****

**CUSTOMER SEGMENTATION USING MACHINE LEARNING IN R**

*Mini Project- II report submitted in partial fulfilment of requirements for the award of the degree of*

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE AND ENGINEERING**

Submitted by

N. Srikar                         Roll. No: 19131A05G1

P. Divyagandh                   Roll. No: 19131A05H7

P. Hema Harsha Vardhan   Roll. No:  19131A05H4

Nitin Ancha Roll. No:  19131A05G0



Under the esteemed guidance of

**Dr. H Prthsarathi Patra**

**(Associate Professor)**

**Department of Computer Science and Engineering**

**GAYATRI VIDYA PARISHAD COLLEGE OF ENGINEERING (AUTONOMOUS)**

(Affiliated to JNTU-K, Kakinada,A.P)

**VISAKHAPATNAM**

**2022 – 2023**

**CERTIFICATE**

This is to certify that the Mini project work titled is **Customer Segmentation Using Machine Learning in R** a Bonafide record of the work done by **N.Srikar(19131A05G1), P. Divyagandh(19131A05H7), P. Hema Harsha Vardhan (19131A05H4), Nitin Ancha(19131A05G0)** to the Department of Computer Science and Engineering, **Gayatri Vidya Parishad College of Engineering (Autonomous), Visakhapatnam,** affiliated to **Jawaharlal Nehru Technological University, Kakinada** inpartial fulfilment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering during** the academic year 2022-2023

**Under the guidance of: Head of the Department:**

**Dr. H Parthsarathi Patra Dr. D.N.D. Harini**

**Associate Professor Associate Professor and**

**Dept. of CSE, Head of the Department**

**GVPCOE(A)** **Dept. of CSE,**

**GVPCOE(A)**

**ACKNOWLEDGEMENTS**

We express our deep sense of gratitude to our esteemed institute **“Gayatri Vidya Parishad College of Engineering (Autonomous)”** which has provided us an opportunity to fulfil the most cherished desire to reach our goal.

We express our profound gratitude and our deep indebtedness to our guide **Dr. H Parthsarathi Patra**, Associate Professor, Department of Computer Science and Engineering, whose valuable suggestions, guidance, and comprehensive assistance helped us a lot.

We express our deep sense of Gratitude to **Dr. D.N.D. HARINI**, Associate Professor and Head of the Department of Computer science and Engineering, Gayatri Vidya Parishad College of Engineering (Autonomous) for giving us an opportunity to do the project in college.

With a great sense of pleasure and privilege, I extend my gratitude and sincere thanks to, **DR. A. BALA KOTESWARA RAO,** Principal, Gayatri Vidya Parishad College of Engineering (Autonomous) for his continuous encouragement during the study.

We also thank our coordinator, **Dr. CH. SITA KUMARI,** Associate Professor, Department of Computer Science and Engineering, for the kind suggestions and guidance for thesuccessful completion of our project work.

Finally, we would also like to thank all the members of the teaching and non-teaching staff of theComputer Science and Engineering Department for all their support in completion of our project.

**Declaration of the student**

We, the undersigned solemnly declare that this report is based on the Mini Project carried out by Us during the study (2021-22) for the partial fulfilment for the award of degree Bachelor of Technology in Engineering.

I further declare that the proofs submitted are genuine to the best of my knowledge

**Signatures of the Students:**

|  |  |
| --- | --- |
| **Name of the Student** | **Signature** |
| N. Srikar |  |
| P. Divyagandh |  |
| P. Hema Harsha Vardhan |  |
| Nitin Ancha |  |

**ABSTRACT**

**CUSTOMER SEGEMENTATION USING MACHINE LEARNING** is a data science model to segregate and find the potential customers for an organization or company. The main aim of our project is to help organizations to find relations and associations of customers’ behaviour. Identification of customers is very important for suggesting business ideals and increasing productivity accordingly. The Output of this Project suggests algorithms and analysing data relativity for the same, increasing the overall performance of the business and profit of the organization.

This project presents a Deep learning framework using unsupervised machine learning algorithms like K-Means algorithm. K-means algorithm is the best algorithm for finding data correlation based on the Euclidian distance between the data points. The steps involved are: first the data of customers is undergone through various pre-processing techniques to reduce the redundancy of the data, then the data is plotted and analysed through various visualization techniques and a general data constraints are defined, The various algorithms are used on the data to form clusters and group the associations of the data. The goal is to find the maximum number of clusters that can optimize the result. Unsupervised Learning Algorithms have many applications such as genetic patterns, data science, media and traffic activity and analysis etc. The project helps to suggest overall profitable ideas which are hidden in the data and are potential to organization and use it to help increase the productivity of the company.

**INDEX**

**CHAPTER 1. INTRODUCTION**

1.1 Objective 8

1.2 About the Algorithm 9

1.3 Purpose 10

1.4 Scope 10

**CHAPTER 2. SRS DOCUMENT**

2.1 Functional Requirements 11

2.2 Non-functional Requirements 11

2.3 Minimum Hardware Requirements 12

2.4 Minimum Software Requirements 12

**CHAPTER 3.** **ALGORITHM ANALYSIS**

3.1 Existing Algorithm 13

3.2 Proposed Algorithm 13

3.3 Feasibility Study 15

3.4 Cost Benefit Analysis 18

**CHAPTER 4. SOFTWARE DESCRIPTION**

4.1 Google Collab 19

4.2 GridExtra 19

4.3 Plotrix 19

4.4 NBClust 20

4.5 FactoExtra 20

4.6 Matplotlib 20

**CHAPTER 5. PROJECT DESCRIPTION**

5.1 Problem Definition 21

5.2 Project Overview 21

5.3 Module Description 22

**CHAPTER 6. SYSTEM DESIGN**

6.1 Introduction to UML 24

6.2 Building Blocks of the ML 25

6.3 UML Diagrams 31

**CHAPTER 7. DEVELOPMENT**

7.1 Sample Code 33

7.2 Results 44

**CHAPTER 8. TESTING**

8.1 Introduction Of Testing 45

.

8.2 Test Cases 49

**CHAPTER 9. CONCLUSION 54**

**CHAPTER 10. FUTURE SCOPE 55**

**CHAPTER 11. REFERENCE LINKS 56**

**1. INTRODUCTION**

**Customers** are an essential part for any business, trade or organization to survive, and the rapid growth of customers also encourages for the computer science engineering to indulge in this field to support the organization to handle, analyse and increase their, profits and productivity. Customer Segmentation deals with handling of large amounts of data on daily basis. Customer Segmentation helps the data to be distributed into various associations that can be used to analyse their behaviours, interests and also predict changes in their flavours, which can help the market to interact with the customers more effectively.

Generally, it is estimated that 200 billion US Dollars of revenue is generated through trade and business, in India. This large amount of transactions attracts more attention towards its maintenance and handling their data.

For this reason, it is extremely important to analyse the customer behaviour. In this work we proposed a solution for clustering data of customers by K-means: Unsupervised Machine Learning algorithm. K-means is most efficient for data segmentations because it effectively forms clusters. And it helps in suggesting the potential ideas. That attract more customers and increase the overall productivity of the Organization.

**1.1 OBJECTIVE**

**“Customer Segmentation using Machine Learning in R”** helps us in identifying the data patterns for various customers used as an approach to business, which will generally result in greater engagement, higher-click through rates, and ultimately better sales. Using

**Unsupervised Learning Algorithms**. K - means algorithm has been used to find the data correlation based on Age, Gender, Annual Income, Spending Score. Implementing customer segmentation leads to plenty of new business opportunities. You can do a lot of optimizations in budgeting, product design, promotion, marketing, customer satisfaction.

**1.2 ABOUT THE ALGORITHM**

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. While using the k-means clustering algorithm, the first step is to indicate the number of clusters (k) that we wish to produce in the final output. The algorithm starts by selecting k objects from dataset randomly that will serve as the initial centres for our clusters. These selected objects are the cluster means, also known as centroids. Then, the remaining objects have an assignment of the closest centroid. This centroid is defined by the Euclidean Distance present between the object and the cluster mean. We refer to this step as “cluster assignment”. When the assignment is complete, the algorithm proceeds to calculate new mean value of each cluster present in the data. After the recalculation of the centres, the observations are checked if they are closer to a different cluster. Using the updated cluster mean, the objects undergo reassignment. This goes on repeatedly through several iterations until the cluster assignments stop altering. The clusters that are present in the current iteration are the same as the ones obtained in the previous iteration. Algorithm steps for K - Means:

1. Average Silhoutte method
2. Gap statistic
3. Elbow method

Diagram

Description automatically generated

FIG 1.1 INTERNAL BLOCK OF K means

**1.3 PURPOSE**

Customer segmentation has the potential to allow marketers to address each customer in the most effective way. Using the large amount of data available on customers (and potential customers), a customer segmentation analysis allows marketers to identify discrete groups of customers with a high degree of accuracy based on demographic, behavioural and other indicators.

Since the marketer’s goal is usually to maximize the value (revenue and/or profit) from each customer, it is critical to know in advance how any particular marketing action will influence the customer. Ideally, such “action-centric” customer segmentation will not focus on the short-term value of a marketing action, but rather the long-term [customer lifetime value (CLV)](https://www.optimove.com/learning-center/customer-lifetime-value/) impact that such a marketing action will have. Thus, it is necessary to group, or segment, customers according to their CLV.

**1.4 SCOPE**

By enabling companies to target specific groups of customers, a customer segmentation model allows for the effective allocation of marketing resources and the maximization of cross- and up-selling opportunities. When a group of customers is sent personalized messages as part of a [marketing mix](https://www.techtarget.com/whatis/definition/Four-Ps) that is designed around their needs, it's easier for companies to send those customers special offers meant to encourage them to buy more products. Customer segmentation can also improve customer service and assist in customer loyalty and retention. As a by-product of its personalized nature, marketing materials sent out using customer segmentation tend to be more valued and appreciated by the customer who receives them as opposed to impersonal brand messaging that doesn't acknowledge purchase history or any kind of customer relationship.

**2. SRS DOCUMENT**

A software requirements specification (**SRS**) is a document that describes what the software will do and how it will be expected to perform.

**2.1 FUNCTIONAL REQUIREMENTS**

A Functional Requirement (FR) is a description of the service that the software must offer. It describes a software system or its component. A function is nothing but inputs to the software system, its behaviour, and outputs. It can be a calculation, data manipulation, business process, user interaction, or any other specific functionality which defines what function a system is likely to perform. Functional Requirements are also called Functional Specification.

* Predict data patterns of customers by observing the customer behaviours with the algorithm.
* Suggesting products to enhance the experience of the customer, after analysing customer behaviour from the given dataset.

**2.2 NON-FUNCTIONAL REQUIREMENTS**

NON-FUNCTIONAL REQUIREMENT (NFR) specifies the quality attribute of a software system. They judge the software system based on Responsiveness, Usability, Security, Portability. Non-functional requirements are called qualities of a system, there are as follows:

* + - **Performance**-The average response time of the system is less.
    - **Reliability -** The system is highly reliable.
    - **Operability -** The interface of the system will be consistent.
    - **Efficiency -** Once user has learned about the system through his interaction, he can perform the task easily.
    - **Understandability**-Because of user friendly interfaces, it is more understandable to the users.

**2.3 MINIMUM HARDWARE REQUIREMENTS**

* **Processor -**Intel Core i5
* **Hard Disk –** 1 TB
* **RAM -** 8GB

**2.4 MINIMUM SOFTWARE REQUIREMENTS**

Python based Computer Vision and Deep Learning libraries will be exploited for the development and experimentation of the project.

* Programming Language – R
* Operating System - Windows 10(64 bit)
* IDE – Google Collab.
* Packages – Plotrix, FactoExtra, GridExtra, NBClust.

**3. ALGORITHM ANALYSIS**

**3.1 EXISTING ALGORITHM**

One of the biggest issues with customer segmentation is data quality. Inaccurate data in source systems will usually result in poor grouping. Larger Inventory is to be maintained by both manufacturer and distributor. It is not Cost effective Expenditure increases based on the market.

If there is a sudden change in the market all the investment made might be Useless

**DRAWBACKS OF EXISTING ALGORITHMS**

* Larger Inventory is to be maintained by both manufacturer and distributor.
* It is not Cost Effective.
* Expenditure increases based on the market.
* If there is a sudden change in the market all the investment made might be Useless.

**3.2 PROPOSED ALGORITHM**

To overcome the drawbacks of the existing applications, the proposed model has been evolved. We built a K - means model based on silhouette algorithm along with some pre-processing techniques, using this model we can build a technique to analyse data patterns. Our project can analyse efficiently and suggest the products based on customer’s behaviour.

**ADVANTAGES OF PROPOSED MODEL**

Implementing customer segmentation leads to plenty of new business opportunities.

You can do a lot of optimizations in:

* Budgeting
* Product design
* Promotion
* Marketing
* Customer satisfaction.

Let us discuss these benefits in more depth.

**Budgeting**

Nobody likes to invest in campaigns that don’t generate any new customers. Most companies don’t have huge marketing budgets, so that money has to be spent right. Segmentation enables you to target customers with the highest potential value first, so you get the most out of your marketing budget.

**Product design**

Customer segmentation helps you understand what your users need. You can identify the most active users/customers, and optimize your application/offer towards their needs.

**Promotion**

Properly implemented customer segmentation helps you plan special offers and deals. Frequent deals have become a staple of e-commerce and commercial software in the past few years. If you reach a customer with just the right offer, at the right time, there’s a huge chance they’re going to buy. Customer segmentation will help you tailor your special offers perfectly.

**Marketing**

The marketing strategy can be directly improved with segmentation because you can plan personalized marketing campaigns for different customer segments, using the channels that they use the most.

**Customer satisfaction**

By studying different customer groups, you learn what they value the most about your company. This information will help you create personalized products and services that perfectly fit your customers’ preferences.

**3.3 FEASIBILITY STUDY**

A feasibility study is an analysis that takes all a project's relevant factors into account including economic, technical, legal, and scheduling considerations to ascertain the likelihood of completing the project successfully. A feasibility study is important and essential to evolute any proposed project is feasible or not. A feasibility study is simply an assessment of the practicality of a proposed plan or project.

**The main objectives of feasibility are mentioned below:**

To determine if the product is technically and financially feasible to develop, is the main aim of the feasibility study activity. A feasibility study should provide management with enough information to decide:

* Whether the project can be done.
* To determine how successful your proposed action will be.
* Whether the final product will benefit its intended users.
* To describe the nature and complexity of the project.
* What are the alternatives among which a solution will be chosen (During subsequent phases)
* To analyse if the software meets organizational requirements.

There are various types of feasibility that can be determined. They are:

**Operational** - Define the urgency of the problem and the acceptability of any solution, includes people-oriented and social issues: internal issues, such as manpower problems, labor objections, manager resistance, organizational conflicts, and policies; also, external issues, including social acceptability, legal aspects, and government regulations.

**Technical**: Is the feasibility within the limits of current technology? Does the technology exist at all? Is it available within a given resource?

**Economic** - Is the project possible, given resource constraints? Are the benefits that will accrue from the new system worth the costs? What are the savings that will result from the system, including tangible and intangible ones? What are the development and operational costs?

**Schedule** - Constraints on the project schedule and whether they could be reasonably met.

**3.3.1 Economic Feasibility:**

Economic analysis could also be referred to as cost/benefit analysis. It is the most frequently used method for evaluating the effectiveness of a new system. In economic analysis the procedure is to determine the benefits and savings that are expected from a candidate system and compare them with costs. Economic feasibility study related to price, and all kinds of expenditure related to the scheme before the project starts. This study also improves project reliability. It is also helpful for the decision-makers to decide the planned scheme processed latter or now, depending on the financial condition of the organization. This evaluation process also studies the price benefits of the proposed scheme. Economic Feasibility also performs the following tasks.

* + - Cost of packaged software/ software development.
    - Cost of doing full system study.
    - Is the system cost Effective?

**3.3.2 Technical Feasibility:**

Large part of determining resources has to do with assessing technical feasibility. It considers the technical requirements of the proposed project. The technical requirements are then compared to the technical capability of the organization. The systems project is considered technically feasible if the internal technical capability is sufficient to support the project requirements. The analyst must find out whether current technical resources can be where the expertise of system analysts is beneficial, since using their own experience and their contact with vendors they will be able to answer the question of technical feasibility.

Technical Feasibility also performs the following tasks.

* + - Is the technology available within the given resource constraints?
    - Is the technology have the capacity to handle the solution
    - Determines whether the relevant technology is stable and established.
    - Is the technology chosen for software development has a large number of users so that they can be consulted when problems arise, or improvements are required.

**3.3.3 Operational Feasibility:**

Operational feasibility is a measure of how well a proposed system solves the problems and takes advantage of the opportunities identified during scope definition and how it satisfies the requirements identified in the requirements analysis phase of system development. The operational feasibility refers to the availability of the operational resources needed to extend research results beyond on which they were developed and for which all the operational requirements are minimal and easily accommodated. In addition, the operational feasibility would include any rational compromises farmers make in adjusting the technology to the limited operational resources available to them. The operational Feasibility also perform the tasks like

* Does the current mode of operation provide adequate response time?
* Does the current of operation make maximum use of resources.
* Determines whether the solution suggested by the software development team is acceptable.
* Does the operation offer an effective way to control the data?

Our project operates with a processor and packages installed are supported by the system.

**3.4 COST BENEFIT ANALYSIS**

The financial and the economic questions during the preliminary investigation are verified to estimate the following:

* The cost of the hardware and software for the class of application being considered.
* The benefits in the form of reduced cost.
* The proposed system will give the minute information, as a result.
* Performance is improved which in turn may be expected to provide increased profits.
* This feasibility checks whether the system can be developed with the available funds.
* This can be done economically if planned judicially, so it is economically feasible.
* The cost of the project depends upon the number of man-hours required.

**4. SOFTWARE DESCRIPTION**

**4.1 Google Collab.**

Google Collab is an open source and free distribution of R and Python programming language for machine learning as well as data science projects. Therefore, it is known as a professional data science platform. It contains a powerful environment manager, which provides a different type of Python and R environment such as a Spyder, Jupyter notebook, and so on. Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository. It is available for Windows, macOS, and Linux.

Google Collab is a free and open-source distributor

* Quickly download 7,500+ Python/R data science packages.
* Manages a lot of libraries
* It provides various environments by virtualization
* Can easily deal with large data computing
* It works properly without the need for any administrative privileges.
* Visualize results with Matplotlib, Bokeh, Datashader , and Holoviews.

**4.2 Plotrix**

Plotrix Provides general plotting functions for R Language. It has lots of Plots, Labelling, axis and colour functions. It is used to display data in charts forms like Boxplot, Bar chart, Line chart, etc. Placing labels for easy identification of various data sets and executing graphical function over the interface.

**4.3 FactoExtra**

FactoExtra is an open-source library written in R. It can run on top of TensorFlow, Microsoft Cognitive Toolkit, Theano, or PlaidML. Designed to enable fast experimentation with large data sets, it focuses on being user-friendly, modular, and extensible. The **FactoExtra** package is an alternative to the graphical functions offered by FactoMineR and Factoshiny. It is an R package that is used to extract and visualize the output of multivariate data analysis that includes Principal Component Analysis (PCA), Multi factor Analysis (MFA). However, the result is presented differently according to the used packages. To help in the interpretation and in the visualization of multivariate analysis - such as cluster analysis and dimensionality reduction analysis - we developed an easy-to-use R package named FactoExtra.

**4.4 GridExtra**

GridExtra is an open-source library written in R. It can run on top of TensorFlow, Microsoft Cognitive Toolkit, Theano, or PlaidML. Designed to enable fast experimentation with large data sets, it focuses on being user-friendly, modular, and extensible. The Grid package provides low-level functions to create graphical objects grobs, and position them on a page in specific viewports. The gtable package introduced a higher-level layout scheme, arguably more amenable to user-level interaction. With the grid arrange package of functions, it builds upon gtable to arrange multiple grobs on a page.

**4.5 NBClust**

The R Library adds cluster forming capabilities to your R interpreter. This library provides extensive file format support, an efficient internal representation, and powerful clustering capabilities. The core data is designed for fast access to data stored in a few basic table formats.It provides 30 indexes for determining the optimal number of clusters in a data set and offers the best clustering scheme from different results to the user.

To form clusters based on the data one should, Import cluster graphics from NBClust library.

**4.6 Matplotlib**

Matplotlib is a plotting library for python. Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was

introduced by John Hunter in the year 2002. Matplotlib comes with a wide variety of plots. Plots help to understand trends, patterns, and to make correlations. It consists of various plots like line, bar, scatter etc. pyplot () is the most important function in matplotlib library, which is used to plot 2D data.

**5. PROJECT DESCRIPTION**

**5.1 PROBLEM DEFINITION**

Customer Segmentation is the present trend of the contemporary world. The huge amounts of data which are being processed, requires the urge to find efficient patterns helpful for the organization.

We use K means algorithms for Customer Segmentation using R programming language. The main objective is to develop a reliable system for finding clusters based on the behavioural data of the customers.

**5.2 PROJECT OVERVIEW**

The main aim of our project is to help companies detect patterns in customer data which is currently potential. Our project comprises of two parts, first one is to visualize the customer data, and the second part is to provide optimal clusters so that one can find better productive outputs. We used **K-means Algorithm** which is an Unsupervised Machine Learning algorithm as it is the best for data grouping. The steps are providing data to the algorithm and then the algorithm shows us important insights on various attributes of the data which helps us to group certain data. It is done by using various R packages.

* The steps involved in our project are:
  1. Collection of datasets.
  2. Data pre-processing.
  3. Data visualization.
  4. Applying various k means algorithm to find the optimal clusters.
  5. Grouping of data and finding important insights.

The output of our project consists of optimal clusters formed through data, visualised through R Programming Packages.

**5.3 MODULE DESCRIPTION**

**5.3.1 MODEL**

K-means Algorithm

* We specify the number of clusters that we need to create.
* The algorithm selects k objects at random from the dataset. This object is the initial

cluster or mean.

* The closest centroid obtains the assignment of a new observation. We base this assignment

on the Euclidean Distance between object and the centroid.

* k clusters in the data points update the centroid through calculation of the new mean values

present in all the data points of the cluster. The kth cluster’s centroid has a length of p that contains means of all variables for observations in the k-th cluster. We denote the number of variables with p.

* Iterative minimization of the total within the sum of squares. Then through the iterative

minimization of the total sum of the square, the assignment stop wavering when we achieve

maximum iteration. The default value is 10 that the R software uses for the maximum

iterations.

### Determining Optimal Clusters

While working with clusters, you need to specify the number of clusters to use. You would like to utilize the optimal number of clusters. To help you in determining the optimal clusters, there are three popular methods –

* Elbow method
* Silhouette method
* Gap statistic

**Elbow Method**

The main goal behind cluster partitioning methods like k-means is to define the clusters such that the intra-cluster variation stays minimum.

**minimize(sum W(Ck)), k=1…k]**

Where Ck represents the kth cluster and W(Ck) denotes the intra-cluster variation. With the measurement of the total intra-cluster variation, one can evaluate the compactness of the clustering boundary. We can then proceed to define the optimal clusters as follows

First, we calculate the clustering algorithm for several values of k. This can be done by creating a variation within k from 1 to 10 clusters. We then calculate the total intra-cluster sum of square (iss). Then, we proceed to plot iss based on the number of k clusters. This plot denotes the appropriate number of clusters required in our model. In the plot, the location of a bend or a knee is the indication of the optimum number of clusters.

#### **Average Silhouette Method**

With the help of the average silhouette method, we can measure the quality of our clustering operation. With this, we can determine how well within the cluster is the data object. If we obtain a high average silhouette width, it means that we have good clustering. The average silhouette method calculates the mean of silhouette observations for different k values. With the optimal number of k clusters, one can maximize the average silhouette over significant values for k clusters. Using the silhouette function in the cluster package, we can compute the average silhouette width using the kmean function.

#### **Gap Statistic Method**

In 2001, researchers at Stanford University – **R. Tibshirani, G.Walther and T. Hastie** published the Gap Statistic Method. We can use this method to any of the clustering method like K-means, hierarchical clustering etc. Using the gap statistic, one can compare the total intracluster variation for different values of k along with their expected values under the null reference distribution of data. With the help of **Monte Carlo simulations**, one can produce the sample dataset. For each variable in the dataset, we can calculate the range between min(xi) and max (xj) through which we can produce values uniformly from interval lower bound to upper bound.

For computing the gap statistics method we can utilize the clusGap function for providing gap statistic as well as standard error for a given output.

**6. SYSTEM DESIGN**

**6.1 Introduction to UML**

Unified Modeling Language **(UML)** is a general-purpose modeling language. The main aim of UML is to define a standard way to **visualize** the way a system has been designed. It is quite like blueprints used in other fields of engineering. UML is not a programming language, it is rather a visual language. We use UML diagrams to portray the behaviour and structure of a system. UML helps software engineers, businessmen and system architects with modeling, design and analysis. The Object Management Group (OMG) adopted Unified Modelling Language as a standard in 1997. It's been managed by OMG ever since. International Organization for Standardization (ISO) published UML as an approved standard in 2005. UML has been revised over the years and is reviewed periodically.

**Why we need UML**

* Complex applications need collaboration and planning from multiple teams and hence require a clear and concise way to communicate amongst them.
* Businessmen do not understand code. So, UML becomes essential to communicate with non-programmers’ essential requirements, functionalities and processes of the system.
* A lot of time is saved down the line when teams can visualize processes, user interactions and static structure of the system.

UML is linked with **object-oriented** design and analysis. UML makes the use of elements and forms associations between them to form diagrams. Diagrams in UML can be broadly classified as:

* **Structural Diagrams –** Capture static aspects or structure of a system. Structural Diagrams include Component Diagrams, Object Diagrams, Class Diagrams and Deployment Diagrams.
* **Behaviour Diagrams –** Capture dynamic aspects or behaviour of the system. Behaviour diagrams include Use Case Diagrams, State Diagrams, Activity Diagrams and Interaction Diagrams.

Building Blocks of the UML Building Blocks of the UML Building Blocks of the UML

Diagram

Description automatically generated

**FIG 6.1 BUILDING BLOCKS IN UML**

**6.2 Building Blocks of the UML**

The vocabulary of the UML encompasses three kinds of building blocks:

* Things
* Relationships
* Diagrams

Things are the abstractions that are first-class citizens in a model; relationships tie these things together; diagrams group interesting collections of things.

**Things in the UML**

There are four kinds of things in the UML:

* Structural things
* Behavioural things
* Grouping things
* Annotational things

These things are the basic object-oriented building blocks of the UML. You use them to write well-formed models.

**Structural Things**

Structural things are the nouns of UML models. These are the mostly static parts of a model, representing elements that are either conceptual or physical. Collectively, the structural things are called classifiers.

A class is a description of a set of objects that share the same attributes, operations, relationships, and semantics. A class implements one or more interfaces. Graphically, a class is rendered as a rectangle, usually including its name, attributes, and operations

**Class** - A Class is a set of identical things that outlines the functionality and properties of an object. It also represents the abstract class whose functionalities are not defined. Its notation is as follows

Table

Description automatically generated with medium confidence

**Interface** - A collection of functions that specify a service of a class or component, i.e.

Externally visible behavior of that class.



**Collaboration** - A larger pattern of behaviors and actions. Example: All classes and behaviors that create the modeling of a moving tank in asimulation.



**Use Case** - A sequence of actions that a system performs that yields an observable result.

Used to structure behavior in a model. Is realized by collaboration.



**Component** - A physical and replaceable part of a system that implements a number of interfaces. Example: a set of classes, interfaces, and collaborations.



**Node** - A physical element existing at run time and represents are source.

A picture containing diagram

Description automatically generated

**Behavioral Things**

Behavioral things are the dynamic parts of UML models. These are the verbs of a model, representing behavior over time and space. In all, there are three primary kinds of behavioral things

* Interaction
* State machine

**Interaction**

It is a behavior that comprises a set of messages exchanged among a set of objects or roles within a particular context to accomplish a specific purpose. The behavior of a society of objects or of an individual operation may be specified with an interaction. An interaction involves a number of other elements, including messages, actions, and connectors (the connection between objects). Graphically, a message is rendered as a directed line, almost always including the name of its operation.



**State machine**

State machine is a behaviour that specifies the sequences of states an object or an interaction goes through during its lifetime in response to events, together with its responses to those events. The behaviour of an individual class or a collaboration of classes may be specified with a state machine. A state machine involves a number of other elements, including states, transitions (the flow from state to state), events (things that trigger a transition), and activities (the response to a transition). Graphically, a state is rendered as a rounded rectangle, usually including its name and its substates.

Letter

Description automatically generated with medium confidence

**Grouping Things**

Grouping things can be defined as a mechanism to group elements of a UML model together. There is only one grouping thing available.

**Package −** Package is the only one grouping thing available for gathering structural and behavioural things.

Diagram

Description automatically generated

**Annotational Things**

Annotational things are the explanatory parts of UML models. These are the comments you may apply to describe, illuminate, and remark about any element in a model. There is one primary kind of annotational thing, called a note. A note is simply a symbol for rendering constraints and comments attached to an element or a collection of elements.

Text

Description automatically generated

**Relationships in the UML**

Relationship is another most important building block of UML. It shows how the elements are associated with each other and this association describes the functionality of an application.

There are four kinds of relationships in the UML:

* Dependency
* Association
* Generalization
* Realization

**Dependency**

It is an element (the independent one) that may affect the semantics of the other element (the dependent one). Graphically, a dependency is rendered as a dashed line, possibly directed, and occasionally including a label.



**Association**

Association is basically a set of links that connects the elements of a UML model. It also describes how many objects are taking part in that relationship.



**Generalization**

It is a specialization/generalization relationship in which the specialized element (the child) builds on the specification of the generalized element (the parent). The child shares the structure and the behavior of the parent. Graphically, a generalization relationship is rendered as a solid line with a hollow arrowhead pointing to the parent



**Realization**

Realization can be defined as a relationship in which two elements are connected. One element describes some responsibility, which is not implemented and the other one implements them. This relationship exists in case of interfaces.



**6.3 UML DIAGRAMS**

UML is a modern approach to modelling and documenting software. It is based on

**Diagrammatic representations** of software components. It is the final output, and the

Diagram represents the system.

**UML includes the following**

* Use case diagram
* Activity diagram
* Flow Chart

Diagram, schematic

Description automatically generated

**FIG 6.3.1 USECASE DIAGRAM**

Diagram

Description automatically generated

**FIG 6.3.3 ACTIVITY DIAGRAM**

Diagram

Description automatically generated

**FIG 6.3.2 FLOWCHART**

**7. DEVELOPMENT**

**7.2 SAMPLE CODE**

**DATASET USED**

* **A DATASET** is a set or collection of data. This set is normally presented in a tabular pattern. Every column describes a particular variable. Each row corresponds to a given member of the data set, as per the given question. This is a part of **DATA MANAGEMENT**.
* We have Collected Mall datasets from **KAGGLE.COM.** Our Datasets consists of 5 attributes and 200 customers data.
* Another data set of 1000 customers is also tested.

Text

Description automatically generated

Graphical user interface, application

Description automatically generated

**Fig 7.2.1 Bar plot for Gender Comparison**

Chart, histogram

Description automatically generated

**Fig 7.2.2 Visualisation based on Age**

Chart, box and whisker chart

Description automatically generated

**Fig 7.2.3 Boxplot for Descriptive Analysis of Age**

**Analysis for Annual Income and Spending Score**

Chart, histogram

Description automatically generated

**Fig 7.2.4 Histogram for Annual Income**

Chart

Description automatically generated

**Fig 7.2.5 Density Plot for Annual Income**

**Silhouette Graphs**

Chart

Description automatically generated

Chart

Description automatically generated

Chart, funnel chart

Description automatically generated

**Scatter Plot based on the Age and Spending Score and Annual Income**

Chart, scatter chart

Description automatically generated

**Fig 7.2.6 Scatter plot for Annual Income**

Chart, scatter chart

Description automatically generated

**Fig 7.2.7 Scatter plot for age and spending Score**

**7.3 RESULT**

**Final Customer segmented Clusters**

Chart, scatter chart

Description automatically generated

**8. TESTING**

**8.1 INTRODUCTION OF TESTING**

SOFTWARE TESTING is defined as an activity to check whether the actual results match the expected results and to ensure that the software system is Defect free. It involves the execution of a software component or system component to evaluate one or more properties of interest. It is required for evaluating the system .This phase is the critical phase of software quality assurance and presents the ultimate view of coding.

**Importance of Testing**

The importance of software testing is imperative. A lot of times this process is skipped, therefore, the product and business might suffer. To understand the importance of testing, here are some key points to explain

* Software Testing saves money
* Provides Security
* Improves Product Quality
* Customer satisfaction

Testing is of different ways The main idea behind the testing is to reduce the errors and do it with a minimum time and effort.

**Benefits of Testing**

* **Cost-Effective:** It is one of the important advantages of software testing. Testing any IT project on time helps you to save your money for the long term. In case if the bugs caught in the earlier stage of software testing, it costs less to fix.
* **Security:** It is the most vulnerable and sensitive benefit of software testing. People are looking for trusted products. It helps in removing risks and problems earlier.
* **Product quality:** It is an essential requirement of any software product. Testing ensures a quality product is delivered to customers.
* **Customer Satisfaction:** The main aim of any product is to give satisfaction to theircustomers. UI/UX Testing ensures the best user experience.

**Different types of Testing**

**Unit Testing:** Unit tests are very low level, close to the source of your application. They consist in testing individual methods and functions of the classes, components or modules used by your software. Unit tests are in general quite cheap to automate and can be run very quickly by a Continuous integration server.

**Integration Testing:** Integration tests verify that different modules or services used by your application work well together. For example, it can be testing the interaction with the database or making sure that microservices work together as expected. These types of tests are more expensive to run as they require multiple parts of the application to be up and running.

**Functional Tests:** Functional tests focus on the business requirements of an application. They only verify the output of an action and do not check the intermediate states of the system when performing that action.

There is sometimes a confusion between integration tests and functional tests as they both require multiple components to interact with each other. The difference is that an integration test may simply verify that you can query the database while a functional test would expect to get a specific value from the database as defined by the product requirements.

**Regression Testing:** Regression testing is a crucial stage for the product & very useful for the developers to identify the stability of the product with the changing requirements. Regression testing is a testing that is done to verify that a code change in the software does not impact the existing functionality of the product.

**System Testing:** System testing of software or hardware is testing conducted on a complete integrated system to evaluate the system’s compliance with its specified requirements. System testing is a series of different tests whose primary purpose is to fully exercise the computer-based system.

**Performance Testing:** It checks the speed, response time, reliability, resource usage, scalability of a software program under their expected workload. The purpose of Performance Testing is not to find functional defects but to eliminate performance bottlenecks in the software or device.

**Alpha Testing:** This is a form of internal acceptance testing performed mainly by the in-house software QA and testing teams. Alpha testing is the last testing done by the test teams at the development site after the acceptance testing and before releasing the software for the beta test. It can also be done by the potential users or customers of the application. But still, this is a form of in-house acceptance testing.

**Beta Testing:** This is a testing stage followed by the internal full alpha test cycle. This is the final testing phase where the companies release the software to a few external user groups outside the company test teams or employees. This initial software version is known as the beta version. Most companies gather user feedback in this release.

**Black Box Testing**:

It is also known as Behavioural Testing, is a software testing method in which the internal structure/design/implementation of the item being tested is **not** known to the tester. These tests can be functional or non-functional, though usually functional.

Graphical user interface

Description automatically generated with low confidence

This method is named so because the software program, in the eyes of the tester, is like a black box; inside which one cannot see. This method attempts to find errors in the following categories:

* Incorrect or missing functions
* Interface errors
* Errors in data structures or external database access
* Behaviour or performance errors
* Initialization and termination errors

**White Box Testing:**

White box testing (also known as Clear Box Testing, Open Box Testing, Glass Box Testing, Transparent Box Testing, Code-Based Testing or Structural Testing) is a software testing method in which the internal structure/design/implementation of the item being tested is known to the tester. The tester chooses inputs to exercise paths through the code and determines the appropriate outputs. Programming know-how and the implementation knowledge is essential. White box testing is testing beyond the user interface and into the nitty-gritty of a system. This method is named so because the software program, in the eyes of the tester, is like a white/transparent box; inside which one clearly sees.

Text

Description automatically generated

**8.3 TEST CASE**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Customer ID** | **Gender** | **Age** | **Annual Income (k$)** | **Spending Score (1-100)** |
| 1 | Male | 19 | 15 | 39 |
| 2 | Male | 21 | 15 | 81 |
| 3 | Female | 20 | 16 | 6 |
| 4 | Female | 23 | 16 | 77 |
| 5 | Female | 31 | 17 | 40 |
| 6 | Female | 22 | 17 | 76 |
| 7 | Female | 35 | 18 | 6 |
| 8 | Female | 23 | 18 | 94 |
| 9 | Male | 64 | 19 | 3 |
| 10 | Female | 30 | 19 | 72 |
| 11 | Male | 67 | 19 | 14 |
| 12 | Female | 35 | 19 | 99 |
| 13 | Female | 58 | 20 | 15 |
| 14 | Female | 24 | 20 | 77 |
| 15 | Male | 37 | 20 | 13 |
| 16 | Male | 22 | 20 | 79 |
| 17 | Female | 35 | 21 | 35 |
| 18 | Male | 20 | 21 | 66 |
| 19 | Male | 52 | 23 | 29 |
| 20 | Female | 35 | 23 | 98 |
| 21 | Male | 35 | 24 | 35 |
| 22 | Male | 25 | 24 | 73 |
| 23 | Female | 46 | 25 | 5 |
| 24 | Male | 31 | 25 | 73 |
| 25 | Female | 54 | 28 | 14 |
| 26 | Male | 29 | 28 | 82 |
| 27 | Female | 45 | 28 | 32 |
| 28 | Male | 35 | 28 | 61 |
| 29 | Female | 40 | 29 | 31 |
| 30 | Female | 23 | 29 | 87 |
| 31 | Male | 60 | 30 | 4 |
| 32 | Female | 21 | 30 | 73 |
| 33 | Male | 53 | 33 | 4 |
| 34 | Male | 18 | 33 | 92 |
| 35 | Female | 49 | 33 | 14 |
| 36 | Female | 21 | 33 | 81 |
| 37 | Female | 42 | 34 | 17 |
| 38 | Female | 30 | 34 | 73 |
| 39 | Female | 36 | 37 | 26 |
| 40 | Female | 20 | 37 | 75 |
| 41 | Female | 65 | 38 | 35 |
| 42 | Male | 24 | 38 | 92 |
| 43 | Male | 48 | 39 | 36 |
| 44 | Female | 31 | 39 | 61 |
| 45 | Female | 49 | 39 | 28 |
| 46 | Female | 24 | 39 | 65 |
| 47 | Female | 50 | 40 | 55 |
| 48 | Female | 27 | 40 | 47 |
| 49 | Female | 29 | 40 | 42 |
| 50 | Female | 31 | 40 | 42 |
| 51 | Female | 49 | 42 | 52 |
| 52 | Male | 33 | 42 | 60 |
| 53 | Female | 31 | 43 | 54 |
| 54 | Male | 59 | 43 | 60 |
| 55 | Female | 50 | 43 | 45 |
| 56 | Male | 47 | 43 | 41 |
| 57 | Female | 51 | 44 | 50 |
| 58 | Male | 69 | 44 | 46 |
| 59 | Female | 27 | 46 | 51 |
| 60 | Male | 53 | 46 | 46 |
| 61 | Male | 70 | 46 | 56 |
| 62 | Male | 19 | 46 | 55 |
| 63 | Female | 67 | 47 | 52 |
| 64 | Female | 54 | 47 | 59 |
| 65 | Male | 63 | 48 | 51 |
| 66 | Male | 18 | 48 | 59 |
| 67 | Female | 43 | 48 | 50 |
| 68 | Female | 68 | 48 | 48 |
| 69 | Male | 19 | 48 | 59 |
| 70 | Female | 32 | 48 | 47 |
| 71 | Male | 70 | 49 | 55 |
| 72 | Female | 47 | 49 | 42 |
| 73 | Female | 60 | 50 | 49 |
| 74 | Female | 60 | 50 | 56 |
| 75 | Male | 59 | 54 | 47 |
| 76 | Male | 26 | 54 | 54 |
| 77 | Female | 45 | 54 | 53 |
| 78 | Male | 40 | 54 | 48 |
| 79 | Female | 23 | 54 | 52 |
| 80 | Female | 49 | 54 | 42 |
| 81 | Male | 57 | 54 | 51 |
| 82 | Male | 38 | 54 | 55 |
| 83 | Male | 67 | 54 | 41 |
| 84 | Female | 46 | 54 | 44 |
| 85 | Female | 21 | 54 | 57 |
| 86 | Male | 48 | 54 | 46 |
| 87 | Female | 55 | 57 | 58 |
| 88 | Female | 22 | 57 | 55 |
| 89 | Female | 34 | 58 | 60 |
| 90 | Female | 50 | 58 | 46 |
| 91 | Female | 68 | 59 | 55 |
| 92 | Male | 18 | 59 | 41 |
| 93 | Male | 48 | 60 | 49 |
| 94 | Female | 40 | 60 | 40 |
| 95 | Female | 32 | 60 | 42 |
| 96 | Male | 24 | 60 | 52 |
| 97 | Female | 47 | 60 | 47 |
| 98 | Female | 27 | 60 | 50 |
| 99 | Male | 48 | 61 | 42 |
| 100 | Male | 20 | 61 | 49 |
| 101 | Female | 23 | 62 | 41 |
| 102 | Female | 49 | 62 | 48 |
| 103 | Male | 67 | 62 | 59 |
| 104 | Male | 26 | 62 | 55 |
| 105 | Male | 49 | 62 | 56 |
| 106 | Female | 21 | 62 | 42 |
| 107 | Female | 66 | 63 | 50 |
| 108 | Male | 54 | 63 | 46 |
| 109 | Male | 68 | 63 | 43 |
| 110 | Male | 66 | 63 | 48 |
| 111 | Male | 65 | 63 | 52 |
| 112 | Female | 19 | 63 | 54 |
| 113 | Female | 38 | 64 | 42 |
| 114 | Male | 19 | 64 | 46 |
| 115 | Female | 18 | 65 | 48 |
| 116 | Female | 19 | 65 | 50 |
| 117 | Female | 63 | 65 | 43 |
| 118 | Female | 49 | 65 | 59 |
| 119 | Female | 51 | 67 | 43 |
| 120 | Female | 50 | 67 | 57 |
| 121 | Male | 27 | 67 | 56 |
| 122 | Female | 38 | 67 | 40 |
| 123 | Female | 40 | 69 | 58 |
| 124 | Male | 39 | 69 | 91 |
| 125 | Female | 23 | 70 | 29 |
| 126 | Female | 31 | 70 | 77 |
| 127 | Male | 43 | 71 | 35 |
| 128 | Male | 40 | 71 | 95 |
| 129 | Male | 59 | 71 | 11 |
| 130 | Male | 38 | 71 | 75 |
| 131 | Male | 47 | 71 | 9 |
| 132 | Male | 39 | 71 | 75 |
| 133 | Female | 25 | 72 | 34 |
| 134 | Female | 31 | 72 | 71 |
| 135 | Male | 20 | 73 | 5 |
| 136 | Female | 29 | 73 | 88 |
| 137 | Female | 44 | 73 | 7 |
| 138 | Male | 32 | 73 | 73 |
| 139 | Male | 19 | 74 | 10 |
| 140 | Female | 35 | 74 | 72 |
| 141 | Female | 57 | 75 | 5 |
| 142 | Male | 32 | 75 | 93 |
| 143 | Female | 28 | 76 | 40 |
| 144 | Female | 32 | 76 | 87 |
| 145 | Male | 25 | 77 | 12 |
| 146 | Male | 28 | 77 | 97 |
| 147 | Male | 48 | 77 | 36 |
|  |  |  |  |  |
| 148 | Female | 32 | 77 | 74 |
| 149 | Female | 34 | 78 | 22 |
| 150 | Male | 34 | 78 | 90 |
| 151 | Male | 43 | 78 | 17 |
| 152 | Male | 39 | 78 | 88 |
| 153 | Female | 44 | 78 | 20 |
| 154 | Female | 38 | 78 | 76 |
| 155 | Female | 47 | 78 | 16 |
| 156 | Female | 27 | 78 | 89 |
| 157 | Male | 37 | 78 | 1 |
| 158 | Female | 30 | 78 | 78 |
| 159 | Male | 34 | 78 | 1 |
| 160 | Female | 30 | 78 | 73 |
| 161 | Female | 56 | 79 | 35 |
| 162 | Female | 29 | 79 | 83 |
| 163 | Male | 19 | 81 | 5 |
| 164 | Female | 31 | 81 | 93 |
| 165 | Male | 50 | 85 | 26 |
| 166 | Female | 36 | 85 | 75 |
| 167 | Male | 42 | 86 | 20 |
| 168 | Female | 33 | 86 | 95 |
| 169 | Female | 36 | 87 | 27 |
| 170 | Male | 32 | 87 | 63 |
| 171 | Male | 40 | 87 | 13 |
| 172 | Male | 28 | 87 | 75 |
| 173 | Male | 36 | 87 | 10 |
| 174 | Male | 36 | 87 | 92 |
| 175 | Female | 52 | 88 | 13 |
| 176 | Female | 30 | 88 | 86 |
| 177 | Male | 58 | 88 | 15 |
| 178 | Male | 27 | 88 | 69 |
| 179 | Male | 59 | 93 | 14 |
| 180 | Male | 35 | 93 | 90 |
| 181 | Female | 37 | 97 | 32 |
| 182 | Female | 32 | 97 | 86 |
| 183 | Male | 46 | 98 | 15 |
| 184 | Female | 29 | 98 | 88 |
| 185 | Female | 41 | 99 | 39 |
| 186 | Male | 30 | 99 | 97 |
| 187 | Female | 54 | 101 | 24 |
| 188 | Male | 28 | 101 | 68 |
| 189 | Female | 41 | 103 | 17 |
| 190 | Female | 36 | 103 | 85 |
| 191 | Female | 34 | 103 | 23 |
| 192 | Female | 32 | 103 | 69 |
| 193 | Male | 33 | 113 | 8 |
| 194 | Female | 38 | 113 | 91 |
| 195 | Female | 47 | 120 | 16 |
| 196 | Female | 35 | 120 | 79 |
| 197 | Female | 45 | 126 | 28 |
| 198 | Male | 32 | 126 | 74 |
| 199 | Male | 32 | 137 | 18 |
| 200 | Male | 30 | 137 | 83 |

**9. CONCLUSION**

In this project, the approach of using K-means algorithm was explored in order to classify and cluster customer data. The complete procedure was described, respectively, from collecting the data and validation for pre-processing and augmentation, finally the procedure of association and grouping for finding potential insights from the data. Different algorithms were executed in order to optimize the accuracy of the created model. As the presented method has not been exploited, as far as we know, in the field of data segmentation, there was no comparison with related results, using the exact technique.

* 1. **FUTURE SCOPE**

This project has a very great potential to be further improved in the future. The work carried out has relevance to real world classification of customer’s behaviour, using different algorithms to get better accuracy rate and it involves both unsupervised learning algorithms for visualization and finding data patterns.

The main goal for the future work will be developing a complete system consisting of server-side components containing a huge data set and an application for smart mobile devices with features such as displaying associations and insights to efficiently cluster the data. By extending this project, we can achieve a valuable impact on sustainable development, impacting the revenue and satisfactions of the customers.

**11. REFERENCE LINKS**

[1] https://neptune.ai/blog/customer-segmentation-using-machine-learning

[2]https://towardsdatascience.com/mall-customers-segmentation-using-machine-learning-274ddf5575d5

[3] https://www.kaggle.com/code/serhanayberkkl/customer-segmentation

[4].https://www.yourarticlelibrary.com/marketing/marketing-management/advantages-and-disadvantages-of-market-segmentation/2795