right time in the right manner. In this paper, various segments of customer segmentation are discussed and different techniques in customer segmentation are presented. Among them, clustering is best and by comparing the techniques of clustering we analyze that K-Means algorithm is the most efficient and it is very simple to use.

1.2 Problem Definition:

We live in a world where large and vast amount of data is collected daily. Analyzing such data is an important need. In the modern era of innovation, where there is a large competition to be better then everyone, the business strategy needs to be according to the modern conditions. The business done today runs on the basis of innovative ideas as there are large number of potential customers who are confounded to what to buy and what not to buy. The companies doing the business are also not able to diagnose the target potential customers. This is where the machine learning comes into picture, the various algorithms are applied to identify the hidden patterns in the data for better decision making. the concept of which customer segment to target is done using the customer segmentation process using the clustering technique. In this paper, the clustering algorithm used is K-means algorithm which is the partitioning algorithm, to segment the customers according to the similar characteristics. To determine the optimal clusters, elbow method is used.

1.3 Objective:

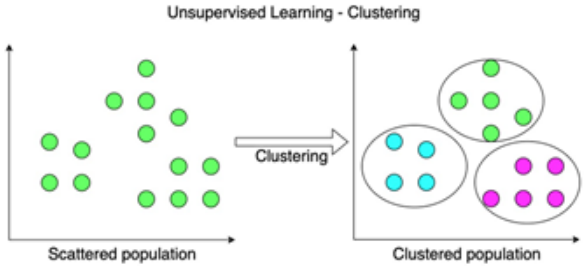
Today's business run on the basis of such innovation having ability to enthrall the customers with the products, but with such a large raft of products leave the customers confounded, what to buy and what to not and also the companies are nonplussed about what section of customers to target to sell their products. This is where machine learning comes into play, various algorithms are applied for unraveling the hidden patterns in the data for better decision making for the future. This elude concept of which segment to target is made unequivocal by applying segmentation. The process of segmenting the customers with similar behaviors into the same segment and with different patterns into different segments is called customer segmentation.

1.4 Methodology:

Clustering algorithms try to find natural clusters in data, the various aspects of how the algorithms to cluster data can be tuned and modified. Clustering is based on the principle that items within the same cluster must be similar to each other. The data is grouped in such a way that related elements are close to each other.

Diverse and different types of data are subdivided into smaller groups

In today's competitive world, it is crucial to understand customer behavior and categorize customers based on their demography and buying behavior. This is a critical aspect of customer segmentation that allows marketers to better tailor their marketing efforts to various audience subsets in terms of promotional, marketing and product development strategies.



We demonstrate the concept of segmentation of a customer data set from an e-commerce site using k-means clustering in python. The data set contains the annual income of ~300 customers and their annual spend on an e-commerce site. We will use the k-means clustering algorithm to derive the optimum number of clusters and understand the underlying customer segments based on the data provided.

1.5 Hardware and Software Tools Used:

HARDWARE REQUIREMENTS:

- Processor : Intel i3 and above
- RAM : 4GB and Higher
- Hard Disk : 500GB: Minimum

SOFTWARE REQUIREMENTS:

- Programming Language: Python
- IDE Jupyter

II. LITERATURE SURVEY

2.1 Related Work:

Agglomerative Clustering-: Agglomerative Clustering is based on forming a hierarchy represented by dendrograms (discussed in later section). Dendrogram acts as memory for the algorithm to tell about how the clusters are being formed. The clustering starts with forming N clusters for N data points and then merging along the closest data points together in each step such that the current step contains one cluster less than the previous one.

Mean shift Clustering-: This clustering algorithm is a non-parametric iterative algorithm functions by assuming the all the data points in the feature space as empirical probability density function. The algorithm clusters each data point by allowing data point converge to a region of local maxima which is achieved by fixing a window around

each data point finding the mean and then shifting the window to the mean and repeat the steps until all the data point converges forming the clusters.

2.2 Existing System:

Customer segmentation is currently performed by processing customer database, i.e. demographic data or purchase history. Several researchers discuss the customer segmentation method on their papers, such as Magento, who used several variables to perform customer segmentation, namely transaction variable, product variable, geographic variable, hobbies variable and page viewed variable

Baer 5 and Colica discuss customer segmentation methods of Business Rule, Quantile membership, Supervised Clustering, Unsupervised Clustering, Customer Profiling, RFM Cell Classification Grouping, Customer Likeness Clustering and Purchase Affinity Clustering. Some of these methods have similarity. Other researchers discuss the implementation of customer segmentation.

2.3 Limitations of Existing System:

- Segmentation also has its limitations as it needs to be implemented in the proper manner. As segmentation is one of the most important processes in the marketing plan or for your business, you need to know the limitations of segmentation and what pitfalls lie ahead if you go wrong with your target market segment.
- Segments are too small – If the chosen segment is too small then you will not have the proper turnover which in turn will affect the total margins and the viability of the business.
- While consumer behavior can be tracked, it is not always easy to pinpoint the motivations behind those behaviors, as they can vary greatly from person to person.
- Behavioral segmentation is often based on complex data constructs that are not always easy to understand.
- Translating attitudinal characteristics into a conventional database model can be challenging, and can sometimes lead to loss of effectiveness in terms of replicability.
- It can be difficult to obtain data for consumers in a given population, because participation in an attitudinal survey is required.
- Consumers are misinterpreted – The right product to the wrong customers. What if your market research says that your customers want a new soap and you come out with a new facial cream. The concept is same, cleanliness. But the concept is completely different.
- Costing is not taken into consideration – Targeting a segment is ok but you also need to know how much you will have to spend to target a particular segment. If it is a Sec A segment and you do not have the budget to be present in the places the the Sec A customer visits, then your segmentation strategy is a failure.
- There are too many brands – Along with segmentation, you also need to check out the competition offered in the same segment from other products. Getting into a segment already saturated will mean higher costs and lesser profit margins.
- Consumer are confused – If the consumer himself doesn't know whether he will be interested in a particular product or not, then that's a sign that you need to get out of that segment / product.
- Product is completely new – If a product is completely new than there is no market research to base your segmentation on. You need to market it to the masses and as acceptance increases, only then will you be able to focus on one particular segment.

2.4 Proposed System:

K-Means clustering is an unsupervised machine learning algorithm that divides the given data into the given number of clusters. Here, the "K" is the given number of predefined clusters that need to be created. It is a centroid based algorithm in which each cluster is associated with a centroid. The main idea is to reduce the distance between the data points and their respective cluster centroid. The algorithm takes raw unlabelled data as an input and divides the dataset into clusters and the process is repeated until the best clusters are found.

K-Means is very easy and simple to implement. It is highly scalable, can be applied to both small and large datasets. There is, however, a problem with choosing the number of clusters or K. Also, with the increase in dimensions, stability decreases. But, overall K Means is a simple and robust algorithm that makes clustering very easy.

III. METHODOLOGY

3.1 **Dataset**:

This data set is created only for the learning purpose of the customer segmentation concepts, also known as market basket analysis. I will demonstrate this by using unsupervised ML technique (KMeans Clustering Algorithm) in the simplest form.

You are owing a supermarket mall and through membership cards , you have some basic data about your

customers like Customer ID, age, gender, annual income and spending score.

Spending Score is something you assign to the customer based on your defined parameters like customer behavior and purchasing data.

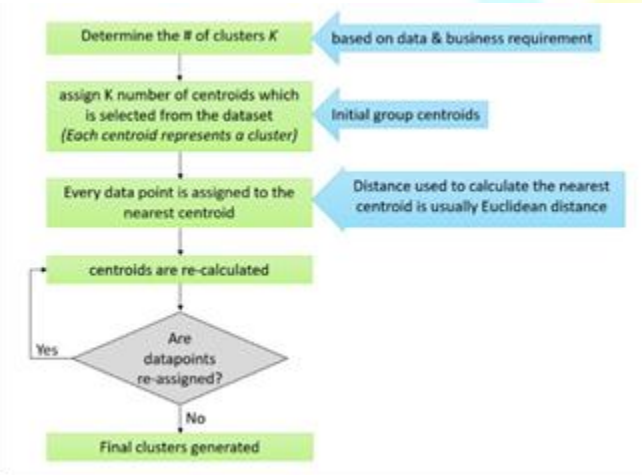
You own the mall and want to understand the customers like who can be easily converge [Target Customers] so that the sense can be given to marketing team and plan the strategy accordingly.

The data includes the following features:

- Customer ID
- Customer Gender
- Customer Age
- Annual Income of the customer (in Thousand Dollars)
- Spending score of the customer (based on customer behaviour and spending nature)

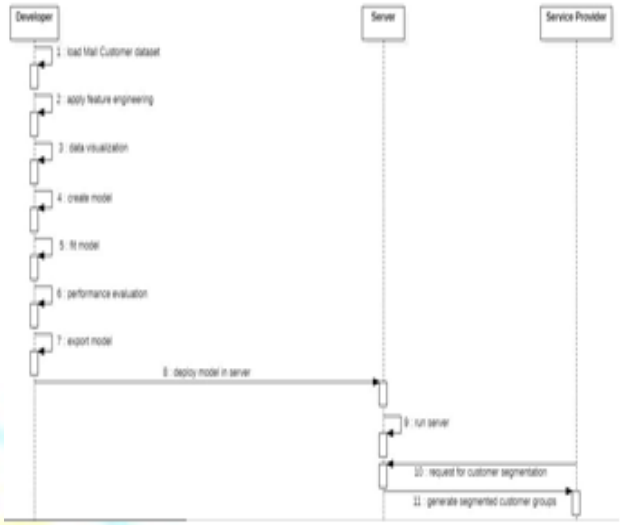
3.2 Architecture:

K-means clustering is an iterative clustering algorithm where the number of clusters K is predetermined and the algorithm iteratively assigns each data point to one of the K clusters based on the feature similarity.



3.3 Sequence diagram:

Sequence Diagrams Represent the objects participating the interaction horizontally and time vertically. A Use Case is a kind of behavioral classifier that represents a declaration of an offered behavior. Each use case specifies some behavior, possibly including variants that the subject can perform in collaboration with one or more actors. Use cases define the offered behavior of the subject without reference to its internal structure. These behaviors, involving interactions between the actor and the subject, may result in changes to the state of the subject and communications with its environment.



IV. TOOL DESCRIPTION

4.1 Hardware Requirements:

Artificial intelligence technologies such as machine learning and deep learning involve use of large amounts of data and complex algorithms that require powerful computation hardware. This makes selecting the best machine for such tasks challenging because you have to consider many factors such as portability, processing speed, and the graphics processing capability among others. This article will help you through the grueling decision-making process.

Important components to consider when selecting a laptop for machine learning

GPU: One of the most important factors to consider when choosing a deep learning machine is the general processing unit (GPU). GPUs are microprocessing chips primarily designed for handling graphics. GPUs have become popular in deep learning field mainly due to their ability to handle simultaneous computations faster than CPUs. Essentially, GPUs have a large number of cores and high memory bandwidth and are thus suited for multiple parallel processing of large amounts of data. This has been boosted by efforts to develop AI-based GPU frameworks such as CuDNN and parallel computation APIs like CUDA by NVIDIA. Such frameworks and APIs allow scientists to leverage GPU parallelism for deep learning tasks.

Here is what to look for in a GPU:

- Opt for a higher memory bandwidth (speed of video RAM) within your budget

- If you will be dealing with large amounts of data, go for a higher number of cores as it dictates the speed of processing data
- Consider the processing power of the GPU if computation time is a factor
- Video RAM size should also be considered for faster processing
- An NVIDIA GPU is preferable because of the available frameworks and APIs (CUDA and CuDNN) compatible with major deep learning frameworks such as TensorFlow and PyTorch. The latest generations of NVIDIA GPUs such as the GeForce RTX based on Turing architecture are AI-enabled with Tensor cores which makes them suitable for deep learning.

RAM: RAM is another important factor to consider when purchasing a deep learning laptop. The larger the RAM the higher the amount of data it can handle, leading to faster processing. With more RAM you can use your machine to perform other tasks as the model trains. Although a minimum of 8GB RAM can do the job, 16GB RAM and above is recommended for most deep learning tasks.

CPU: When it comes to CPU, a minimum of 7th generation (Intel Core i7 processor) is recommended. However, getting Intel Core i5 with Turbo Boosts can do the trick. If one opts for a desktop then selecting the right combination of CPU and motherboard that match your GPU specifications is recommended. In that case, the choice of the number of PCIe lanes (PCIe lanes determine the speed of transferring data from CPU RAM to GPU RAM) should also be taken into consideration (4-16 PCIe lanes is best for most deep learning tasks).

Storage: Storage is also an important factor, specifically due to the increasing size of deep learning datasets requiring higher storage capacity. For example, Imagenet, one of the most popular datasets for deep learning, is 150 GB in size and consists of more than 14 million images across 20,000 categories. Although SSD is recommended for its speed and efficiency, you can get an HDD at a relatively cheaper price to do the job. However, if you value speed, price and efficiency then a hybrid of the two is the best option.

4.2 Software Requirements:

Python:

The Python programming language is an Open Source, cross-platform, high level, dynamic, interpreted language.

The Python 'philosophy' emphasizes readability, clarity and simplicity, whilst maximizing the power and expressiveness available to the programmer. The ultimate compliment to a Python programmer is not that his code is clever, but that it is elegant. For these reasons Python is an excellent 'first language', while still being a powerful tool in the hands of the seasoned and cynical programmer.

Python is a very flexible language. It is widely used for many different purposes. Typical uses include:

- Web application programming with frameworks like Zope, Django and Turbogears
- System administration tasks via simple scripts
- Desktop applications using GUI toolkits like Tkinter or wxPython (and recently Windows Forms and IronPython)
- Creating windows applications, using the Pywin32 extension for full windows integration and possibly Py2exe to create standalone programs
- Scientific research using packages like Scipy and Matplotlib
- Good to know
- The most recent major version of Python is Python 3, which we shall be using in this tutorial. However, Python 2, although not being updated with anything other than security updates, is still quite popular.
- In this tutorial Python will be written in a text editor. It is possible to write Python in an Integrated Development Environment, such as Thonny, Pycharm, Netbeans or Eclipse which are particularly useful when managing larger collections of Python files.
- Python Syntax compared to other programming languages
- Python was designed for readability, and has some similarities to the English language with influence from mathematics.
- Python uses new lines to complete a command, as opposed to other programming languages which often use semicolons or parentheses.
- Python relies on indentation, using whitespace, to define scope; such as the scope of loops, functions and classes. Other programming languages often use curly-brackets for this purpose.

Jupyter notebook:

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

Jupyter Notebooks are a powerful way to write and iterate on your Python code for data analysis. Rather than writing and re-writing an entire program, you can write lines of code and run them one at a time. Then, if you need to make a change, you can go back and make your edit and rerun the program again, all in the same window.

Jupyter Notebook is built off of IPython, an interactive way of running Python code in the terminal using the REPL model (Read-Eval-Print-Loop). The IPython Kernel runs the computations and communicates with the Jupyter Notebook front-end interface. It also allows Jupyter Notebook to support multiple languages. Jupyter Notebooks extend IPython through additional features, like storing your code and output and allowing you to keep markdown notes.

If you'd rather watch a video instead of read an article, please watch the following instructions on how to use a Jupyter Notebook. They cover the same information.

LAUNCH A NOTEBOOK

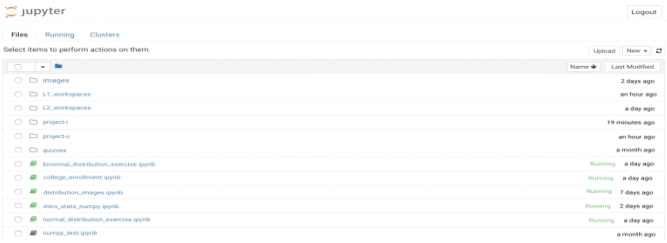
To launch a Jupyter notebook, open your terminal and navigate to the directory where you would like to save your notebook. Then type the command jupyter notebook and the program will instantiate a local server at localhost:8888 (or another specified port).

```
[9] -> jupyter notebook
[I 18:31:51.264 NotebookApp] Serving notebooks from, local directory: /Users/janedoe
[I 18:31:51.264 NotebookApp] 0 active kernels
[I 18:31:51.264 NotebookApp] The Jupyter Notebook is running at:
[I 18:31:51.264 NotebookApp] http://localhost:8888/?token=f9b639294933fa68c64c78effb9b65
19f9d8c45c903baa43
```

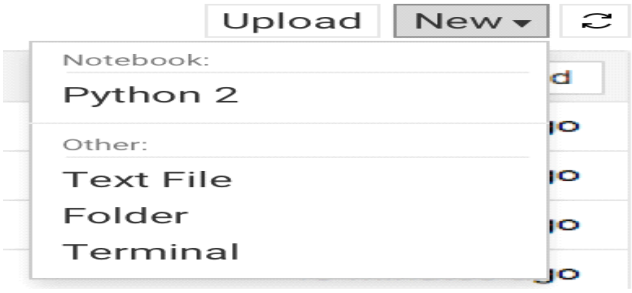
A browser window should immediately pop up with the Jupyter Notebook interface; otherwise, you can use the address it gives you. The notebooks have a unique token since the software uses pre-built Docker containers to put notebooks on their own unique path. To stop the server and shutdown the kernel from the terminal, hit control-C twice.

JUPYTER INTERFACE

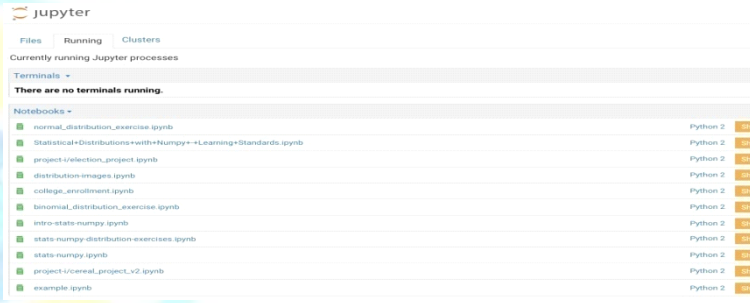
Now you're in the Jupyter Notebook interface, and you can see all of the files in your current directory. All Jupyter Notebooks are identifiable by the notebook icon next to their name. If you already have a Jupyter Notebook in your current directory that you want to view, find it in your files list and click it to open.



To create a new notebook, go to New and select Notebook - Python 2. If you have other Jupyter Notebooks on your system that you want to use, you can click Upload and navigate to that particular file.

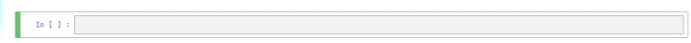


Notebooks currently running will have a green icon, while non-running ones will be grey. To find all currently running notebooks, click on the Running tab to see a list.



INSIDE THE NOTEBOOK

When you open a new Jupyter notebook, you'll notice that it contains a cell.



Cells are how notebooks are structured and are the areas where you write your code. To run a piece of code, click on the cell to select it, then press SHIFT+ENTER or press the play button in the toolbar above. Additionally, the Cell dropdown menu has several options to run cells, including running one cell at a time or to run all cells at once.

V. IMPLEMENTATION

Data Collection: The dataset has been taken from a local retail shop consisting of two features, average number of visits to the shop and average amount of shopping done on yearly basis.

This project is a part of the Mall Customer Segmentation Data competition held on Kaggle.

Feature Scaling: The data has been scaled using Standard Scaler, by applying standard scaler the data gets centred around 0 with standard deviation of 1.

K means Clustering: Choosing the optimal number of clusters: Elbow method is applied to calculate value of K for the dataset.

- **Step-1:** Run the algorithm for various values of k i.e making the k vary from 1 to 10.
- **Step-2:** Calculate the within cluster squared error.
- **Step-3:** Plot the calculated error, where a bent elbow like structure will form, will give the optimal value of clusters.

Algorithm:

- **Step-1:** Initialize the K (= 5) clusters.
- **Step-2:** Assign the data point that is closest to any particular cluster.
- **Step-3:** Recalculate the centroid position based on the mean of the cluster formed
- **Step-4:** Repeat step 2 and 3 until the centroid position remains unchanged in the previous and current iteration.

Libraries:

Numpy: NumPy is a Python package which stands for 'Numerical Python'. It is the core library for scientific computing, which contains a powerful n-dimensional array object, provide tools for integrating C, C++ etc. It is also useful in linear algebra, random number capability etc. NumPy array can also be used as an efficient multi-dimensional container for generic data. Now, let me tell you what exactly a python numpy array is.

To install Python NumPy, go to your command prompt and type "pip install numpy". Once the installation is completed, go to your IDE (For example: PyCharm) and simply import it by typing: "import numpy as np".

Pandas: Pandas are an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures. The name Pandas is derived from the word Panel Data – an Econometrics from Multidimensional data.

In 2008, developer Wes McKinney started developing pandas when in need of high performance, flexible tool for analysis of data.

Prior to Pandas, Python was majorly used for data munging and preparation. It had very little contribution towards data analysis. Pandas solved this problem. Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data — load, prepare, manipulate, model, and analyze.

Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc.

Standard Python distribution doesn't come bundled with Pandas module. A lightweight alternative is to install NumPy using popular Python package installer, pip. pip install pandas

Matplotlib: Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib produces publication-quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shell, web application servers, and various graphical user interface toolkits.

Seaborn: Seaborn is a library for making statistical graphics in Python. It is built on top of matplotlib and closely integrated with pandas data structures.

Here is some of the functionality that seaborn offers:

- A dataset-oriented API for examining relationships between multiple variables
- Specialized support for using categorical variables to show observations or aggregate statistics
- Options for visualizing univariate or bivariate distributions and for comparing them between subsets of data
- Automatic estimation and plotting of linear regression models for different kinds dependent variables
- Convenient views onto the overall structure of complex datasets
- High-level abstractions for structuring multi-plot grids that let you easily build complex visualizations
- Concise control over matplotlib figure styling with several built-in themes
- Tools for choosing color palettes that faithfully reveal patterns in your data

Seaborn aims to make visualization a central part of exploring and understanding data. Its dataset-oriented plotting functions operate on dataframes and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots.

Sklearn: Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

VI. RESULTS AND ANALYSIS

6.1 Result Discussion:

We have taken two internal clustering measure, silhouette score and Calinski-Harabasz index. Silhouette Score: It is a way of measuring how well the data point has been clustered into the correct cluster.

First Step-: a = Average distance between the centroid of a cluster and the data points embroiled into it.

Second Step: b = Average distance between the data point and the closest cluster data points.

Third Step: Silhouette Score For the data point to be well grounded in its cluster 'b' needs to be large and 'a' needs to be small so that difference between the two is as large as possible. 'max (b,a)' is added to normalized the silhouette score. Higher the score better the data point belongs to that cluster.

6.2 Comparison with Previous Studies:

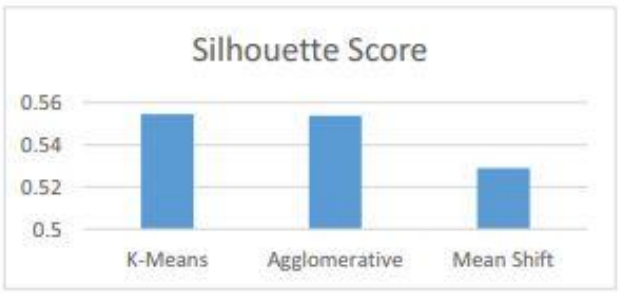


Figure above displays the silhouette score for the three algorithms applied in this paper, the graph shows there is not much significant difference in K-means and Agglomerative clustering. Hence, these two algorithms were able to cluster our data well than Mean shift algorithm as displayed by the low value of silhouette score.

6.3 Analysis:

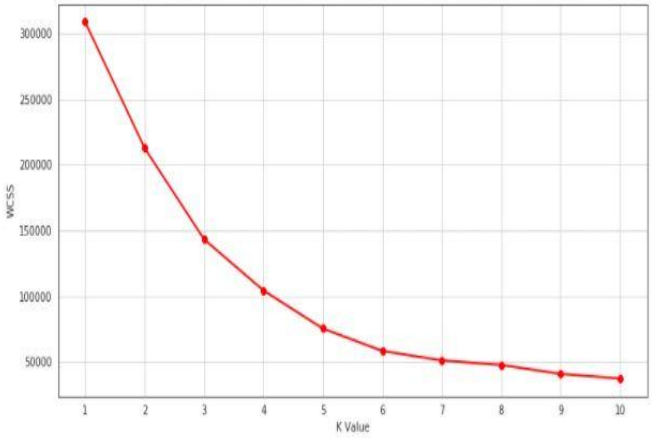
K means clustering is one of the most popular clustering algorithms and usually the first thing practitioners apply when solving clustering tasks to get an idea of the structure of the dataset. The goal of K means is to group data points into distinct non-overlapping subgroups. One of the major application of K means clustering is segmentation of customers to get a better understanding of them which in turn could be used to increase the revenue of the company.

- **The Elbow Method**

Calculate the Within Cluster Sum of Squared Errors (WSS) for different values of k, and choose the k for which WSS first starts to diminish. In the plot of WSS-versus k, this is visible as an elbow.

The steps can be summarized in the below steps:

- Compute K-Means clustering for different values of K by varying K from 1 to 10 clusters.
- For each K, calculate the total within-cluster sum of square (WCSS).
- Plot the curve of WCSS vs the number of clusters K.
- The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.



VII. CONCLUSIONS AND FUTURE SCOPE

Conclusion:

As our dataset was unlabelled, in this paper we have opted for internal clustering validation rather than external clustering validation, which depend on some external data like labels. Internal cluster validation can be used for choosing clustering algorithm which best suits the dataset and can correctly cluster data into its opposite cluster.

From Results it can be observed that Cluster 1 denotes the customer who has high annual income as well as high yearly spend. Cluster 2 represents the cluster having high annual income and low annual spend. Cluster 3 represents customer with low annual income and low annual spend. Cluster 5 denotes the low annual income but high yearly spend.

Cluster 4 and cluster 6 denotes the customer with medium income and medium spending score.

We used K-Means clustering to understand customer data. K-Means is a good clustering algorithm. Almost all the clusters have similar density. It is also fast and efficient in terms of computational cost.

Future Scope:

Further research can be added with several criteria so that the system can be better and right on target. Development of supporting applications that use other tools and methods can be used as comparisons to the system that have been developed.

The proposed system can be developed in many different directions which have vast scope for improvements in the system. These includes: 1. Increase the accuracy of the algorithms. 2. Improvising the algorithms to add more efficiency of the system and enhance its working. 3. Working on some more attributes so to tackle diabetes even more.

REFERENCES

[1]. Jayant Tikmani, Sudhanshu Tiwari, Sujata Khedkar "Telecom customer segmentation based on cluster analysisAn Approach to Customer Classification using k-means", IJIRCCE,Year: 2015.

[2]. Chinedu Pascal Ezenkwu, Simeon Ozuomba, Constance kalu Electrical/Electronics & Computer Engineering Department, University of Uyo, Uyo, Akwa Ibom State, Nigeria "Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services", IJARAI,Year: 2015.

[3]. T.NelsonGnanarajDr.K.Ramesh Kumar N.Monica"Survey on mining clusters using new k-mean algorithm from structured and unstructured data", IJACST,Year: 2014.

[4]. Yogita Rani and Dr. Harish Rohil"A Study of Hierarchical Clustering Algorithm", IJICT,Year: 2013.

[5]. Omar Kettani, FaycalRamdani, BenaissaTadili"An Agglomerative Clustering Method for Large Data Sets", IJCA,Year: 2014.

[6]. Snekha, ChetnaSachdeva, Rajesh Birok"Real Time Object Tracking Using Different Mean Shift Techniques–a Review", IJSCE,Year: 2013.

[7]. SulekhaGoyat"The basis of market segmentation: a critical review of literature", EJBM,Year: 2011.

[8]. Vaishali R. Patel and Rupa G. Mehta "Impact of Outlier Removal and Normalization Approach in Modified k-Means Clustering Algorithm", IJCSI,Year: 2011.

[9]. Scikit-learn: https://scikit-learn.org

[10]. Tanupriya Choudhury, Vivek Kumar, Darshika Nigam, Intelligent Classification & Clustering Of Lung & Oral Cancer through Decision Tree & Genetic Algorithm, International Journal of Advanced Research in Computer Science and Software Engineering,2015