VISVESVARAYA TECHNOLOGICAL UNIVERSITY

"Jnana Sangama", Belgavi-590 018, Karnataka, India

An Internship Report
On

CUSTOMER SEGMENTATION

Submitted in Partial Fulfillment of the requirement for the award of the degree of

BACHELOR OF ENGINEERING
IN
COMPUTER SCIENCE AND ENGINEERING

Submitted By

Kushal S **1SJ18CS049**

Carried out at
Exposys Data Labs
(Software Company)
Yelahanka, Bangalore

Under the guidance of

Internal Guide	External Guide
Prof. Divakar K M	**Y Vishnuvardhan**
Assistant Professor	**Founder & CEO**
Dept. Of CSE, SJCIT	**Exposys Data Labs**

S J C INSTITUTE OF TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
CHIKKABALLAPUR-562101
2021-2022

||Jai Sri Gurudev||

Sri Adichunchanagiri Shikshana Trust®

S.J.C INSTITUTE OF TECHNOLOGY, Chikkballapur - 562101

Department of Computer Science and Engineering

CERTIFICATE

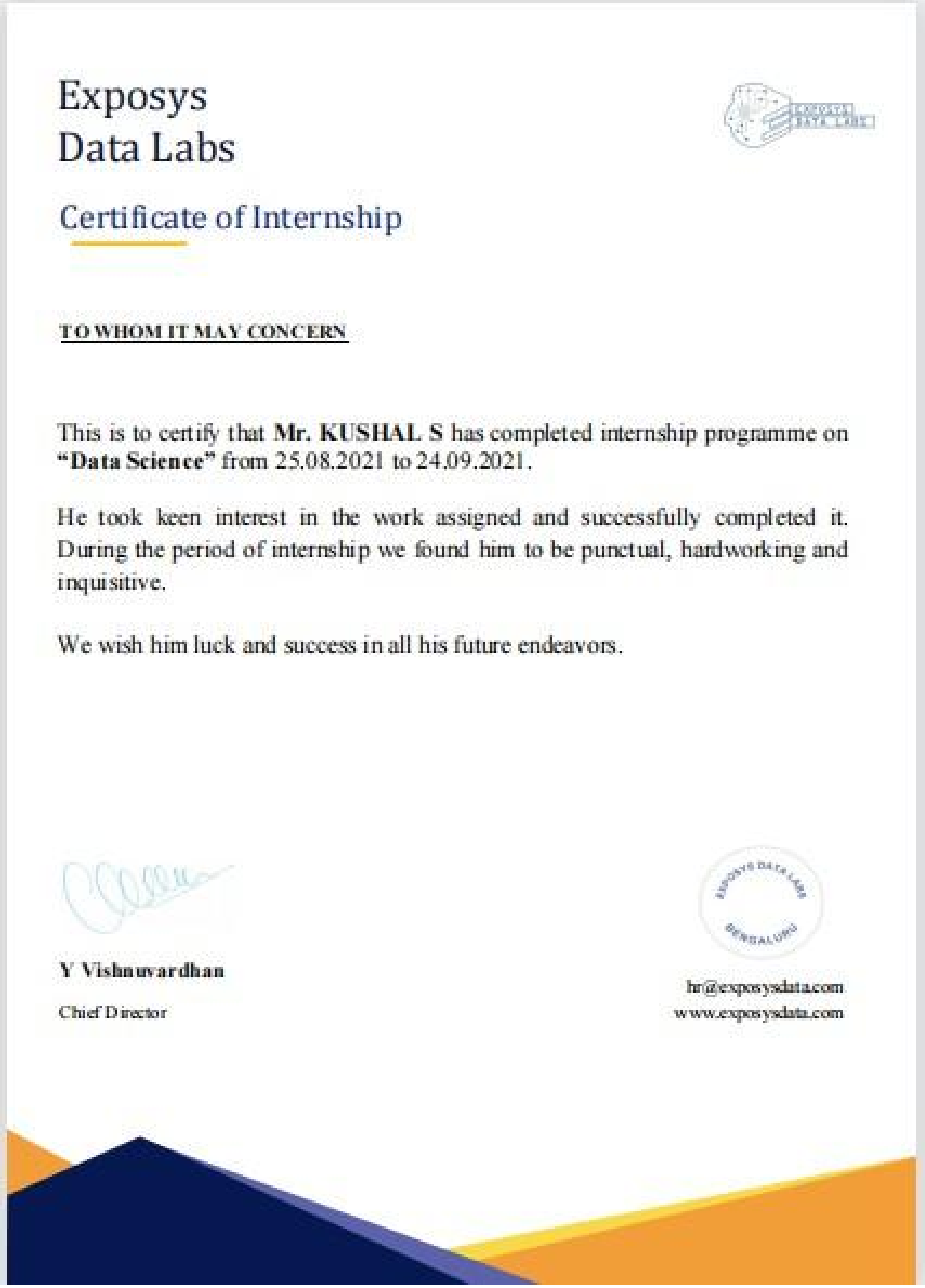
This is to certify that the Internship work entitled **"Customer Segmentation"** carried out by **KUSHAL S** bearing USN:1S18CS049 a bonafide student of Sri Jagadguru Chandrashekaranatha Institute of Technology in partial fulfilment for the award of **Bachelor of Engineering in Computer Science and Engineering of Visvesvaraya Technological University, Belgaum** during the year **2021-22**. It is certificated that all corrections / suggestions indicated for internal assessment have been incorporated in the report deposited in the departmental library. The Internship report has been approved as it satisfies the academic requirements in respect of Internship work prescribed for the said Degree.

Signature of Guide	Signature of HOD	~~Signature of Principal~~
Prof. Divakar K M	**Dr. Manjunath Kumar B H**	**Dr. G T Raju**
Assistant Professor	Professor & HOD,	Principal, SJCIT,
Dept. of CSE,SJCIT	Dept. of CSE,SJCIT	Chikkballapur

External Examiners:

Name of the Examiners	**Signature with Date**
1. H Swasamy	
2. Ajay.N	Ajay 25/7/22

COMPANY CERTIFICATE



Exposys
Data Labs

Certificate of Internship

TO WHOM IT MAY CONCERN

This is to certify that **Mr. KUSHAL S** has completed internship programme on **"Data Science"** from 25.08.2021 to 24.09.2021.

He took keen interest in the work assigned and successfully completed it. During the period of internship we found him to be punctual, hardworking and inquisitive.

We wish him luck and success in all his future endeavors.

EXPOSYS DATA LABS
BENGALURU

Y Vishnuvardhan
Chief Director

hr@exposysdata.com
www.exposysdata.com

DECLARATION

I, **KUSHAL S**, student of VIII semester B.E in Computer science & Engineering at S J C Institute of Technology, Chickballapur, hereby declare that the Internship work entitled “CUSTOMER SEGMENTATION” has been independently carried out by me under the supervision of **Divakar K M,** Assistant Professor, and the coordinator **Swetha T** Assistant Professor, submitted in partial fulfillment of the course requirement for the award of degree in **Bachelor of Engineering** in **Computer Science & Engineering** of **Visveswaraya Technological University, Belgavi** during the year 2021-2022. I further declare that the report has not been submitted to any other University for the award of any other degree.

PLACE:CHIKKABALLAPUR	**KUSHAL S**
Date:13-05-2022	**1SJ18CS049**

ABSTRACT

Looking around and finding that most companies are now data-driven. They make strategic decisions based on data analysis, enabling them to examine and organize their data for better service. There has always been a lot of competition in the market as to who can provide the best customer experience, attract new customers based on their needs, and satisfy their demands, enhancing their profit and growth. However, this is not very easy and calls for various data mining techniques and algorithms.

Machine learning can help them target potential customers. The algorithms deep dive into the data pool to extract hidden treasures and patterns that can bring wonderous profits to many organizations and better decision making. Customer segmentation is one such beautiful concept. Customer segmentation finds its use in many sectors. For example, in Netflix, it can be used as a recommendation system to find a group of similar users and use it for filtering, categorizing, or recommending movies. Banks or insurance companies use it for fraud detection or to evaluate certain insurance risks to segmented customers.

Will be using Customer Segmentation in the retail industry, a Mall, to segment customers into various groups and target potential. The industry can then work towards attractive ideas to sell products and services inclined towards these specific customers.

ACKNOWLEDGEMENT

With reverential pranam, we express my sincere gratitude and salutations to the feet of his holiness **Byravaikya Padmabhushana Sri Sri Sri Dr. Balagangadharanatha Maha Swamiji,** & his holiness **Jagadguru Sri Sri Sri Dr. Nirmalanandanatha Swamiji** of Sri Adichunchanagiri Mutt for their unlimited blessings. First and foremost we wish to express my deep sincere feelings of gratitude to our institution, **Sri Jagadguru Chandrashekaranatha Swamiji Institute of Technology.** For providing me an opportunities for completing my internship work successfully.

I extend deep sense of sincere gratitude to **Dr. G T Raju**, **Principal, S J C Institute of Technology, Chickballapur**, for providing an opportunity to complete the Internship Work.

I extend special in-depth, heartfelt, and sincere gratitude to our HOD **Dr. Manjunatha Kumar B H, Professor and Head of the Department, Computer Science and Engineering, S J C Institute of Technology, Chickballapur,** for her constant support and valuable guidance of the Internship Work.

I convey our sincere thanks to Internship Internal Guide **Guide Name, Designation, Department of Computer Science and Engineering, S J C Institute of Technology,** for his/her constant support, valuable guidance and suggestions of the Internship Work.

I am thankful to Internship External Guide **Y Vishnuvardhan, Founder & CEO, Exposys Data Labs, Bengaluru** for providing valuable guidance and encouragement of the Internship Work.

I also feel immense pleasure to express deep and profound gratitude to our Internship Coordinator **Prof. Swetha T, Assistant Professor, Department of Computer Science and Engineering, S J C Institute of Technology,** for his guidance and suggestions of the Internship Work.

Finally, I would like to thank all faculty members of Department of Computer Science and Engineering, S J C Institute of Technology, Chickballapur for their support.

I also thank all those who extended their support and co-operation while bringing out this Internship Report.

Kushal S(1SJ18CS049)

CONTENTS

LIST OF FIGURES

CHAPTER - 1

COMPANY PROFILE

1.1 History of the Organization

Exposys Data Labs is a world leader in Robotics, Universe Intelligence (UI), Artificial Intelligence (AI) research and its applications that directly impact Planet Earth and human life.

1.1.1 Objectives

Exposys Data Labs aims to Solve real world business problems like Automation, Big Data and data Science. our core team of experts in various technologies help businesses to identify issues,oppurtunities and prototype solutions using trending technologies like AI, ML, Deep Learning and Data Science. we follow a human-focussed and not technology driven approach to achieve success in our clients endeavours.

1.1.2 Operation of the Organization

We are based in Bengaluru India. Exposys Data Labs aims to Solve real world business problems like Automation, Big Data and data Science.

1.2 Major Milestones

- Realistic, Scalable Marketing Strategy
- Profitable Business Model
- Hire and Train a Solid Team
- Gain Authority in Your Industry

1.3 Structure of the Organization

Y Vishnuvardhan.
Founder & CEO
Dr Aravind Kumar.
Research Scientist

1.4 Services Offered

- Software Development
- Web Application Development
- IT Out Source Services
- Internships
- Trainings

CHAPTER – 2

ABOUT THE DEPARTMENT

2.1 Specific Functionalities of the Department

- Exposys Data Labs aims to Solve real world business problems like Automation, Big Data and data Science.
- Our core team of experts in various technologies help businesses to identify issues,opportunities and prototype solutions using trending technologies like AI, ML, Deep Learning and Data Science. we follow a human-focused and not technology driven approach to achieve success in our clients endeavours.

2.2 Process Adopted

The department aims to first understand the user requirements. Further on, a basic structure of the product that needs to be built is drawn and understood. Eventually, the technologies that would best help in developing the product are understood. If the product has database requirements, the schema and the database design are worked upon. The department believes in "Think before you code"- the requirements and logics are first understood over a paper and then are moved to a code form. Agile processes generally promote a disciplined project management process that encourages frequent inspection and adaptation, a leadership philosophy that encourages teamwork, self-organization and accountability, a set of engineering best practices intended to allow for rapid delivery of high-quality software, and a business approach that aligns development with customer needs and company goals. Agile development refers to any development process that is aligned with the concepts of the Agile Manifesto. The Manifesto was developed by a group fourteen leading figures in the software industry, and reflects their experience of what approaches do and do not work for software development.

2.3 Testing

Testing was done according to the Corporate Standards. As each component was being built, Unit testing was performed in order to check if the desired functionality is obtained. Each component in turn is tested with multiple test cases to verify if it is properly working. These unit tested components are integrated with the existing built components and then integration testing is performed. Here again, multiple test cases are run to ensure the newly built component runs in co-ordination with the existing components. Unit and Integration testing are iteratively performed until the complete product is built. Once the complete product is built, it is again tested against multiple test cases and all the functionalities.

The product could be working fine in the developer’s environment but might not necessarily work well in all other environments that the users could be using. Hence, the product is also tested under multiple environments (Various operating systems and devices). At every step, if a flaw is observed, the component is rebuilt to fix the bugs. This way, testing is done hierarchically and iteratively.

2.4 Structure of the Department

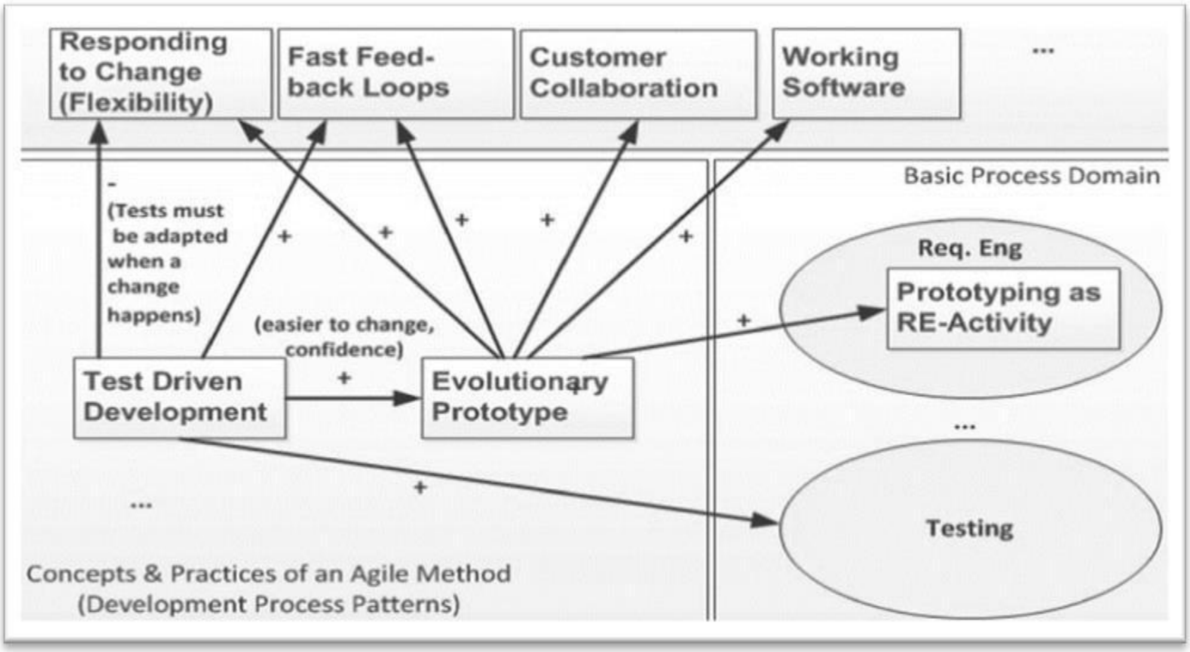

Responding to Change (Flexibility)
Fast Feed-back Loops
Customer Collaboration
Working Software
...
Basic Process Domain
- (Tests must be adapted when a change happens)
+
+
+
+
+
(easier to change, confidence)
+
Req. Eng
Prototyping as RE-Activity
+
Test Driven Development
Evolutionary Prototype
...
+
...
Testing
Concepts & Practices of an Agile Method (Development Process Patterns)

Figure 2.4.1 Structure of the Department

2.5 Roles and Responsibilities of Individuals

Since the internship was remotely conducted by the company, to ensure easy onboarding of interns, the company had individuals who took care of the smooth run of online training.

- Operation and Strategy Head- Ensured there were no difficulties for interns while onboarding. Best of mentors and doubt clarifying sessions were arranged too.
- Technical Lead- Ensured the technicalities of online training to be smooth. Bestplatforms were arranged for our meetings and trainings.
- Mentors- They have helped us to understand the concepts, gave us tasks to get practical take a way and clarified doubts to the best.
- Interns- Worked through the tasks given either individually or in a group.

CHAPTER – 3

TASK PERFORMED

Internship was on Data Science.

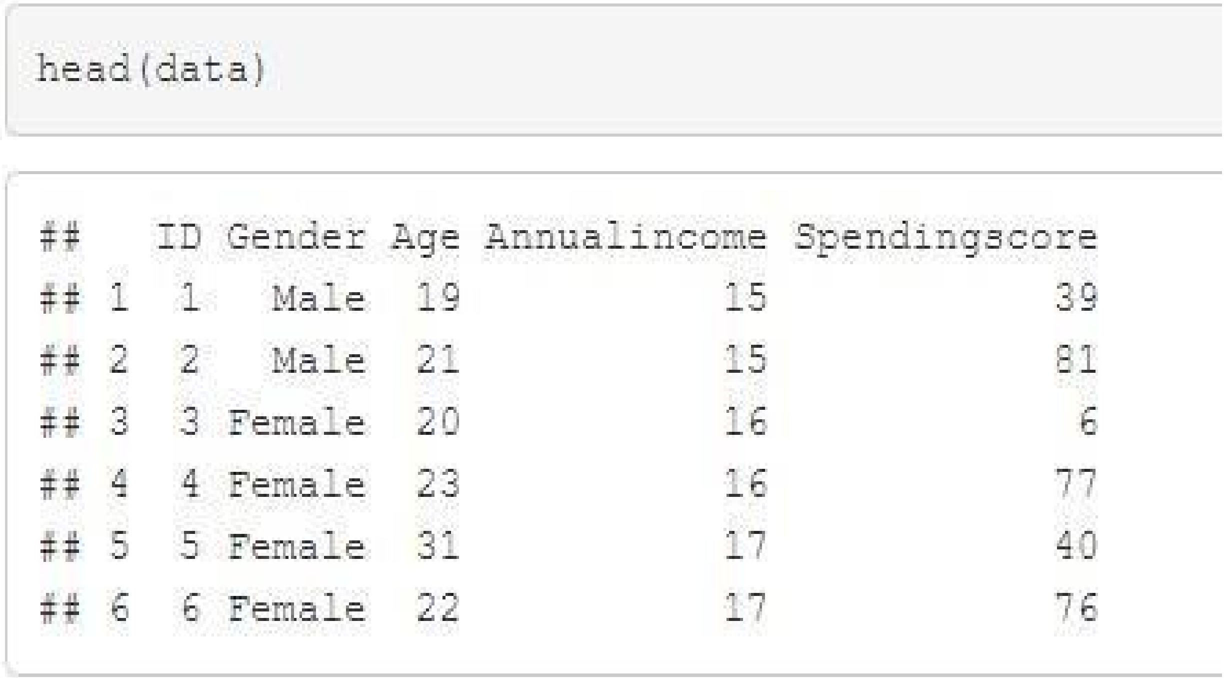
Training Program

The internship is a platform where the trainees are assigned with the specific task. In the initial days of the internship, I was trained on the following:

- Python Programming
- Machine Learning Algorithms

DATA SET:

For this project we have used Mall Customer Dataset, our main objective is to divide customers into groups according to common characteristics.



```
head(data)
```

```
##   ID Gender Age Annualincome Spendingscore
## 1  1   Male  19           15            39
## 2  2   Male  21           15            81
## 3  3 Female  20           16             6
## 4  4 Female  23           16            77
## 5  5 Female  31           17            40
## 6  6 Female  22           17            76
```

Table: Fields in Mall Customer CSV File

DATASET EXTRACTION AND TRANSFORMATION:

We imported the dataset and using Elbow Method and K-means. we identified the segments of customers to target the potential user base. By observing these customers are divided into groups according to common characteristics.

Elbow Method:

In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use. The same method can be used to choose the number of parameters in other data-driven models, such as the number of principal components to describe a data set.

K-Means:

K-Means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. k-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. For instance, better Euclidean solutions can be found using k-medians and k-medoids.

The problem is computationally difficult (NP-hard); however, efficient heuristic algorithms converge quickly to a local optimum. These are usually similar to the expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both k-means and Gaussian mixture modeling. They both use cluster centers to model the data; however, k-means clustering tends to find clusters of comparable spatial extent, while the Gaussian mixture model allows clusters to have different shapes.

The unsupervised k-means algorithm has a loose relationship to the k-nearest neighbor classifier, a popular supervised machine learning technique for classification that is often confused with k-means due to the name. Applying the 1-nearest neighbor classifier to the cluster centers obtained by k-means classifies new data into the existing clusters. This is known as nearest centroid classifier or Rocchio algorithm.

CHAPTER – 4

REFLECTION NOTES

4.1 Experience

As per our experience during the internship, Exposys Data Labs follows a good work culture and it has friendly employees, starting from the staff level to the management level. The trainers are well versed in their fields and they treat everyone equally. There is no distinguishing between fresher graduates and corporates and everyone is respected equally. There is a lot of teamwork followed in every task, be it hard or easy and there is a very calm and friendly atmosphere maintained at all times. There is a lot of scope for self-improvement due to the great communication and support that can be found. Interns have been treated and taught well and all our doubts and concerns regarding the training or the companies have been properly answered. All in all, Knowledge Solutions India was a great place for a fresher to start career and also for a corporate to boost his/her career. It has been a great experience to be an intern in such a reputed organization.

4.2 Technical Outcomes

4.2.1 System Requirements and Specification

HARDWARE REQUIREMENTS:

- Processor : x86 or x64
- Hard Disk : 256 GB or more.
- Ram : 2 GB or more

SOFTWARE REQUIREMENTS:

- Operating System : Windows or Linux
- Tools used : Anaconda, Google Collab

4.3 System Analysis and Design

4.3.1 Existing System

In the existing system the analysis and segmentation of the customers was done by visualizing the graphs by clustering the two or more feature. Using the basic graphs analysis can be done up to some data.

4.3.2 Disadvantages of the Existing System

- Data Points are overlapped.
- The data does not hold some patterns..
- Data clusters were Inappropriate for accurate considerations.

4.3.3 Proposed System

In the proposed system to get accurate result we used K-Means Algorithm was applied. In K-Means the data points were grouped accurately depending on the value of k the no. of clusters were identified. To find optimized k value Elbow method is used.

4.3.4 Advantages of the Proposed System

- Determine appropriate product pricing.
- Develop customized marketing campaigns.
- Design an optimal distribution strategy.
- Choose specific product features for deployment.
- Prioritize new product development efforts.

4.4 System Architecture

4.4.1 Data Flow Diagram

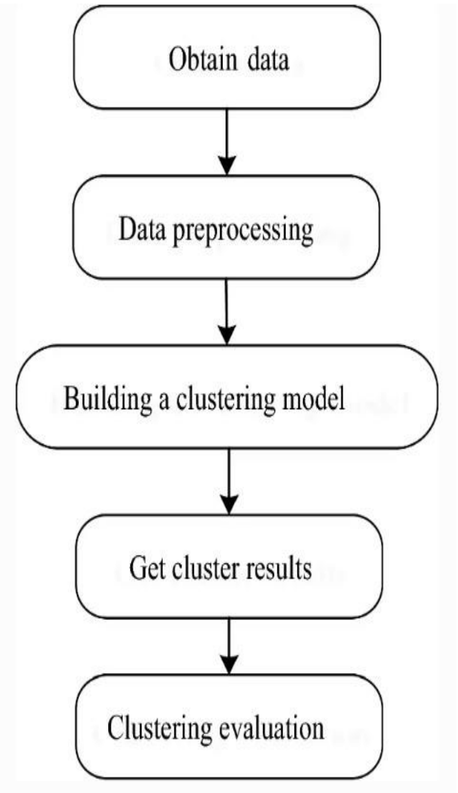


Figure 4.4.1.1 Data Flow Diagram

4.5 Implementation

Customer segmentation

Customer segmentation is partitioning a customer database into group of people with similar characteristics. It is an application of unsupervised learning. It is a business strategy that allows targeting a specific group of customers and effectively allocate marketing resources. For such large datasets, we need an analytical approach like clustering to make customer segments.

There are four major ways of segmentation, i.e., geographical, economic, demographic, and behavioural patterns.

In this project, we divide a Mall customer's dataset based on gender, age, income, spending habits, etc. We also visualize gender and age distributions and analyse their annual incomes and spending scores to target the potential user base. The method used is K-means clustering.

Language: Python

Supervised and Unsupervised Learning

Supervised learning is training a model with labelled data. There are two types regression and classification. Regression is the process of predicting a continuous value as opposed to predicting an absolute value in classification. In classification, the class is predefined and predict categorial classed labels. Classification approaches include decision trees, logistic regression to predict the default value of the new entry.In unsupervised learning, the model discovers information on its own. There is no prior information on the data or the outcomes of the analysis. Dimension reduction, density estimation, market basket analysis, and clustering are the most widely used unsupervised machine learning techniques. Generally, clustering is used for exploratory data analysis, summarisation, dimension reduction, outlier detection, and other such data mining tasks.

In comparison to supervised learning, unsupervised learning has fewer models and fewer evaluation methods that can be used to ensure that the outcome of the model is accurate. As such, unsupervised learning creates a less controllable environment as the machine is creating outcomes for us.

Clustering

Clustering can group data unsupervised solely based on similarities to each other. It will partition customers into mutually exclusive groups aka clusters. Having the result would help understand and predict customer preferences and differences, thus making the company deliver personalised experiences for each group of customers.

Types of Clustering:

Partition-based clustering is a group of clustering algorithms that produces sphere-like clusters, such as; K-Means, K-Medians or Fuzzy c-Means. These algorithms are relatively efficient and are used for medium to large- sized databases.

Hierarchical clustering algorithms produce trees of clusters, such as agglomerative and divisive algorithms. This group of algorithms are very intuitive and are generally suitable for use with small-size datasets.

Density-based clustering algorithms produce arbitrary-shaped clusters. They are outstanding when dealing with spatial clusters or noise in the data set, for example, the DB scan algorithm.

4.6 Screen Shots

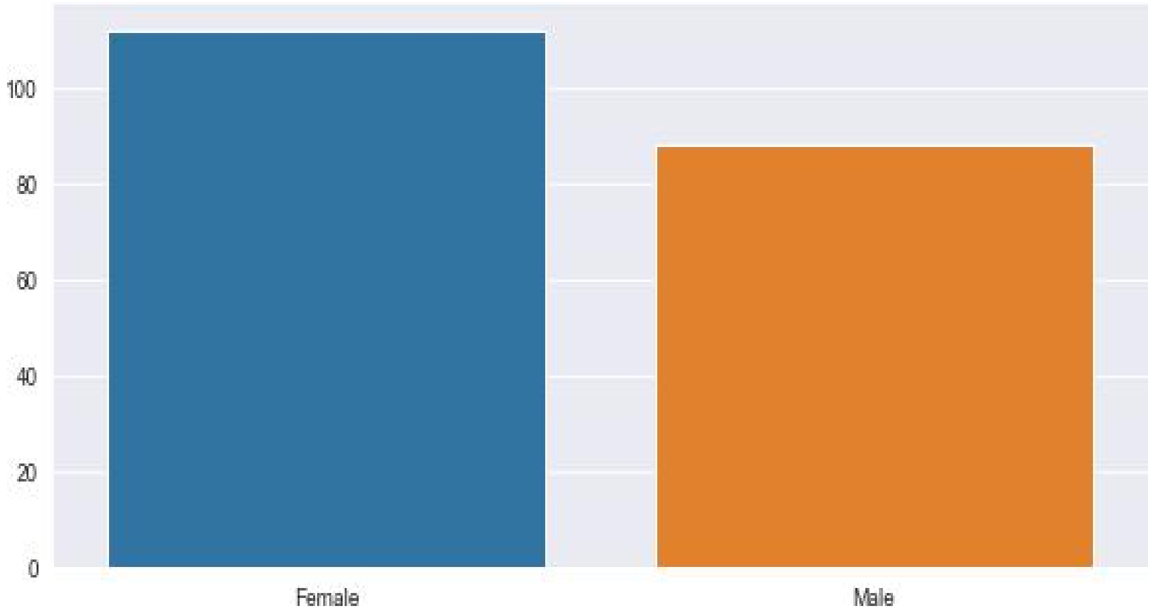


Figure 4.6.1 Gender Distribution

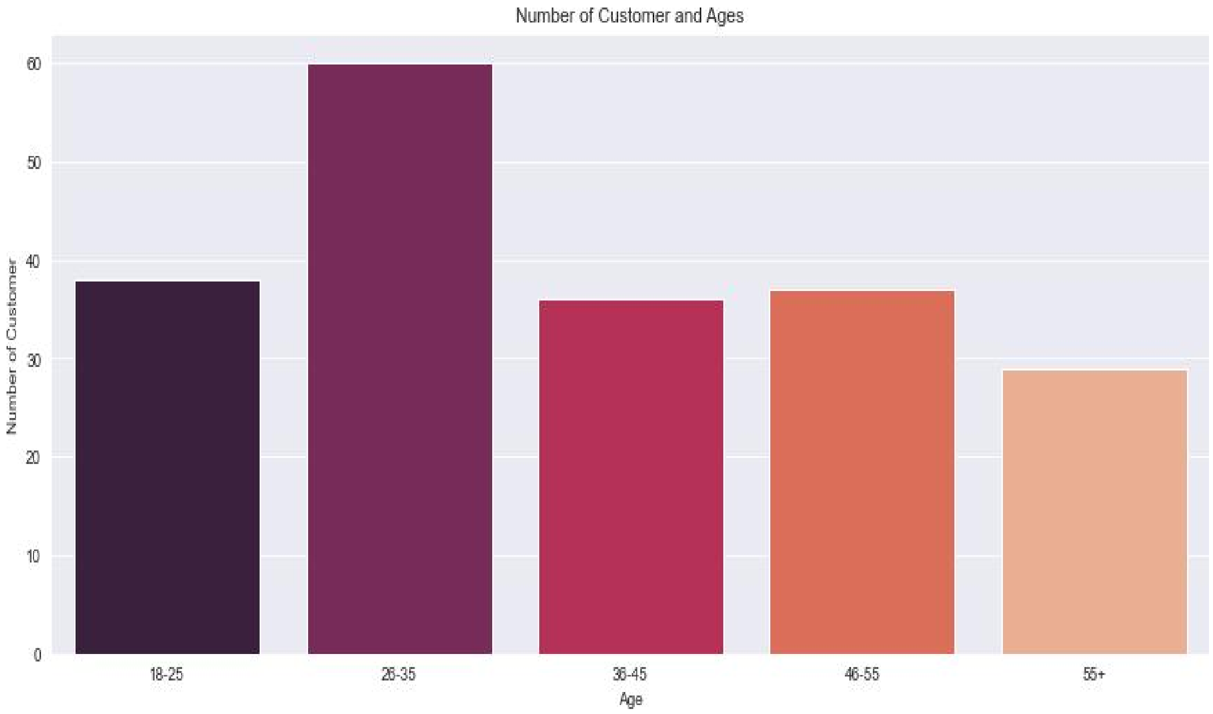


Figure 4.6.2 Age Distribution

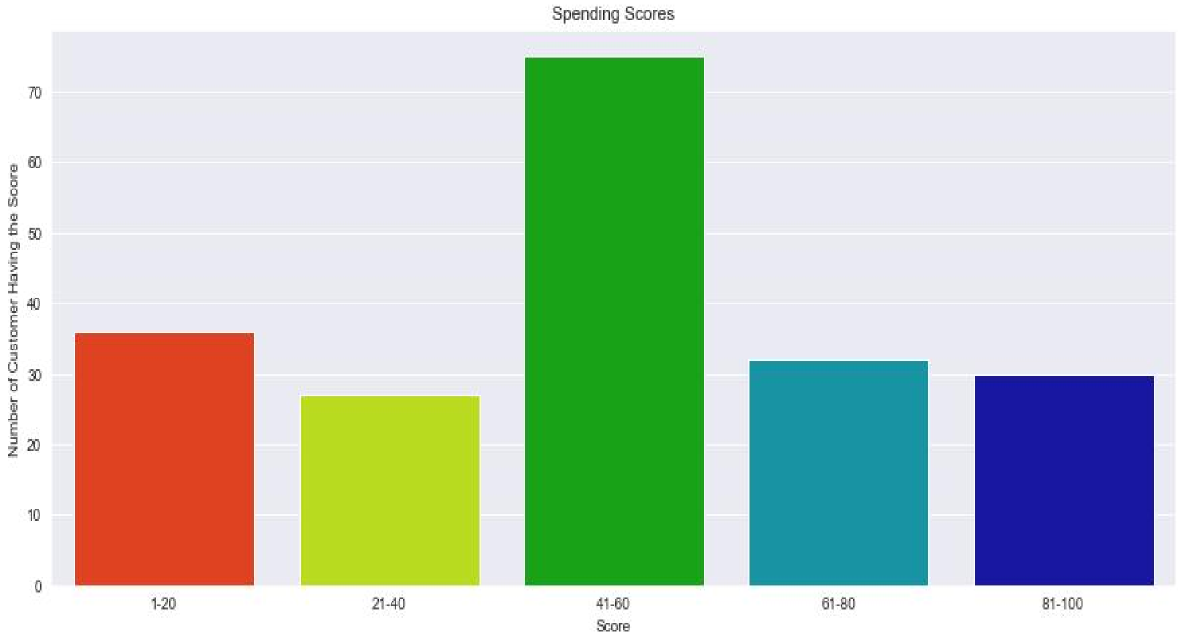


Figure 4.6.3 Annual Income Distribution

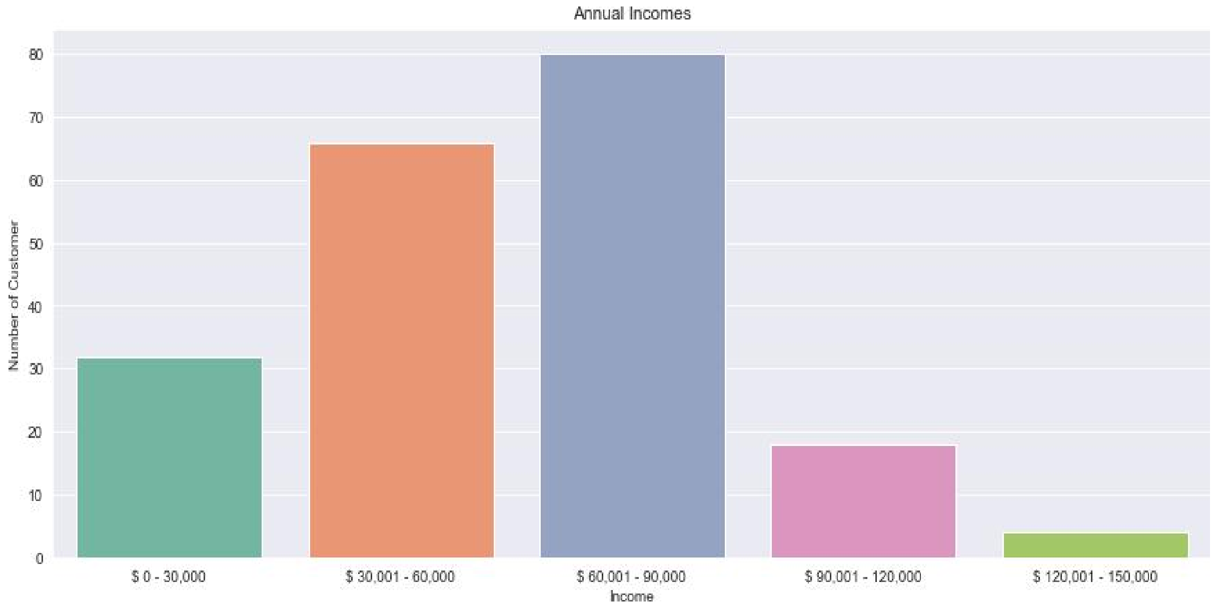


Figure 4.6.4 Spending score Distribution

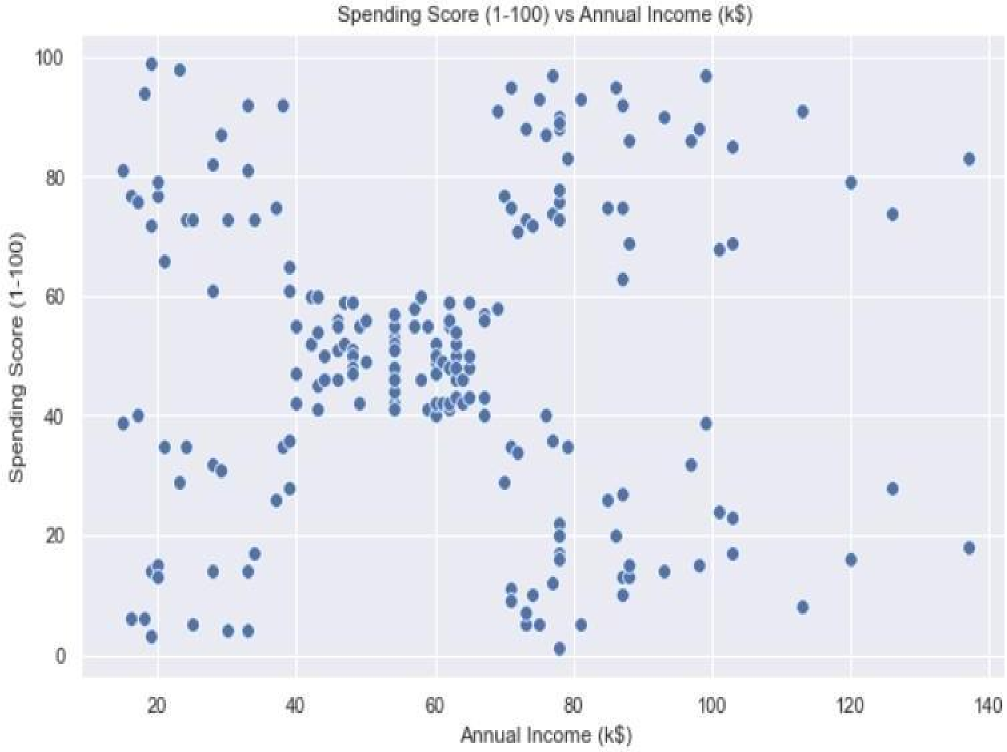


Figure 4.6.5 Annual Income vs Spending Score

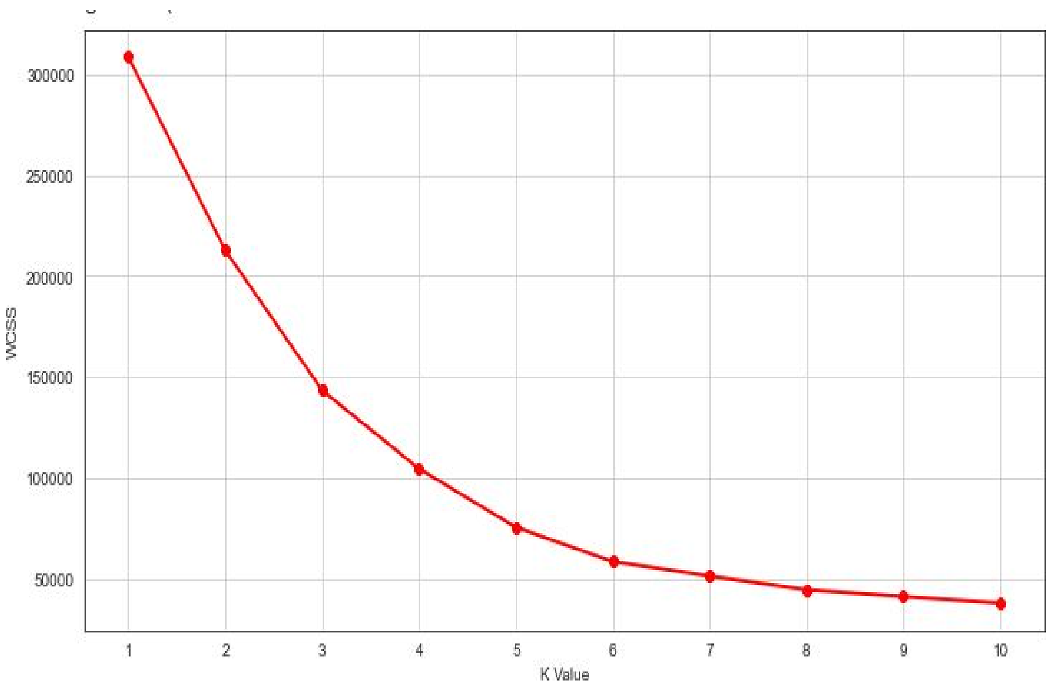


Figure 4.6.6 Elbow Graph

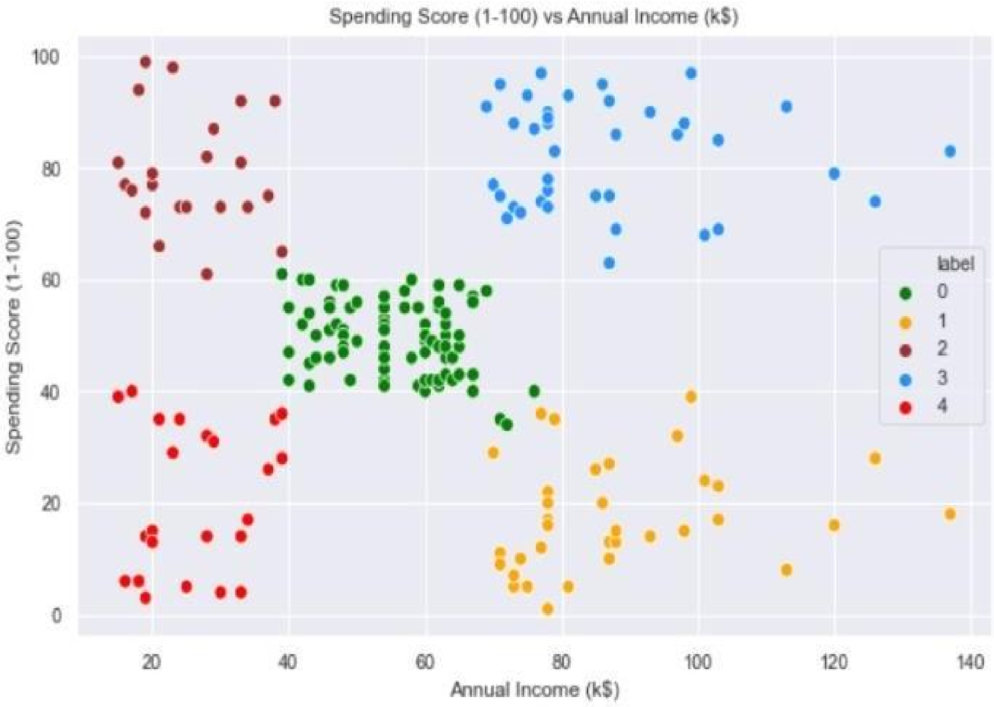


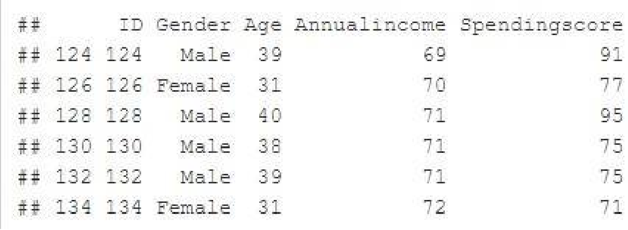
Figure 4.6.7 Graph after applying K-Means

CHAPTER – 4

CONCLUSION

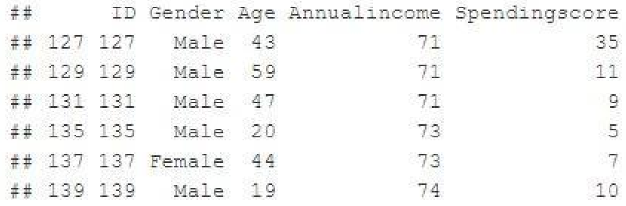
The goal of K means is to group data points into distinct non-overlapping subgroups.

Cluster 3: high spending scores and high-income; alert them with new arrivals as they are potential customer for increase in revenue.



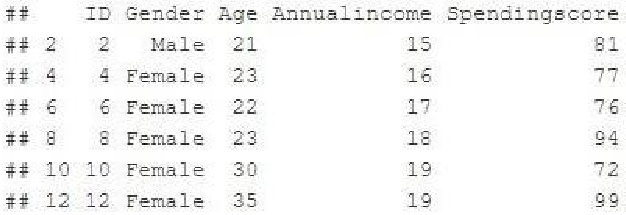
```
##       ID Gender Age Annualincome Spendingscore
## 124 124   Male  39           69            91
## 126 126 Female  31           70            77
## 128 128   Male  40           71            95
## 130 130   Male  38           71            75
## 132 132   Male  39           71            75
## 134 134 Female  31           72            71
```

Cluster 1: high income and low spending score; ask them for feedback and advertise them with new products that might attract them, they have the potential to convert into cluster 4.



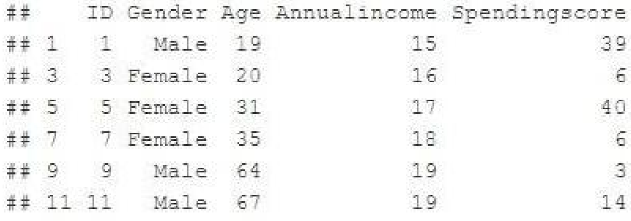
```
##       ID Gender Age Annualincome Spendingscore
## 127 127   Male  43           71            35
## 129 129   Male  59           71            11
## 131 131   Male  47           71             9
## 135 135   Male  20           73             5
## 137 137 Female  44           73             7
## 139 139   Male  19           74            10
```

Cluster 2: low income and high spending scores; can help them by providing new deals and sales offers so that despite low income they still remain loyal.



```
##     ID Gender Age Annualincome Spendingscore
## 2    2   Male  21           15            81
## 4    4 Female  23           16            77
## 6    6 Female  22           17            76
## 8    8 Female  23           18            94
## 10  10 Female  30           19            72
## 12  12 Female  35           19            99
```

Cluster 4: low income and low spending score; it won’t be beneficial to both the parties to target these customers.



```
##     ID Gender Age Annualincome Spendingscore
## 1    1   Male  19           15            39
## 3    3 Female  20           16             6
## 5    5 Female  31           17            40
## 7    7 Female  35           18             6
## 9    9   Male  64           19             3
## 11  11   Male  67           19            14
```

Rest are average and the company can use them according to market conditions.

BIBLIOGRAPHY

[1] Al-Qaed F, Sutcliffe A. Adaptive Decision Support System (ADSS) for B2C E Commerce. 2006 ICEC Eighth Int Conf Electron Commer Proc NEW E-COMMERCE Innov Conqu Curr BARRIERS, Obs LIMITATIONS TO Conduct Success Bus INTERNET. 2006:492-503.

[2] Mobasher B, Cooley R, Srivastava J. Automatic Personalization Based on Web Usage Mining. Commun ACM. 2000;43(8).

[3] Cherna Y, Tzenga G. Measuring Consumer Loyalty of B2C e-Retailing Service by Fuzzy Integral: a FANP-Based Synthetic Model. In: International Conference on Fuzzy Theory and Its Applications iFUZZY.; 2012:48-56.

[4] Magento. An Introduction to Customer Segmentation. 2014. info2.magento.com/.../ An_Introduction_to_Customer_Segmentation..

APPENDIX

Appendix A: Abbreviation

AI: Artificial intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. The term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving.

ML: Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being

explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values.