**GROUPING CUSTOMERS USING CLASSIFICATION ALGORTHMS**

**A FINAL YEAR CAPSTONE DESIGN PROJECT**

**(Phase-II)**

***Submitted by***

**RAVI TEJA YAMSANI - 9918004165**

***In partial fulfillment for the award of the degree***

***Of***

**BACHELOR OF TECHNOLOGY**

## IN

COMPUTER SCIENCE AND ENGINEERING



**SCHOOL OF COMPUTING**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING KALASALINGAM ACADEMY OF RESEARCH AND EDUCATION KRISHNANKOIL 626 126**

**MAY 2022**

## DECLARATION

I hereby certify that the work which is being presented in the B.Tech. Community Service Project Report entitled **“GROUPING CUSTOMERS USING CLASSIFICATION ALGORTHMS** in partial fulfillment of the requirements for the award of the **Bachelor of Technology in Computer science and engineering** and submitted to the Department of Computer science and Engineering of Kalasalingam Academy of Research and Education (Deemed to be University) – Tamil Nadu is an authentic record of my own work carried out during a period from February 2021 to May 2021 under the supervision of **RAVI TEJA YAMSANI (9918004165)** of CSE

Department. The matter presented in this thesis has not been submitted by me for the award of any other degree elsewhere.

*Signature of Candidate*

## Y. RAVI TEJA (9918004165)



**BONAFIDE CERTIFICATE**

Certified that this project report “**GROUPING CUSTOMERS USING CLASSIFICATION ALGORTHMS”** is the bonafide work of “**RAVI TEJA YAMSANI (9918004165)** who carried out the project work under my supervision.

## SUPERVISOR HEAD OF THE DEPARTMENT

Dr. A. Saravanan Dr.P. Deepalakshmi

Associate professor Professor/Head/Dean

Computer Science and Engineering Computer Science and Engineering Kalasalingam Academy of Research and Kalasalingam Academy of Research and Education and Education

Krishnankoil 626126 Krishnankoil 626126

Submitted for the Project Viva-voce examination held on 10/05/2022.

**Internal Examiner External Examiner**

# ACKNOWLEDGEMENT

First and foremost, we thank the ‘Supreme Power’ for the immense grace showered on us which enabled us to do this project. We take this opportunity to express by sincere thanks to the late**,**

**“Kalvivallal” Thiru T. Kalasalingam, Chairman, Kalasalingam Group of Institutions, “Illayavallal” Dr. K. Sridharan,Ph.D., Chancellor, Dr. S. Shasi Anand, Ph.D., Vice President**, who is the guiding light for all the activities in our University.

We thank our Vice chancellor **Dr.R. Nagaraj, Ph.D.,** for guiding every one of us and infusing us the strength and enthusiasm to work over successful.

We wish to express our sincere thanks to our respected Head of the Department **Dr. P. Deepalakshmi, Professor&Dean**, whose moral support encouraged us to process through our project work successfully.

We offer our sincerest gratitude to our Supervisor, **Dr. A. Saravanan,** for his patience, motivation, enthusiasm and immense knowledge.

We are extremely grateful to our Overall Project Coordinator, **Dr. R. Murugeswari** for constant

encouragement in the completion of the Community Service Project.

Finally, we thank all, our Parents, Faculty, Non-Teaching Faculty and our friends for their moral support.



# School of Computing

**Department of Computer Science and Engineering Project Summary**

|  |  |  |
| --- | --- | --- |
| Project Title | Cardiac disease analysis using Machine learning  Machine learning | |
| Project Team Members (Name with Register No) | Ravi teja Yamsani (9918004165) | |
| Guide Name/Designation | Dr. A. Saravanan , Associate Professor, Department of Computer  Science and Engineering | |
| Program Concentration Area | Online business data analysis and visualization | |
| Technical Requirements | Jupyter notebook | |
| **Area** | **Codes & Standards / Realistic Constraints** | **Tick**  ✓ |
| Economic | This project is used for segmentation of the people into groups and increase the business sales by suitable recommendation | ✓ |
| Environmental |  |  |
| Social |  |  |
| Ethical |  |  |
| Health and Safety |  |  |
| Manufacturability |  |  |
| Sustainability |  |  |

**Realistic Constraints:** Real time data and privacy invasion are the constraints for this project. In real time large amount of data is being processed and implemented for which we need a good data base and a dedicated server for maintaining all records.

## Economic:

The economy of the companies are increased dramatically after introduction of the recommending systems and as well as this grouping of the customers made the company to focus on their clients with their suitable interests and which helped them to gain a lot more new customers.

## Engineering standards:

IEEE 3652.1-2020 - Guide for Architectural Framework and Application of Machine Learning.

# ABSTRACT

Meeting of many competitors at once entrepreneurs have created a major conflict between competitors businesses to find new customers and keep old ones. As a result formerly, the need for unique customer service becomes relevant regardless of the size of the business. In addition, the ability of any business to understand the needs of each customer will supply a larger customer support for the provision of targeted customer services and development customized customer service plans. This understanding possible through systematic customer service. Each part has customers who share similar market features. Big data ideas and machine learning have encouraged greater acceptance of automatic customer segregation methods in favor traditional market statistics that often do not work there the customer base is very large. Customer segregation is important in customer relationship management software. The most common way to differentiate one customer from another is to develop a customer group according to their interests and factors like demographic factors, Psychological factors, Behavioral factors, and geographical factors . In this work, the company's hand-separated customer data is analyzed. As machine learning methods are useful for solving problems with data management and data processing, the solution is searched within machine learning methods. The different classification methods are used to differentiate the people into different groups and help with suitable solutions. For increasing the efficiency and precision we used all the classification algorithms as a ensemble algorithm and created a model.

# TABLE OF CONTENTS

## CHAPTER NO. TITLE PAGE NO. ABSTRACT

**LIST OF TABLES LIST OF FIGURES**

## LIST OF SYMBOLS (if required) LIST OF ABBREVIATIONS

**CHAPTER 1 INTRODUCTION 1**

* 1. General
  2. Existing System
  3. Proposed System

CHAPTER 2 LITERATURE SURVEY ……………….…………………………………. 2

CHAPTER 3 OBJECTIVES AND SCOPE OF THE PROJECT… 3

CHAPTER 4 PROJECT PLAN……………………………………………………………..4

CHAPTER 5 PROBLEM DEFINATION…………………………………………………..5

CHAPTER 6 REQUIREMENTS…………………………………………………………...7

* 1. Requirement description
  2. Hardware requirements
  3. Software requirements

## CHAPTER 7 SYSTEM DESIGN 8

* 1. Proposed Model for customer grouping
  2. Data collection and pre-processing
  3. Classification Algorithm

## CHAPTER 8 MODULE IMPLEMENTATION 12

## CHAPTER 9 CONCLUSION 24

**CHAPTER 10 REFERENCES 26**

# LIST OF TABLES

|  |  |  |
| --- | --- | --- |
| **S. NO** | **TITLE** | **PAGE NO** |
| 7.1 | SYSTEM DESIGN TABLE 1 | 9 |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **S. NO.** | **TITLE** | **PAGE NO** |
| 3.1 | OBJECTIVES AND SCOPE OF THE PROJECT FIGURE 1 | 3 |
| 5.1 | PROBLEM DEFINATION FIGURE 1 | 6 |
| 5.2 | PROBLEM DEFINATION FIGURE 2 | 6 |
| 7.1 | SYSTEM DESIGN FIGURE 1 | 8 |
| 7.2 | SYSTEM DESIGN FIGURE 2 | 9 |
| 8.1  8.2  8.3  8.4  8.5  8.6  8.7  8.8  8.9  8.10  8.11  8.12  8.13  8.14  8.15  8.16  8.17  8.18  8.19  8.20  8.21  8.22  8.23  8.24  9.1 | MODULE IMPLEMETATION FIGURE 1  MODULE IMPLEMETATION FIGURE 2  MODULE IMPLEMETATION FIGURE 3  MODULE IMPLEMETATION FIGURE 4  MODULE IMPLEMETATION FIGURE 5  MODULE IMPLEMETATION FIGURE 6  MODULE IMPLEMETATION FIGURE 7  MODULE IMPLEMETATION FIGURE 8  MODULE IMPLEMETATION FIGURE 9  MODULE IMPLEMETATION FIGURE 10  MODULE IMPLEMETATION FIGURE 11  MODULE IMPLEMETATION FIGURE 12  MODULE IMPLEMETATION FIGURE 13  MODULE IMPLEMETATION FIGURE 14  MODULE IMPLEMETATION FIGURE 15  MODULE IMPLEMETATION FIGURE 16  MODULE IMPLEMETATION FIGURE 17  MODULE IMPLEMETATION FIGURE 18  MODULE IMPLEMETATION FIGURE 19  MODULE IMPLEMETATION FIGURE 20  MODULE IMPLEMETATION FIGURE 21  MODULE IMPLEMETATION FIGURE 22  MODULE IMPLEMETATION FIGURE 23  MODULE IMPLEMETATION FIGURE 24  CONCLUSION FIGURE 1 | 12  12  13  13  14  14  15  15  16  16  17  17  18  18  19  19  20  20  21  21  22  22  23  23  25 |
|  |  |  |

# LIST OF ABBREVIATIONS

|  |  |
| --- | --- |
| **KNN -** | **K-Nearest Neighbour** |
| **SVM-** | **Super vector machine** |
| **DT-** | **Decision tree** |
| **EDA-** | **Exploratory Data analysis** |
| **CART-** | **Classification and regression tree algorithm** |

## CHAPTER 1

## INTRODUCTION

* 1. **General**

From many years, Due to the rise in competitive nature among the businesses and the use of the past data has motivated in implementation of data mining techniques for finding the critical and useful insights from the information. Data mining procedure is helpful for extraction of the logical information from the dataset of the company and project in a human understanding manner such that human can take required pre requisite to solve the problem.

.

## Existing System

In the existing system only KNN algorithm is used for the project which cannot define all the essential factors or features for efficient precision.

Recently, Big Data research has gained momentum. Where big data describes a large number of formal and informal data, which cannot be analyzed using traditional methods and algorithms.

No proper data visualization is used in the existing system such that only tech geek can understand the results from the implementation.

## Proposed System

In Proposed system we Classify customers into segments and Anticipate the purchases that will be made by a customer, during the following year.For classification, I took data from Kaggle. After pre processing the data we will visualize the data and its factors in a graph model using matploit and seaborn libraries. To handle the big data problem I used all the classification algorithms and selected best classifiers in each algorithm and made the prediction by combining all the parameters

For better understanding even for the non tech geek I used word cloud and made the visualizations of each group as samples.In this project large data set is used such that the results which are produced are not biased and the system which is proposed can be used for real time purpose. Ensembling the algorithms and producing the results make the prediction more accurate. The geographical parameters are also took into consideration before classification.

## CHAPTER 2

## LITERATURE SURVEY

* In literature survey we discussed various machine learning techniques have been proposed by the scientists for identification and grouping the customers into different clusters. This research study shows some existing machine learning based techniques in order to explain importance of the implemented work.
* The market separation or grouping is based on only few factors for determining the customer group
* Over the years, the commercial world has become more competitive, as organizations such as these have to meet the needs and desires of their customers, attract new customers, and thus improve their businesses.
* For amazon, flipkart, and Netflix the customer grouping is done by item to item and only built by the data received by their website. All their systems algorithms are confidential and not shared with the all but only the basis of their fundamental are shared for privacy concern of the customers.
* Few data scientists did the project using KNN algorithm for separation but the data considered for the project is likely small and the accuracy is about 78.3%.
* Grouping also includes following factors like Demographic, geographic, technographic and behavioral segmentation.
* Data collection is part of the research in all fields of study including natural and social sciences, people and businesses. The purpose of all data collection to obtain quality evidence leading to analysis creates straightforward and misleading answers to questions presented.
* Integration is the process of combining information into a Database based on similarities. There are several algorithms, which can be used in data sets based on given condition. [7] However, there is no general consolidation. An algorithm exists, so it is important to choose appropriate methods of integration
* Customer segmentation used to be done manually and wasn’t too precise. They would manually create and populating different data tables, and analyze the data.

## CHAPTER 3

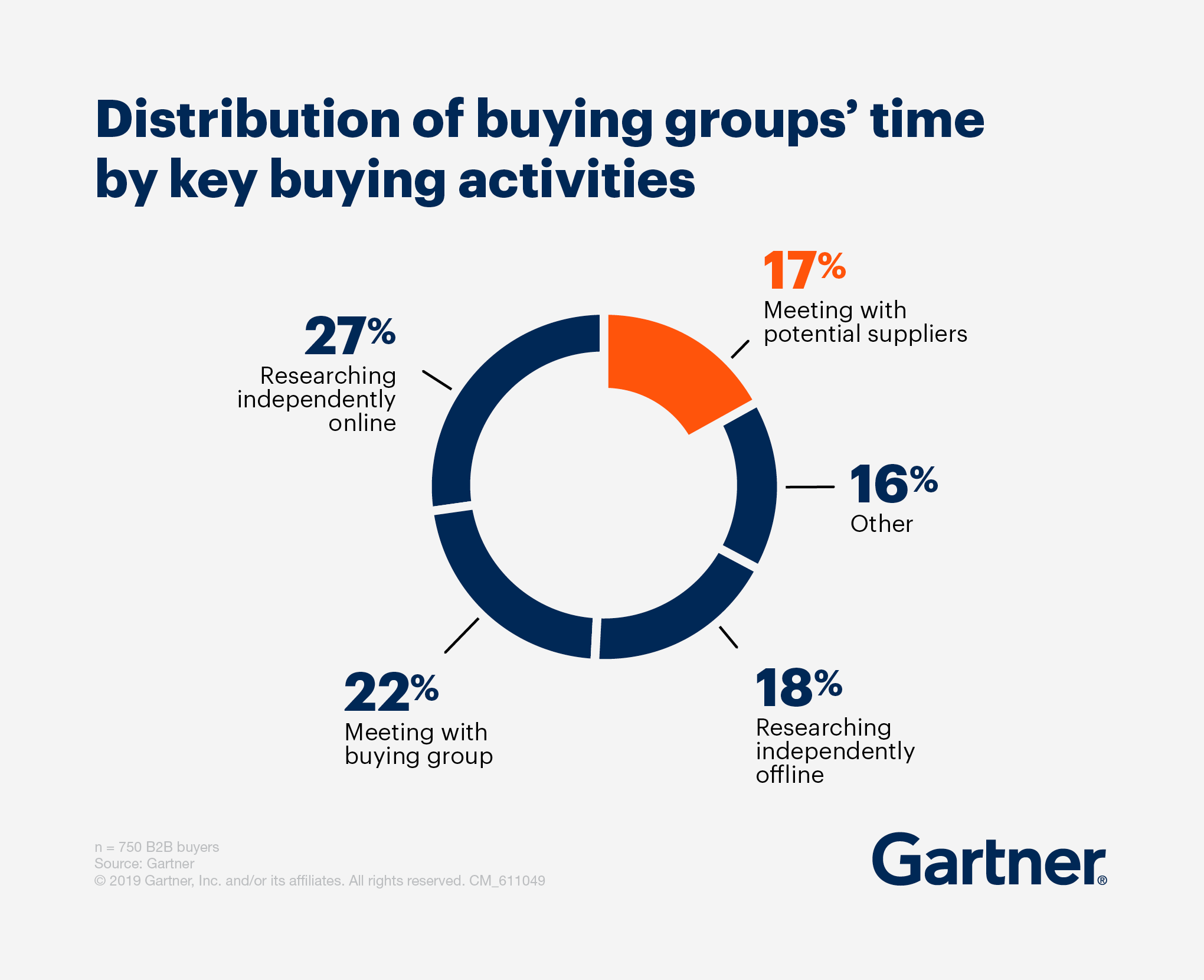
**OBJECTIVES AND SCOPE OF THE PROJECT**

## Objective:

* + This project objective is dividing a company’s customers into groups that reflect similarity among customers in each group.
  + In addition, we can use the project demographics for building the recommendation system and can increase the potentiality of business.
  + The goal of grouping customers is to decide how to relate to customers in each segment in order to maximize the value of each customer to the business.
  + The predicted accuracy of the testing phase is around 89 %.

## Scope:

Project scope is to build an eco-system for market specialization and help businesses to increase their efficiency and implementation. This traditionally is manually implemented but now can be automated using machine learning algorithm.



**Figure 3.1**- Gartner company proves that segregating of customers brings 2nd highest sales.

## CHAPTER 4

## PROJECT PLAN

|  |  |
| --- | --- |
| **Assigned work** | **Date/Time** |
| Data preprocessing and Exploring | JANUARY |
| Constructing the ML Model | FEBRUARY |
| Ensemble model and Testing phase | MARCH |

**Data preprocessing and Exploring** : In data preprocessing the unfilled data is filled with the considerate values such that there will be no underfitting and overfitting of the ML model. After processing the data and we get into exploration section where we describe what are all the columns present in the data frame and understand the number of wanted and unwanted data for example unwanted data like cancelled order are removed because the customer didn’t purchased the product in first place and hence, such type of data is unnecessary for our training model.

**Constructing the ML model:** In this phase we are dealing with the product unique codes and description. Thereafter understanding all the products we need to define the product categories for further implementation. For this categories we need to create the clusters for each one here, I created 5 such categories clusters and visualized using the word cloud so that we can easily understand what type of interests do the customer have in each particular cluster. The clusters model to be trained using all the classification algorithms.

**Ensemble model and Testing phase**: In this final stage, we train our data with the more precise parameters of each classification algorithm i.e ensemble model. In which we train with the training data and can see the huge increase in the precision of algorithm. Next, We test our machine leaning model with the testing data. This testing data consists of last 2 months shopping transactions and hence it helps us segregate the customers into the suitable groups.

## CHAPTER 5

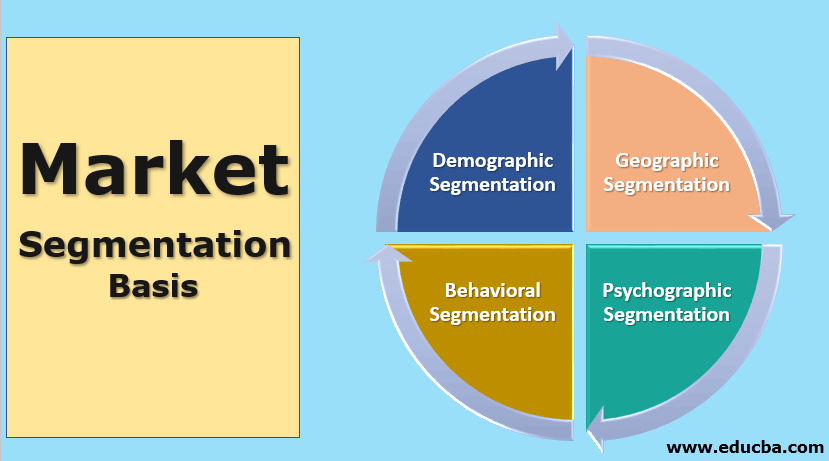
## PROBLEM DEFINITION

The marketing methods able to get results by using the traditional methods by the assumption that every customer buy whatever they need without going to the other shops and such that only such type of customers are attracted to the business and the products availability will be limited.

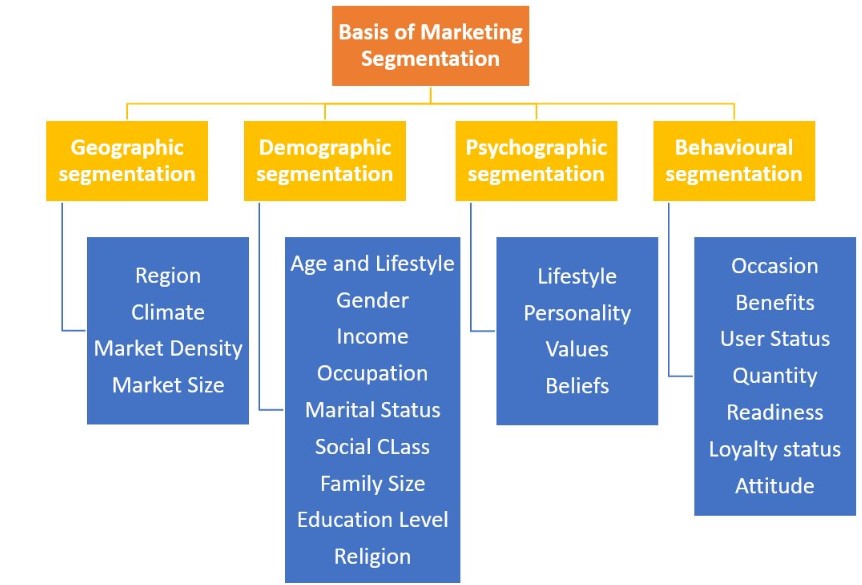
For instance if we make groups and assign customers into those categories such that where one size fits method. This can be obtained through only the successful segmentation. The segmented customers are grouped into clusters where all of them almost likes the same things and think similarly, This helps to drive dynamic content and personalization tactics for increase in the business.

However, in order for the division to be used effectively, it needs to take into account that different customers are buying for different reasons, and retailers need to use wisely a lot of considerations that could affect their buying decisions. A Harvard Business School professor even said that of the 30,000 new products introduced each year, 95% fail because of poor market performance.

There are numerous reasons why the company fail in selecting their target group. It is a wildguess that segmentation is only done on the one factor like demographics or the cluster is too broad to be under the one roof. There may be no strategic goal, such as lead discovery or customer retention. Or, it may be that a lack of One Customer View and aggregated data has led to a misunderstanding or misinterpretation of your customer base. So for segmentation in this project we use almost all factors like demographic, geographic, psychographic and Behavioral.



**Figure 5.1**: Market Grouping Factors.



**Figure 5.2 :** Market segmentation factors brief explanation.

## 

## CHAPTER 6

## REQUIREMENTS

1. **Requirement description**

The type of requirements we need to complete the project are categorized into 2 following section. As this is a machine learning algorithm we need to make use of a optimized system to get good and accurate results.

## Hardware requirements

* GB RAM
* 1 TB HDD
* Intel core i7

## Software requirements

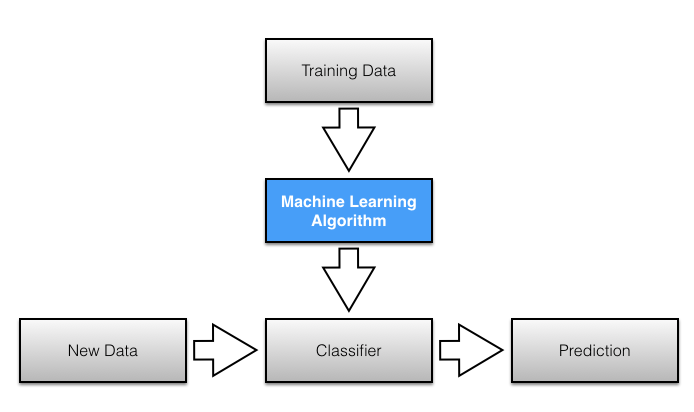
* Windows 7/10 (or any operating system)
* Python
* Jupyter
* Kaggle /Google collab

## CHAPTER 7

## SYSTEM DESIGN

* 1. **Proposed Model for Customer Grouping**

The proposed model helps us in grouping only after providing the suitable data for retrieving the information. In this model the company provides the data to the model and then perform the ensemble model on the data for segregating the customers for increasing their sales.



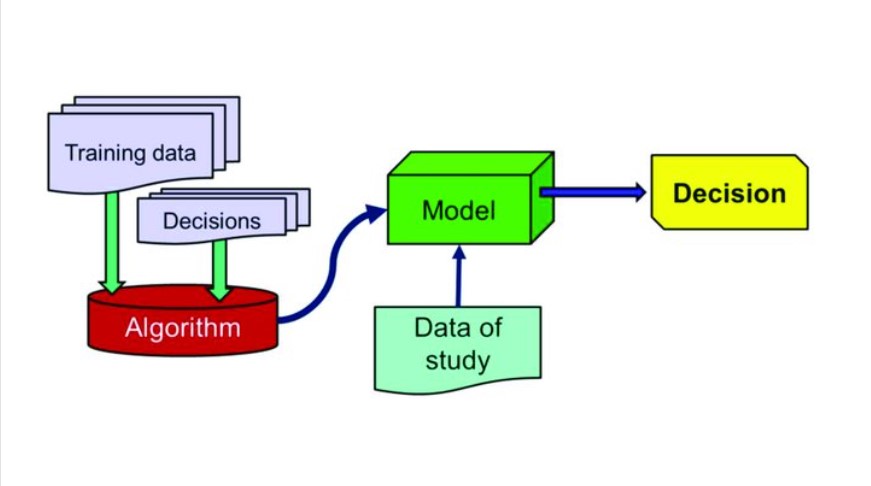
**Figure 7.1** : Flow chart of Classification algorithm

## Data collection and pre-processing

The database used here for machine learning comes from sales of the UK retailer business. The goal of segregation is to increase the sales of the company and attract new customers by showing the their needs in the websites. The data set consists of above 5lakh rows and 8 columns.

|  |  |
| --- | --- |
| **Invoice Number** | It is six digits code which is assigned to each product and can be easily identified by scanning the barcode. Usually this code is present in each transaction. If this code has letter c in the starting it states that it is a cancelled order. |
| **Stock Code** | It is generally called as Product or code, and has a serial number a 5-digit number and uniquely assigned every product. |
| **Description** | It is the explanation of the product that customer brought. |
| **Quantity** | It is a numeric value. This explains the number of products the same product is been sold. |
| **Invoice Date**: | The date and time when the transaction had took place. This is also a numeric value. |
| **Unit Price** | This say about the price in sterling for unit product. |
| **Customer ID** | It is a unique code which is assigned for the each customer during the account creation. This can be used as primary key in SQL query incase of data verification. |
| **Country** | This column say about the identity of the persons nationality. Such that we can use this geographical factor for our ML model. |

**Table 7.1:** Variables of the Machine Learning model



**Figure 7.2**: Training and Decision process of ML Model.

## Classification Algorithm

The attributes mentioned in Table 1 are presented as inputs to various ML algorithms such as Random Forest, Decision Tree, Logistic Retreat, Naive Bayes, XG Boost and Light GBM split strategies. The set of input data is divided into 80% of the training database and the remaining 20% of the test database. Database training is a database used to train a model. The test database is used to test the performance of a trained model. In each algorithms the performance is calculated and analyzed based on the various metrics used such as accuracy, precision, recall and F scoring scores as described further. The various algorithms tested in this paper are listed below.

The ML process starts in the pre-processing phase of data followed by feature selection and feature rating and we will use it with the following algorithms such as KNN, SVM, Decision trees, XGBOOST , Support vector Machine and Random Forest algorithms.

**Logistic Regression :** Regression usually refers to the analysis of segmentation problems, despite your regression name in the segmentation algorithm. It can be used for both binary classification and multiple classification. The predictable rate in depreciation is by category. The sigmoidal function is used to reduce the value from a range of 0 to 1.

**XGBOOST:** The gradient boosted trees technique is implemented in XGBOOST, a popular and efficient open-source implementation. Gradient boosting is a supervised learning approach that combines the estimates of a collection of smaller, weaker models to attempt to accurately predict a target variable. The data is built in a sequential manner, with each successive data aiming to minimise the entire data's faults. Extreme Gradient Boosting is abbreviated as XGBoost. The word xgboost, on the other hand, alludes to the technical objective of pushing the computational resources for boosted tree algorithms to their limits. XGBoost is a software library that you may download and install on your computer, then use from a number of different interfaces.

**Support vector machines:** SVM selects extreme points / vectors that help create hyperplane. These extreme cases are called supporting vectors, which is why the algorithm is called Vector Support Machine. Consider the diagram below in which there are two distinct categories divided by resolution or hyperplane.

**Random Forest:** Random forest are built from decision trees. Generally, tree works well with the data they have or familiar but fails to classify new samples. Random forest adds up the simplicity of the decision tree with flexibility resulting in the large improvement and in the accuracy. At each time we take the bootstrapped data and consider only a subset of the variables at each step which results in a wide variety of trees by applying the decision tree approach. The final result depends on the forests (multiple trees) that we created with the variety of trees. The remaining data after bootstrapping at each step is called out-of-bag data. The correct classification of the out-of-bag data is solemnly responsible for the accuracy of the random forest algorithm. It is an example of the bagging technique.

**KNN:** K-means the algorithm is one of the most popular segmentation algorithms. This integration algorithm depends on it centro, where each data point is placed in one of scattered, pre-filtered in the K-algorithm. Creating collections corresponding to hidden patterns in data that provides the information needed to help determine execution. process.

**Decision Tree**: In Decision tree algorithm the data is split by using the conditions and create the classes for each branch. These trees can continuously take the real number values it can also called as regression trees. In this algorithm the identification of the root node parameter is difficult. The process of identifying is called attribute selection. For this we need to take 2 factors under consideration they are information gain and Gini index.

**CHAPTER 8**

**MODULE IMPLEMENTATION**

## 

## Figure 8.1: Importing the essential dependencies

## 

## Figure 8.2: Data preprocessing and data cleaning

## 

## Figure 8.3: Exploring the data and geographical purchases

## 

## Figure 8.4: Flow chart of the transactions and order of amounts.

## 

## Figure 8.5: Product Description and clusters are created by these keywords

## 

## Figure 8.6: Silhouette score of the cluster groups are determined.

## 

## Figure 8.7: All the customers are classified by words and visualized

## 

## Figure 8.8: After the customers are grouped now we should group the prodcuts

## 

## Figure 8.9 : Next I created the combinations of the customers transactions made\

## 

## Figure 8.10 : Data encoding turning words into numerical values for processing

## 

## Figure 8.11 : Customer categories created.

## 

## Figure 8.12 : Customer categories are visualized

## 

## Figure 8.13 : Program for creating the categories display and enhancement

## 

## Figure 8.14 : Displaying the confuse matrix of the MLmodel

## 

## Figure 8.15 : Accuracy of the SVC.

## 

## Figure 8.16 : Accuracy of the Logistic regression

## 

## Figure 8.17 : Accuracy of the KNN

## 

## Figure 8.18 : Accuracy of the Decision tree.

## 

## Figure 8.19: Accuracy of the Random Forest

## 

## Figure 8.20 : Accuracy of the Ada boost

## 

## Figure 8.21 : Accuracy of the XGBOOST

## 

## Figure 8.22 : Accuracy of the Ensemble model in the training phase

## 

## Figure 8.23 : Testing phase where the data is took from the last 2 months transactions of dataset

## 

**Figure 8.24 :** Accuracy of the all the algorithms and the ensemble model in the

testing phase.

## CHAPTER 9

## CONCLUSION

## Since our database was inconsistent as it is a real time data , we need to have a good computing power to run the billions of data. At last we provided the testing phase results where the ML model achieved about 90% accuracy. This is achieved due to the ensemble model learning where it chose the best parameters from each classification algorithm and used as the parameter. The visualization of the clusters is done by the word clouds in the boxes such all the customers who belong to one category has similar interests for the products. In the training phase we can observe that XG boost gave the more accuracy comparatively to other algorithms.

## 

## Figure 9.1: Displaying all the categories customers.

## CHAPTER 10 REFERENCES

* + [1] 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 [2] Griva, A., Bardaki, C., Pramatari, K., Papakiriakopoulos, D. (2018). Sales business analysis: Customer categories use market basket data. Systems Expert Systems, 100, 1-16.
  + [3] 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.
  + [4] 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
  + [5] 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.
  + [6] 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.
  + [7] 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
  + [8] 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.
  + [10] McKinsey Global Institute. Big data. The next frontier is creativity, competition and productivity. 2011. Accessed at: www.mckinsey.com/mgi on July 14, 2015.
  + [11] Jean Yan. - Big Data, Big Opportunities- Domains of Data.gov: Promote, lead, contribute, and collaborate in the big data era. 2013. Retrieved from http://www.meritalk.com/pdfs/bdx/bdxwhitepaper-090413.pdf July 14, 2015.
  + [12] A.K. Jain, M.N. Murty and P.J. Flynn.‖ Data Integration: A Review‖. ACM Computer Research. 1999. Vol. 31, No. 3.
  + [13] Vishish R. Patel1 and Rupa G. Mehta. MpImpact for External Removal and Standard Procedures for JCSI International International Science Issues Issues, Vol. 8, Appeals 5, No 2, September 2011 ISSN (Online): 1694-0814 7
  + [14] Jayant Tikmani, Sudhanshu Tiwari, Sujata Khedkar "Telecom Customer Classification Based on Group Analysis of K-methods", JIRCCE, Year: 2015.
  + [15] Vaishali R. Patel and Rupa G. Mehta “Impact of Outlier Removal and Normalization Approach in Modified k-Mean