A Project Report

on

CUSTOMER SEGMENTATION

Submitted in partial fulfillment of the

requirement for the award of the degree of

Bachelor of Technology in Computer Science and Engineering

Under The Supervision of

Dr. Sudeept Singh Yadav

Associate Professor

Department of Computer Science and Engineering

Submitted By

ABHIMANYU KUMAR - 18SCSE1180023

DEVESH KUMAR SINGH - 18SCSE1180022

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

GALGOTIAS UNIVERSITY, GREATER NOIDA

INDIA

May 2022

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE'S DECLARATION

We hereby certify that the work which is being presented in the project, entitled **"CUSTOMER SEGMENTATION"** in partial fulfillment of the requirements for the award of the **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING** submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of **January 2022 to May 2022**, under the supervision of **Dr. Sudeept Singh Yadav [Associate Professor], Department of Computer Science and Engineering** of School of Computing Science and Engineering, Galgotias University, Greater Noida.

The matter presented in the project has not been submitted by me/us for the award of any other degree of this or any other places.

Abhimanyu Kumar - 18SCSE1180023

Devesh Kumar Singh - 18SCSE1180022

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Supervisor

Dr. Sudeept Singh Yadav

Associate Professor

CERTIFICATE

The Final Project Viva-Voce examination of **Abhimanyu Kumar - 18SCSE1180023 and Devesh Kumar Singh - 18SCSE1180022** has been held on_____and his work is recommended for the award of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING**

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Project Coordinator **Signature of Dean**

Date: May 2022

Place: Greater Noida

Acknowledgment

We take this opportunity to express our sincere gratitude to **Dr. Sudeept Singh Yadav [Associate Professor],** Department of COMPUTER SCIENCE AND ENGINEERING GALGOTIAS UNIVERSITY, GREATER NOIDA. Deep Knowledge & keen interest of our supervisor in the field of "Machine Learning" to carry out this project. His endless patience, scholarly guidance, strong motivation, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to other faculty members and the staff of the COMPUTER SCIENCE AND ENGINEERING department of GALGOTIAS UNIVERSITY, GREATER NOIDA to finish our project.

We would like to thank our entire course mate in GALGOTIAS UNIVERSITY, GREATER NOIDA, who took part in this discussion while completing the course work.

Abstract

The goal of this study is to use survey data to do a segmentation analysis and categorize members of a target segment. The analysis of this exploration is to validate the usage of data science methods in marketing and sales by analyzing customer survey data. This study is expected to lead to a better understanding of data-driven corporate strategy. The development objective was to use current data science methods to first segment clients and then identify members of the target group. Marketing ideas such as segmentation and targeting form the theoretical underpinning for the analysis of customer survey data. Furthermore, the creator's own insight with marketing and B2B and B2C sales tasks was advantageous. Since the beginning of this research, several algorithms for data science methods have been investigated. This study's research methods are quantitative analytics methods from the discipline of data science. Predicting target members necessitated the use of supervised learning algorithms such as an ensemble approach and machine learning.

Table of Contents

List of Tables

STUDENT NAME	ADMISSION NUMBER
ABHIMANYU KUMAR	18SCSE1180023
DEVESH KUMAR SINGH	18SCSE1180022

GUIDE NAME	**Dr. Sudeept Singh Yadav**
REVIEWER NAME	

List of Figures

Acronyms

SVM	Support Vector Machine
ML	Machine Learning
NN	Nearest Neighbor
SVC	Support Vector Machine Classifier
SOM	Sorting Map
CRM	Customer Relation Management

CHAPTER-1

Introduction

1.1 Introduction

Businesses ' environments and industries are becoming rapidly reliant on data collection and analysis. As a result, several organizations made decisions based on what they learned from the analysis. Because of this, there emerged a new tendency in a lexicon termed "data-driven." This paper was started as a way to reflect on and track the business megatrend of information-driven navigation. Furthermore, the case dataset was selected for data analysis in support of a business research project. To secure the achievement of a new product in a market, its entails' moving extraordinary thoughts onto things' and 'identifying buyers.' This second point is emphasized in this thesis through segmentation and targeting. If a company identifies a certain clientele, it may provide more specialized products or services. Furthermore, this can help the organization use marketing assets more effectively, resulting in higher profitability and customer satisfaction.

The survey dataset utilized in this thesis is made up of customers' replies to questions on their social, environmental, and moral obligations. This might be a wonderful hotspot for figuring out expected clients for an organization that values corporate responsibility. When making speeches, certain assumptions are formed, according to this thesis. This was unavoidable since the author designed the problems to be answered in order for them to be applicable to real-world circumstances. This study becomes more focused on specific problems as a result of these assumptions.

The significance of customer segmentation includes, but is not limited to, a business's capacity to tailor market strategy to each section of its customers; Assistance with hazardous business decisions, such as credit agreements with customers; Determine items associated with certain components and methods for managing demand and supply electricity; The interdependence and interaction of consumers, goods, and customers are highlighted

which the association might be uninformed about; the capacity to assess client decreases and figure out which clients are probably going to have issues; The capacity to create more market research questions and provide insights for solution discovery. This type of learning is known as supervised learning. There are several integration algorithms available, including the K-means algorithm, the K-nearest algorithm, and the Sorting Map (SOM). These algorithms may identify groups in data without prior knowledge of the data by continually looking at input patterns if a static aptitude for the subject matter or process is achieved in training samples. Each set comprises data points that are very similar while also being notably distinct from the data points in the other groups.

1.2 Formulation of Problem

Customer Segmentation is an increasingly significant issue in today's competitive commercial area. The zeitgeist of modern era is innovation, where everyone is embroiled into competition to be better than others.

Today's business run on the basis of such innovation having ability to enthral 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.

1.3 Tool and Technology Used

- **Python (3.7.4 used)**

  Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991. Python has a dynamic type system and memory management that is automated.

  It contains a wide standard library and supports several programming paradigms, including object-oriented, imperative, functional, and procedural. For a wide range of operating systems, Python interpreters are available.

  C Python, like nearly all of Python's other implementations, is open-source software with a community-based development strategy. The Python Software Foundation, a non-profit organization, oversees Python and C Python.

- **R programming Language**

  R is a programming language for statistical computing and graphics supported by the R Core Team and the R Foundation for Statistical Computing. Created by statisticians Ross Ihaka and Robert Gentleman, R is used among data miners and statisticians for data analysis and developing statistical software.

- **Anaconda**

  Anaconda is a Python and R programming language distribution for scientific computing that promises to make package management and deployment easier. Data-science packages for Windows, Linux, and macOS are included in the release.

- **IDE (Jupyter used)**

  IDE stands for Integrated Development Environment. It's a coding tool that allows you to write, test, and debug your code more efficiently.

  Jupyter Notebook was born out of IPython in 2014. It is a web application based on the server-client structure, and it allows you to create and manipulate notebook documents - or just "notebooks".

  > *"Jupyter Notebook should be an integral part of any Python data scientist's toolbox. It's great for prototyping and sharing notebooks with visualizations."*

PYTHON LIBRARIES:

- **Numpy**

  NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object and tools for working with these arrays.

  It is the fundamental package for scientific computing with Python. It contains various features including these important ones:

  a. A powerful N-dimensional array object
  b. Sophisticated (broadcasting) functions
  c. Tools for integrating C/C++ and Fortran code
  d. Useful linear algebra, Fourier transform, and random number capabilities

- **Matplotlib**

  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.

One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram, etc.

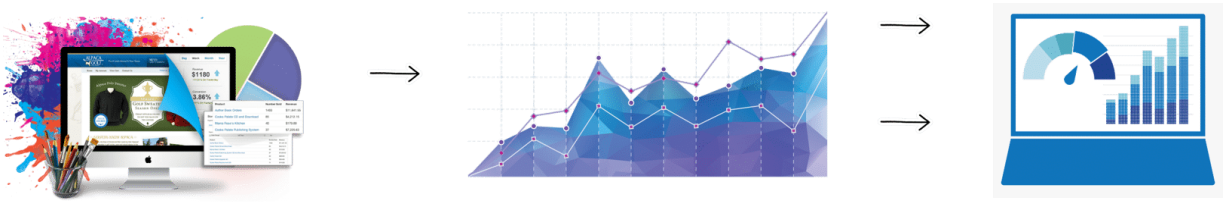
- **Pandas**

  pandas is a software library written in Python and is the basis for data manipulation and analysis in the language. Its name comes from "panel data," an econometrics term for datasets that include observations over multiple time periods for the same individuals. pandas offer a collection of high-performance, easy-to-use, and intuitive data structures and analysis tools that are of great use to marketing analysts and data scientists alike. It has the following two primary object types: DataFrame: This is the fundamental tabular relationship object that stores data in rows and columns (like a spreadsheet). To perform data analysis, functions and operations can be directly applied to DataFrames. Series: This refers to a single column of the DataFrame. The value can be accessed through its index. As Series automatically infers a type, it automatically makes all DataFrames well-structured.

- **Seaborn**

  Seaborn is one of the world's most regarded Python libraries that is purpose-built to create beautiful-looking visualizations. It can be considered as an extension of another library called Matplotlib as it is built on top of that.

  Data visualization is easily performed in Seaborn, and this is how the workflow looks like:



Data from various sources: The data that is needed to perform visualizations and analytics can come into the architecture from a variety of sources, such as a local storage unit, server, cloud structure, etc.

Data visualization: This is where the data is transformed from its number-state into an aesthetically pleasing visual counterpart. Seaborn plays the main role here.

Data Analytics: The result of data visualization is to take a look at the data in a way you have not done before. Analysis helps doing just this to reveal insights and trends that could not have been spotted otherwise.

- **Sklearn**

  Scikit-learn is a Python-based machine learning library that is available for free. It includes support vector machines, random forests, gradient boosting, k-means, and DBSCAN, among other classification, regression, and clustering techniques, and is designed to work with the Python numerical and scientific libraries NumPy and SciPy.

R LIBRARIES:

- **Plotly**

  Plotly's R graphing library makes interactive, publication-quality graphs. Examples of how to make line plots, scatter plots, area charts, bar charts, error bars, box plots, histograms, heatmaps, subplots, multiple-axes, and 3D (WebGL based) charts.

- **Ggplot2**

  ggplot2 package in R Programming Language also termed as Grammar of Graphics is a free, open-source, and easy-to-use visualization package widely used in R. It is the most powerful visualization package written by Hadley Wickham.

- **Plotrix**

  The plotrix R package contains tools for the plotting of data in R. A large number of specialized plots and accessory functions like color scaling, text placement and legends.

- **Purrr**

  Purrr enhances R's functional programming (FP) toolkit by providing a complete and consistent set of tools for working with functions and vectors. If you've never heard of FP before, the best place to start is the family of map() functions which allow you to replace many for loops with code that is both more succinct and easier to read. The best place to learn about the map() functions is the iteration chapter in R for data science.

- **Cluster**

  Clustering in R is an unsupervised learning technique in which the data set is partitioned into several groups called as clusters based on their similarity. Several clusters of data are produced after the segmentation of data.

- **gridExtra**

  gridExtra is a very useful package with two functions for showing multiple ggplot2 plots: arrangeGrob and grid.arrange. However, using these functions inside a package has proven to be difficult because of the way gridExtra handles

namespaces.

- **Grid**

  Grid package in R Programming Language has been removed from CRAN repository as it is now available as the base package. This package is a basis for all other higher graphical functions used in other packages such as lattice, ggplot2, etc. Also, it can manipulate the lattice outputs. Being a base package, there is no requirement to install it. It is automatically installed when R is installed.

CHAPTER-2 LITERATURE REVIEW

A. Customer Segmentation in the business world has been increasingly serious throughout time, as enterprises have been compelled to address the issues and wants of their clients, draw in new clients, thus grow their activities. In a business, it is difficult to recognize and meet the needs and expectations of each customer. This is because clients' needs, desires, demographics, size, taste, and other factors vary. Treating all customers equally in business is a terrible strategy as it is. This problem has given rise to the concept of customer segmentation or market segmentation, which separates customers into subcategories or segments, with members of each subcategory displaying similar market behaviors or qualities. Customer segmentation, on the other hand, refers to the practice of partitioning a market into native gatherings.

B. Big data is defined as a huge volume of structured and unstructured data that can't be handled utilizing standard techniques and algorithms. Businesses keep billions of records on their clients, providers, and activities, and a huge number of inside associated sensors give data to this present reality through devices like mobile phones and cars, detecting, assembling, and communications data. Capability to anticipate better, save money, further develop products, and work on a few components of life such as traffic management, climate gauging, debacle anticipation, finance, extortion control, deals, public safety, schooling, and medical care. Volume, variability, and velocity are the three Vs of big data. Additional 2Vs are available, increasing the total to 5V (authenticity and cost).

C. According to Sulekha Goyat, the Repository of data assortment is the act of gathering and analyzing data on changes to a setup framework to answer pertinent questions and assess the outcomes. Data collection is a necessary component of examination in many disciplines, including the physical and social sciences, the humanities, and business. All data collection is conducted with the goal of obtaining top-notch proof that will enable the examination to develop concrete and deceptive replies to the questions posed.

D. Clustering of data the technique of categorizing data in a dataset considering shared qualities is known as clustering. A variety of algorithms may be employed to analyze datasets based on the criterion provided. However, because there is no universal clustering algorithm, selecting the appropriate clustering strategies is crucial. In this study, we used the Python scalar library to create three clustering algorithms.

E. K-Means that a classification algorithm is one of the most widely used classification algorithms. The K-algorithm is used to allocate each data point to one of the pre-sorted overlapping groups in this clustering algorithm, which is based on Centro. Clusters are produced that compare to underlying patterns in the data, providing vital knowledge to aid in the decision-making process. There are numerous ways to construct K-means; in this case, we will use the elbow strategy.
The Basic Machine Learning structure is:

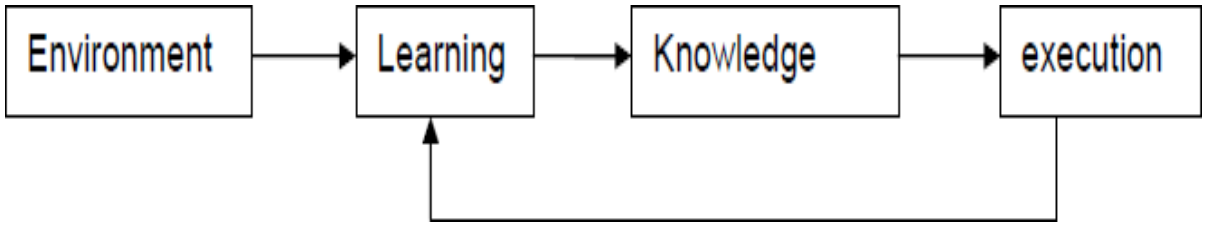


Fig. 1: ML Structure

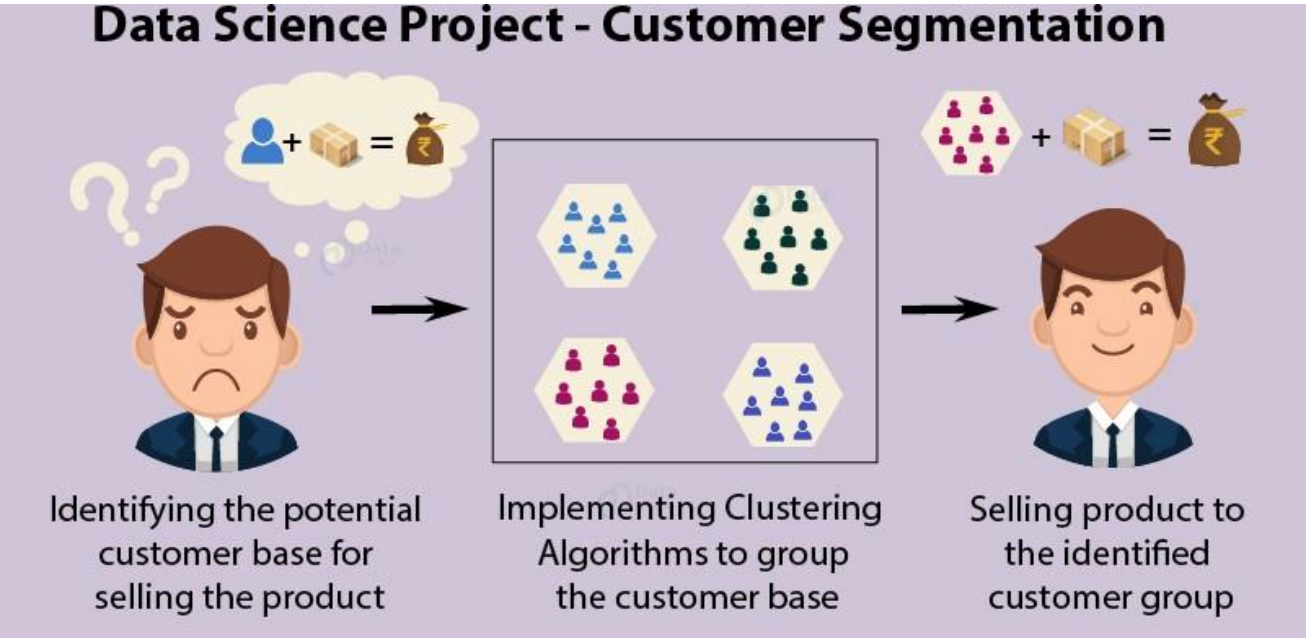


Fig. 2: Basic Customer Segmentation Model

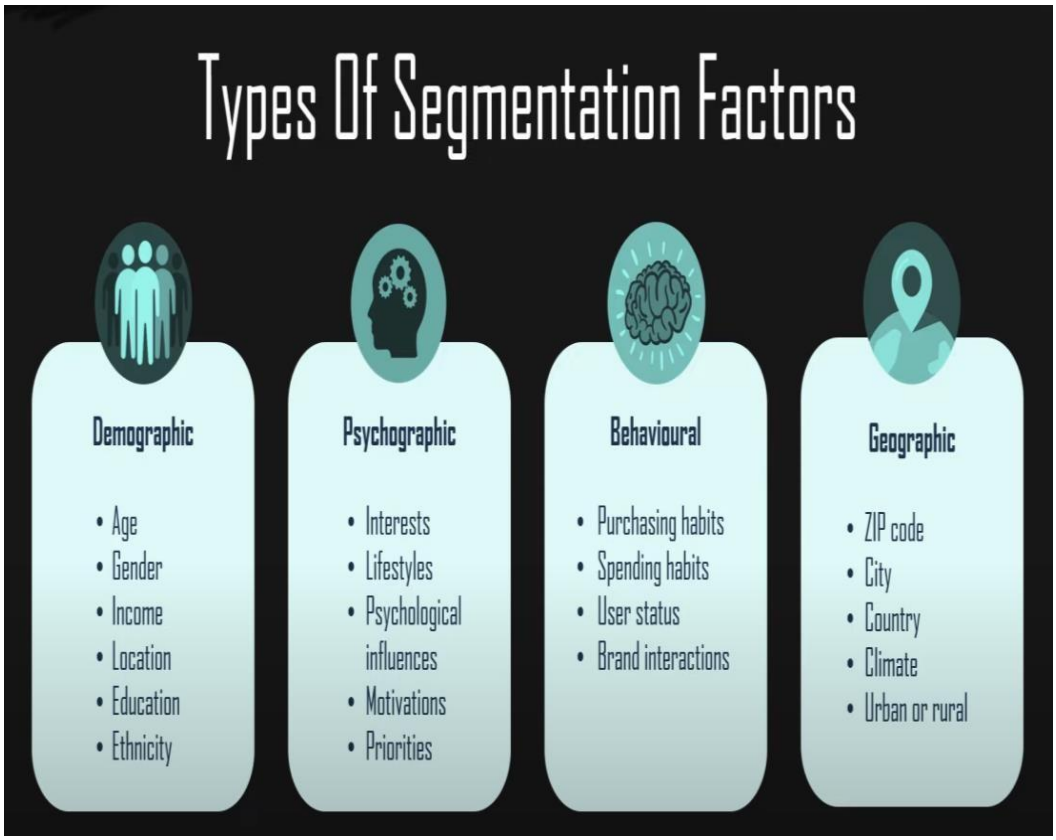


Fig. 3: Types of Customer

CHAPTER-3 FUNCTIONALITY/WORKING OF PROJECT

Machine Learning (ML) is a set of data-handling frameworks that are created and used to replicate the human cerebrum. The fundamental purpose of machine learning analysis is to build a computing device that can mimic the brain and do various computing tasks at a faster rate than earlier systems. Machine learning algorithms execute a wide range of tasks, including pattern matching and classification, optimization, and data clustering. These tasks are extremely difficult for conventional PCs, which are more efficient at algorithmically logging responsibilities and juggling precise numbers. ML assembles a huge number of highly interconnected processing components known as hubs, units, or neurons, which usually function in parallel and are organized in standard models. An association interface connects one neuron to the next. Weights convey information about the linked info flag and are connected with each association interface. This information is required by the neuron net in order to handle a specific issue. Machine learning collective behavior is defined by their ability to learn, review, and summarize by preparing examples or information, much like the human mind.

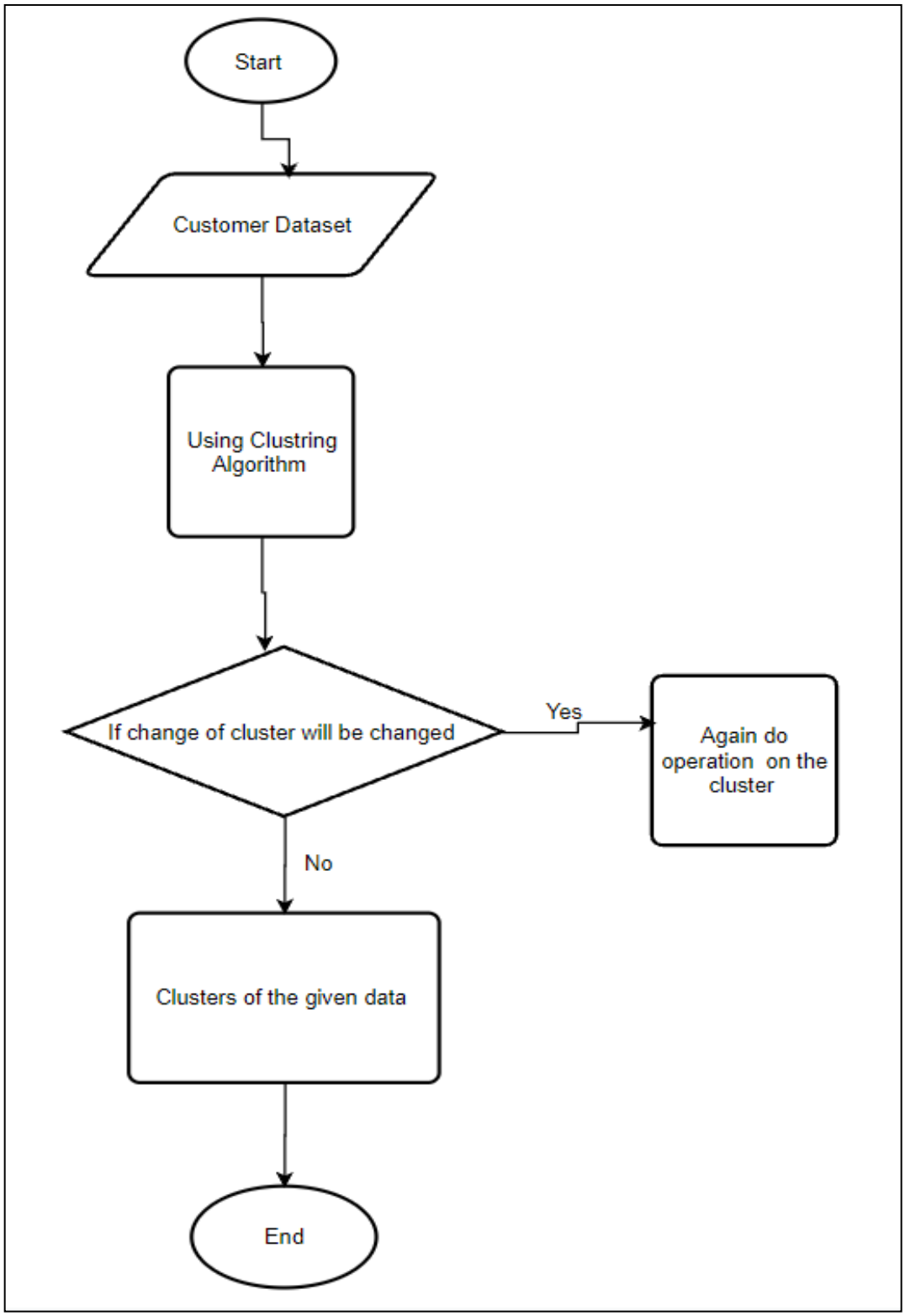


Fig. 4: Flow Chart of Proposed Methodology

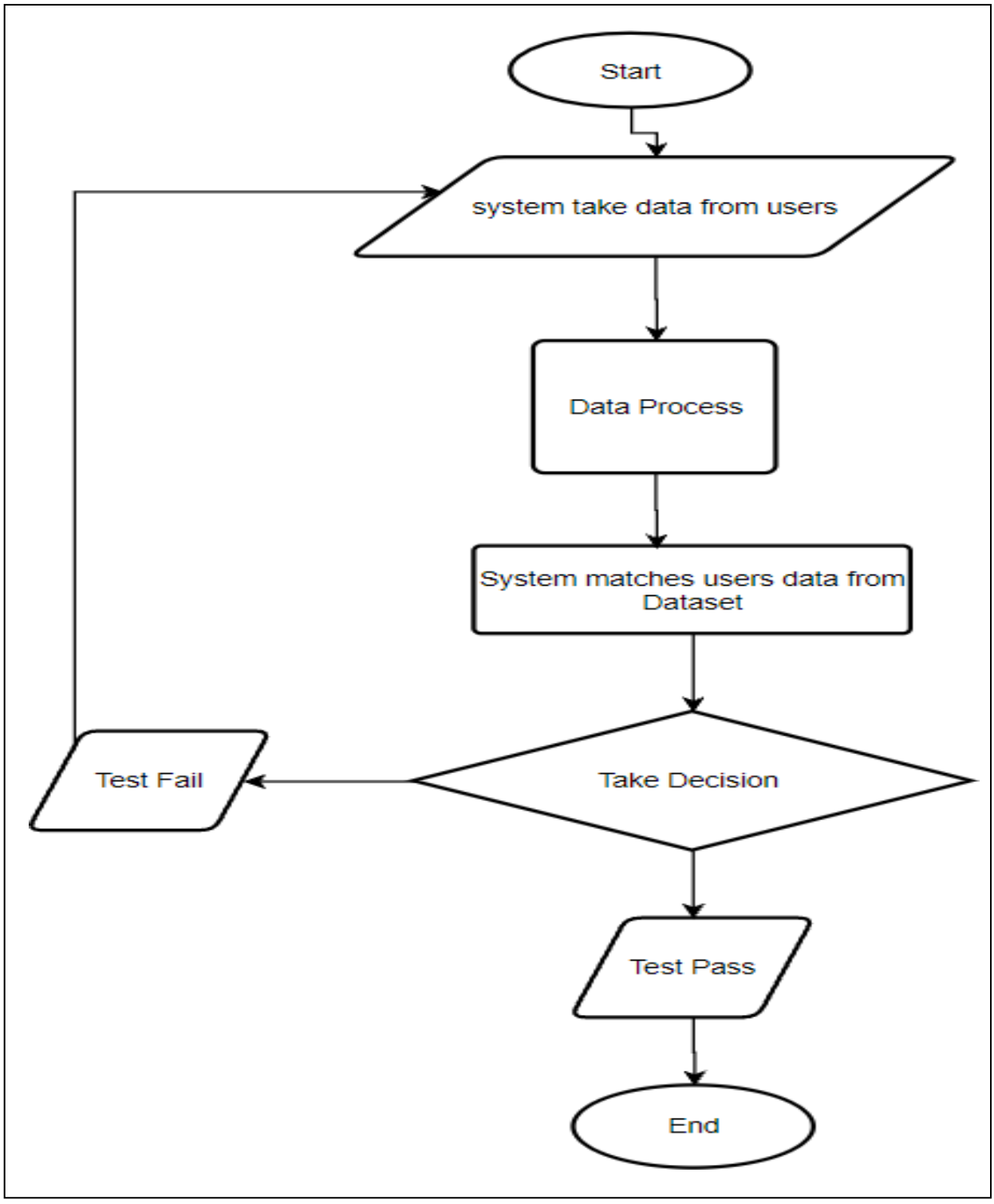


Fig. 5: Data flow of Proposed Methodology

Learning

The main property of an ML is its capability to learn. Learning or preparing is a procedure by methods for which a neural system adjusts to a boost by making legitimate parameter modifications, bringing about the generation of wanted reaction. Learning in an ML is chiefly ordered into two classes as.

- Supervised learning
- Unsupervised learning

Supervised Learning

Supervised learning is a type of machine learning method in which we provide sample labeled data to the machine learning system in order to train it, and on that basis, it predicts the output. The system creates a model using labeled data to understand the datasets and learn about each data, once the training and processing are done then we test the model by providing a sample data to check whether it is predicting the exact output or not.

Supervised learning is a learning in which we teach or train the machine using data which is well labelled that means some data is already tagged with correct answer. After that, machine is provided with new set of examples so that supervised learning algorithm analyses the training data and produces a correct outcome from labelled data.

Basically, they can apply what has been learned in the past to new data using labelled examples to predict future events. Starting from the

analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.

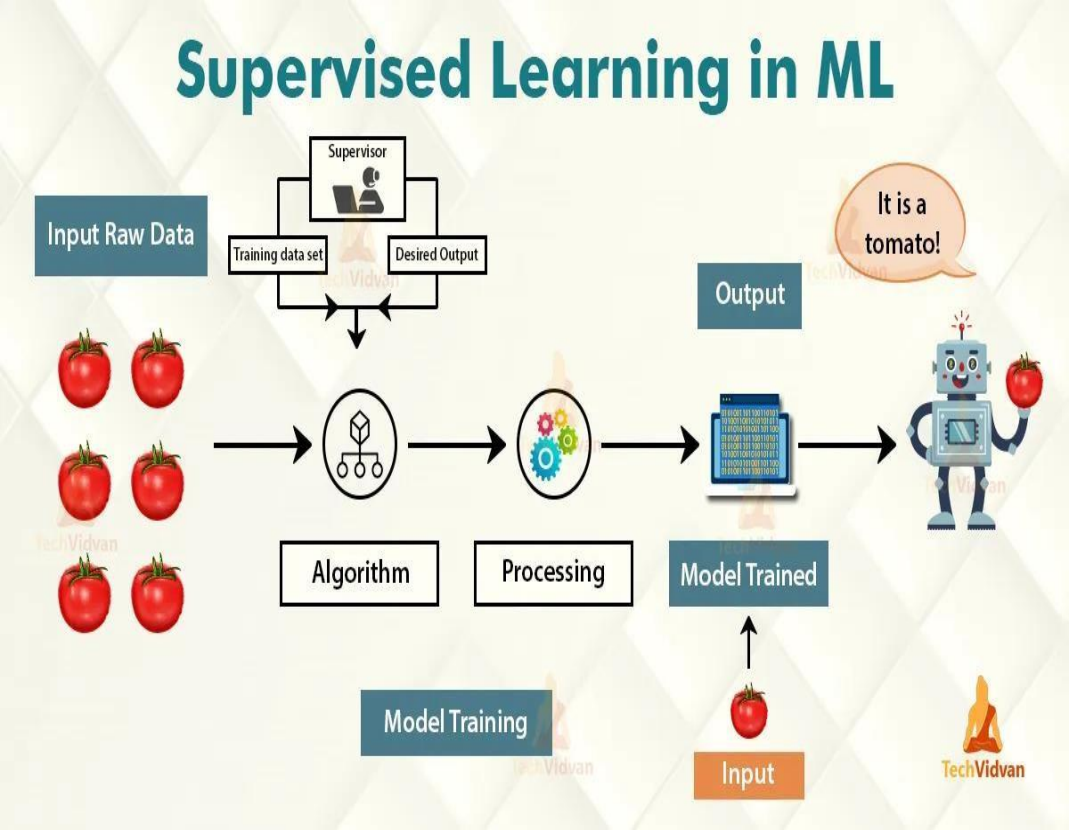


Fig. 6: Supervised Learning

Unsupervised learning

Unsupervised learning is a learning method in which a machine learns without any Supervision. The training is provided to the machine with the set of data that has notbeen labeled, classified, or categorized, and the algorithm needs to act on that data without any supervision. The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns.

Data mining is a highly efficient area that makes use of techniques like association rule learning. In the process of creating massive databases, data mining employs interesting machine-learning algorithms such as inductive rule learning and decision tree construction. Data mining techniques are applied in large, interesting organizations and vast data investigations. To extract meaningful information from continuous data streams, many data mining techniques include classification-based algorithms.

Unsupervised learning is also known as undirected data mining. In this all variables (explanatory and dependent) are treated in the similar way with no distinction between them. However, there are still some targets to be achieved. The targets can be general or specific based on data reduction or clustering respectively. It can be used to categorise or explain some particular target field.

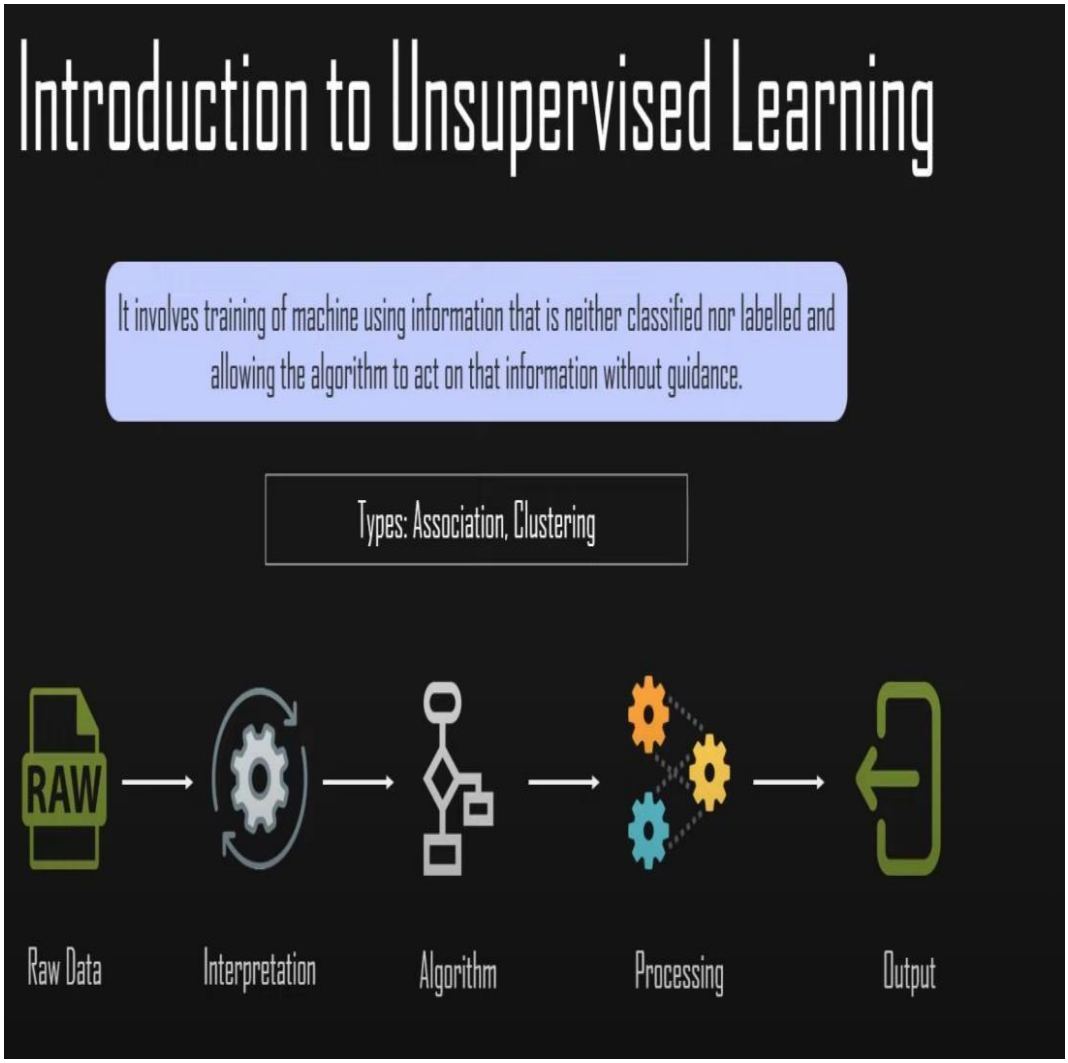


Fig. 7: Unsupervised Learning

Nearest Neighbors Algorithm

The Nearest Neighbor (NN) rule discriminates the classification of an obscure data item based on the class of its nearest neighbor. The nearest neighbor is determined using an estimate of k, which indicates the number of nearest neighbors used to describe the data point class. It employs multiple nearest neighbour to decide the class to which the provided data point belongs, which is referred to as KNN. Memory-based techniques need the data samples to be stored in memory at run time.

The training points are allocated weights based on their distances from the

sample data point. However, the computational complexity and memory requirements remained a key issue. For addressing the memory utilization problem, the size of data gets minimized. The repeated patterns without additional data are removed from the training data set.

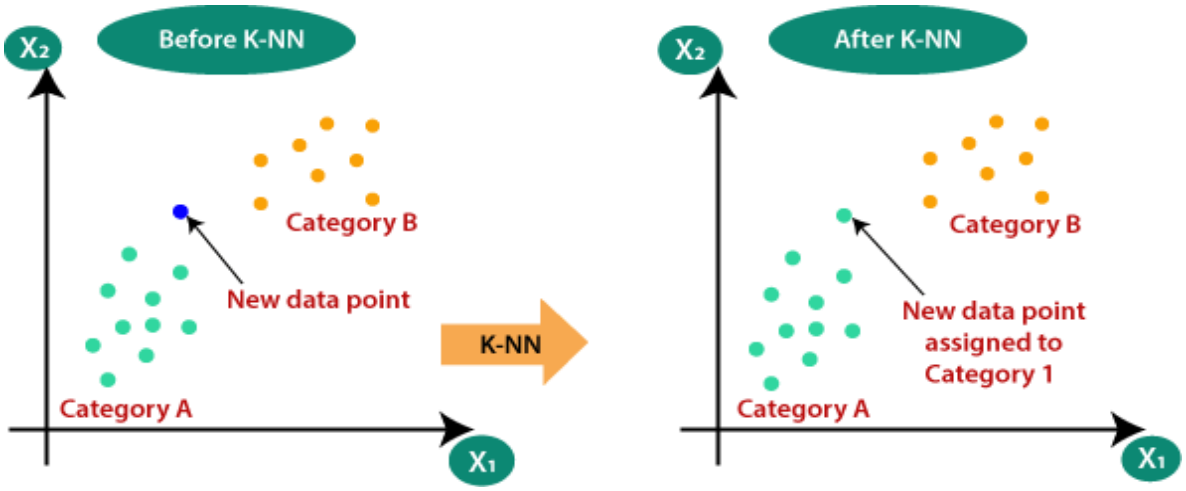


Fig. 8: K-NN Algorithm

Naive Bayes Classifier

Based on the Bayesian theorem, the Naive Bayes Classifier technique is used. When the input has a high degree of dimension, the designed technique is applied. The Bayesian classifier is used to determine the probable outputs given an input. It is feasible to add more raw data during runtime. When the class variable is known, a Naive Bayes classifier indicates the presence (or lack) of a class feature (attribute) that is independent of the presence (or absence) of some other feature. Naïve Bayesian Classification Algorithm was introduced by Shinde S.B and Amrit Priyadarshi (2015) that denotes statistical method and supervised learning

method for classification. A naive Bayesian Algorithm is used to predict heart disease. A raw hospital dataset is employed. After that, the data gets preprocessed and transformed. Finally, by using the designed data mining algorithm, heart disease was predicted, and accuracy was computed.

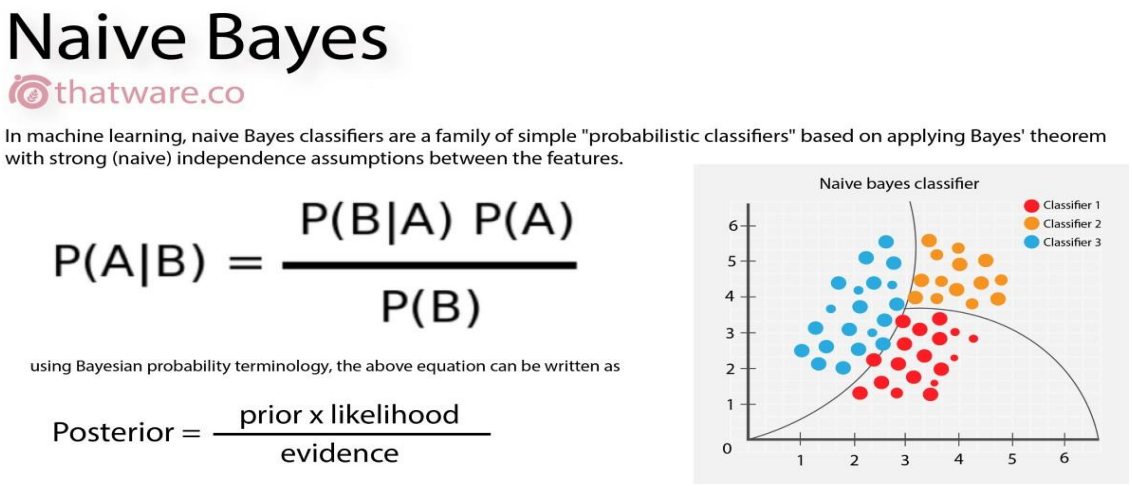


Fig. 9: Naive Bayes Classifier

Support Vector Machine

SVM is used in many applications like medical, military for classification purposes. SVM is employed for classification, regression, or ranking functions. SVM is based on statistical learning theory and the structural risk minimization concept. SVM estimates the position of decision boundaries known as hyperplanes for optimum class separation, as shown in Figure 3 Margin maximization is used to minimize the upper bound on projected generalization error by increasing the distance between the separating hyperplane and the instances on either side. The dimension of the recognized objects has little effect on the SVM's classification accuracy. SVM uses convex quadratic programming to analyze data. Quadratic programming methodologies are expensive because they necessitate large matrix

operations and extensive numerical computations.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:

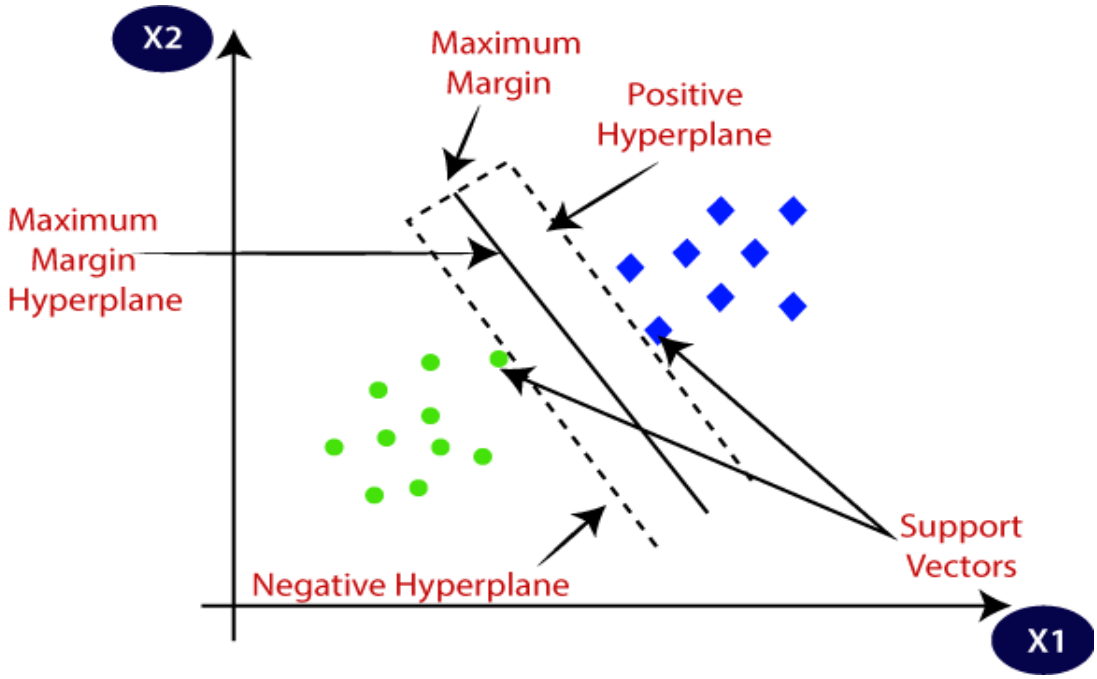


Fig. 10: SVM

EXCUTION OF PROJECT USING R PROGRAMMING

First, we will do data exploration as the initial phase in our data science project. We'll import the necessary packages for this role before reading our data. Finally, we'll look over the input data to get the information we want.

Dataset link: https://drive.google.com/file/d/19BOhwz52NUY3dg8XErVYglctpr5sjTy4/view

```
customer_data = read.csv("Mall_Customers.csv")
str(customer_data)
```

```
'data.frame':   200 obs. of  5 variables:
 $ CustomerID            : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Gender                : Factor w/ 2 levels "Female","Male": 2 2 1 1 1 1 1
1 2 1 ...
 $ Age                   : int  19 21 20 23 31 22 35 23 64 30 ...
 $ Annual.Income..k..    : int  15 15 16 16 17 17 18 18 19 19 ...
 $ Spending.Score..1.100.: int  39 81 6 77 40 76 6 94 3 72 ...
```

Customer Gender Visualization

We'll make a barplot and a piechart to demonstrate the gender distribution in our customer data dataset in this section.

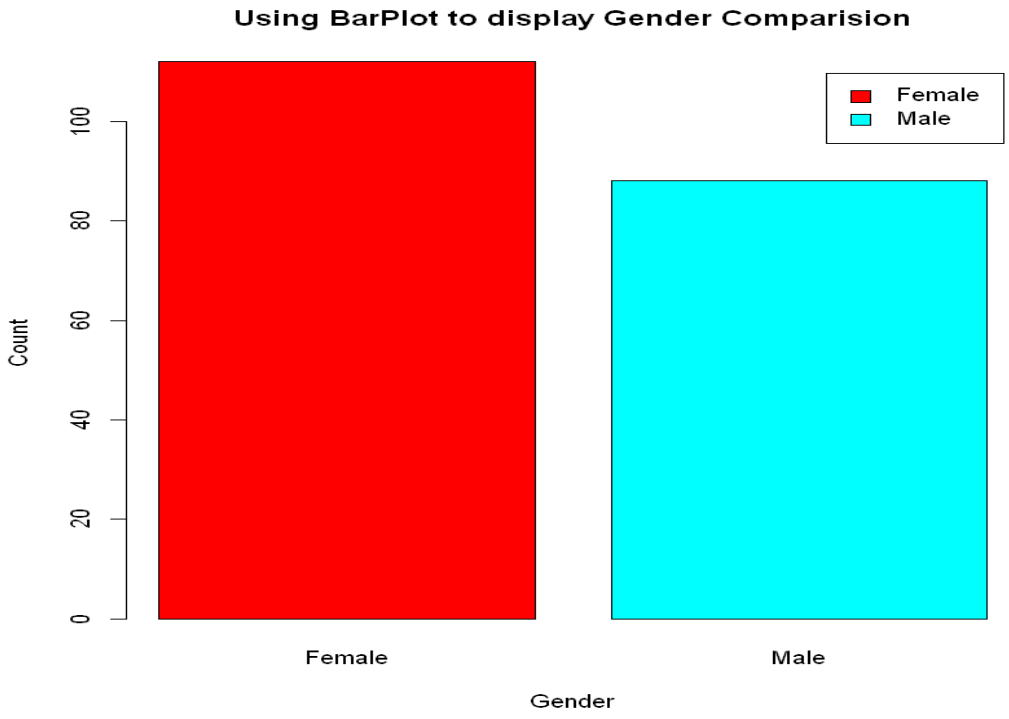


Fig. 11: Barplot to display Gender Comparision

We can see from the barplot above that the amount of women is greater than the number of males. Let's create a pie chart to display the male-to-female distribution ratio

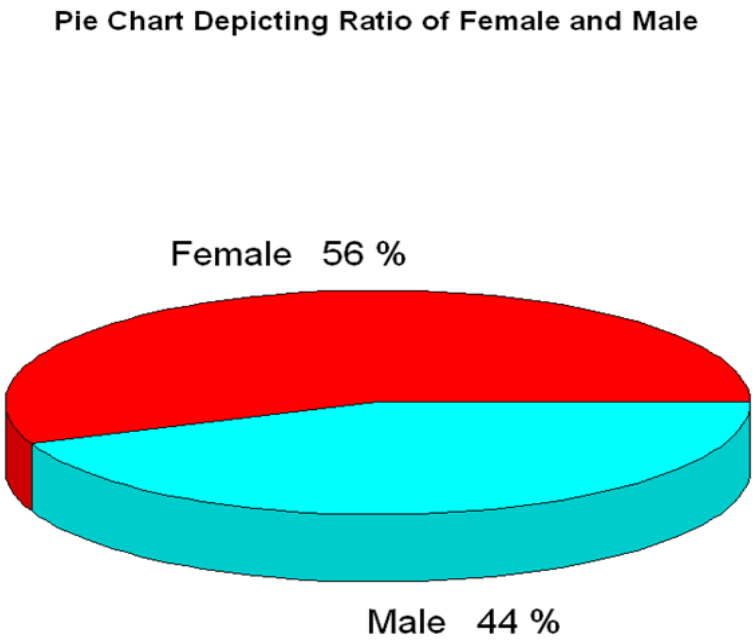


Fig. 12: Pie Chart Ratio of Female and Male

Visualization of Age Distribution

Allow us to plot a histogram to see the circulation to plot the recurrence of customers ages. We will initially continue by taking a rundown of the Age variable.

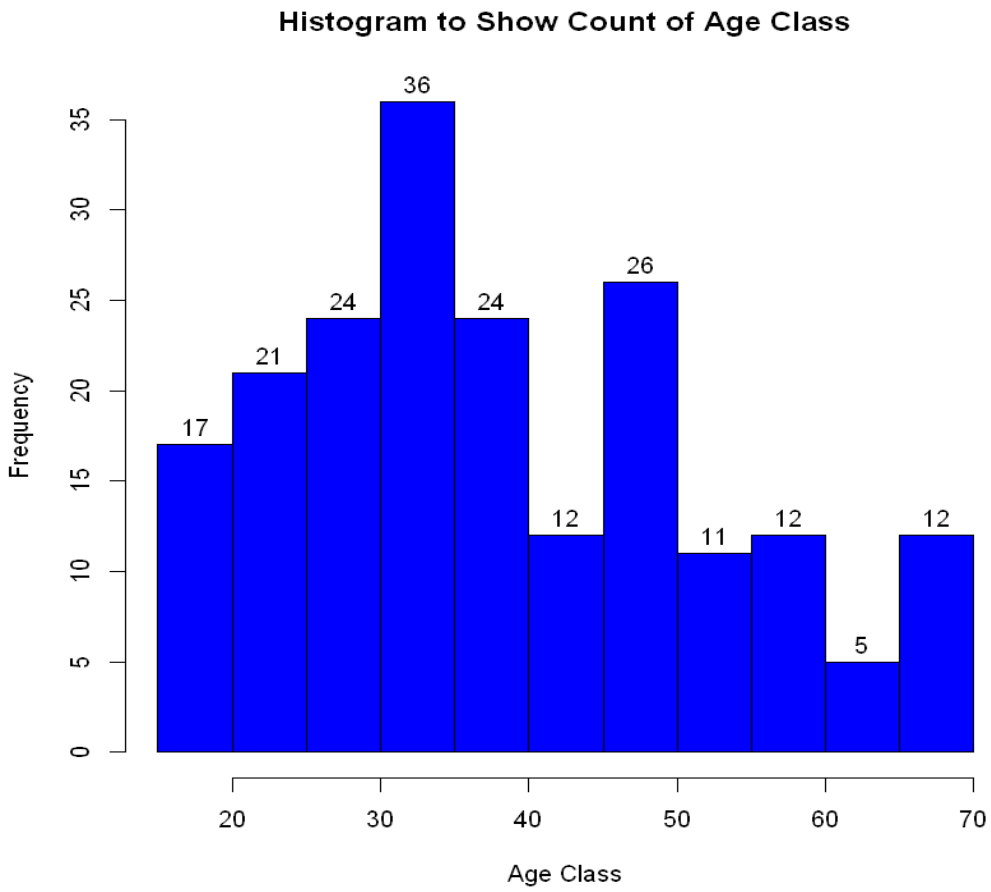


Fig. 13: Histogram to Show Count Age

Visualization of Age Distribution Allow us to plot a Boxplot

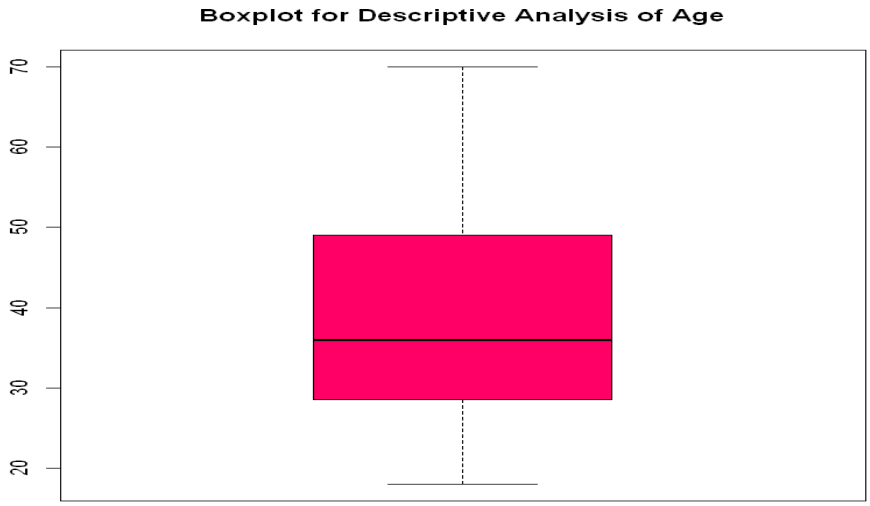


Fig. 14: Boxplot for Analysis of Age

From the over two Visualization, we infer that the greatest customers ages are somewhere in the range of 30 and 35. The lowest customers age is 18, while, the greatest age is 70.

Analysis of the Annual Income

We will make visualizations to examine the annual pay of the customers. We will plot a histogram and afterward, we will continue to analyze this information utilizing a thickness plot.

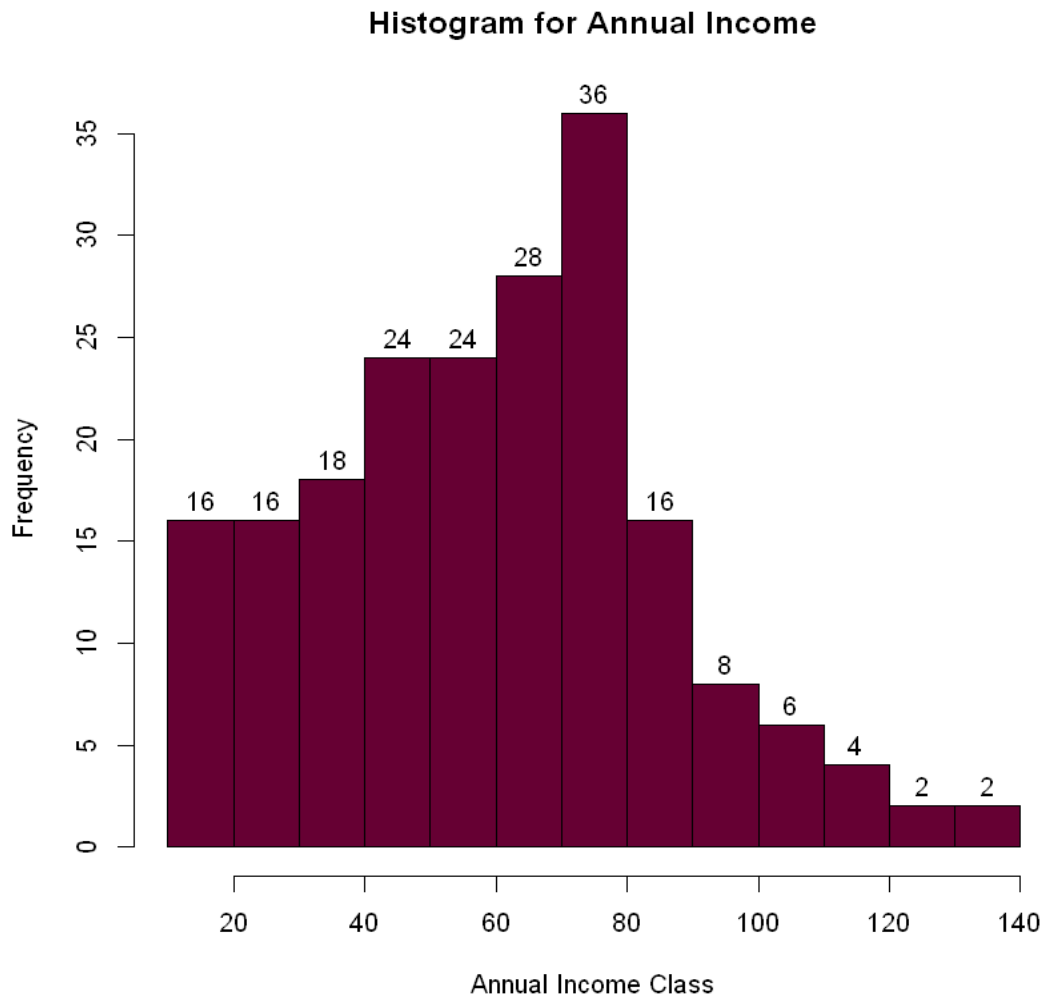


Fig. 15: Histogram for Annual Income

Now we see the Density plot of income and then we compare

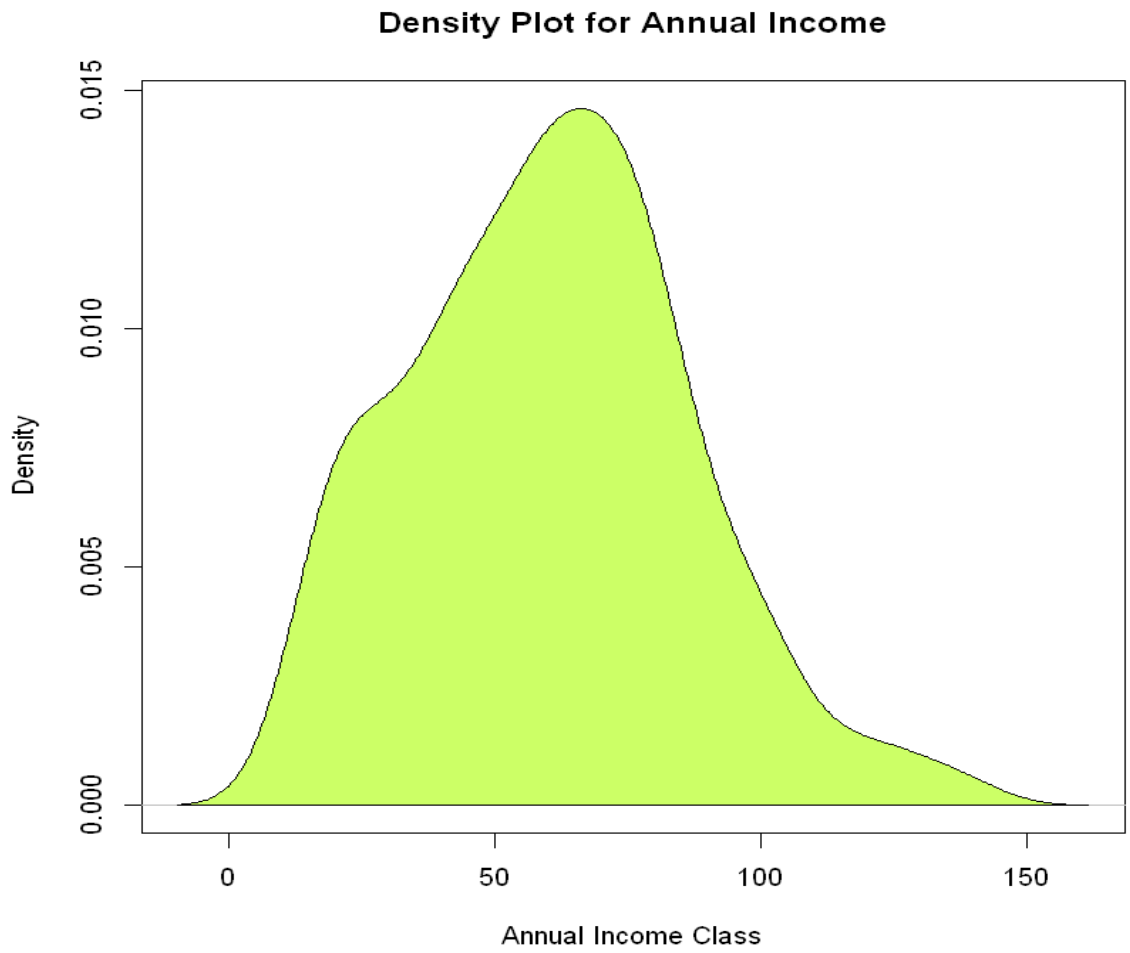


Fig. 16: Density plot for Annual Income

From the above descriptive analysis, we conclude that the minimum annual income of the customers is 15 and the maximum income is 137. People earning an average income of 70 have the highest frequency count in our histogram distribution. The average salary of all the customers is 60.56. In the Kernel Density Plot that we displayed above, we observe that the annual income has a normal distribution.

Analyzing Spending Score

With respect of plotting a graph using Boxplot

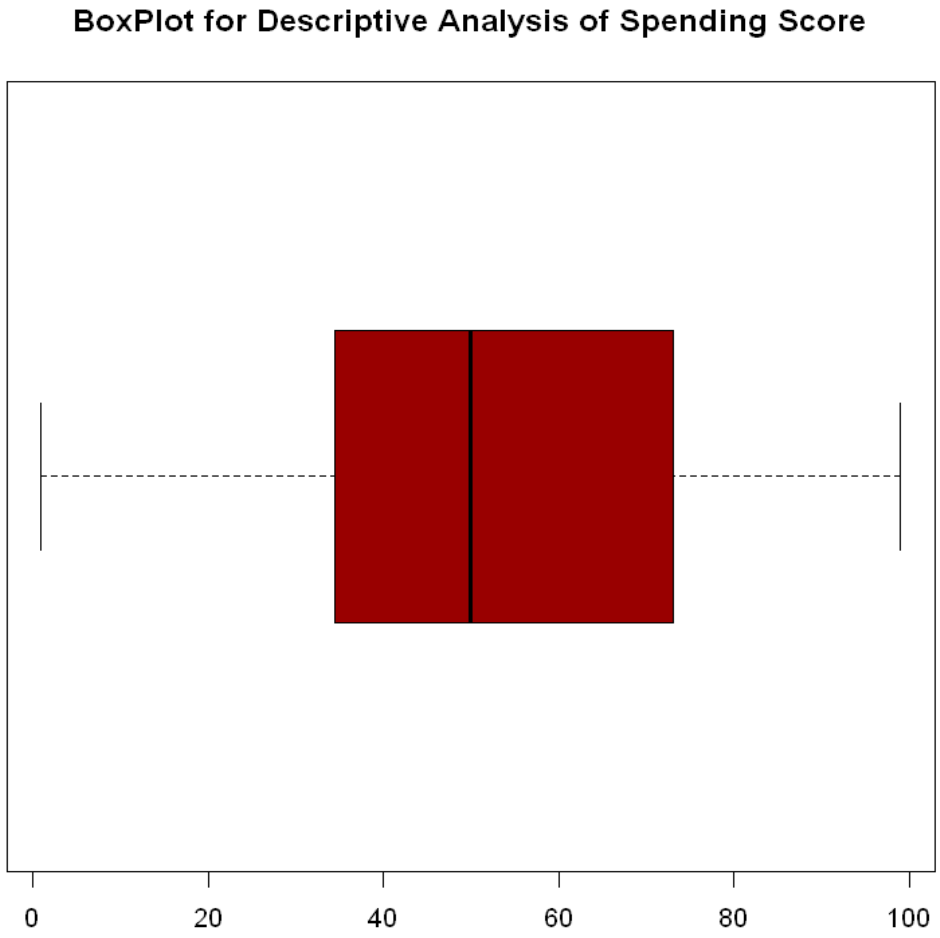


Fig. 17: Boxplot for Analysis of Spending Score

Now we are plotting a graph using Histogram and then compare the result

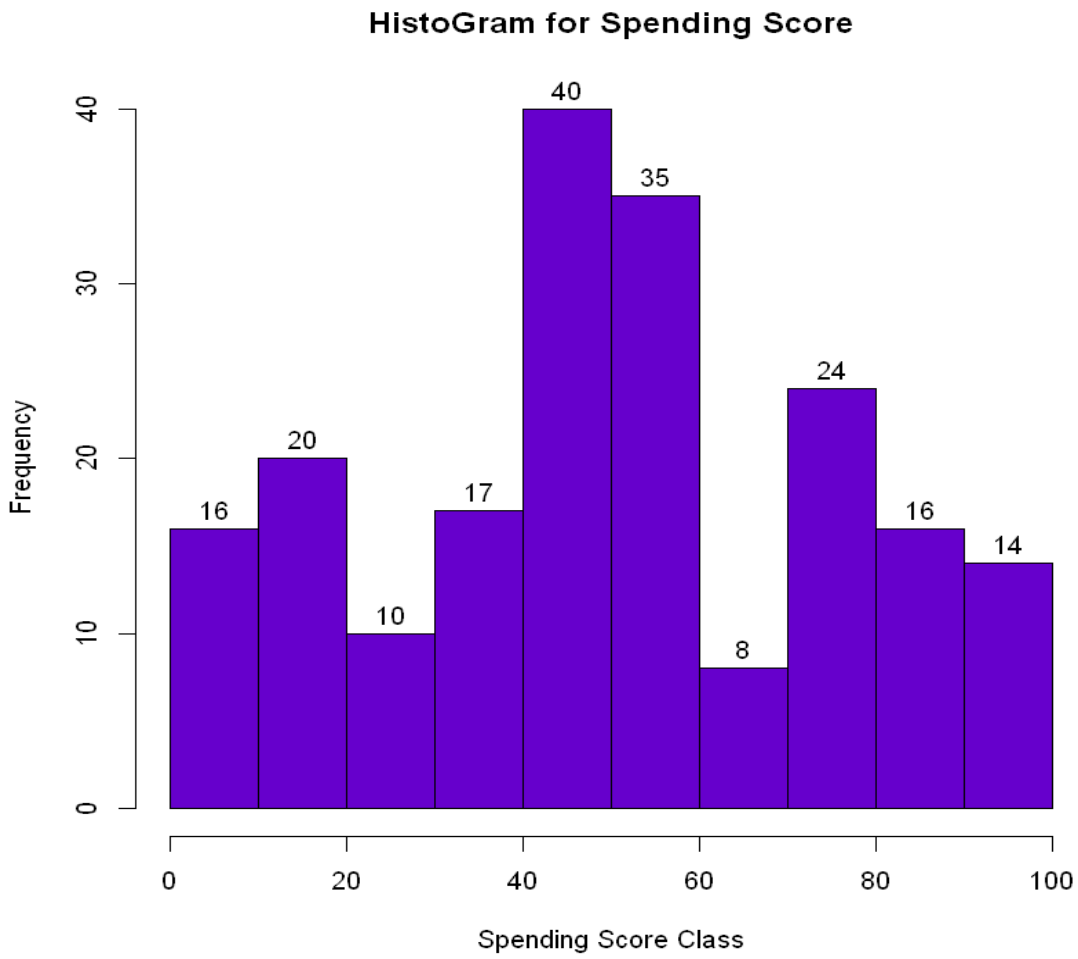


Fig. 18: Histogram for Spending Score

The minimum spending score is 1, maximum is 99 and the average is 50.20. We can see Descriptive Analysis of Spending Score is that Min is 1, Max is 99 and avg. is 50.20. From the histogram, we conclude that customers between class 40 and 50 have the highest spending score among all the classes.

K-means Algorithm

It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.

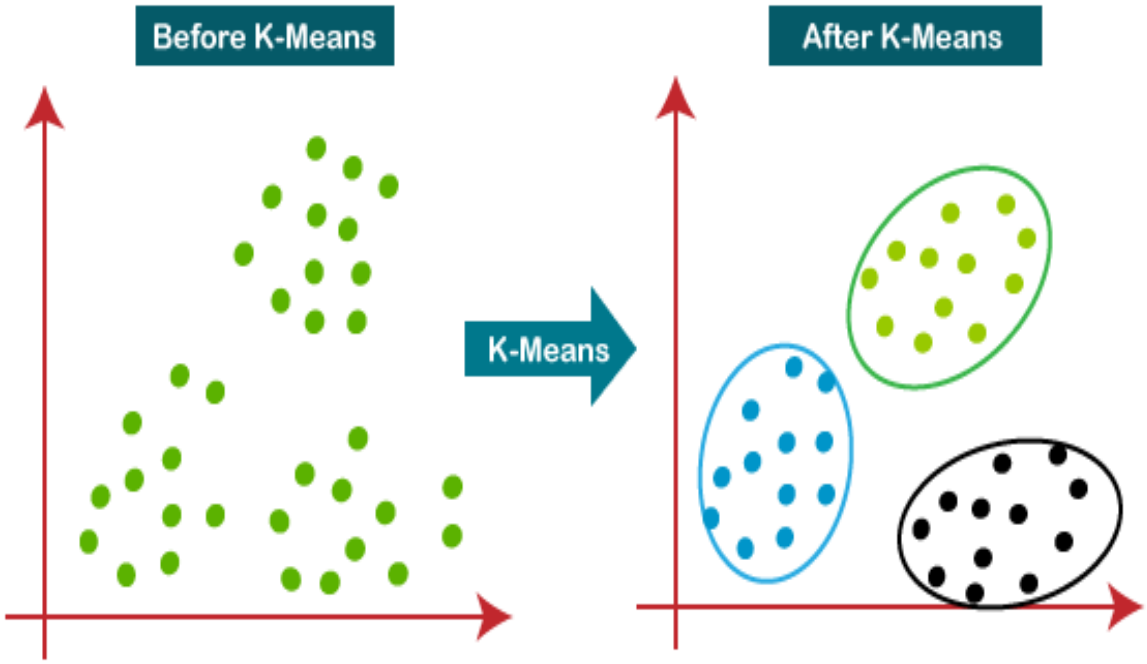


Fig. 19: K-Means

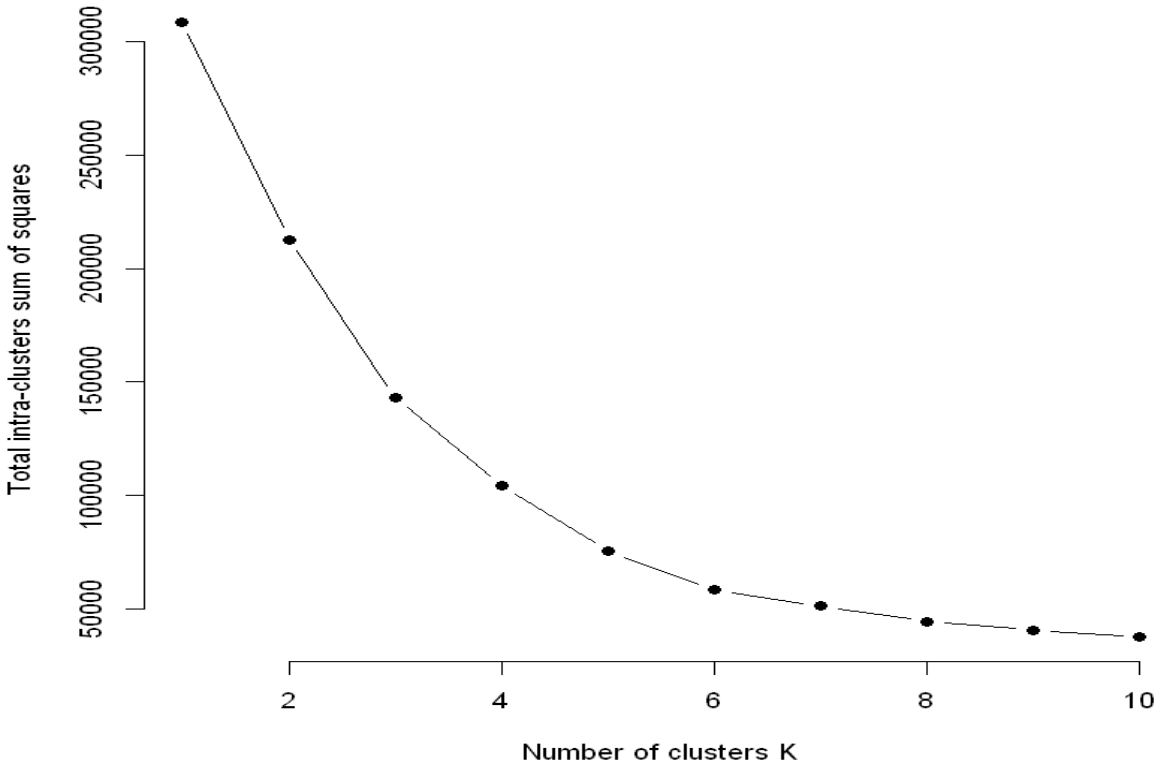


Fig. 20: Graph for K-Means Algorithm

From the above graph, we conclude that 4 is the appropriate number of clusters since it seems to be appearing at the bend in the elbow plot.

EXCUTION OF PROJECT USING PYTHON

Dataset link: https://drive.google.com/file/d/19keV8Y0SULl62WPjopeunOlxujV2Iby6/view?usp=shar

Data preparation

As an initial step, I load all the modules that will be utilized in this notebook.

1. Investigating the content of variables

This data frame contains 8 variables that relate to.

Invoice Number: Nominal, a 6-digit integral number uniquely allotted to each transaction. If this code starts with the letter 'c', it demonstrates a cancellation.

Stock Code: Product (item) code. Nominal, a 5-digit integral number uniquely allocated to each particular product.

Description: Product (item) name.

Quantity: The quantities of each product (item) per transaction. Numeric.

Invoice Date: Invoice Date and time. Numeric, the day and time at the point when each transaction was produced.

Unit Price: Numeric, Product price per unit in sterling.

Customer ID: Customer number. Nominal, a 5-digit integral number uniquely allocated to each customer.

Country: Country name. Nominal, the name of the country where each customer lives in.

2. Insight on Product Categories

The Stock Code variable is used to uniquely identify products in the data frame. The Description variable contains a brief description of the products. I aim to utilize the content of this later variable in this section to categorize the products.

2.1 Products Description

As an initial step, I remove critical data from the Description variable. This is accomplished by employing the following function:

- Extract the names (proper, common) showing up in the products description
- For each name, I extract the root of the word and total the arrangement of names associated with this root.
- Count the number of times each root appears in the data frame.
- When many words are listed for the same root, I consider the shortest name to be the keyword associated with that root (this method selects the singular when there are single/plural variations).

2.2 Defining product categories

The obtained list has over 1400 words, with the most commonly used occurring in over 200 different commodities. However, following closer inspection of the list's content, I note that a few of the names are nonsensical. Colors, for example, do not transmit information. As a result, these terms are excluded from the following analysis, and I also limit myself to terms that appear more than 13 times.

3. Customer categories

3.1 Formatting data

The past area categorizes the numerous components into five groups. A first step in introducing the rest of the study is to prepare this data into the data frame. Now establish the categorical variable categ product, which contains the cluster to which each product belongs.

4. Classification of customers

The objective of this section will be to fine-tune a classifier that will characterize customers into the various customer gatherings created in the past area. The goal is to have this categorization available on the primary visit. To accomplish this objective, I will put numerous scikit-learn classifiers to the test. To begin, I design a class that permits connecting some of the functions shared by these several classifiers in order to facilitate their use.

4.1 Support Vector Machine Classifier (SVC)

The SVC classifier is the first classifier I utilize. I use it by instantiating the Class Fit class and then doing a grid search (). I supply the following parameters to this method when I call it: The hyperparameters for which the best optimal value will be sought the number of cross-validation folds that should be used

```
svc.grid_predict(X_test, Y_test)
```

```
Precision: 80.75 %
```

Confusion matrix

A confusion matrix is a table that outlines different predictions and test results and contrasts them with real-world values. Confusion matrices are used in statistics, data mining, machine learning models and other artificial intelligence (AI) applications. A confusion matrix can also be called an error matrix.

Confusion matrices are used to make the in-depth analysis of statistical data faster and the results easier to read through clear data visualization. The tables can help analyze faults in statistics, data mining, forensics and medical tests.

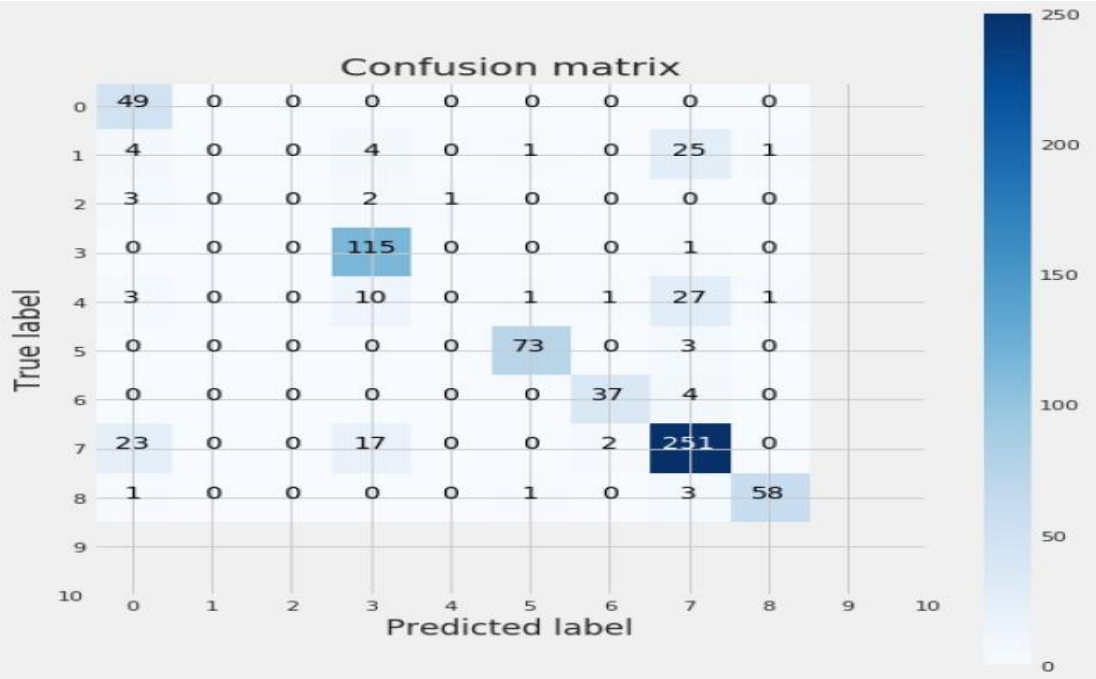


Fig. 21: Confusion matrix

Learning curve

A learning curve is a correlation between a learner's performance on a task and the number of attempts or time required to complete the task; this can be represented

as a direct proportion on a graph.

The learning curve theory proposes that a learner's efficiency in a task improves over time the more the learner performs the task.

Creating a learning curve is a typical strategy for assessing the quality of a match. This type of curve allows for the discovery of possible model defects such as over- or under-fitting. Furthermore, this indicates how much the mode would benefit from a larger data set.

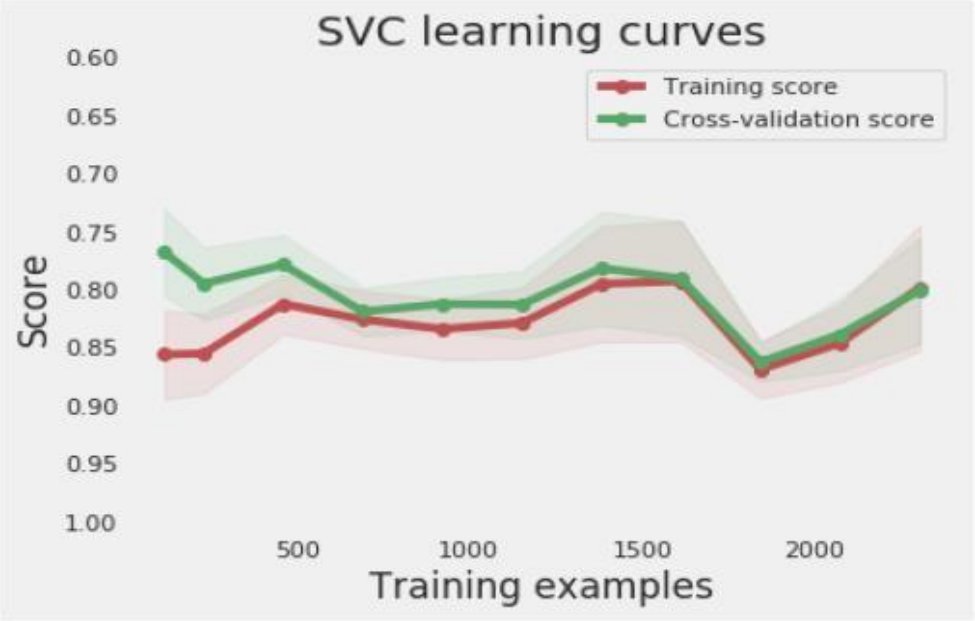


Fig. 22: SVC learning curve

Logistic Regression

It is a classification not a regression algorithm. It is used to estimate discrete values (Binary values like 0/1, yes/no, true/false) based on given set of independent variable(s). In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function. Hence, it is also known as **logit regression**. Since, it predicts the probability, its output values lies between 0 and 1 (as expected).

I'm going to look at the logistic regression classification technique now. As before, I instantiate the Class Fit class, use the training data to adjust the model, and compare the predicted and actual values:

```
lr = Class_Fit(clf = linear_model.LogisticRegression)
lr.grid_search(parameters = [{'C':np.logspace(-2,2,20)}], Kfold = 5)
lr.grid_fit(X = X_train, Y = Y_train)
lr.grid_predict(X_test, Y_test)
```

```
Precision: 86.29 %
```

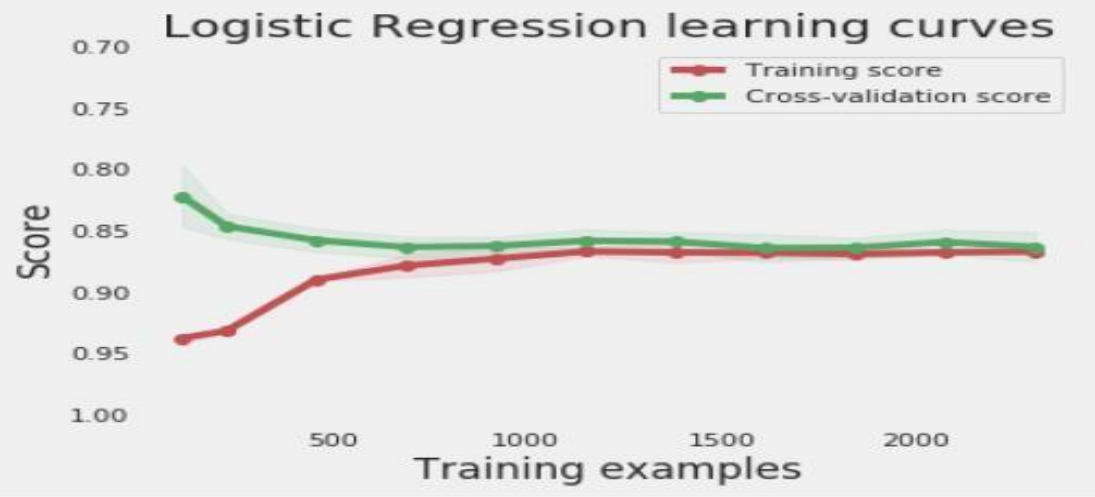


Fig. 23: Logistic Regression learning curves

k-Nearest Neighbors

It can be used for both classification and regression problems. However, it is more widely used in classification problems in the industry. K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases by a majority vote of its k neighbors. The case being assigned to the class is most common amongst its K nearest neighbors measured by a distance function.

These distance functions can be Euclidean, Manhattan, Minkowski and Hamming

distance. First three functions are used for continuous function and fourth one (Hamming) for categorical variables. If K = 1, then the case is simply assigned to the class of its nearest neighbor. At times, choosing K turns out to be a challenge while performing kNN modeling.

```
knn = Class_Fit(clf = neighbors.KNeighborsClassifier)
knn.grid_search(parameters = [{'n_neighbors': np.arange(1,50,1)}], Kfold = 5)
knn.grid_fit(X = X_train, Y = Y_train)
knn.grid_predict(X_test, Y_test)
```

```
Precision: 79.78 %
```

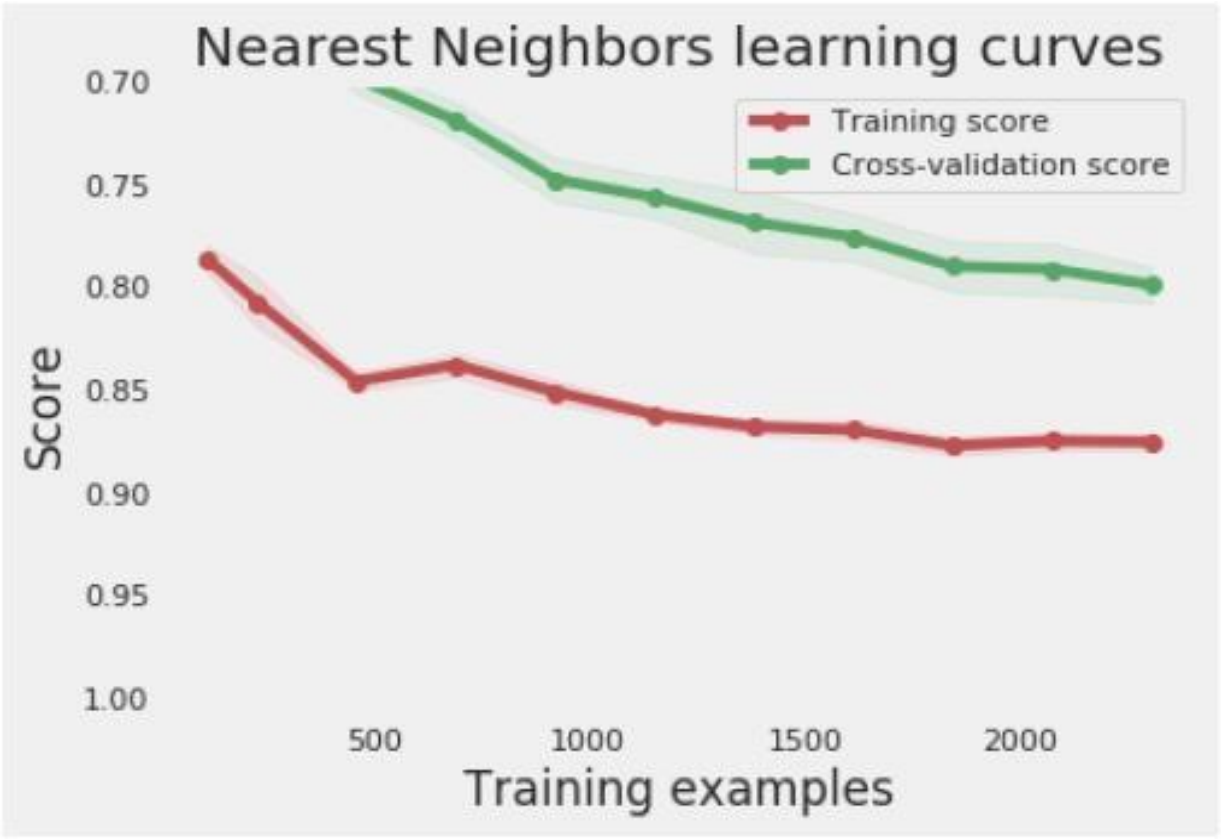


Fig. 24: Nearest neighbors learning curves

Decision Tree

This is one of my favorite algorithm and I use it quite frequently. It is a type of supervised learning algorithm that is mostly used for classification problems. Surprisingly, it works for both categorical and continuous dependent variables. In this algorithm, we split the population into two or more homogeneous sets. This is done based on most significant attributes/ independent variables to make as distinct groups as possible.

```
tr = Class_Fit(clf = tree.DecisionTreeClassifier)
tr.grid_search(parameters = [{'criterion' : ['entropy', 'gini'], 'max_features' :['sqrt', '
tr.grid_fit(X = X_train, Y = Y_train)
tr.grid_predict(X_test, Y_test)
```

```
Precision: 83.24 %
```

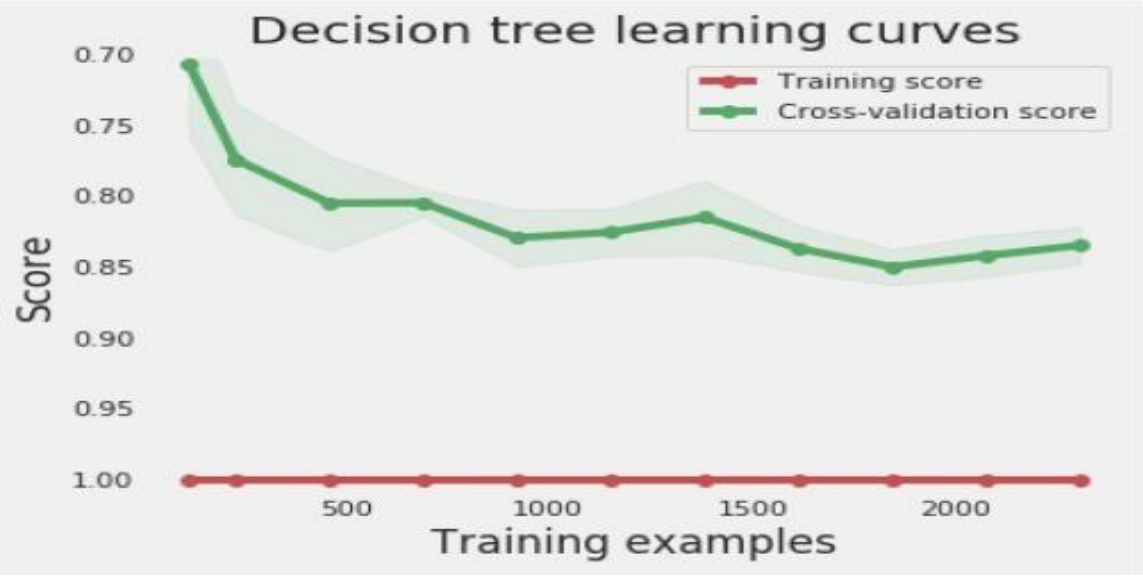


Fig. 25: Decision tree learning curves

Random Forest

Random Forest is a trademark term for an ensemble of decision trees. In Random Forest, we've collection of decision trees (so known as "Forest"). To classify a new object based on attributes, each tree gives a classification and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

Each tree is planted & grown as follows:

1. If the number of cases in the training set is N, then sample of N cases is taken at random but *with replacement*. This sample will be the training set for growing the tree.
2. If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
3. Each tree is grown to the largest extent possible. There is no pruning.

```
rf = Class_Fit(clf = ensemble.RandomForestClassifier)
param_grid = {'criterion' : ['entropy', 'gini'], 'n_estimators' : [20, 40, 60, 80, 100],
              'max_features' :['sqrt', 'log2']}
rf.grid_search(parameters = param_grid, Kfold = 5)
rf.grid_fit(X = X_train, Y = Y_train)
rf.grid_predict(X_test, Y_test)
```

```
Precision: 89.61 %
```

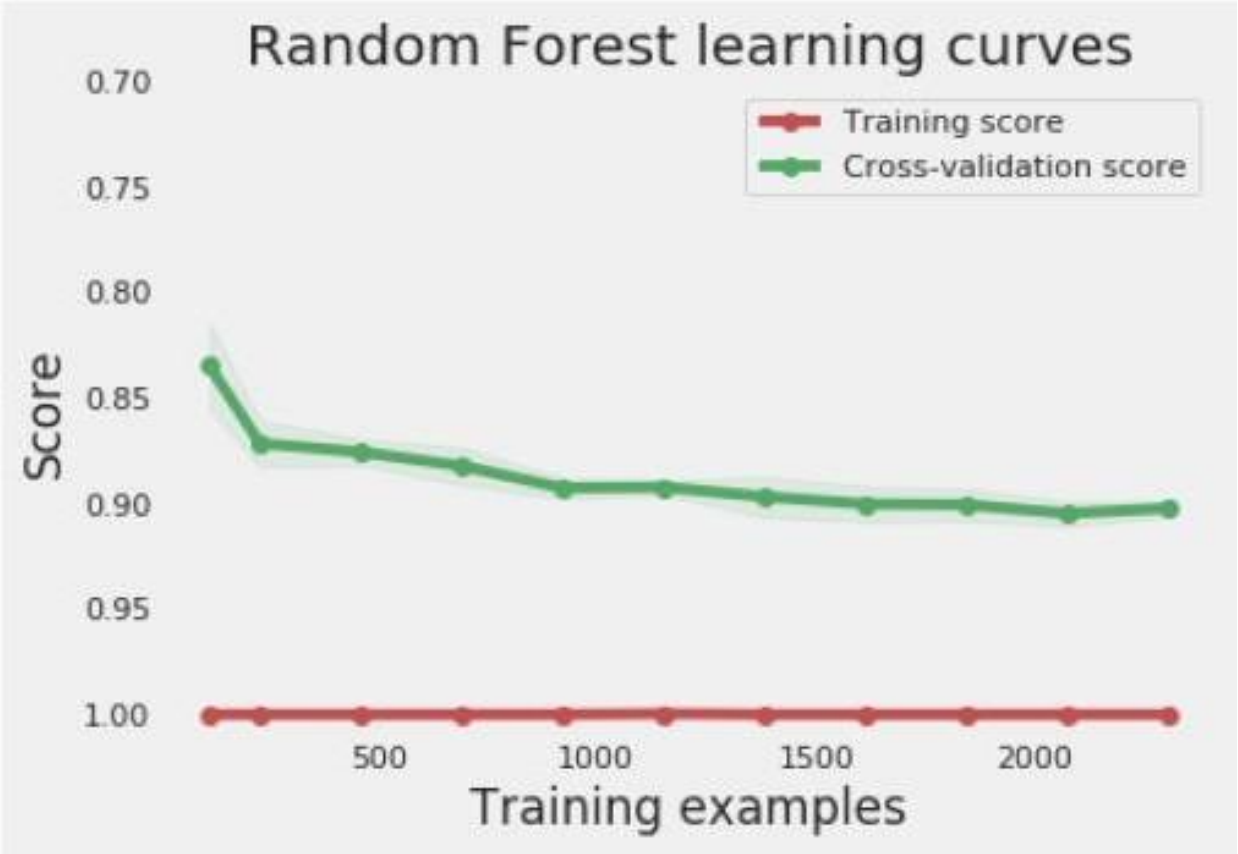


Fig. 26: Random Forest learning curves

AdaBoost Classifier

In machine learning, boosting originated from the question of whether a set of weak classifiers could be converted to a strong classifier. A weak learner or classifier is a learner who is better than random guessing. This will be robust in over-fitting as in a large set of weak classifiers, each weak classifier being better than random. As a weak classifier, a simple threshold on a single feature is generally used. If the feature is above the threshold than predicted, it belongs to positive otherwise belongs to negative.

AdaBoost stands for ‘Adaptive Boosting’, which transforms weak learners or predictors to strong predictors in order to solve problems of classification.

```
ada = Class_Fit(clf = AdaBoostClassifier)
param_grid = {'n_estimators' : [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]}
ada.grid_search(parameters = param_grid, Kfold = 5)
ada.grid_fit(X = X_train, Y = Y_train)
ada.grid_predict(X_test, Y_test)
```

```
Precision: 54.57 %
```

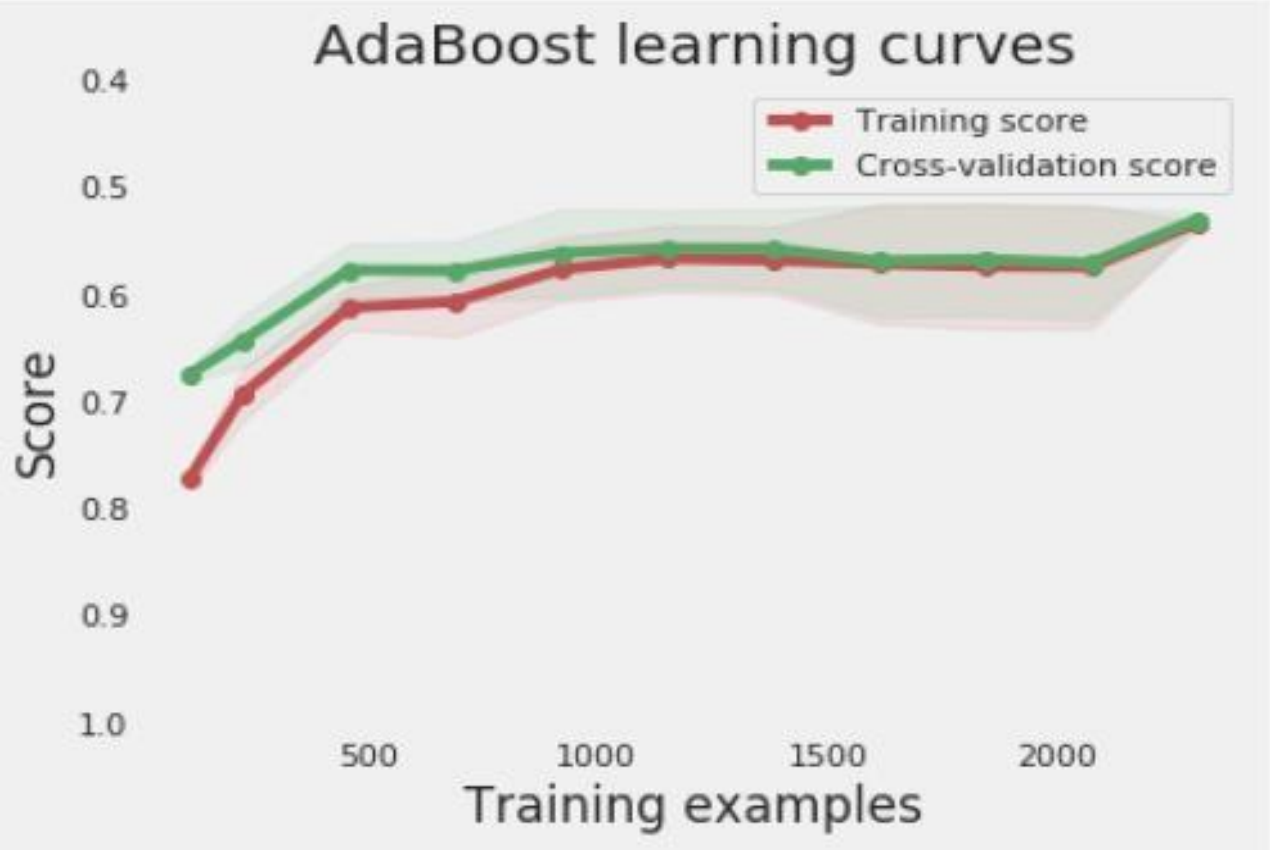


Fig. 27: AdaBoost learning curves

Gradient Boosting Classifier

GBM is a boosting algorithm used when we deal with plenty of data to make a prediction with high prediction power. Boosting is actually an ensemble of learning algorithms which combines the prediction of several base estimators in order to improve robustness over a single estimator. It combines multiple weak or average predictors to a build strong predictor. These boosting algorithms always work well in data science competitions like Kaggle, AV Hackathon, CrowdAnalytix.

```
gb = Class_Fit(clf = ensemble.GradientBoostingClassifier)
param_grid = {'n_estimators' : [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]}
gb.grid_search(parameters = param_grid, Kfold = 5)
gb.grid_fit(X = X_train, Y = Y_train)
gb.grid_predict(X_test, Y_test)
```

```
Precision: 89.47 %
```

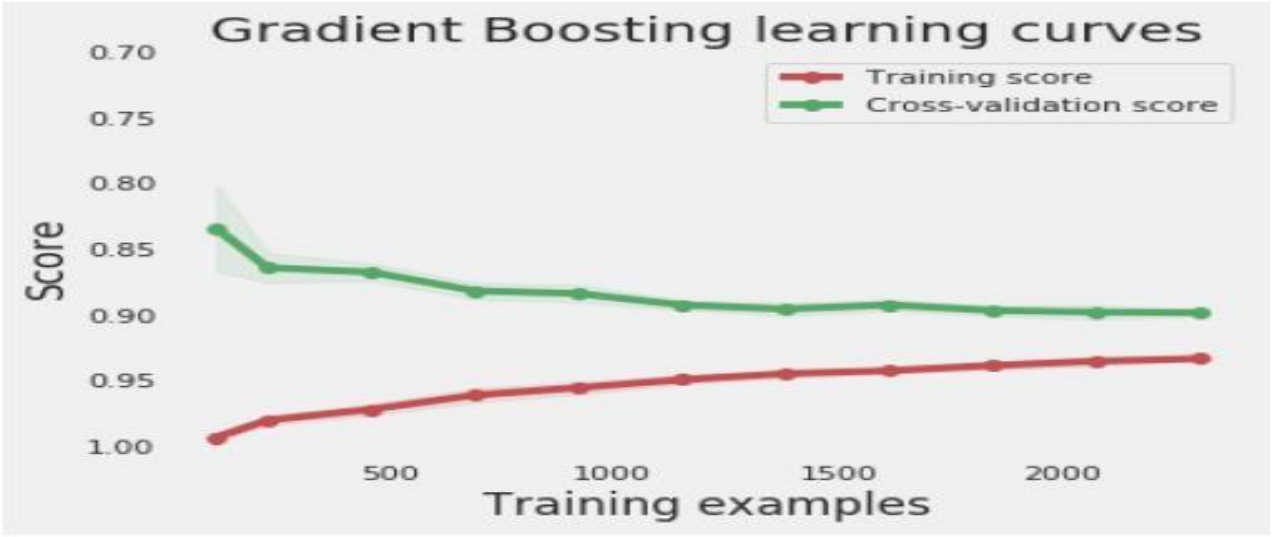


Fig. 28: Gradient boosting learning curves

CHAPTER-4 RESULT AND DISCUSSION

The results of the study showed that data science is a useful technique for performing customer segmentation based on behavioral data. Even in the case where the classes are not naturally divisible, the application provides valuable insights on user behaviour that can help the company become more data-driven.

We trained a few classifiers to categorize customers in the previous section. And we get the result like this

```
_____________________________________
Support Vector Machine
Precision: 65.93 %

_____________________________________
Logostic Regression
Precision: 71.34 %

_____________________________________
k-Nearest Neighbors
Precision: 67.58 %

_____________________________________
Decision Tree
Precision: 71.38 %

_____________________________________
Random Forest
Precision: 75.38 %

_____________________________________
Gradient Boosting
Precision: 75.23 %
```

According to the study, it is analysed that cluster analysis give very productive results to the marketers in analysing their customers. Data available with the organisation need to be segmented by cluster analysis in order to achieve

organizational goals. Cluster analysis group the customer in order to give clear picture to the marketers which customers to focus and which not to.

Market strategies like better schemes/promotional offers, sales or discounts, best everyday price for common merchandise may be adopted by the malls and supermarkets in order to attract more customers and making businesses profitable. Level of satisfaction can be enhanced by the businesses targeting married, educated customers who are either in service/business/retired with age group of 26–60 years. By targeting such customers marketers can increase their sales and build long-term relationships and thus make a successful CRM.

To check the stability of the clusters, the sample data was first split into two parts and was checked that whether similar stable and distinct clusters emerged from both the sub-samples. These analyses at the end provided further illustrations of using cluster method for market segmentation for forecasting. Computing based system developed was an intelligent and it automatically presented results to the mangers to infer for quick and fast decision making process. The simulation tests were also computed for cluster brands and other characteristics of the cluster representing a particular class of people. The future work will involve more trials and automation of the market forecasting and planning.

CHAPTER-5 CONCLUSION AND FUTURE SCOPE

5.1 CONCLUSION

It is vital to determine the proper target in business since a firm can only efficiently service consumers if it knows who its clients are among a big number of people. Customer segmentation is the practice of categorizing consumers so that a company may focus on offering personalized products or services to a certain group of people. A corporation can better allocate marketing efforts by segmenting and targeting a single or a few categories. Above all, a corporation that does segmentation analysis is more likely to get a higher return. Additionally, it increases the likelihood of developing a long-term customer connection. The purpose of this research was to determine whether data science methods can be studied to segmentation and target problems when accompanied by relevant business theories. Customer survey analysis and research problems were handled through the application of varied quantitative analytics methods in data science.

5.2 FUTURE SCOPE

- clustering is applied by combining Internet Banking transaction data, Socio-demographic information, and product ownership data.
- Create a rule of customer behavior using classification methods.

Customers have different consumption customs in different industries. Therefore, when the model is applied to customer segmentation, large amounts of historical

data must be analyzed first so that the dimensions of consumption levels and consumption fluctuation can be found out. The actual value of the model depends on the amount of data and the length of time-series data. The larger the amount is and the longer the time series are, the more actually the behavior mode is reflected by the model.

The purpose of building a customer segmentation model is to segment customers into different groups so that enterprises can sell their products according to the different needs of the segmented customers, which means establishing different marketing schemes for different targeted groups of customers according to the subsections of consuming time, the different levels of customers, the combination of products and service, and the emphasis on market positioning.

The findings in the paper may help the marketers in analysing their customers and to predict their behaviour. The study proposed that cluster analysis can be used to predict which segment of customers need to target. It may influence the managers to align the businesses with the services provided to the customers in building a successful CRM. High predictor value may influence the marketers to target the customers in increasing their profit. Marketers can focus on these segments to increase the level of satisfaction between the customer and organisation. They can also build long-term relationships which will result in successful and profitable CRM.

REFERENCES

[1] Blanchard, T., Behera, D. & Bhatnagar, P. 2019. Data Science for Marketing Analytics: Achieve Your Marketing Goals with the Data Analytics Power of Python. Birmingham: Packt Publishing. Book from library.

[2] Hague, P., Hague, N. & Morgan, C. 2013. Market Research in Practice: How to Get Greater Insight From Your Market. Second edition.

[3] Saunders, M., Lewis, P. & Thornhill, A. 2015. Research Methods for Business Students. Seventh edition. London: Pearson Education UK. Book from ebrary.

[4] Scikit-learn. 2019. Topic extraction with Non-negative Matrix Factorization and Latent Dirichlet Allocation.

[5] Sunil, R. 2017. Commonly used Machine Learning Algorithms (with Python and R Codes). Analytics Vidhya.

[6] Blanchard, Tommy. Bhatnagar, Pranshu. Behera, Trash. (2019). Marketing Analytics Scientific Data: Achieve your marketing objectives with Python's data analytics capabilities. S.l: Packt printing is limited.

[7] Griva, A., Bardaki, C., Pramatari, K., Papakiriakopoulos, D. (2018). Sales business analysis: Customer categories use market basket data. Systems Expert Systems, 100, 1-16.

[8] Hong, T., Kim, E. (2011). It separates consumers from online stores based on factors that affect the customer's intention to purchase. Expert System Applications, 39 (2), 2127-2131.

[9] Hwang, Y. H. (2019). Hands-on Advertising Science Data: Develop your machine learning marketing strategies… using python and r. S.l: Packt printing is limited.

[10] Puwanenthiren Premkanth, - Market Classification and Its Impact on Customer Satisfaction and Special Reference to the Commercial Bank of Ceylon PLC.‖ Global Journal of Management and Business Publisher Research: Global Magazenals Inc. (USA). 2012. Print ISSN: 0975-5853. Volume 12 Issue 1.

[11] Sulekha Goyat. "The basis of market segmentation: a critical review of the literature. European Journal of Business and Management www.iiste.org. 2011. ISSN 2222-1905 (Paper) ISSN 2222-2839 (Online). Vol 3, No.9, 2011.

[12] By Jerry W Thomas. 2007. Accessed at: www.decisionanalyst.com on July 12, 2015. [9] T.Nelson Gnanaraj, Dr.K.Ramesh Kumar N.Monica. AnuManufactured cluster analysis using a new algorithm from structured and unstructured data. International Journal of Advances in Computer Science and Technology. 2007. Volume 3, No.2.

[13] McKinsey Global Institute. Big data. The next frontier is creativity, competition, and productivity. 2011. Accessed at: www.mckinsey.com/mgi

on July 14, 2015.

[14] Roung-Shiunn Wu; Po-Hsuan Chou (2011). Customer segmentation of multiple category data in e-commerce using a soft-clustering approach., 10(3), 331–341. DOI: 10.1016/j.elerap.2010.11.002

[15] Wu, Jing; Lin, Zheng (2005). [ACM Press the 7th international conference - Xi'an, China (2005.08.15-2005.08.17)] Proceedings of the 7th international conference on Electronic commerce – ICEC'05-Research on customer segmentation model by clustering., (), 316–. doi:10.1145/1089551.1089610