Finding Ideal Geographical Location for Businesses using Machine Learning Technique

by

Mir Ibtid Mahmud
17101351
Onez Chowdhury
21341063
Yashwant Alvee
16341010
Tawsif Sadman
17101107
Imrul Haque Shaon
17301045

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
September 2021

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Ibtid

Mir Ibtid Mahmud
17101351

Onez

Onez Chowdhury
21341063

Yashwant

Yashwant Alvee
16341010

Tawsif

Tawsif Sadman
17101107

Shaon

Imrul Haque Shaon
17301045

Approval

The thesis titled “Finding Ideal Geographical Location for Businesses using Machine Learning Technique” submitted by

1. Mir Ibtid Mahmud (17101351)
2. Onez Chowdhury (21341063)
3. Yashwant Alvee (16341010)
4. Tawsif Sadman (17101107)
5. Imrul Haque Shaon (17301045)

Of Summer, 2021 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on September 26, 2021.

Examining Committee:

Supervisor:
(Member)

Faisal Bin Ashraf
Lecturer
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Dr.Md.Golam Rabiul Alam
Associate Professor
Department of Computer Science Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi
Chairperson, School of Data and Sciences
Department of Computer Science and Engineering
Brac University

Abstract

The world has become a place where the economy is at the epicenter of it all. World economic growth has paved the way for people to enrich their lives with all sorts of blessings. A major chunk of this shift in the world's treasury all comes from the tireless endeavors of affluent and flourishing businesses. The more a business thrives, the more economic sustainability it brings upon the society. And one of the key factors of building up a thriving business is what motivated us to forgo on our research. Location analytics in recent times plays an important role in making a sustainable and profitable business. Very often trade and commerce rely on uninformed struggles to analyze the perfect location for their establishment. Hence, we used unsupervised learning to evaluate a dataset and create a decision making model to accurately investigate whether a location will be befitting for a particular business model based on customer behaviour and interests. As such we built a dataset concentrating on questionnaire responses taken via online survey and tested our model on the gathered data. We exercised the analyzed dataset in different clustering algorithms such as kmeans, minibatch kmeans and hierarchical clustering. Finally, using a decision tree model, we were able to extract an explanatory rule in terms of quantitative values, which were further discerned in making an elaborate assumption. Hence, based on consumer habits and interests, we concluded that we could analyze which location or marketplace would be better suited for which accessories a seller is selling, and hence suggest the perfect location for his business to thrive upon.

Keywords: Machine Learning; Data Analysis; Geo Analytics; Decision tree; Linear Regression Analysis; K-Means Clustering

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption. Secondly, to our advisor Mr. Faisal Bin Ashraf sir for his kind support and advice in our work. He helped us whenever we needed help. And finally to our parents without their throughout sup-port it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

Table of Contents

Declaration **i**

Approval **ii**

Abstract **iii**

Acknowledgment **iv**

Table of Contents **v**

List of Figures **vii**

List of Tables **1**

1 Introduction **2**

1.1 Background Information 2

1.2 Problem Statement 3

1.3 Research Motivation 4

1.4 Research Objectives 4

2 Related Work **6**

2.1 Literature review 6

3 Proposed Methodology **10**

3.1 Research Framework 10

3.2 Machine Learning 11

3.3 Unsupervised Learning/Clustering 11

3.3.1 K-Means Clustering 12

3.3.1.1 Principal Component Analysis 13

3.3.1.2 Elbow Method 14

3.3.2 Mini-batch K-Means Clustering 15

3.3.3 Hierarchical Clustering 16

3.4 Clustering Algorithm Evaluation Method 17

3.4.1 Davies Bouldin Score 17

3.4.2 Silhouette Score 18

3.4.3 Davies Bouldin VS Silhouette Score 20

3.5 Cluster Analysis 20

4 Dataset & Experiment **23**
4.1 Data Collection 23
4.1.1 Collection Procedure 23
4.1.2 Dataset Features 24
4.1.3 Data Analysis 25
4.2 Data Initialization 30

5 Experimental Result & Analysis **32**
5.1 Elbow Method Result 32
5.2 Data Representation 32
5.2.1 K-Means Clustering 33
5.2.2 Mini-batch K-Means Clustering 33
5.2.3 Hierarchical Clustering 34
5.3 Results - Clustering Comparison 35
5.4 Results - Cluster Analysis 36
5.5 Performance Evaluation 40

6 Conclusion & Future Works **41**
6.1 Conclusion 41
6.2 Future Works 41

Bibliography **43**

List of Figures

3.1 Research Framework . . . 10
3.2 Clustering Example . . . 12
3.3 How K-Means Works . . . 13
3.4 PCA Function . . . 14
3.5 Graph of Elbow Method . . . 15
3.6 Mini-Batch K-Means Clustering . . . 16
3.7 Hierarchical Clustering . . . 17

4.1 Gender . . . 25
4.2 Income Range . . . 26
4.3 Consumer Taste On The Type of Products . . . 26
4.4 Medium of Shopping . . . 26
4.5 Budget of Respondents . . . 27
4.6 Preferred Areas to Shop Most Often . . . 27
4.7 Ambience Rating . . . 27
4.8 Transportation Convenience . . . 28
4.9 Time Preference for Shopping . . . 28
4.10 Impact of Specific Brands . . . 29
4.11 Agreement of Annoyance on not Finding Specific Brands . . . 29

5.1 Graph of Elbow Method . . . 32
5.2 3D Representation of K-Means . . . 33
5.3 3D Representation of Mini-Batch K-Means . . . 34
5.4 Dendrogram Graphical Representation of Hierarchical Clustering Segmentation . . . 35

List of Tables

3.1 Difference between Davies Bouldin and Silhouette Score 20

4.1 Age Range 30
4.2 Gender 30
4.3 Monthly Income 30
4.4 Preferred Medium 31
4.5 Spend 31
4.6 Shopping Partner Preference 31
4.7 Nearby or Look Elsewhere 31
4.8 Shopping or Not 31
4.9 Acceptable Time 31

5.1 Silhouette and Davies Bouldin Scores for K-Means Clusters 33
5.2 Silhouette and Davies Bouldin Scores for Mini-Batch K-Means Clusters 34
5.3 Scores for Hierarchical Clustering 35
5.4 Comparing Three Clustering Algorithms Based on Their Davies Bouldin and Silhouette Scores 35
5.5 Rule List 37
5.6 Preferred Shopping Mall 38

Chapter 1

Introduction

1.1 Background Information

In 1854 the first instance of maps being used for geospatial analysis was recorded. A map was developed by John Snow in which he used geo-location to assess the origin of London's cholera outbreak. His map shows the location of a well where sick local people went to fetch water and how all of them were related to a single water pump. The locations indicated that cases were clustered around certain intersections-thus exposing both the problem and the solution[1].This is one of the early examples where location analysis was used to figure out a major problem troubling the human society-out of the box solution never approached before. With time, research and development progressed which introduced new techniques to solve traditional problems increasing efficiency in human life. Machine Learning is an important tool frequently deployed in business, research works. Using algorithms, neural networks, it helps computers to improve their output performance. Machine learning algorithms construct a mathematical model using training data. The ultimate aim is to take decisions without being programmed to make them. Machine learning is based on the part of a model of cell and brain interaction. This model was created by Donald Hebb in 1949. In the 1950s, Arthur Samuel from IBM made a computer program that was used for playing numerous mechanisms which allowed the program to become better[2]. He is hailed in the industry as the pioneer of the tool designed. During his IBM days he made computer learning programs-the first to do so. They were made to run the game called checkers. His program was unique because every time checkers was played, the computer would get better at rectifying mistakes that were made earlier and sought efficient paths to win. The data stored earlier helped it. This is an earlier example of using machine learning.

Since then machine learning has become one of the most significant inventions made. As businesses became more competitive, machine learning is helping companies to undergo massive changes slowly taking steps as the dawn of automation unveils ahead. The impact machine learning algorithms displayed so far is well defined as more businesses use them at large scale operations. Today most applications and software use it. Artificial intelligence includes subcategories which consist of machine learning ,deep learning. Deep Learning is a special version of machine learning. It uses more complex processes to solve difficult tasks. Machine learning is probabilistic. On the other hand Deep learning can be stated to be deterministic.

When machine learning was not prominent, artificial intelligence was used to do low-level tasks like automation, classification etc. Machine Learning empowered computers to get better with each iteration with the capability to evolve. Machine learning algorithms process large amounts of data and then bring out the necessary relevant information. Machine learning algorithms are trained through three main methods mostly nowadays. These are: supervised learning, unsupervised learning, and reinforcement learning. Machine learning has two kinds of data known as labeled/unlabeled data. Labeled data has both input, output in machine-readable form but requires major human involvement. Unlabeled data only has maximum one parameter which the machine can read, hence mitigating the need for human intervention demanding complex results. Many machine learning algorithms are available to be used for specific tasks. For our research we shall be focusing only on unsupervised machine learning[3]

Unsupervised Learning is one of the machine learning methods where we do not have to supervise. The system works by itself through finding patterns and discovering information previously not recognized. Unsupervised learning deals with unlabeled data. It also allows users to do complicated processing tasks. Unsupervised learning in comparison to other natural learning methods is more unpredictable. In 1930 Harold Driver and Alfred Kroebar (American Anthropologists) collected statistical data from ethnographic analysis that was carried out on Polynesian cultures. Clustering algorithm was described in their book named “Quantitative Expression of Cultural Relationships”. Following next ten years two psychologists named Joseph Zubin and Robert Tryon introduced clustering analysis in the field of psychology and it was deployed to classify traits of personality[4]. Unsupervised learning models can be used for many tasks such as association, clustering and reduction. Our work is a case of clustering.

Clustering could be a significant unsupervised learning issue. Here the work is to find structure in unlabeled data. Hence clustering could be defined to organize objects into groups solely based on similarity[5]. It seeks similarity through features like shape or size or color or perhaps behavior. Clustering is majorly used for statistics and data evaluation. Clustering profile attributes very quickly and easily. This enables users to sort data and pin down their relevant groups. It also helps companies to approach customers with specific products that aligns with their taste hence maximizing profits.

Location analytics is a modern technique that combines geographical and statistical tools to find the best possible location for a business. In our research we used unsupervised machine learning to solve the problem of finding the ideal geographical location for entrepreneurs.

1.2 Problem Statement

Entrepreneurs use demographical data traditionally to plan for a new store location. This selection is a decision problem where the main aim is to maximize profit, reduce cost through monotonous statistical forecasting. In many situations, these forecast results are not reliable due to measurement anomalies. Adding to that is the expense the process requires. Today, Machine Learning makes it possible to bring out intelligent outcomes from complex data. Machine Learning tools help to collect, store, analyze and visualize data. The purpose of our research is to use these tools to yield more efficient and accurate decision-making outcomes by using Clustering. The end goal is to present a successful model to predict new store locations. We can represent our work through the following function:

$$f(C) = R \epsilon Z \tag{1.1}$$

Where,
C = Set of Features
Z = Set of Rules

The function shown above precisely projects our research objective. The categorical features we derive from individuals through survey shall be used as inputs to find out the best possible location for entrepreneurs that sell relevant commodities. Machine learning techniques will be used to solve the problem of finding the best location for them. Traditional methods are time consuming and involves higher costs.

1.3 Research Motivation

In a world where economic stability is highly dependent on thriving businesses, we were very disappointed to realize that most businesses hindered the possibility of a stable profit margin due to improper decision making when it came to setting up a physical establishment for their enterprises. The most common example came to our view when we surveyed and found out that most of the shops in the big shopping malls of our country were barely able to make rent from their businesses despite being in a prominent location. Less customer traffic and marginal sales played a role even when their shop was situated in some of the biggest shopping hubs of the country. We talked to shop managers and learnt that even when having quality merchandise, they experienced less customer walkthroughs. In conclusion, we found out that some shops in big shopping malls were just there - sitting idle and occupying space in a prominent place without adequate return. We could not imagine the economic downfall this might bring when more shops start to open up in locations where they experience the same lack of customer interaction and sales, especially in a post pandemic world where physical shops will soon be opening up again. Hence we became determined to find a solution that will help better the decision making when trying to figure out a location for setting up shop for businesses. We urged ourselves to prevent any further economic downplay in our society, even if the outcome is marginal. Because every small step taken is a step taken for a bigger cause.

1.4 Research Objectives

Our research work has objectives to obtain. They are pointed below:

- **Mitigates the need for fieldwork.**
  Earlier location selection involved traversing through places to gather information and then find a pattern manually which is taxing.

- **Optimizing the procedure of site selection for businesses.**
  Using unsupervised clustering we were able to find a model that optimizes site selection more effectively and at a lower cost.

- **Reduce cost for Entrepreneurs.**
  As explained above the process to travel and gather information requires hiring of more labour.This increases operational costs of entrepreneurs.

- **Increase statistical efficiency from traditional systems.**
  Machine learning techniques significantly increase the quality of statistical outcome with better insights on locations to select.Computers outperform humans on specific unitasks.

- **Faster prediction hence time efficiency.**
  All the points discussed above indicate that as machine learning techniques are used to find location the task takes less time than traditional methods.

Chapter 2

Related Work

2.1 Literature review

The large and comprehensive study on Location Analytics undertaken by scholars all around the world is extensive and comprehensive.

A study paper on making site selection using machine learning was released by students from Sinhgad Institute of Technology. Their goal was to develop a model that would help businesspeople in a huge country like India invest in a specific place. A restaurant, a clothing store, and a grocery shop have been used as examples. They developed an algorithm to forecast a site's ranking based on profitability. According to this article, finding a good location for company expansion is challenging owing to fast urban population growth, which causes unplanned sprawl, inadequate housing amenities, traffic congestion, and other issues. Tools like GIS and AHP were used to conduct the site suitability analysis. The Machine Learning Model will next compute the likelihood Y of profit for the site by calculating the weights of various parameters. It is vital for firms to locate their operations in the best possible location. Choosing the ideal place necessitates extensive fieldwork. The researchers used AHP to present a model for predicting hospital site selection in their survey (a theory of measurement through pairwise comparisons). AHP is a model for making decisions. Various systems were addressed, including Google API, GIS, geographical data, meteorological observations, and so on. They presented a system that consisted of a blend of diverse algorithms, such as AHP and Random Forest, after exhaustive literature research. In addition, a multilinear classifier and a decision tree were deployed. Surveys from companies were used to collect data. The model only used data that has already been validated as input. They used AHP to prepare the dataset for Model Analysis. The researchers next did a Model Comparison. They employed three different algorithms: Random Forest, Multi-Linear Regression, and Decision Tree. Scikit-learn was used to code. To sum up, the paper puts forward a system that provides the degree of profitability of a given business location. The best Machine Learning Algorithms were used to predict the degree of profitability of the business site. Future scopes and limitations were discussed[6]

Jesse K. Pearson worked on a site location feasibility using the GIS. This article used Geographic Information Systems (GIS) to conduct a site-location study to determine where future Kowalski shops may be placed. The prospective market research was based on Kowalski's Markets' Demographic data from their two most successful retail locations. A comparison was done between the final GIS analysis and the gravity model, in which both location analysis approaches were employed to give findings and site selection suggestions for places with the highest market potential. The purpose of the site selection investigation was to offer Kowalski with four or five potential store locations. GIS was utilized in the initial portion of the research to find possible store sites based on important demographic data from two successful stores. The gravity model is then compared to the final GIS. To summarize their findings, the GIS Analysis at the conclusion proposes that certain site choices would be acceptable for the construction of new stores. The GIS and gravity model findings were used to determine the final suitability score[7]

IMT Institute (Italy), Cambridge University, and Queen Mary University collaborated on another groundbreaking study on Geo-spotting. They mine characteristics based on two factors: geography and user mobility. It can be observed that when numerous characteristics are included in supervised learning algorithms, performance increases dramatically, implying that the success of a retail firm may be determined by a variety of things. The authors define this topic as a data-mining job in which a set of characteristics is extracted and used to evaluate the retail quality of a geographic region. Popularity of retail stores in terms of location and mobility. Finally they merged diverse data in a series of supervised learning algorithms projecting that combining geographic and mobility information may better describe location popularity[8].

Tsinghua University students did research on demand-driven retail location selection using diverse spatial-temporal data. Traditional techniques for selecting a desirable site are time-consuming and ineffective for the changing market. They mined data from Baidu Maps and they suggested methodology in this study. Their concept blends the geographic distribution of customer needs with each location's attractiveness and economic characteristics. Their research is based on fieldwork in Changchun, Jiangsu Province, where they planned and collected data. The GIS platform is used in this article to investigate issues such as population density, transit availability, current rivals, and alternative sites. To do so, they employ the Huff model[9].

Aditi has written another essential study about reserving a hotel online, which is a difficult undertaking given the hundreds of hotels available in any destination. They are in charge of making hotel recommendations to users. They utilized Expedia's hotel suggestion dataset, which includes a number of factors that assist users in selecting their favourite hotels. Their goal is to forecast and propose five hotel clusters to a user out of a hundred different clusters that the user is more likely to book[10].

Luyao Wang, Hong Fan, and Yankan Wang have written an important work on site selection based on availability of space. They deal with the issue of site selection. They suggest a two-step hybrid approach for small retail businesses in the study, which includes geographic accessibility evaluation and market potential calculation. Comparing the PCA-BP model to other existing regression techniques, they found it to be adequate. Their proposed strategy aids retail chains in better arranging their company locations[11].

Ashok Kumar and Shiva Shankar G co-authored a significant work in which they developed an algorithm for determining the optimal location to launch a business where demand is strong and supply is limited. They used average service time to gauge the quality of recommendations and developed an algorithm that outperforms the KNN method[12].

An overview on analysis for hotel recommendation systems using Machine Learning was conducted by the students of University of Birmingham, United Kingdom with rapid development of advanced machine learning programs, which is being used for tourists and hotels business which is considered one of the most powerful industries nowadays. Since the majority of labor is done online, this scientific technique of 'recommender systems' has emerged.

Customers may use this technology to discover hotels based on their preferences and previous experiences. Machine learning is used to provide reliable suggestions for future potential consumers using diverse and large-scale data. Their system works by sifting through a large group of people and finding a smaller group of clients with similar likes to a given client. This study also considers the location, the types of visitors, and the surrounding environment, as well as methods such as lexical analysis, syntactic analysis, semantic analysis, and NLTK for determining the polarity of textual reviews. The suggested hotels recommender system is written in Python and Java code and uses Smarty formats to get the best choices for the best hotels. The program may be accessed using any web browser, and the client may get information online from any device; it is built on robust open-source technologies. Recommender systems can boost their clients' loyalty by improving the client experience and recommending additional programs to them. Thus, numerous organizations are putting forth a lot of effort to set up and improve their own business recommender systems[13].

Our research work proposes a model based on unsupervised machine learning tasks called clustering.Several research papers have harnessed this technique to solve problems across different domains which are discussed below.

A collaborative research has been conducted by researchers from Brac university, Southeast university and University of North Texas, USA on climate data analysis. The main aim of analyzing climate data is to understand conditions and phenomenon concerning atmosphere interconnecting various aspects of the environment. As humanity experiences frequent weather changes this research is significant as it provides insights on how to keep our socioeconomic value intact. The researchers propose a model here in order to scrutinize climate data from Bangladesh over a timeline comprising six decades that were gathered from 11 weather stations spread across the country.The objective is to find precious patterns of climate change of different areas in Bangladesh. The proposed structure consists of correlation calculation, hierarchical clusters formations across the areas under contemplation, measurement of change for individual areas and finally correlation calculation to conclude with validity.The paper also uses data mining methods to provide result analysis to wrap it up[14].

Students from Brac University has used machine learning techniques in social media domain to detect fake profile with the help of image processing. As social media platforms like Instagram, Facebook etc has taken over consumers by storm connecting the world together unfortunately some misuse it. Accounts are made using other individual's information with harassment motives, widespread of fake information etc resulting panic among masses. In order to identify fake accounts and also halt users to make new fake accounts the researchers used a deep learning algorithm. K-Means algorithm was applied for identifying fake data inputs in a dataset[15].

Yasmeen Farouk and Sherine Rady, faculties from Computer and Information Sciences of Ain Shams University has conducted a research on early diagnosis of alzheimer's disease by using unsupervised clustering.As we know alzheimer's disease is a brain disorder that is progressive in nature combined with dementia.Early diagnosis of the disease by undergoing MRI helps patients to halt further brain deterioration.This research work uses unsupervised clustering to detect early AD signs.They also address the issue of inaccuracies from labeled data that creates a problem. Here K-Means and K-medoids are compared with Voxel Based features extraction from MRI images.At the end they were able to show that their proposed model managed to detect AD early with 76% precision[16].

An important research was carried out by researchers from University of Joseph Fourier, France on detection and localization of three-dimensional objects based on audio-visual aspects. The paper works on the issues that concerns detecting and pinpoint objects that is seen or can be heard in a scene. They sort the problem by clustering observations into coherent and relevant groups.The researchers propose a model that is based on probability and goes on to establish relation between audio and video. They use a version of the expectation-maximization algorithm[17].

Fidelia Orji and Julita Vassileva from University of Saskatchewan conducted work on the relation between student engagement and student performance using machine learning.It is a known fact that student engagement is a significant factor that affects student performance. This is an experimental study on finding ways to bring more students engaging in an online platform with the help of data interventions. Supervised Random Forest and Unsupervised Clustering machine learning were used by the researchers for the work.They were able to identify pattern on student engagement and depicted that engagement and assessment are intertwined to good academic performance[18].

K Kumar and S Kumanan have done significant work on decision making for selecting locations by implementing an amalgamated approach consisting of clustering and TOPSIS. Finding a suitable facility location is a big problem in the supply chain. The aim of this research is to build a decision supporting model to seek the best location. They incorporate Fuzzy C-Means, K-Means Clustering techniques. They found set of locations and then went on to choose the most suitable one. In the end, it is seen that the decision-making model they proposed is highly effective[19].

Chapter 3

Proposed Methodology

It is important for any research work to demonstrate a proper working plan that outlines how the study is being conducted ensuring validity of outcomes while addressing the objectives concerned. The route to find our desired output is based on the methods described underneath. Providing cheaper alternatives with better forecasting results has been our principal aim of work. We are required to mitigate the need for fieldwork which surely reduces the cost for entrepreneurs. Machine learning provides the required assistance to accomplish our task. With the world almost healing from the COVID-19 pandemic, shopping malls will soon start to open up again and fashion accessory sellers will once again come face to face with their ever evolving client base - or will they? With improper decision making during setting up a shop at a particular location, shop owners often face a lack of customer store traffic and experience marginal sales of their products. We hope to diminish this factor and build a decision model that provides sellers with adequate information about customer interests and behaviors along with location analysis. Our wish is that our research may help build a better understanding of where to open up a shop in order to avoid such a scenario in which economic stability is hindering.

3.1 Research Framework

Summarized into three major parts - data collection, clustering and cluster analysis, the backbone of our entire research can be visualized into the following chart.

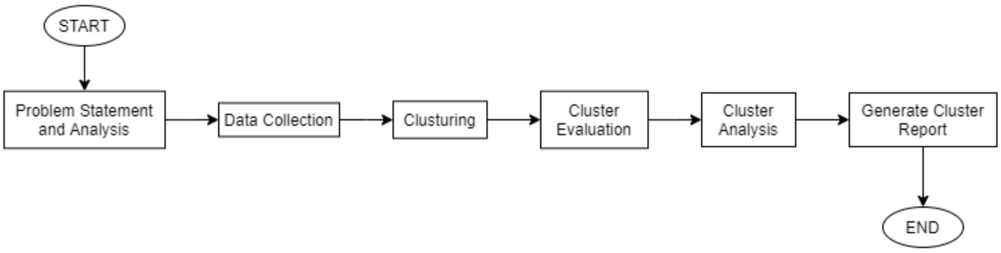

Figure 3.1: Research Framework

3.2 Machine Learning

Machine learning is an area of artificial intelligence (AI) and computer science that focuses on using data and algorithms to mimic the way humans learn, with the goal of steadily improving accuracy.

Machine learning has a long history at IBM. Arthur Samuel, is credited with coining the term ”machine learning.” In 1962, Robert Nealey, competed against an IBM 7094 computer in a game of checkers. This achievement may appear insignificant in comparison to what is possible now, yet it is regarded as a key milestone in the field of artificial intelligence. Technological advancements in storage and computing power will enable some revolutionary items that we know and love today during the next few decades, Netflix's recommendation engine or self-driving cars are the examples.

Machine learning is a crucial part of the rapidly expanding discipline of data science. Algorithms are trained to generate classifications or predictions using statistical approaches, revealing crucial insights in data mining initiatives. Following that, these insights drive decision-making within applications and enterprises, with the goal of influencing important growth KPIs. As big data expands and grows, the demand for data scientists will rise, necessitating their assistance in identifying the most relevant business questions and, as a result, the data needed to answer them[20].

3.3 Unsupervised Learning/Clustering

Unsupervised Learning is a widely used machine learning technique where human supervision is not required to train any model. The model figures out its own to identify patterns, information that probably went previously under the radar. This technique deals with unlabeled data. Their algorithms basically empowers users to do complicated tasks when compared with the supervised learning. Unsupervised learning is also unpredictable in comparison with other machine learning ways. Some examples are: clustering, association, neural networks, etc[21]. For our research work we shall be focusing on clustering.

Clustering is a vital machine learning tool for identifying structures in datasets. It works on datasets where there are no outcomes or nothing is known about the relationship between the unlabeled data. It is a data mining method that groups unlabeled data on the basis of similarities, differences etc. Clustering algorithms are of few types like exclusive, overlapping, hierarchical, probabilistic etc. An example of using clustering can be finding similar purchases made by consumers.

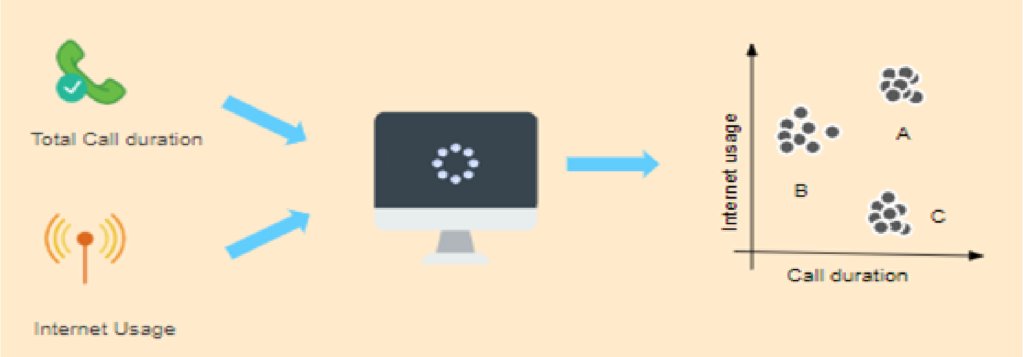

Figure 3.2: Clustering Example

Figure 3.2 provides a scenario where a telecom company sets an aim to mitigate its customer churn rate and plans to provide personalized calls and data plans. They study consumer behaviour and a model is built that manages to segment consumers having similar traits. Then they devise several ways to minimize the rate and maximize profit with the help of relevant promotions.

The right side of the image shows where customers are grouped. Group A consumers have higher call durations and data. Group B consumers include high end users on the other hand Group C consumers have high call duration. Thus Group B requires more data benefit plans, Group C requires cheaper call rates and group A has both benefits[22].

3.3.1 K-Means Clustering

Unsupervised learning method K-Means Clustering, it's used to handle clustering issues in machine learning, data science, and other fields. The K Means algorithm is an iterative technique that tries to divide a dataset into K non-overlapping clusters with each data point belonging to just one of them. The basic goal is to make intra-cluster data points as comparable as feasible while keeping the clusters as distinct as feasible. The technique distributes data points to clusters in such a manner that the sum of the squared distance between data points and the arithmetic mean of all data points owned by the cluster is as little as possible. The less variance there is within clusters, the more similar the data points are and the more clusters they belong to.

K means algorithm works in the following way:

- State precisely number of clusters K.
- Start centroids by shuffling the dataset first and then select randomly K data points for the centroids without replacing.
- Repeat until the centroids do not change. Calculate the total of all data points' squared distances from all centroids.
- Assign every data point to the closest centroid.
- Finally, take the average of all data points in each cluster to get the centroids for each cluster.

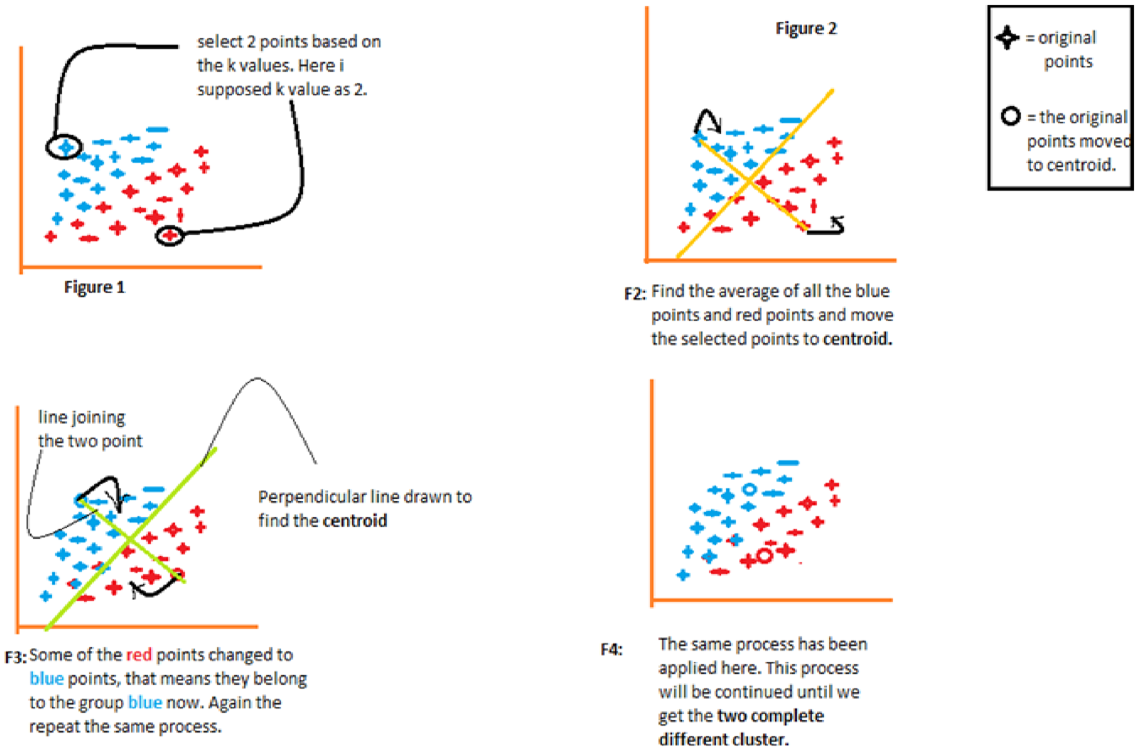
select 2 points based on the k values. Here i supposed k value as 2.
Figure 1
Figure 2
✚ = original points
O = the original points moved to centroid.
F2: Find the average of all the blue points and red points and move the selected points to centroid.
line joining the two point
Perpendicular line drawn to find the centroid
F3: Some of the red points changed to blue points, that means they belong to the group blue now. Again the repeat the same process.
F4: The same process has been applied here. This process will be continued until we get the two complete different cluster.

Figure 3.3: How K-Means Works

Let's analyze the steps projected in Figure 3.3

- First image represents data of two separate items. 1st item is in blue color while the other item is in red color. We select the K value to be 2(random). K value can be determined by various other methods.
- Two points are joined in the second image. Next we find a centroid drawing a perpendicular line. Points gradually move towards it. Some red points have moved towards blue points. They belong to the blue group now.
- Process repeated in the third image. More red points are now in the blue group.
- Process is repeated till we get two different clusters[23].

3.3.1.1 Principal Component Analysis

Principal Component Analysis is a technique for unsupervised learning. In the realm of machine learning, it is used to reduce dimensionality. Through orthogonal transformation, a statistical method turns observations of correlated characteristics into a set of linearly uncorrelated components. The Principal Components are the name for these new features. It is one of the most widely used machine learning algorithms. It may be used for both exploratory and predictive data analysis. It's a technique for decreasing variations and uncovering hidden patterns in a dataset. PCA is a feature extraction approach that is modeled after. PCA seeks to portray high-dimensional data with lower-dimensional surfaces. The variances of each feature are determined through PCA. High variance features project the separation between the classes and, as a result, reduce the dimensional aspect.

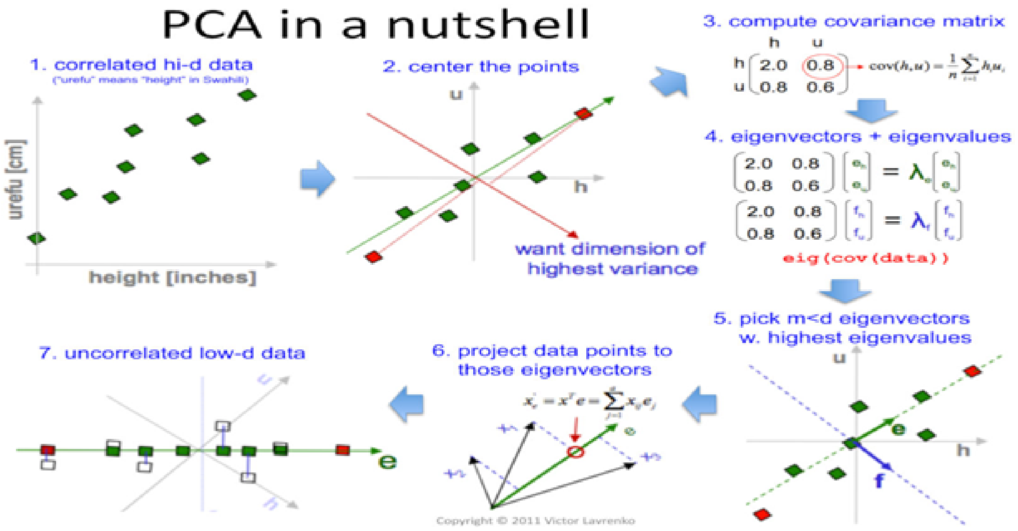
PCA in a nutshell
1. correlated hi-d data
("urefu" means "height" in Swahili)
urefu [cm]
height [inches]
2. center the points
u
h
want dimension of highest variance
3. compute covariance matrix
h u
h 2.0 0.8
u 0.8 0.6
cov(h,u) = 1/n Σ h_i u_i
4. eigenvectors + eigenvalues
2.0 0.8 0.8 0.6 = λ_e
2.0 0.8 0.8 0.6 = λ_f
eig(cov(data))
5. pick m<d eigenvectors w. highest eigenvalues
u
h
e
f
6. project data points to those eigenvectors
x_e = x^T e = Σ x_j e_j
e
Copyright © 2011 Victor Lavrenko
7. uncorrelated low-d data
e

Figure 3.4: PCA Function

Figure 3.4 demonstrates that PCA Algorithms work by the following ways [27] -

- First collect the dataset
- Arrange data into give a structure
- Normalize then the given data
- Calculate the Covariance Z
- The EigenValues and EigenVectors must then be determined.
- Next sort calculated EigenVectors
- Assess the new features
- Finally, remove any characteristics from the new dataset that aren't needed.

3.3.1.2 Elbow Method

The elbow technique is a technique for analyzing the clusters created from our dataset and determining the number of ideal clusters in the dataset. It entails looping the algorithm with an increasing number of cluster choices and then graphing the clustering score as a function of the number of clusters. The ideal number of clusters is picked by selecting the elbow slope of the curve.

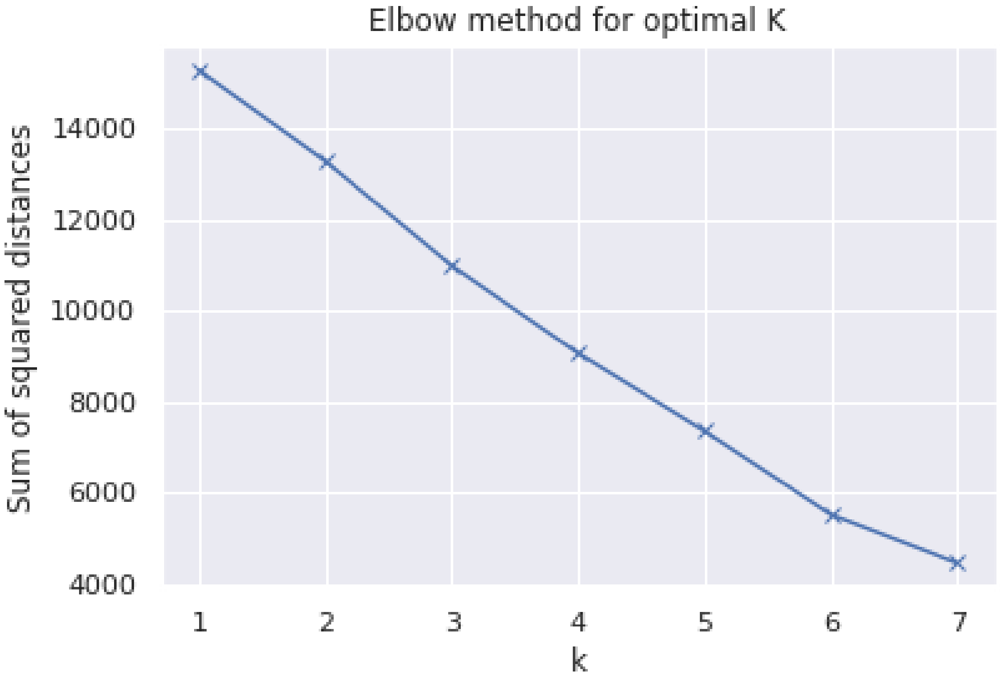

Figure 3.5: Graph of Elbow Method

Figure 3.5 demonstrates the graph made with the function of the elbow method. We can see that the elbow slope of the curve is found at $k = 4$. So the ideal number of clusters to use for the clustering algorithm from Figure 3.5 will be 4.

3.3.2 Mini-batch K-Means Clustering

The fundamental concept is to store tiny random batches of data in memory by using tiny random batches of data of a defined size. Each iteration obtains a fresh random sample from the dataset and uses it to update the clusters, and the process is continued until convergence. Each micro batch updates the clusters using a convex mixture of the prototype values and the data, with a decreasing learning rate as the number of iterations increases. This learning rate is proportional to the amount of data clusters created during the procedure. Because the influence of incoming data diminishes as the number of iterations grows, convergence can be observed when no changes in the clusters occur for multiple iterations in a row[23]. It's generally useful in web exercises where the measure of data can be huge, and the time available for clustering possibly limited.

The Mini-batch K-means clustering algorithm may be a version of the quality K-means algorithm in machine learning. It uses small, random, fixed-size batches of knowledge to store in memory, then with each iteration, a random sample of the info is collected and want to update the clusters[23].

K-Means is one among the favored clustering algorithms, mainly due to its blast performance. When the dimensions of the info set increases, K-Means will end in a memory issue since it needs the whole dataset. For those reasons, to scale back the time and space complexity of the algorithm, an approach called Mini-Batch K-

Means was proposed.

The Mini-Batch K-Means algorithm tries to suit the info within the main memory in a way where the algorithm uses small batches of knowledge that are of fixed size chosen randomly . Here are a few of points to notice about the Mini-Batch K-Means:

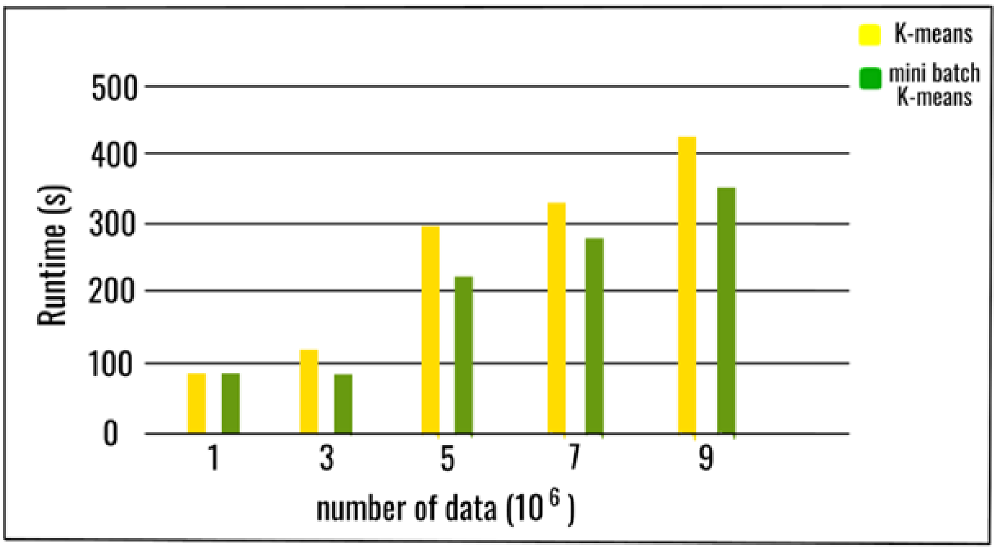

Figure 3.6: Mini-Batch K-Means Clustering

From figure 3.6, we see clusters are updated (depending on the previous location of cluster centroids) in each iteration by obtaining new arbitrary samples from the dataset, and these steps are repeated until convergence. Some of the research suggests that this method saves significant computational time with a trade-off, a touch moment of loss in cluster quality. But intense research has not been done to quantify the quantity of clusters or their size which can impact the cluster quality. The location of the clusters is updated and supports the new points from each batch. The update made is the gradient descent update, which is notably faster than normal batch K-Means[24].

3.3.3 Hierarchical Clustering

Hierarchical clustering is a form of analysis. It is mainly an algorithm that seeks to group the same objects into particular sets,also known as clusters. Finally we get a group of clusters each is distinct from another.Objects that fall in the same cluster are similar.

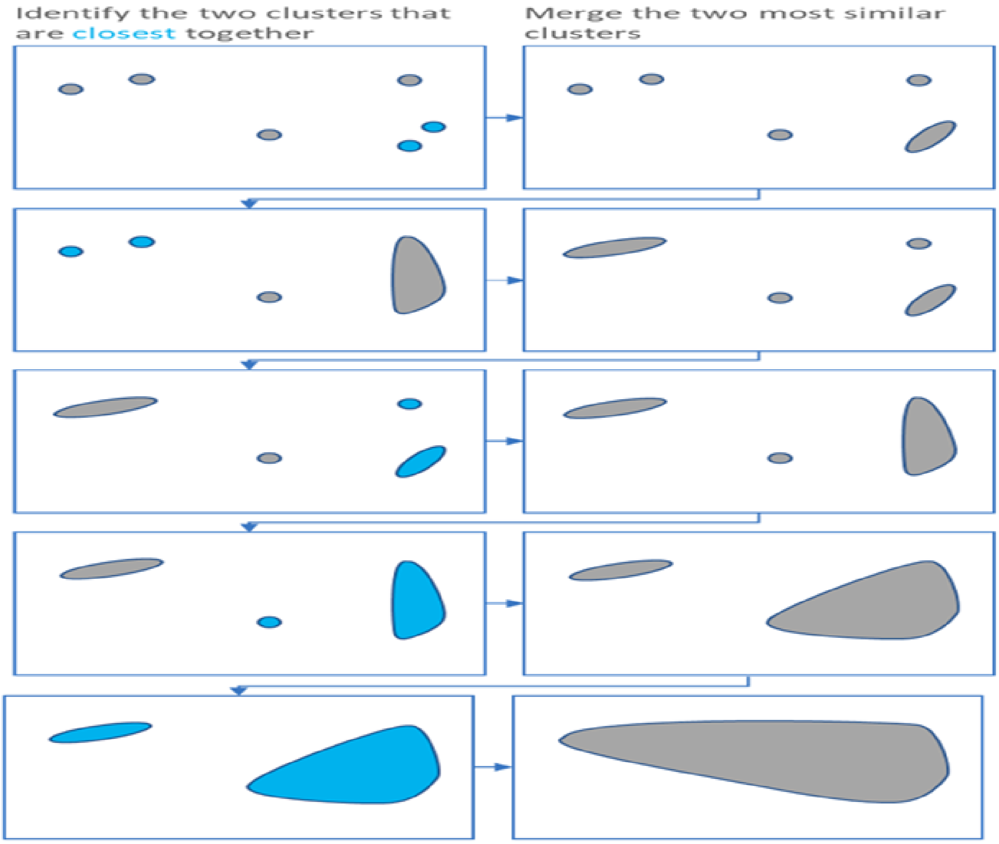
Identify the two clusters that are closest together
Merge the two most similar clusters

Figure 3.7: Hierarchical Clustering

From Figure 3.7, we see that hierarchical clustering works by taking individually observed sets as distinct clusters. Then it runs the following steps: (1) Find two clusters nearest to one another, Then (2) amalgamate two closest clusters. Process is iterative till all clusters combine[25].

3.4 Clustering Algorithm Evaluation Method

Clustering provides an analysis into data representation by grouping data with similar characteristics into the same groups. This grouping of the dataset can be done with a number of different clustering algorithms. Similarly, to evaluate how the clusters made with a certain algorithm are performing is another important part of clustering. For this, certain metrics are used that determine how the groups of data are better or worse off from one another. These metrics are defined by their own rules and they work differently for each clustering algorithm. We have used two cluster evaluation methods for our algorithms. These are the Davies Bouldin Score metric and Silhouette Score metric.

3.4.1 Davies Bouldin Score

The ratio is between the dispersion and separation of the cluster. Essentially, it is a balance between inter-cluster distances and inter-point lengths. To determine the best settings such that clusters are clustered more closely and have more separation from each other (i.e. distance between two clusters is high). As a result, a smaller

Davies Bouldin index value will indicate stronger grouping.

To get the Davies Bouldin Index, the following equation must be used:

$$DB = \frac{1}{n_c} \sum_{i=1}^{n_c} R_i \tag{3.1}$$

Where, $R_i = \max_{j=1...n_c, i \neq j} R_{ij}, \quad i = 1...n_c$
$R_{ij} = \frac{s_i + s_j}{d_{ij}}$
$d_{ij} = d(v_i, v_j), s_i = \frac{1}{||c_i||} \sum_{x \epsilon c_i} d(x, v_i)$
Where, d(x,y) is the euclidean distance between x and y.
c_i is the cluster i.
v_i is the centroid of cluster c_i.
$| \mid c_i \mid |$ refers to the norm of c_i.

3.4.2 Silhouette Score

Each sample's Silhouette Score is derived by taking the mean of (a) intra-cluster distance (how far apart each sample is inside the sample cluster) and (b) nearest-cluster distance (how far apart each sample is to the closest sample in its sample cluster). $\frac{(b-a)}{max}$ for a sample (a, b). The first measure is the distance from a given sample point to its centroid, and the second measurement (which is really composed of two parts) is the distance from the sample point to the closest cluster that it is not a member of. In order to do that, we want the silhouette score to be high. So, we need to locate the greatest that may be found globally for this technique.

There is a significant change in the appearance of the Silhouette coefficient as compared to the bend of the elbow method. It's much simpler to understand and grasp. As shown by the shape coefficients that are around +1, the sample is far from adjacent clusters. Samples are on or very near to the decision border, which is described by a value of 0. This implies that they may have been allocated to the incorrect cluster.

The average distance from the point I to all other points belonging to the same cluster Ci. is calculated as $x(i) = \epsilon \frac{j(dist(i,j)}{len(C_i)}$ for each data point i. In the case of Large $A(I)$, it may be deduced that the point I is outside of its cluster. In layman's terms, the average distance of a point in the zero cluster (denoted as I with all the other points in the zero cluster is equal to the distance between point I and zero.

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \epsilon C_i, i \neq j} d(i,j) \tag{3.2}$$

Furthermore, we describe:

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \epsilon C_k} d(i,j) \tag{3.3}$$

This is the distance from point I to the next closest cluster's farthest point Big B means that I is unique and different from its neighboring cluster.

The process requires two stages:

- Identify all of the distance from point I to points in other clusters (different from C_i).
- The lowest of the averages is what you should aim for.

Suppose the point I is part of the zero cluster. You calculate the average distance between point I and all points in cluster 1, then you do the same for points in cluster 2, and so on. After you figure out all those values, you'll choose the minimal distance.

To round up the discussion, we've given a definition for the silhouette score of data point I as:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{3.4}$$

- $-1 \leq s(i) \leq 1$
- If $s(i)$ is almost equal to 1, then $a(i)$ is considerably closer to its own cluster than to its neighbor, and the point is in the right cluster.
- In cases where $s(i) \approx -1$, then $a(i) \gg b(i)$ the point is more similar to its neighbour than to its allocated cluster; it should thus be in a different cluster.
- $S(i) = 0$ if $|Ci| = 1$. (i.e. the silhouette score of a point in a single-element cluster is 0).
- If the point is almost equal to zero, it's perhaps in another cluster.

Global silhouette score is define as:

$$S = \frac{1}{N} \sum_i s_i \tag{3.5}$$

In general, we use the silhouette analysis to search for a value of k that meets certain conditions:

- Very well-defined silhouettes.
- Because it indicates improper clustering and a terrible choice of k, clustering must be avoided if the greatest silhouette score is lower than the average.
- Data clusters with well-composed members, where the silhouettes are homogenous in size and shape.

3.4.3 Davies Bouldin VS Silhouette Score

Silhouette Score	Davies Bouldin
The range for the score is from -1 (inaccurate clustering) to +1 (very dense clustering). Scores in the negative numbers indicate that the clusters overlap.	Is easier to compute than Silhouette scores. The lowest possible score is zero. The better the partition, the closer the values are to zero.
Densely clustered and well-separated clusters get better scores. These ideas both refer to the conventional notion of a cluster.	All the data for the index must be included inside the dataset itself.
Clusters created using DBSCAN tend to have a lower Silhouette Coefficient, as do more broad cluster definitions.	Convex clusters typically have higher Davies-Bouldin indices than density-based clusters like those produced using DBSCAN.

Table 3.1: Difference between Davies Bouldin and Silhouette Score

3.5 Cluster Analysis

Clustering has always been the answer for any unsupervised learning problem. It has always helped create ways for us to gain knowledge on how data are grouped together based on their inherent similarities. What remained a challenge was to properly express these clusters or groups of data. In some cases, statistical analysis and visual representation are not enough. In order to fully understand it, we needed some explanation in a more simplistic term. We knew that the instances that were clustered together had definite similarities or patterns amongst them which defined them to be in the same cluster, but there is no definite way of knowing what these qualities or characteristics are. Hence we went with an approach that would help us analyze the different clusters in terms of a set of rules. These rules will allow us to determine why particular instances are grouped together during clustering.

Algorithm 1: Creating Rules List

```
Input  : Decision Classifier Tree, Feature Names
Output: Dictionary of Rules
Set decision classifier tree as global inner tree
Get classes from tree
Create class rules dictionary
Get Rules tree from dfs function and set it on class rules dictionary
return class rules dictionary
```

To interpret the clusters in terms of these rules, we firstly used a decision tree model. The algorithm would automatically generate rules whilst it trained itself using original features from our dataset and the clustering result as labels. The labels used here are the clustering result generated from the previous k-means clustering. We chose to work with k-means because of the optimal analysis it showed during cluster comparison.

```
Algorithm 2: Creating Decision Tree
Input  : Node Id, Current Rule List
Output: Rules Tree
Get split feature from global inner tree
if $split\_feature \neq \_tree.\_TREE\_UNDEFINED$ then
    Set left rule which contains column names $\leq$ threshold
    Calling tree dfs function recursively with left children of inner tree and left
      rule as parameter
    Set left rule which contains column names $>$ threshold
    Calling tree dfs function recursively with right children of inner tree and
      right rule as parameter
else
    if length of current rule is 0 then
        Set rule string as 'ALL'
    else
        Set rule string while joining current rule with 'AND'
    end
    Append rule string and probability of class in class rules dictionary
end
return class rules dictionary
```

Secondly, the generated list of rules were passed on as inputs with individual id of the tree into another algorithm and as a result an entire rules tree was generated. This tree will be the rule inscribed decision tree from which the rule sets of different clusters would be extracted.

```
Algorithm 3: Generate Cluster Report
Input  : Dataset, Clusters, Minimum Sample Leaf, Pruning Level
Output: Cluster Report
Create DecisionTreeClassifier Model
Fit the model
Set columns name as feature names
Get the class rule dictionary by calling get_class_rule function
foreach class name in key of class rules dictionary do
    Get rule list from value of class name in class rule dictionary
    foreach rule in rule list do
        Get combined string from 0th and 1st index of rule
    end
    Append class name and combined string in report class list
end
Get cluster instances by making a series of clusters
Get report by merging cluster instances, report class list
return cluster report
```

Finally, the cluster report algorithm was used to structurally visualize everything.

Taking the clusters and dataset as input and fitting them against the rule tree, we called upon a function that wrapped around the decision tree and extracted the rules in a tabular form.

Chapter 4

Dataset & Experiment

We carefully calculated the proper utilization of unsupervised learning in order to bring out the best result that helped optimize our findings, and hence were able to properly establish an insight for businessmen (fashion accessory sellers) so that they may have the opportunity to incorporate location based knowledge while opening up a shop where customer traffic and sales are always optimal. We hope that through our analysis, we were able to find a solution to help bring a proper foundation of economic stability in the society. For in a world where businesses are at the heart of the national treasury, we hope that these records will aid businessmen in their endeavours for that thriving economical benefit they hope to bring to the table.

4.1 Data Collection

The initial raw dataset[26] has 22 columns that carry the features (inputs) of 988 respondents who are displayed through the rows.

4.1.1 Collection Procedure

The goal of any machine learning problem is heavily dependent on its raw initial dataset. For this, we had to be extra careful in gathering our information. We figured out that for our particular problem, we had to go forward with data that was not only authentic, but also versatile enough to get the proper information from relevant sources. Hence we opted to build our dataset by dispatching a survey questionnaire among targeted audiences. Firstly, we built the questionnaire with inquiries that were made after keen observation. Studying the behavior from our intended audience, we came up with factual and informative queries that generated appropriate data. Our prime audience were those who were highly into buying fashion accessories and hence we had created a set of questionnaires that emphasized on questions related to that field. We had a total of twenty one well revised questions that gave us an overview of all answers that were going to be needed within the proximity of our system. As such, we had the opportunity to gather over half a thousand authentic leads and concluded upon our initial raw dataset accordingly.

The final data count came down to 988 respondents.

Link to the form: Shopaholic Survey

4.1.2 Dataset Features

The 22 features (columns) of the initial raw data are described as follows:

Timestamp - Real time clock and calendar input of respondents

1. Your Name - Name of respondents
2. Your Age - Age group of respondents
3. Gender - Gender of respondents
4. What is your monthly income? - Approx. monthly income of respondents
5. Which fashion accessory do you most frequently buy? (select multiple if needed) - Most frequently bought fashion accessories of respondents
6. What is your most preferred medium when you are shopping for fashion accessories? - Preferred medium for purchasing fashion accessories of the respondents
7. How much do you usually spend on shopping (at one go) for fashion accessories? - Spending (at one go) on shopping for fashion accessories by the respondents
8. Do you prefer shopping alone or with company? - Shopping preference of the respondents
9. Are you a fan of window shopping? - Opinion regarding window shopping by the respondents
10. Around which area do you live? - Living area of the respondents
11. Do you shop for fashion accessories near your house or look elsewhere? - Shopping area of the respondents
12. Which is your most preferred shopping mall to go to when you go shopping for fashion accessories? - Preferred shopping mall of the respondents
13. Which of the following shopping malls have you most often visited? (select multiple if needed) - Most often visited shopping mls by the respondents
14. Why are you a fan of the preferred malls you have selected? (select multiple if needed) - Reason behind preference of aforementioned shopping mall by the respondents
15. Rate the ambience of your most preferred shopping mall. - Ambience of preferred shopping mall by the respondent
16. How would you rate the convenience for transportation availability of your most preferred shopping mall? - Transportation availability for preferred shopping mall of the respondent

17. Have you ever gone into a shopping mall and came out without shopping for anything? - Asking respondents how often they shop when they go to the shopping mall.

18. What is your most acceptable time to go shopping? - Preferred time to go shopping by the respondents

19. Would you prefer a 24/7 shopping mall? - Respondent's opinion on a day-night shopping mall

20. Are you a brand specific person when shopping for fashion accessories? - Respondent's preference on brands while shopping

21. Do you feel annoyed when you cannot find the store of your desired brand in a shopping mall? - Asking how the respondent feels when they cannot find their desired brand while shopping.

4.1.3 Data Analysis

A holistic information of consumers is essential to derive desired outcomes for our research work. The questions asked in our survey intend to decipher relevant information that will help an entrepreneur to choose the right place for operating business.

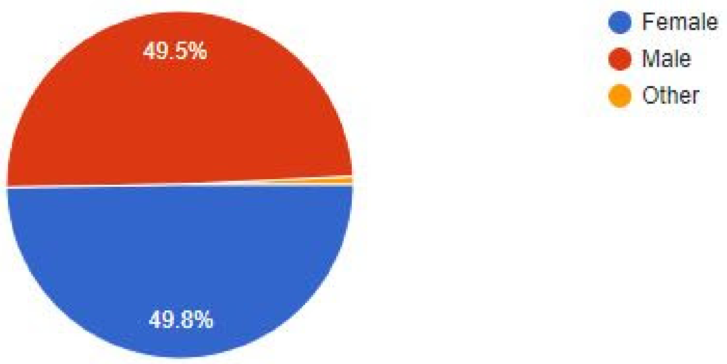

Figure 4.1: Gender

From figure 4.1, it shows that our dataset is contributed equally among people of both male and female genders.

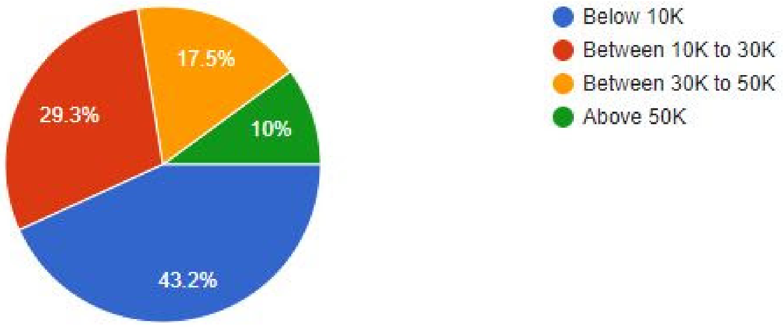

Figure 4.2: Income Range

Figure 4.2 provides us insight into consumer's income projecting an economical status of the respondents.

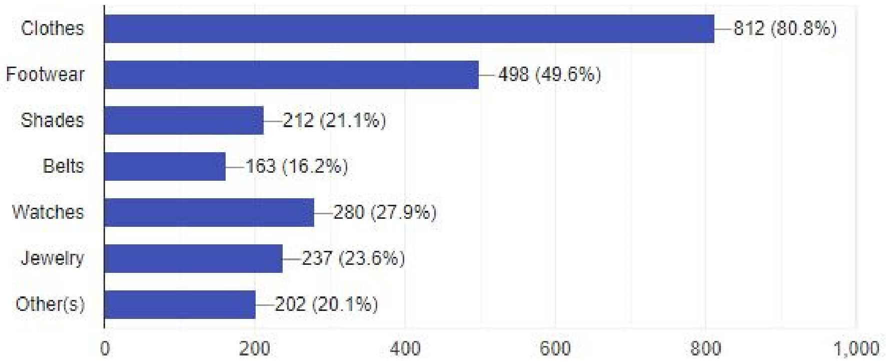

Figure 4.3: Consumer Taste On The Type of Products

Figure 4.3 gives us an idea of consumer taste on the type of products they choose to buy, suggesting which commodities are more on demand.

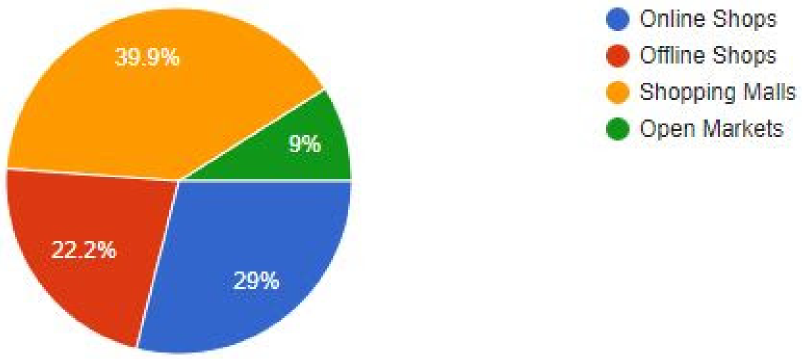

Figure 4.4: Medium of Shopping

Figure 4.4 provides clues to the medium of shopping consumers prefer the most.

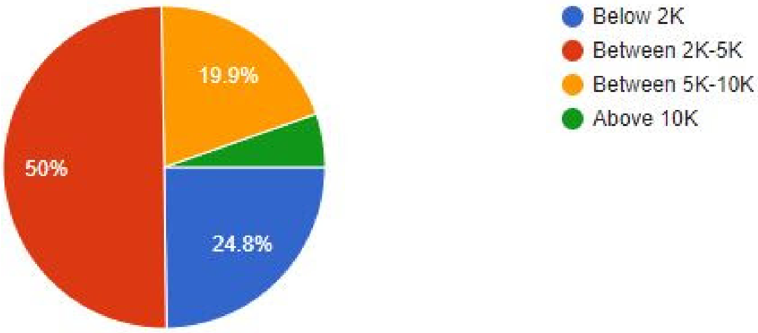

Figure 4.5: Budget of Respondents

Figure 4.5 gives us information on the budget of respondents.

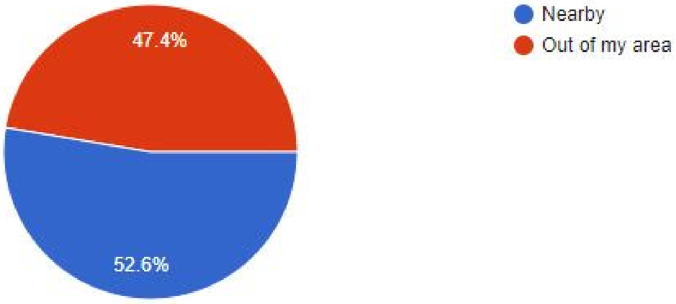

Figure 4.6: Preferred Areas to Shop Most Often

Figure 4.6 contains information on the preferred areas that consumers shop most often.

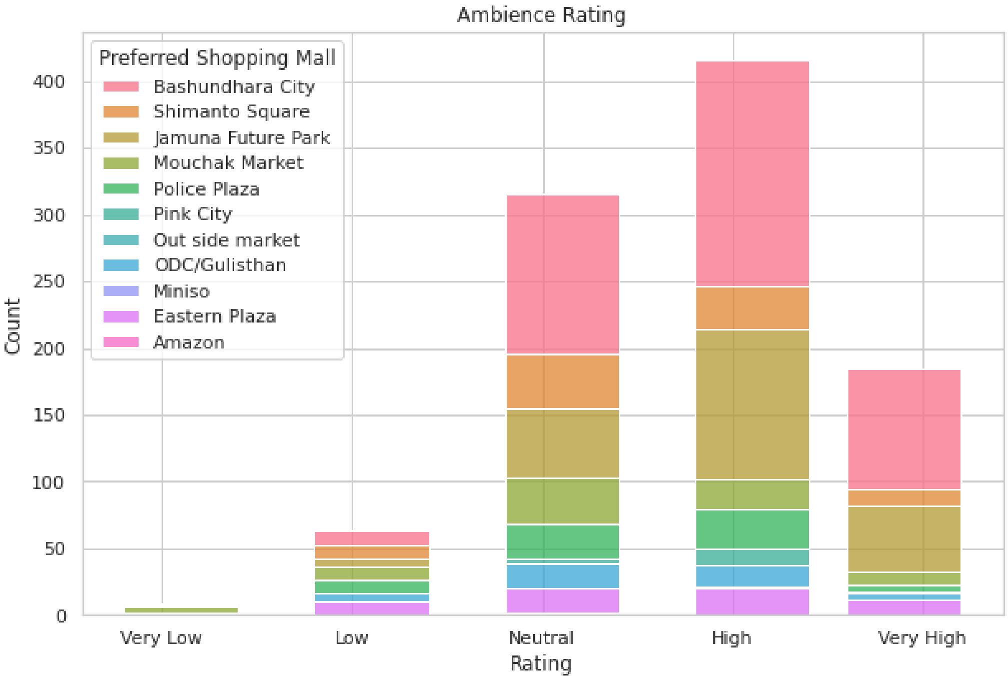

Figure 4.7: Ambience Rating

Figure 4.7 projects the ambience rating against each preferred shopping mall. The horizontal labels go from very low to very high, which suggests how much consumers are satisfied with the ambience of their preferred shopping mall.

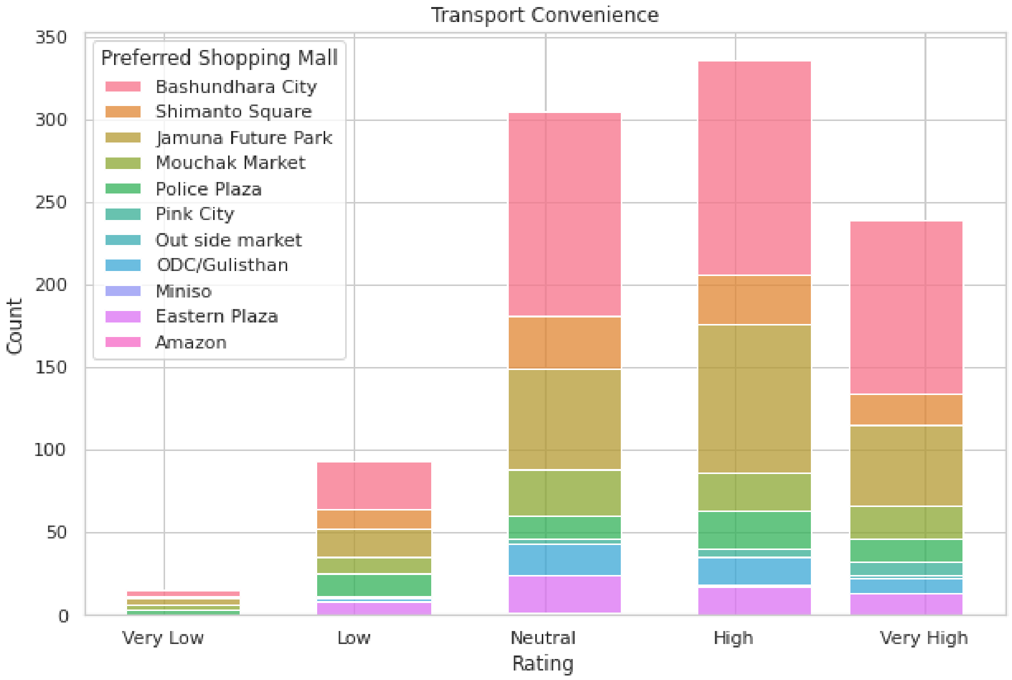

Figure 4.8: Transportation Convenience

Figure 4.8 provides information on the transportation convenience near a consumers preferred shopping mall. The level of convenience is measured from very low to very high.

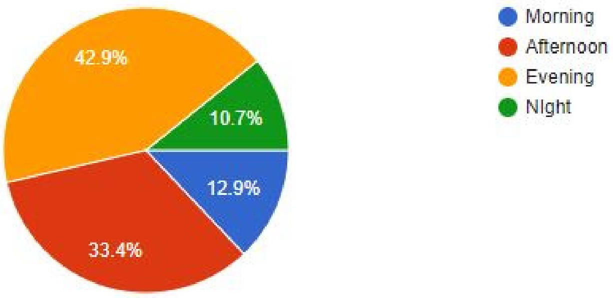

Figure 4.9: Time Preference for Shopping

Figure 4.9 gives on the timing respondents prefer for shopping.

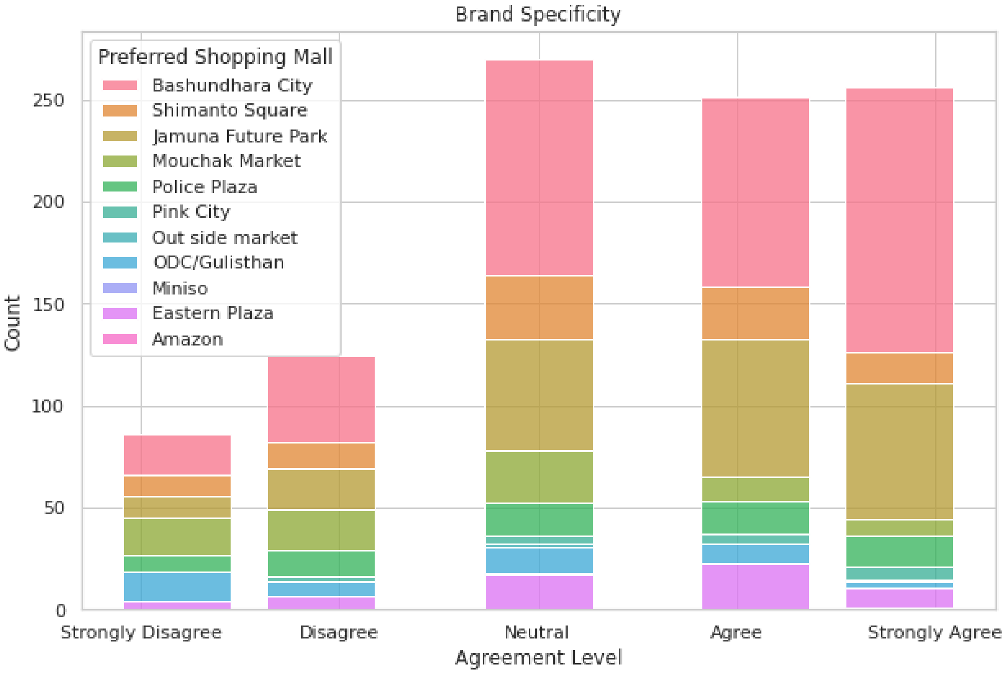

Figure 4.10: Impact of Specific Brands

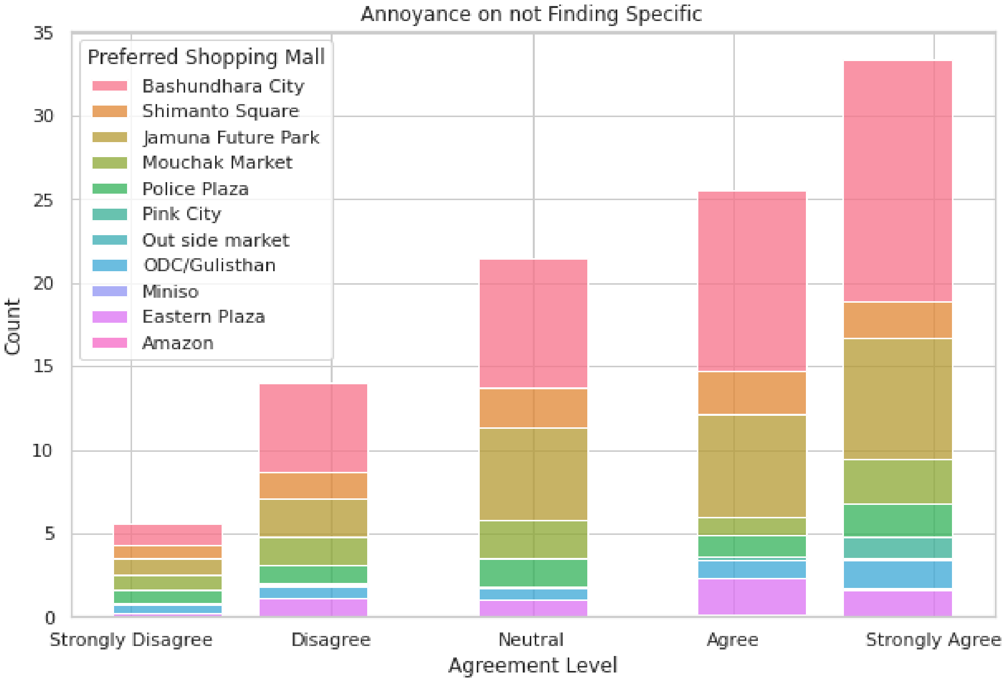

Figure 4.11: Agreement of Annoyance on not Finding Specific Brands

From figure 4.10 and figure 4.11 highlights the impact brands have on consumer purchases, integrated with their preferred shopping malls. On a scale of strongly disagreeing to strongly agreeing, figure 4.10 dissipates whether the consumer is a brand specific person or not. And figure 4.11 displays the agreement of their level of annoyance when they cannot find their specific brands while shopping for fashion accessories.

4.2 Data Initialization

In order to ready the dataset in terms of a system that is going to emphasize learning and optimization, we need to carefully sort out, plan and encode our entire dataset into machine readable numerical variables. Hence we need to remove null values (if any), replace string inputs with numerics, remove or drop unwanted features and so on. An initialized or encoded dataset is one that is at the end used to run the final algorithms on.

After successfully gathering the initial raw dataset, we set out straight into sorting it out. Raw datasets are a good source for initial data mining, but eventually for a learning/optimization problem, we have to trim and clean the raw dataset. We do so because we want data with no holes and that can help us get desired and accurate results. So data preprocessing is done in order to enhance the accuracy of success for the system. After careful examination of the dataset, we started to trim it down into a more presentable manner. As we take in a lot of string input (categorical data) into the dataset, we have to encode these inputs into numerical data to make it machine readable. Although this may sometimes increase the size of the dataset, it also helps to bring greater working efficiency for the system. The results of the data pre-processing stage will lead to a completely workable, neat and organized data, which would become the foundation for our system to operate.

Some encoded features with equivalent numerical labels are stated below:

Label	**Feature Value**
0	Above 50 years
1	Below 20 years
2	Between 20 to 35 years
3	Between 35 to 50 years

Table 4.1: Age Range

Label	**Feature Value**
0	Female
1	Male
2	Other

Table 4.2: Gender

Label	**Feature Value**
0	Above 50k
1	Below 10k
2	Between 10k to 30k
3	Between 30K to 50K

Table 4.3: Monthly Income

Label	**Feature Value**
0	Offline Shops
1	Online Shops
2	Open Markets
3	Shopping Malls

Table 4.4: Preferred Medium

Label	**Feature Value**
0	Above 10k
1	Below 2K
2	Between 2k-5k
3	Between 5k-10k

Table 4.5: Spend

Label	**Feature Value**
0	Alone
1	Company

Table 4.6: Shopping Partner Preference

Label	**Feature Value**
0	Out of My Area
1	Nearby

Table 4.7: Nearby or Look Elsewhere

Label	**Feature Value**
0	No
1	Yes

Table 4.8: Shopping or Not

Label	**Feature Value**
0	Afternoon
1	Evening
2	Morning
3	Night

Table 4.9: Acceptable Time

Chapter 5

Experimental Result & Analysis

5.1 Elbow Method Result

We executed the elbow technique with our dataset to find the optimal number of clusters we wanted to make with the clustering algorithms.

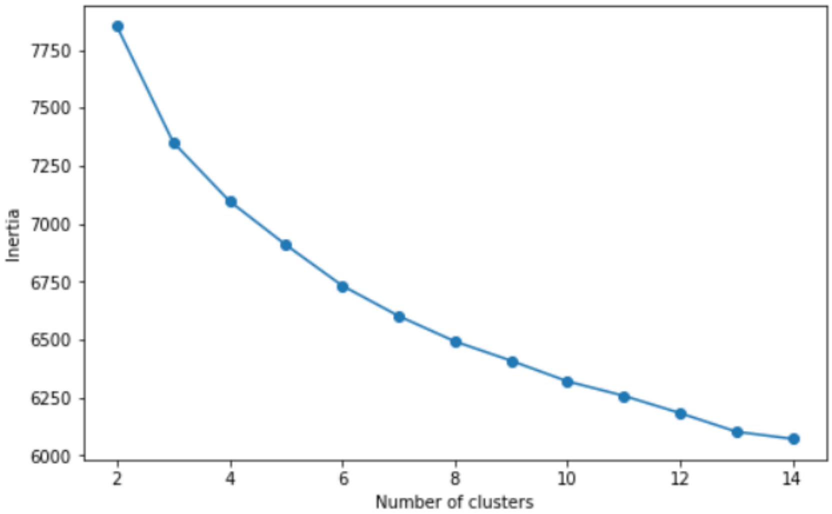

Figure 5.1: Graph of Elbow Method

From the above graph we can see that the elbow slope is made at the number of clusters = 7. Hence we used $k = 7$ for all clustering algorithms.

5.2 Data Representation

Among the clustering algorithms explained earlier, we have experimented our dataset on three types:

- K-Means Clustering
- Mini-batch K-Means Clustering
- Hierarchical Clustering

Furthermore, in order to evaluate in between the three clustering algorithms we had selected, we used two metrics - silhouette score and davies bouldin score. We found these scores for each of the three algorithms - k-means, mini-batch k-means and hierarchical. Later on we created a comparison table to compare between the scores for all three algorithms.

5.2.1 K-Means Clustering

Using k-means clustering, we were able to transform our dataset in terms of clusters having similar qualities. We ran the algorithm and created 7 different clusters. In order to view these clusters in a more visually present manner, we created a 3D space where we highlighted the clusters with a different shade of hue.

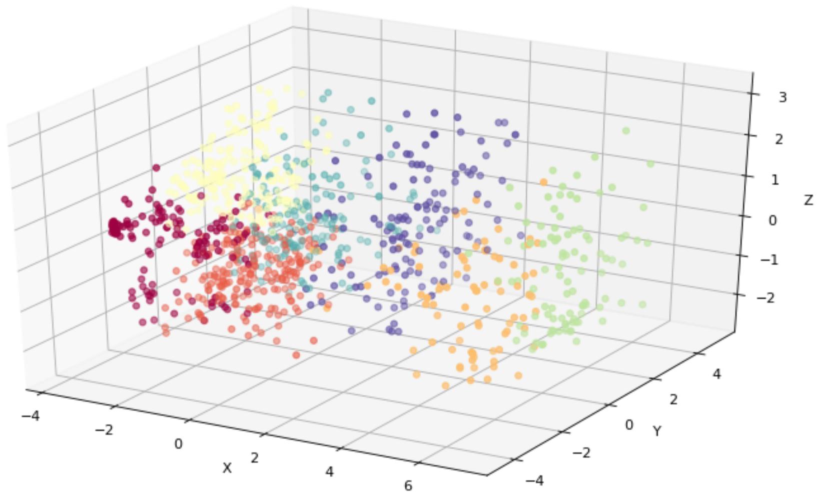

Figure 5.2: 3D Representation of K-Means

After representing the k-means clustering, we found out the silhouette and davies bouldin score for the $n = 7$ clusters.

Cluster Number	Silhouette Scores	Davies Bouldin Scores
2	0.303	1.387
3	0.174	2.067
4	0.12	2.332
5	0.113	2.482
6	0.106	2.431
7	0.09	2.522
8	0.091	2.442

Table 5.1: Silhouette and Davies Bouldin Scores for K-Means Clusters

5.2.2 Mini-batch K-Means Clustering

Similarly we ran the clustering algorithm of mini-batch k-means with $n = 7$ cluster number and created different clusters. Just as before, we viewed the clusters in a

3D space where we highlighted the clusters with a different shade of hue.

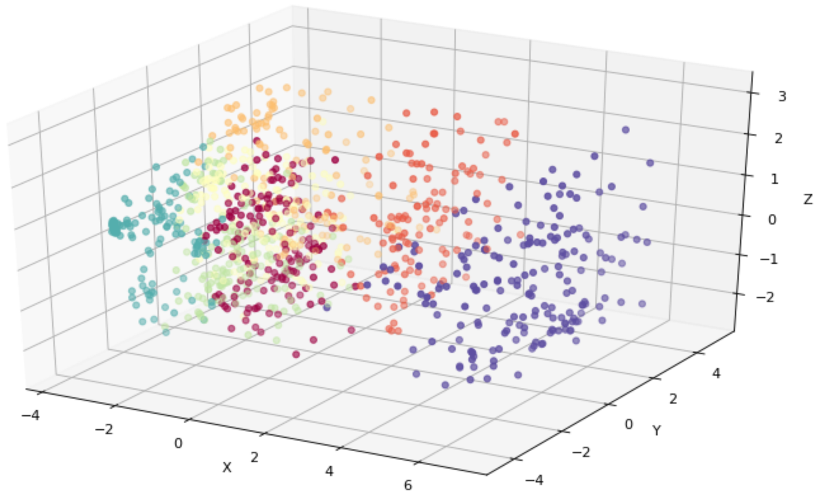

Figure 5.3: 3D Representation of Mini-Batch K-Means

After representing the mini-batch k-means clustering, we found out the silhouette and davies bouldin score for the $n = 7$ clusters.

Cluster Number	**Silhouette Scores**	**Davies Bouldin Scores**
2	0.305	2.741
3	0.174	2.741
4	0.156	2.741
5	0.109	2.741
6	0.094	2.741
7	0.096	2.741
8	0.084	2.741

Table 5.2: Silhouette and Davies Bouldin Scores for Mini-Batch K-Means Clusters

5.2.3 Hierarchical Clustering

For a third clustering method, we went with hierarchical clustering and represented the cluster in terms of a dendogram.

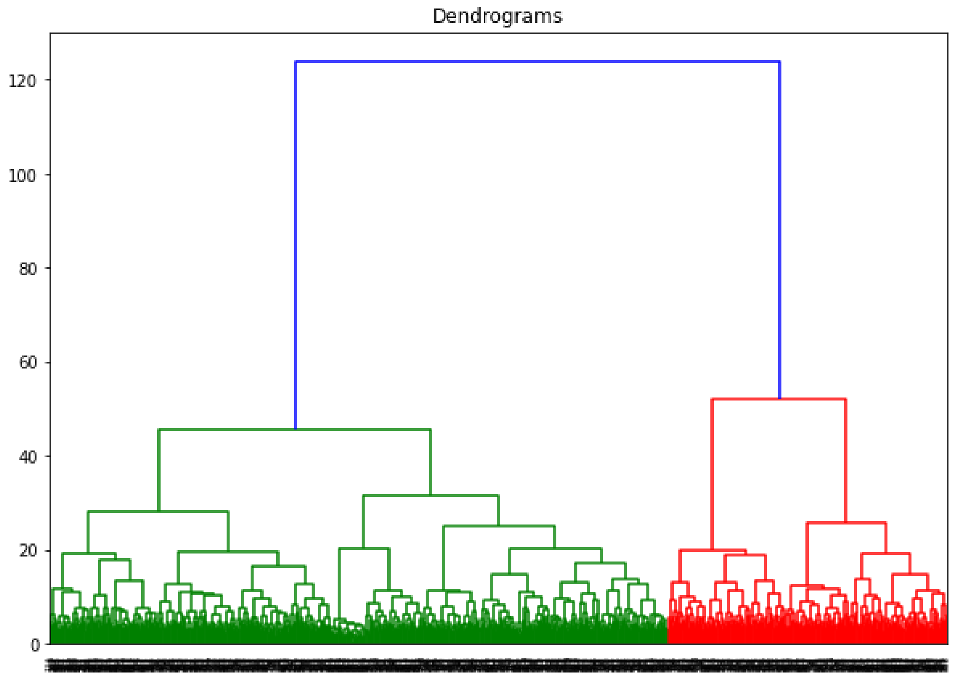

Figure 5.4: Dendrogram Graphical Representation of Hierarchical Clustering Segmentation

After representing the hierarchical clustering, we found out the silhouette and davies bouldin score for the hierarchical clusters.

Method Name	Score
Silhouette	0.058
Davies Bouldin	2.797

Table 5.3: Scores for Hierarchical Clustering

5.3 Results - Clustering Comparison

With silhouette and davies bouldin scores of three different clustering algorithms in hand, we created a table taking the average of all the evaluation values and represented them side by side.

Algorithms	**Davies Bouldin Scores**	**Silhouette Scores**
K-Means	2.52	0.09
Mini-Batch K-Means	2.65	0.09
Hierarchical Clustering	2.80	0.05

Table 5.4: Comparing Three Clustering Algorithms Based on Their Davies Bouldin and Silhouette Scores

From the table 5.4, we came to the conclusion that out of the three clustering algorithms we tested our dataset against, we can say that k-means give the most optimum value because of its respective silhouette and davies bouldin score. We know that higher the silhouette score and lower the davies bouldin score helps to give the optimum clusters. From previous cluster comparison, we found out that k-means algorithm displayed the lowest davies bouldin score with 2.522004 and has the average value for silhouette with 0.090342. These values make k-means clustering the optimal clustering method to use against our dataset. Hence we moved onto the cluster analysis stage where we analyzed the different clusters generated by k-means clustering.

5.4 Results - Cluster Analysis

After completing cluster analysis, we extracted a rule tree that gave us an insight into why the instances of a particular cluster where grouped together, From there we can visualize the following rule set table:

Cluster Number	**Class Name**	**Instance Count**	**Rule List**
0	0	316	[**0.78**] (Shimanto Square ≤ 0.5) and ((Jewelery) ≤ 0.5) and (Mouchak Market ≤ 0.5) and (Eastern Plaza ≤ 0.5) and (Preferred Shopping Mall ≤ 4.0) and ((Shades) ≤ 0.5)
1	6	218	[**0.67**] (Shimanto Square ≤ 0.5) and ((Jewelery) ≤ 0.5) and (Mouchak Market ≤ 0.5) and (Eastern Plaza > 0.5) [**0.69**] (Shimanto Square > 0.5) and (Jamuna Future Park ≤ 0.5)
2	5	191	[**0.41**] (Shimanto Square ≤ 0.5) and ((Jewelery) ≤ 0.5) and (Mouchak Market ≤ 0.5) and (Eastern Plaza ≤ 0.5) and (Preferred Shopping Mall ≤ 4.0) and ((Shades) > 0.5) [**0.75**] (Shimanto Square ≤ 0.5) and ((Jewelery) > 0.5) and (20. Are you a brand specific person when shopping for fashion accessories? > 3.5) [**0.35**] (Shimanto Square > 0.5) and (Jamuna Future Park > 0.5
3	1	186	[**0.72**] (Shimanto Square ≤ 0.5) and ((Jewelery) ≤ 0.5) and (Mouchak Market > 0.5) [**0.56**] (Shimanto Square ≤ 0.5) and ((Jewelery) > 0.5) and (20. Are you a brand specific person when shopping for fashion accessories? ≤ 3.5)
4	9	68	[**0.35**] (Shimanto Square ≤ 0.5) and ((Jewelery) ≤ 0.5) and (Mouchak Market ≤ 0.5) and (Eastern Plaza ≤ 0.5) and (Preferred Shopping Mall > 4.0)
5	8	3	NaN
6	4	3	NaN

Table 5.5: Rule List

Label	Preferred Shopping Mall
0	Amazon
1	Basundhara City
2	Eastern Plaza
3	Jamuna Future Park
4	Miniso
5	Mouchak Market
6	New Market
7	Out Side
8	Pink City
9	Police Plaza
10	Shimanto Square

Table 5.6: Preferred Shopping Mall

From the generated table 5.5 and table 5.6, we can come to the following conclusions:

1. For cluster 0 with instance count 316 - those who do not often visit Shimanto Square (Shimanto Square ≤ 0.5), Mouchak Market (Mouchak Market ≤ 0.5) and Eastern Plaza (Eastern Plaza ≤ 0.5) and with preferred shopping malls (while shopping for fashion accessories) as Bashundhara City, Eastern Plaza and Jamuna Future Park (Preferred Shopping Mall ≤ 4.0) tend to not shop for shades ((Shades) ≤ 0.5) and jewelries ((Jewelery) ≤ 0.5). Hence we can conclude that if you own a fashion accessory store that primarily focuses on selling shades or jewelries, then it is better to avoid Bashundhara City, Jamuna Future Park and Eastern Plaza while setting up your shop. Because chances are you will not be attracting customers who shop for fashion accessories in those regions/shopping malls. This is true for 77% of the total instance count of this cluster.

2. For cluster 1 with instance count 218 - those who do not often visit Shimanto Square (Shimanto Square ≤ 0.5) and Mouchak Market (Mouchak Market ≤ 0.5) but often visits Eastern Plaza (Eastern Plaza > 0.5) are not a fan of frequently buying jewellery accessories. Hence for a jewellery seller, setting up shop in Eastern Plaza is not ideal, as the customer base who frequently visits the mall are not fans of jewellery products.This is true for 67% of the total instance count of this cluster.

   Again the second rule for this cluster states that those who often visit Shimanto Square (Shimanto Square > 0.5) do not often visit Jamuna Future Park (Jamuna Future Park ≤ 0.5). So this points to the fact that people who shop around the Dhanmondi area do not usually go over to the Panthapath area. Hence we can keep the trade off in mind that if we were to set up shop in either one of these malls, then we would have to give up the customer base for the other and plan our sale campaigns accordingly. This is true for 69% of the total instance count of this cluster.

3. For cluster 2 with instance count 191 - those who do not often visit Shimanto Square (Shimanto Square ≤ 0.5), Mouchak Market (Mouchak Market ≤ 0.5)

and Eastern Plaza (Eastern Plaza ≤ 0.5) and with preferred shopping malls (while shopping for fashion accessories) as Bashundhara City, Eastern Plaza and Jamuna Future Park (Preferred Shopping Mall ≤ 4.0) tend to not shop for jewelries ((Jewelery) ≤ 0.5) but rather shop for shades ((Shades) > 0.5). So if you own a fashion accessory store that primarily focuses on selling jewelry, then it is better to avoid Bashundhara City, Jamuna Future Park and Eastern Plaza while setting up your shop. But on the other hand, if you have a shop that focuses on selling shades, then Bashundhara City, Jamuna Future Park and Eastern Plaza should be your go to place while setting up shop. Because chances are you will be attracting customers who do not shop for jewelry but shops for shades in those regions/shopping malls. This is true for 40% of the total instance count of this cluster.

For the second rule of this cluster we see that brand specific persons (20. Are you a brand specific person when shopping for fashion accessories? > 3.5) tend to be frequent buyers of jewellery products ((Jewelery) > 0.5) and do not often visit Shimanto Square (Shimanto Square ≤ 0.5). This is a very prominent conclusion. We can verify from this that if you own a store that sells high end products like jewellery, then it is easier to attract customers who shop according to brands. Considering jewellery to be a high end product in terms of pricing, if you're already selling this fashion accessory, then it would be wise to sell products of specific or notable brands that make jewellery items. This would hugely attract the brand specific consumer base who are always monotonous in searching for jewellery products of the same brand.Furthermore, you can avoid selling your product in Shimanto square because this same group of people do not often visit this region/shopping mall. This is true for 74% of total instance count of this cluster.

Finally the third rule for this cluster states that those who often visit Shimanto Square (Shimanto Square > 0.5) also often visit Jamuna Future Park (Jamuna Future Park > 0.5). So this points to the fact that people who shop around the Dhanmondi area also go over to the Panthapath area. Hence we can keep in mind that if we were to set up shop in either one of these malls, we could attract the customer base of both these regions/shopping malls. This is true for 34% of the total instance count of this cluster.

4. For cluster 3 with instance count 186 - those who do not often visit Shimanto Square (Shimanto Square ≤ 0.5) but often visit Mouchak Market (Mouchak Market > 0.5) tend to not shop for jewelries ((Jewelery) ≤ 0.5). So if you own a jewelry shop then it is better to avoid setting up shop in Mouchak Market because people who often visit there do not usually shop for jewelry. This is true for 72% of the total instance count of this cluster.

   The second rule of this cluster states that people who are not brand specific shoppers (20. Are you a brand specific person when shopping for fashion accessories? ≤ 3.5), do not often go to Shimanto Square (Shimanto Square ≤ 0.5) and also frequently buys jewelry ((Jewelery) > 0.5). We can conclude from this that if you are a shop owner who sells brand specific items, especially jewelry, then it would be better to set up shop in Shimanto Square as it is not a hub for consumers who are not brand specific. This is true for 56% of the total instance count of this cluster.

5. For cluster 4 with instance count 68 - people who do not often visit Shimanto Square (Shimanto Square ≤ 0.5), Mouchak Market (Mouchak Market ≤ 0.5), Eastern Plaza (Eastern Plaza ≤ 0.5) but rather has preferred shopping malls (while shopping for fashion accessories) as Pink City, Mouchak Market, New Market, Police Plaza and Shimanto Square, are not a fan of frequently buying jewellery accessories ((Jewelery) ≤ 0.5). Hence if you are a seller of jewellery accessories it is better to avoid Pink City, Mouchak Market, New Market, Police Plaza and Shimanto Square as people who prefer to go to these regions/shopping malls for buying fashion accessories do not frequently buy jewelry accessories. This is true for 34% of the total instance count of this cluster.

This analysis of the rule set for each cluster helps to better understand why the instances of a particular cluster were grouped together. And from that we can come to our concluding remarks on which argument is beneficial for the instances within the same cluster.

5.5 Performance Evaluation

The clustering comparison gave us that out of our chosen three clustering algorithms, k-means clustering was performing optimal. We evaluated this through the silhouette and davies bouldin score analysis. Lower value of davies bouldin score and higher value of silhouette score provided us with the optimal method. As the dataset evolves, we hope that the scores for davies bouldin and silhouette will evaluate the clusters with even more optimal values.

As the cluster analysis is carried out, we notice that out of the 7 created clusters, 5 of them were able to generate a rule list. If we start from cluster 0, we observe that the instance count was 316 with 77% of it following the extracted rule set for that cluster. Similarly, instance counts of 218, 191, 186 and 68 with 67%, 74%, 56% and 34% following the respective rule list were derived from clusters 1, 2, 3 and 4 respectively. And finally clusters 5 and 6 with instance counts of 3 each were unable to generate a rule list. If we observe here, we can notice that the instance count of each cluster gradually went down along with the percentage of instances that followed the rule set of that cluster. So we can say that as we went from cluster to cluster, the efficiency gradually decreased until finally in the last two clusters the instance count was so low that the rule set for those clusters failed to generate. However this does not compromise the fact that the initial clusters displayed good performance levels. We hope that as the dataset evolves and more data is accumulated, the instance count of each cluster will increase which will enhance their performance level. We further hope that greater instance count with more percentage selection will also help individual clusters to extract better rule lists from which we can evaluate them with more accuracy.

Chapter 6

Conclusion & Future Works

6.1 Conclusion

We started off with addressing one of the most unanswered questions - why do shops experience less returns even when being set up in one of the country's best shopping malls, and concluded with a comprehensive analysis on solving the matter. In order to block off an economic downfall in the post pandemic era, we created a model using machine learning algorithms to evaluate how particular shops can have hefty sales returns depending on customer interest and behaviours when grouped together with location analytics.

We were able to gather a database of almost 1000 instances which we trimmed and encoded to make it more re-presentable for our model. We later grouped the encoded dataset into clusters using different clustering algorithms and made a comparison that deduced which clustering method was working best for our dataset. The clusters of the optimal method were then analyzed and trained upon through a decision tree in order to generate a set of rules by which we may decipher why the instances of a particular cluster were grouped together. Through comprehensive analysis on these sets of rules, we were able to make our findings and conclusions.

Our findings of research were adequate enough to explain the outcome we were looking for. By using clustering methods on a collected dataset, and later going into in depth analysis of individual clusters, we were able to determine how fashion accessory sellers could benefit from our generated rule list. If they can work their way around the particulars we were able to derive, then we would be able to suggest where and how their shops could benefit from.

6.2 Future Works

We sincerely believe that our work will pave the way for one of the most controversial theories ever, that setting up your shop in the biggest shopping mall in the world will not always guarantee sales and customer foot traffic. We aim to utilize this primarily on our domestic platform because it is here in our own backyard that we first realized this major problem. But to do so, we will need to accumulate a huge database. Our dataset was sufficient enough to get us results that explained how to diminish the problem, but to do this on a larger scale with more in depth analysis, we must bring the instance count up drastically. If we were to do so, then we would find more comprehensive rule sets that will allow us to solve the question in hand with much more efficiency and accuracy. With that hope we wish to conclude that we are extremely proud of the problem outcome that has come to light through our machine learning model and we take pride in these few small steps. Because as we stated earlier, we believe every small step taken is a step taken for a bigger cause.

Bibliography

[1] S. Rogers, “John snow’s data journalism: the cholera map that changed the world,” 03 2013.

[2] K. D. Foote, “A brief history of machine learning - dataversity,” 03 2019.

[3] A. V K, “What is machine learning: Definition, types, applications and examples,” 12 2019.

[4] T. Wood, “Unsupervised learning,” 05 2019.

[5] S. Mishra, “Unsupervised learning and data clustering,” 05 2017.

[6] J. Bhole, S. Nandiyawar, S. Pawar, and P. Vora, “Smart site selection using machine learning,” *International Research Journal of Engineering and Technology (IRJET)*, vol. 07, pp. 3012–3015, 05 2020.

[7] J. K. Pearson, “A comparative business site-location feasibility analysis using geographic information systems and the gravity model,”

[8] D. Karamshuk, A. Noulas, S. Scellato, V. Nicosia, and C. Mascolo, “Geo-spotting,” *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’13*, 2013.

[9] M. Xu, T. Wang, Z. Wu, J. Zhou, J. Li, and H. Wu, “Store location selection via mining search query logs of baidu maps,” *arXiv:1606.03662 [cs]*, 06 2016.

[10] A. A. Mavalankar, A. Gupta, C. Gandotra, and R. Misra, “Hotel recommendation system,” *arXiv:1908.07498 [cs, stat]*, 2019.

[11] L. Wang, H. Fan, and Y. Wang, “Site selection of retail shops based on spatial accessibility and hybrid bp neural network,” *ISPRS International Journal of Geo-Information*, vol. 7, p. 202, 05 2018.

[12] A. K. P, S. S. G, P. K. R. Maddikunta, T. R. Gadekallu, A. Al-Ahmari, and M. H. Abidi, “Location based business recommendation using spatial demand,” *Sustainability*, vol. 12, p. 4124, 05 2020.

[13] B. Ramzan, I. S. Bajwa, N. Jamil, R. U. Amin, S. Ramzan, F. Mirza, and N. Sarwar, “An intelligent data analysis for recommendation systems using machine learning,” *Scientific Programming*, vol. 2019, pp. 1–20, 10 2019.

[14] F. B. Ashraf, M. R. Kabir, M. S. R. Shafi, and J. I. M. Rifat, “Finding homogeneous climate zones in bangladesh from statistical analysis of climate data using machine learning technique,” p. 1–6, 12 2020.

[15] S. Sen, M. I. Islam, S. S. Azim, F. A. Norin, and S. T. Shuha, "Fake profile detection in social media using image processing and machine learning," 2021.

[16] Y. Farouk and S. Rady, "Early diagnosis of alzheimer's disease using unsupervised clustering," *International Journal of Intelligent Computing and Information Sciences*, vol. 20, pp. 112–124, 12 2020.

[17] V. Khalidov, F. Forbes, M. Hansard, E. Arnaud, and R. Horaud, "Detection and localization of 3d audio-visual objects using unsupervised clustering," *Proceedings of the 10th international conference on Multimodal interfaces - IMCI '08*, 2008.

[18] F. Orji and J. Vassileva, "Using machine learning to explore the relation between student engagement and student performance," *2020 24th International Conference Information Visualisation (IV)*, 09 2020.

[19] K. , K. , Kumanan, and S. , "Decision making in location selection: An integrated approach with clustering and topsis," *IUP Journal of Operations Management*, vol. xi, pp. 1–14, 09 2012.

[20] I. C. Education, "What is machine learning?," 07 2020.

[21] "Supervised and unsupervised learning in (machine learning)," 2021.

[22] T. Bock, "What is hierarchical clustering? — displayr.com," 2018.

[23] A. Kharwal, "Mini-batch k-means clustering in machine learning," 09 2021.

[24] A. CR, "Exploring clustering algorithms: Explanation and use cases," 09 2021.

[25] A. Kumar Pandey, "A simple explanation of k-means clustering and its adavantages," 10 2020.

[26] M. I. Mahmud, Y. Alvee, O. Chowdhury, T. Sadman, I. H. Shaon, and F. B. Ashraf, "Find ideal location for business in bangladesh," *data.mendeley.com*, vol. 2, 09 2021.