CBEBirr Customer Segmentation Using Machine Learning in Commercial Bank of Ethiopia

Thesis Prepared

By

Yigeremu Yohanes Gelaw

To

The Faculty of Informatics

Of

St. Mary's University

In Partial Fulfilment of the Requirements

For the Degree of Master of Science

In

Computer Science

March, 2024

ACCEPTANCE

CBEBirr Customer Segmentation Using Machine Learning in Commercial Bank of Ethiopia

By

Yigeremu Yohanes Gelaw

Accepted by the Faculty of Informatics, St. Mary's University, in partial fulfilment of the requirements for the degree of Master of Science in Computer Science

Thesis Examination Committee:

Internal Examiner

Alembante Mulu (PhD)

External Examiner

Million Meshesha (PhD)

Dean, Faculty of Informatics

Alembante Mulu Kumlign (PhD)

February, 2024

DECLARATION

I, the undersigned, declare that this thesis work is my original work, has not been presented for a degree in this or any other University, and all sources of materials used for the thesis work have been duly acknowledged.

Yigeremu Yohanes Gelaw

_____

Signature

Addis Ababa

Ethiopia

This thesis has been submitted for examination with my approval as advisor.

Tessfu Geteye Fantaye (PhD)

_____

Signature

Addis Ababa

Ethiopia

February 13, 2024

Acknowledgments

Above all and foremost, may all the prize go to Almighty God who has always been there with me in all difficulties. Any attempt at any level cannot be completed without the support and guidance of my advisor Tessfu Geteye (PhD) who helped me to stay on track to do this thesis on the topic CBEBirr customer segmentation using machine learning in Commercial Bank of Ethiopia. I would like to express my special thanks and gratitude to my sister Elsabet Demissie and my colleagues Demeke Admasu, Hailu Bekele and Getachew Bualew for helping me throughout my research. Finally, I am thankful in all humbleness and gratefulness to acknowledge my depth to all those who have helped me to put these ideas, well above the level of simplicity and into something concrete.

TABLE OF CONTENTS

List of Figures

List of Tables

Lists of Abbreviations

ATM	Automated Teller Machine
CBE	Commercial Bank of Ethiopia
CBEBirr	Commercial Bank of Ethiopia's Mobile Money
CPU	Computer Processing Unit
CRISP-DM	Cross-Industry Standard Process for Data Mining
CRM	Customer Relationship Management
CVA	Cash and Voucher Assistance
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
GDP	Gross Domestic Product
HAC	Hierarchical Agglomerative Clustering
IMF	International Monetary Fund
M-BIRR	Mobile Money provided by several International Monetary Fund
NBE	National Bank of Ethiopia
PCA	Principal Component Analysis
RAM	Random Access Memory
ROE	Return on Equity
SSD	Solid State Drive
USSD	Unstructured Supplementary Service Data
WWW	World Wide Web

Abstract

Customer segmentation helps organizations group similar customers, aiding in tailored marketing strategies. Mobile money services, like CBEBirr by the Commercial Bank of Ethiopia, are widely used in Ethiopia, with 10.2 million customers utilizing CBEBirr for services such as cash in, cash out, send money, buy airtime, pay bills, buy goods, and other financial services. Previously, CBEBirr customers were not segmented to get information. Thus, our study explores using unsupervised machine learning to segment CBEBirr customers at the Commercial Bank of Ethiopia. In this study, CBEBirr customers are segmented according to their similarities based on demographics, including age, gender, and data. CBEBirr customers are recruited by agents, branches, and merchants. Besides demographic data, behavioural data such as the number of cash-in transactions, cash-out transactions, send money transactions, buy air time transactions, pay bill transactions, and buy goods transactions of the customer were also used. The segmentation model is done using four unsupervised machine learning algorithms: K-means clustering, agglomerative clustering, density-based spatial clustering of applications with noise (DBSCAN), and mean shift, using 170,012 CBEBirr customers' data gathered from Commercial Bank of Ethiopia. To evaluate the performance of the developed model, the two most popular evaluation metrics for clustering algorithms, the silhouette coefficient, and the Davies-Bouldin index, were used. We obtained silhouette scores of 0.792, 0.676, -0.129 and 0.792 and Davies Bouldin Scores of 0.291,0.290,1.306, and 0.290 for K-means clustering, agglomerative clustering, DBSCAN, and the mean shift algorithm, respectively. Hence, this concludes that considering those evaluation metrics among the four algorithms we used to cluster our CBEBirr customer’s data, the mean shift algorithm is better than agglomerative clustering, DBSCAN, and the K-means algorithm, as displayed by the high value of the silhouette score and the low value of the Davies Bouldin score. The study will support Commercial Bank of Ethiopia in gaining knowledge about its CBEBirr customers, specifically which services of CBEBirr are more commonly used by the customers, and the bank will formulate marketing strategies accordingly, in turn achieving its goal of making a cashless society.

Keywords: ***CBEBirr, Customer Segmentation, Machine Learning, K-means, DBSCAN, HAC, and Mean Shift***

CHAPTER ONE

INTRODUCTION

1.1. Background of the study

Nowadays new technologies are being developed and e-commerce is growing, any company needs to develop new strategies that help them to win the competitive environment. Mobile Money is a mobile phone-based technology that provides financial transactions more safely and quickly across a vast geographical space [1]. During their research, Omigie, et al. [2] pointed out that the mobile financial service market is expanding and is replacing traditional financial services. It is necessary therefore to understand the fundamental service value that determines customer choice behavior to use mobile financial for market success and sustainability. Mobile Money services possess both advantages and disadvantages, along with prospects and risks. [3].

The global banking sector has undergone significant transformations in its operational landscape. The advent of electronic banking has streamlined business operations, facilitating easier capture of transactional data, resulting in a substantial increase in data volume. Analyzing this vast pool of raw data surpasses human capacity, hindering effective transformation of data into actionable insights for organizations [4].In the current competitive landscape, it is essential for survival to effectively target the appropriate customers with tailored offers at precisely the opportune moments. Machine learning emerges as a vital tool for organizations, enabling them to address various business challenges by uncovering patterns, associations, and correlations within data. Applications of machine learning encompass real-time fraud detection, targeted customer outreach, compliance adherence, risk evaluation, customer service enhancement, and analysis of purchasing behaviors [5].

A major challenge in marketing banking services lies in the substantial variation in customer preferences from one individual to another. Banks rarely have the ability to customize their services to cater to each customer's unique needs. To overcome this obstacle, banks must employ "market segmentation" techniques, leveraging the wealth of customer data at their disposal to target and sell to them in a manner akin to the strategies employed by tech giants and social media platforms [6]. Through customer segmentation, marketers can adopt a more methodical approach to future planning, thereby optimizing the utilization of marketing resources and refining their marketing initiatives. This segmentation enables organizations to categorize the market into subgroups of customers with similar everyday needs, interests, and priorities, allowing for the design and implementation of tailored strategies. By understanding the fundamental characteristics shared among their customer base, segmentation can be the difference between an underperforming service and one that resonates effectively with customers.

The Commercial Bank of Ethiopia offers multiple banking channels, including mobile banking, branch networks, online banking, and its mobile money service called CBEBirr. CBEBirr operates through mobile devices and involves the selection, training, and approval of agents who provide banking services on behalf of the bank via mobile phones. It serves as a means to extend financial services to the unbanked population and does not include savings or checking account features [7]. According to the IMF's financial access survey, mobile money refers to a digital payment method that can be accessed through a mobile money account and is supported by a network of mobile money agents. This service, provided by cellular network operators or other entities associated with them, operates independently of traditional banking networks and does not require a bank account; a simple mobile phone is sufficient. Conversely, mobile banking involves the use of applications on mobile devices to access and perform banking services such as check deposits, balance inquiries, and remittances. When mobile phones are utilized solely as a channel for accessing traditional banking products, it is classified as mobile banking rather than mobile money.

With CBEBirr, customers no longer have to travel long distances to access CBE branches; instead, they can utilize services offered by nearby CBE agents. Through CBEBirr, customers can easily deposit, withdraw, transfer funds, make payments, purchase mobile airtime, and settle bills using a simple and convenient mobile phone interface. Recruitment of CBEBirr customers is conducted by agents, branches, and merchants, and the bank has currently enlisted 10.2 million CBEBirr customers. The growth of a company's customer base is essential for achieving profitability in a competitive market environment.

Aside from the essential goals of retaining existing customers and acquiring new ones, a pivotal aspect of successful business management involves increasing profitability per customer. Various models assist business managers in implementing strategic initiatives tailored to individual customer preferences, although each model has its constraints. To elevate business performance further, an approach to customer segmentation based on diverse customer behaviours is necessary to yield greater benefits for the company [8].

The key to this paper is to use a machine learning approach to develop a model to segment CBEBirr customers based on their demographic and behavioural data in the Commercial Bank of Ethiopia. Behavioural data may be the habits of customers with the product or services: actions or inactions, spending/consumption habits, feature use, session frequency, and browsing history.

1.2. The Motivation of the Study

The Commercial Bank of Ethiopia aims to collaborate with its CBEBirr clientele to bolster the utilization of Mobile Money services, thereby offering straightforward and transparent solutions and contributing to the establishment of a cashless society. Among various electronic payment channels, mobile money stands out as particularly promising for facilitating humanitarian aid in low and middle-income nations like Ethiopia, owing to its minimal infrastructure requirements and widespread mobile phone penetration. The primary mobile money services in Ethiopia include M-BIRR (offered by six MFIs), HelloCash (offered by three banks and one MFI), CBEBirr (provided by the state-owned

Commercial Bank of Ethiopia), and Amole Mobile Money (provided by Dashen Bank S.C).

Furthermore, mobile technologies are reshaping economic activities in developing nations, where a considerable number of individuals are utilizing cell phones for various financial transactions, including sending and receiving money transfers. These mobile money platforms are accessible through basic mobile phones, with low associated transaction costs. Globally, the main channels for financial transactions include direct payments into bank accounts, mobile money, prepaid or debit cards issued by banks, and e-voucher systems. Mobile money services cater to individuals who are geographically isolated and/or have limited incomes, facilitating cashless payments, reducing dependence on cash, and enabling the tracking of transaction records [9].

CBEBirr, provided by the state-owned Commercial Bank of Ethiopia, stands as the sole mobile money service offered by a major state-owned bank in Ethiopia. The bank is concentrating its efforts on enhancing its CBEBirr customer base to expand the number of mobile money users and advance the transition towards a cashless society. To achieve this objective, the bank is segmenting its existing CBEBirr customers based on their usage patterns. This segmentation strategy aids the bank in refining its marketing tactics and promotional endeavors, ultimately driving towards the realization of its vision for a cashless society and extending financial access to the unbanked population. Consequently, the bank is motivated to develop a model for CBEBirr customer segmentation through the utilization of machine learning approaches, as this facilitates a deeper understanding of the behaviors exhibited by existing CBEBirr customers.

1.3. Statements of the Problem and Justification of the Study

Customer segmentation is a crucial aspect of business intelligence, involving the categorization of consumers based on similar demographic, geographic, behavioural, and psychographic characteristics. Clustering methods identify groups that exhibit internal homogeneity but external diversity. Given the variations in behaviour, needs, desires, and traits among customers, the primary objective of clustering techniques is to discern distinct

customer types and segment the customer base into clusters with comparable profiles, facilitating more efficient target marketing [10].

Mobile money users are highly increasing entire the country as several banks and other institutions are providing the services or products. Everyone here and there is using mobile money as it makes life easier. Mobile money customers were largely segmented primarily by transaction type runs, which were then divided into segments such as acceptors, bulk mailers, depositors and withdrawers, broadcast time merchants, and service providers or agents [11]. This segmentation model, which primarily depends on behavioural data or transaction run falls in obtaining a 360-degree view of the consumer, just like every other segmentation model used about a product or group of items. Whereas the majority of market segments are developed based on demographic data like gender, location, age, occupation and other elements impacted by technological innovation are emerging that need a shift in consumer behaviour.

CBEBirr, a mobile money service offered by the Commercial Bank of Ethiopia, boasts a customer base of 10.2 million. Currently, CBEBirr predominantly facilitates services like airtime purchases, goods purchases, bill payments, cash-in, cash-out, among others. Understanding customer behaviours and the utilization patterns of CBEBirr services is paramount for the Commercial Bank of Ethiopia to effectively manage its CBEBirr clientele. However, there lacks an approach to gather crucial information from CBEBirr customers. Consequently, this study aims to segment CBEBirr customers using unsupervised machine learning techniques to obtain essential insights and differentiate between the utilization of CBEBirr services.

Numerous studies have investigated customer segmentation in foreign banks and other institutions, recognizing its importance in Customer Relationship Management (CRM) and customer recruitment [12]. To achieve this, banks must invest resources in comprehensively understanding existing and potential customers.

Additionally, recent local studies [13], [14], [15] have examined challenges, practices, and operational barriers associated with CBEBirr services, employing qualitative and

quantitative approaches. These studies utilized primary data collected from CBEBirr customers through questionnaires sourced from various channels.

This study endeavours to segment CBEBirr customers based on demographic and behavioural similarities using four unsupervised machine learning algorithms: K-means, DBSCAN, Agglomerative Clustering (bottom-up), and Means-Shift. The findings will aid the Commercial Bank of Ethiopia in understanding its CBEBirr clientele, retaining existing customers, focusing on targeted segments, and enhancing its competitive edge. Furthermore, the bank can formulate marketing strategies accordingly to advance its goal of fostering a cashless society. Hence, this research investigates machine learning techniques for customer segmentation among CBEBirr customers of the Commercial Bank of Ethiopia, addressing the following research questions:

- What are the features and characteristics of the CBEBirr customers that can be used for CBEBirr customer segmentation?
- Which machine learning techniques are best performed to Segment CBEBirr customers?

1.4. Objective of the study

1.4.1. General Objective

The general objective of this research is to develop a model for segmenting customers based on their behavioural and demographic features in Commercial Bank of Ethiopia using a machine learning approach.

1.4.2. Specific Objectives

The following specific objectives are set to meet the general objective of the study:

- To review related works done to acquire a concept, principle, and techniques of customer segmentation.
- To collect a dataset from CBE, identify the features for modelling CBEBirr customer behaviour, perform data pre-processing tasks, and prepare the dataset for segmentation.

- To develop machine learning models for CBEBirr customer segmentation
- To evaluate the performance of the developed models.

1.5. Significance of the Study

This study will investigate the applicability of machine learning techniques in the Commercial Bank of Ethiopia to build models that can segment CBEBirr customers. Accordingly, the motivation behind this study is to use the unsupervised AI approach to track down the characteristics and behaviour of CBEBirr customers in the Commercial Bank of Ethiopia. As stated earlier, one of the ways of getting exact customer information is through segmentation and the benefits of segmentation are extensive. Customer segmentation will enhance the marketing strategy or plan based on the type of segmented customers.

This study is important because:

- Useful in Retaining CBEBirr customer
- Ensure efficient resource allocation for all marketing components
- Concentrated on the target customers' specific needs Prioritisation of the most profitable present and potential customer groupings, including the finding of potential new growth of customer
- It will try to identify the groups of customers that will require an advertisement and information concerning the use of CBEBirr service.
- The study will have an important role for CBE to achieve its goal of creating a cashless society.
- Supports the bank to strengthen the competitive position of the company.

Generally, the study will support the Commercial Bank of Ethiopia to maximize the value of CBEBirr customers in their marketing strategies to achieve its goal of making a cashless society.

1.6. Methodology

The following set of guidelines, methods, and processes were used to achieve the overall objectives of this research.

Literature Review: A review of the literature was conducted to understand various components of data categorization. Specifically, we reviewed different literature in the area of customer segmentation which is explored locally and globally.

Data collection: We collected CBEBirr customer data from the commercial bank of Ethiopia Datawarehouse. The data was organized and structured in a way that they are easy for experimentation and testing.

Data preprocessing: We cleaned the data by removing duplicates, handling missing values, and transforming variables.

Feature selection: We selected the most relevant features from the collected data (such as Age, Gender, recruited by, Number of Buy Air Time Transaction, Number of Send Money Transaction, Number of Pay Bill Transaction, Number of Buy Goods Transaction, Number of Cash-in Transaction, and Number of Cash-Out Transaction) that can differentiate customer behaviour.

Experimentation: We experimented with the available tools (Jupyter Notebook version 6.5.4) and engaged in programming to complete the research's goals. We also utilized a high-performance laptop equipped with 16 GB of RAM, a 3.0 GHz CPU, and a 1TB SSD hard disk for our analysis needs.

Testing and Model Evaluation: The system was put to the test using collected data from the company to assess the performance of the suggested solution. By using evaluation methodologies which are the Silhouette Coefficient and Davies-Bouldin index, the outcome was evaluated.

1.7. Scope and Limitation of the Study

The research primarily focuses on segmenting CBEBirr customers using an unsupervised machine learning approach within the Commercial Bank of Ethiopia. This approach is chosen for its ability to uncover hidden patterns and insights within the dataset, distinguishing it from other machine learning methodologies. Four clustering algorithms, namely K-means, DBSCAN, Agglomerative Clustering (bottom-up), and Means-Shift, are

utilized in this study. The study utilizes secondary data obtained from the Commercial Bank of Ethiopia, considering historical data spanning six months. The selection of this data aligns with the literature that initiated the research problem. Segmentation is conducted based on demographic factors such as age and gender, as well as customer behavioural data, including the number of cash-in, cash-out, send money, buy goods, pay bill, and buy airtime transactions of CBEBirr customers.

The main tasks undertaken during the study include data collection, data preparation, model construction, model evaluation, and selection of the optimal model. The research is conducted within an academic framework, excluding the use of demographic data such as economic characteristics, race, income level, and education level of CBEBirr customers, as well as behavioural data like feature usage and duration. Additionally, the deployment of the models falls outside the scope of this study.

1.8. Organization of the Thesis

This thesis is structured into five chapters. The first chapter introduces the background, outlines the problem statement, and delineates both the general and specific objectives of the study. Additionally, it addresses the scope and limitations of the research, presents the methodology employed, and highlights the contribution of the study. Chapter two centres on the literature review and previous related works, specifically focusing on the fundamental concepts of customer segmentation. The third chapter provides a comprehensive overview of the materials, major techniques, and methodologies utilized in the study, along with presenting the proposed system architecture. Chapter four offers an in-depth analysis of the findings, results, and data analysis. Lastly, chapter five presents the conclusions drawn from the study and offers recommendations based on the research findings.

CHAPTER TWO

LITERATURE REVIEW

2.1. Introduction

In numerous fields of study, the categorization of data into cohesive groups based on shared information holds significant importance. Clustering stands out as a widely utilized technique across various sectors, serving numerous purposes such as document organization, classification, collaborative filtering, and customer segmentation among others [16]. Customer segmentation involves the division of customers into groups based on their attributes, behaviors, or preferences. This process enables businesses to gain deeper insights into their customer base, tailor marketing strategies accordingly, and enhance customer satisfaction and loyalty. While segmentation methods encompass demographic, geographic, psychographic, and behavioral criteria, they often rely on predefined rules or assumptions, potentially overlooking the intricate nuances and diversity within customer data. To address this limitation, machine learning techniques have seen growing application in customer segmentation tasks, as they possess the capability to autonomously identify patterns and clusters within large and heterogeneous datasets. In this chapter, we provide an extensive review of relevant literature concerning segmentation, emphasizing its importance across various domains. Specifically, we delve into the Ethiopian banking sector, offering background insights into the Commercial Bank of Ethiopia (CBE). Additionally, we conduct thorough examinations of fundamental concepts related to customer data records within the banking industry, highlighting the challenges associated with data collection in this context. Furthermore, we explore diverse clustering techniques employed in similar scenarios and discuss pertinent model evaluation metrics pertinent to segmentation endeavors within the banking sector.

2.2. Bank Sector in Ethiopia

The Ethiopian banking sector comprises 30 privately-owned banks and one state-owned entity, the Commercial Bank of Ethiopia [17] . These banks collectively amassed 1.7 trillion birrs in deposits from 33 million savers during the 2022 fiscal year. Notably, the Commercial Bank of Ethiopia extended credit amounting to 890.1 billion birrs, equivalent to 67 percent of the country's GDP. Moreover, the sector contributed nearly $10 billion to annual foreign exchange trading. Employment within the banking industry stands at approximately 90,000 individuals, with an annual net increase of around 10,000 jobs over the past five years, offering attractive returns to approximately 115,000 shareholders. Notably, one-fifth of the government's annual income tax revenue, totalling 13 billion birrs, is sourced from the banking sector. These statistics underscore the significant role played by the banking sector in the country's economy.

Over the previous decade, both the Commercial Bank of Ethiopia (CBE) and other commercial banks in Ethiopia have witnessed consistent and substantial profits and expansion. Data indicates an annual growth rate of 28 percent for deposits, 31 percent for loans, and 22 percent for profits [18]. Private commercial banks in Ethiopia exhibit notable distinctions compared to their counterparts in other economies within Africa and Asia. They operate within stringent regulatory frameworks, maintain comparatively smaller scales of operation, prioritize lending activities over alternative revenue streams, experience lower levels of non-performing loans, incur reduced overhead costs, and demonstrate slightly higher profitability in terms of return on equity (ROE).

2.3. Commercial Bank of Ethiopia

The origins of the Commercial Bank of Ethiopia (CBE) trace back to the establishment of the State Bank of Ethiopia in 1942. CBE was formally incorporated as a joint-stock company in 1963. In 1974, it merged with the privately owned Addis Ababa Bank. Since then, CBE has played a pivotal role in the country's development. Presently, CBE boasts more than 40.3 million account holders and is actively involved in providing a range of

commercial banking services to its clientele, including digital banking, checking accounts, savings accounts, and loans [19].

To facilitate the operational activities of the bank, the Commercial Bank of Ethiopia (CBE) has organized its management into thirty districts, with ten located in Addis Ababa and the remaining twenty distributed across various regions. The highest authority within the bank's structure lies with the board of directors, comprised of a chairperson and nine additional members. Furthermore, overseeing the bank's core operations is the executive council, which serves as the top administrative level responsible for managing key activities [19].

In its endeavor to position itself as a leading commercial bank driving Ethiopia's financial future, CBE has embarked on various initiatives, recognizing the strategic importance of information technology (IT) services in today's banking landscape. Technological infrastructures have become indispensable to the success and performance of companies, including CBE. The continuous availability of these services is a key feature enabling the bank to maintain its leading position in the sector. To attract and retain customers, CBE has implemented progressive measures such as extended business hours, ATM services, internet banking, and enhanced banking hall facilities, all aimed at improving customer convenience. CBE is making significant investments in diverse banking technologies to enhance its service delivery and overall performance. Consequently, numerous IT projects have been deployed, encompassing core banking, Enterprise Resource Planning (ERP), card banking, Call Centre, and IT infrastructure, leveraging advanced and sophisticated high-end servers. Additionally, several projects are currently in progress. The emergence of new private banks has further enriched the industry, prompting CBE to continuously adapt and enhance its services to remain competitive. The primary drivers behind CBE's adoption of technologically advanced services include evolving customer requirements and preferences, as well as aspirations to enhance organizational performance. Despite encountering challenges across various sectors, such as improving customer engagement, reducing transaction costs, expanding geographical coverage, building organizational reputation, and enhancing customer satisfaction, CBE has embraced electronic payment

systems to streamline business processes, reduce paperwork, transaction costs, and labor costs. The aggressive expansion efforts of the bank from multiple perspectives have led to a substantial increase in transaction volumes and customer numbers, consequently generating vast amounts of data. Therefore, the digital transformations undertaken by the bank are aimed at managing this large volume of data efficiently.

CBE is actively enhancing its digital payment solutions, notably through the introduction of the CBEBirr mobile money service on December 11, 2017. This digital payment channel aims to extend access to banking services to previously unbanked and underserved segments of society by enabling customers to conduct transactions via their mobile devices through contractual agreements with customers, agents, and merchants. Presently, the bank serves a substantial customer base of 10.2 million individuals, facilitated by a network of over 43,622 agents nationwide, and more than 56,831 merchants who accept payments through CBEBirr. CBEBirr (*847#) functions as a mobile banking service provided by the Commercial Bank of Ethiopia through designated banking agents, who are selected, trained, and authorized by the bank to deliver banking services on its behalf via mobile phones. Through CBEBirr, users can conveniently make payments for various goods and services, including groceries, online purchases, and flight tickets, among others [20].

2.4. Customer Segmentation

According to Kolter et al. [21], customer segmentation refers to the practice of categorizing a company's customers into groups that share similarities, aiming to convert customer data into actionable insights. This segmentation aids businesses in maximizing the value derived from each customer by guiding how they interact with customers within each segment. Customers represent the most valuable asset for any organization, necessitating the development of specific strategies for customer management. The concept of market segmentation emerged in the 1950s alongside product differentiation, serving as a key marketing strategy. Subsequently, in the 1970s and 1980s, market segmentation evolved into a strategy providing a competitive advantage. By the 1990s, target or direct marketers began employing sophisticated techniques, including market

segmentation, to offer customers the best-customized offerings possible [22]. The primary objective across industries is to comprehend each customer individually and leverage this understanding to facilitate seamless transactions, fostering customer loyalty and retention. Customer segmentation involves dividing customers into distinct and meaningful groups based on shared attributes. This segmentation enables marketers to efficiently tailor their approaches to each customer group. By harnessing a wealth of available customer data, segmentation analysis allows marketers to accurately identify discrete customer groups based on demographic, behavioral, and other relevant indicators.

Since the primary aim of marketers is typically to optimize the value, whether in terms of revenue or profit, derived from each customer, it becomes crucial to anticipate how specific marketing actions will impact customers. Ideally, such "action-centric" customer segmentation should not solely focus on the immediate value of marketing activities but rather consider the long-term impact on customer lifetime value (CLV) that these actions will bring about. Traditional variables used for market segmentation can generally be categorized into five main groups: Demographic, Geographic, Psychographic, Behavioral, and Product-Related Factors. These categories are elaborated below according to [23] [23].

Demographic segmentation involves dividing the market into segments based on objective and factual data that differentiates them. Variables for demographic segmentation encompass age, gender, income, occupation, marital status, family size, race, religion, and nationality. Due to the ease of measuring these variables, demographic segmentation is a widely used method for segmenting customer markets.

Geographic segmentation entails identifying potential markets based on their geographical location. This segmentation considers variables such as climate, terrain, natural resources, and population density, among other geographic factors. Geographic segmentation is crucial because these variables often distinguish customers across different regions.

Psychographic segmentation involves grouping the customer market based on psychological variables such as values, lifestyle, attitudes, interests, and personality traits.

A psychographic approach to segmentation can be employed independently or in conjunction with other segmentation variables. Different consumers may respond differently to marketing efforts based on their psychological characteristics.

Behavioral segmentation categorizes individuals into distinct groups that share specific behavioral patterns. These patterns may include similar lifecycle stages, previous purchase behaviors, or consistent responses to marketing messages.

Product-related segmentation relies on variables and their associations with the product, such as segmenting based on sought-after product benefits, usage rates, or brand loyalty.

While various segmentation methods exist, for a segmentation strategy to be successful, the market segment should ideally possess the following four parameters.

a) Measurability: It should be feasible to quantify the size of the segment.
b) Substantiality: This refers to the extent to which segments are sufficiently profitable to warrant pursuit through customized marketing initiatives.
c) Accessibility: It pertains to the extent to which segments can be reached and served effectively.
d) Actionability: This concerns the extent to which effective strategies can be developed to attract and serve relevant segments.

2.4.1. Customer segmentation in the financial industry

In the era of technology, information has become the most valuable asset for financial institutions. It offers valuable insights into product effectiveness, customer behavior, market dynamics, and other crucial aspects. Customer segmentation is pivotal in leveraging this data, enabling a deeper understanding of individual customer needs and preferences. This, in turn, facilitates more effective targeting strategies aimed at boosting revenue.

Currently, within the realm of commercial banking, customer relationship management (CRM) stands as a central aspect of customer-centric bank management. Success lies in the adept utilization of machine learning for accurately categorizing and processing customer data within commercial banks. Subsequently, segmentation of customers enables timely adjustments and enhancements to address market shifts and pursue

additional business opportunities. In the landscape of commercial economics, machine learning has emerged as a critical foundation for customer segmentation within commercial banks.

Effective business growth strategies invariably commence with thorough customer segmentation. This practice allows organizations to pinpoint the distinct attributes of their product and service consumers, facilitating the formulation of efficient business strategies. Within the banking industry, customer segmentation has evolved into a valuable asset for acquiring new customers and maximizing the value derived from existing ones.

The banking sector increasingly acknowledges the significance of the wealth of information it possesses about its clientele. This encompasses a vast array of customer data, spanning demographics and transactional records. Banks must allocate resources toward gaining deeper insights into both their current and potential customer base. Leveraging appropriate machine learning tools enables banks to explore various avenues within the industry. These include customer segmentation and profitability analysis, credit scoring and approval processes, prediction of payment defaults, targeted marketing initiatives, detection of fraudulent activities, efficient cash management, operational forecasting, optimization of investment portfolios, and prioritization of investments. Furthermore, machine learning can aid banks in identifying their most lucrative credit card customers or assessing the risk associated with loan applicants [24].

In ensuring the success of bank service offerings, segmentation practices significantly contribute to the overarching marketing strategy. This involves the division, identification, and assessment of market segments by marketing managers, enabling them to tailor a marketing mix that resonates with each segment's specific characteristics.

The banking sector is increasingly acknowledging the significance of its wealth of customer information. Undoubtedly, it possesses one of the most extensive and comprehensive repositories of customer data, encompassing demographics, transaction records, credit card usage patterns, and more. Given that banking operates within the service industry, the imperative of maintaining robust and efficient customer relationship

management (CRM) cannot be overstated. Below are some of the key roles that machine learning plays in the banking industry:

One of the most commonly utilized applications of machine learning in the banking sector is in marketing. By leveraging machine learning, the bank's marketing department can analyze customer databases comprehensively. Machine learning conducts various analyses on the collected data to discern consumer behavior pertaining to product preferences, pricing dynamics, and distribution channels. Furthermore, it allows for gauging customer responses to both existing and new products, thereby enabling banks to tailor promotional efforts, enhance product and service quality, and gain a competitive edge. Bank analysts can also utilize machine learning to scrutinize past trends, ascertain current demand levels, and forecast customer behaviors across different products and services, thereby capitalizing on additional business opportunities and pre-empting behavioral patterns. Additionally, machine learning aids in identifying profitable customers while distinguishing them from less profitable ones. Techniques in machine learning can also predict customer reactions to changes in interest rates and assess the risk profile of specific customer segments regarding loan defaults [24].

Machine learning serves as a prevalent tool for risk management within the banking sector. Bank executives face the imperative of determining the reliability of their customers. Initiatives such as extending credit card offers to new customers, increasing credit limits for existing customers, and approving loans all entail inherent risks if banks lack comprehensive insights into their customers. When assessing loan applications, banks scrutinize various factors such as loan amount, interest rates, repayment terms, type of collateral, customer demographics, income levels, and credit histories. Customers who have maintained long-standing relationships with the bank and belong to higher income brackets typically find it easier to secure loans. Despite the cautious approach adopted by banks in loan provisioning, there remains a risk of default by customers. Machine learning techniques aid in discerning borrowers who exhibit prompt repayment behaviour from those who do not, thus assisting banks in mitigating loan default risks [24].

Another prominent domain where machine learning finds application in the banking sector is in fraud detection. The ability to identify fraudulent activities is becoming increasingly crucial for businesses, and with the aid of machine learning, more instances of fraud can be identified and addressed. Financial institutions have developed two distinct approaches to detect patterns of fraud. In the first approach, a bank leverages the data repository of a third party and employs machine learning algorithms to uncover fraudulent patterns. Subsequently, the bank cross-references these patterns with its internal database to identify any indications of internal irregularities. Conversely, the second approach relies solely on the bank's internal data to identify fraud patterns. Many banks opt for a "hybrid" approach, combining elements of both strategies to enhance fraud detection capabilities [24].

In today's fiercely competitive landscape, customers are deemed paramount. Machine learning proves valuable across all stages of the customer relationship cycle: Customer Acquisition, Increasing Customer Value, and Customer Retention. Particularly in the banking industry, customer acquisition and retention are pivotal concerns. Machine learning methodologies aid in discerning loyal customers from those inclined to switch banks for improved services elsewhere. By analysing customer behaviours, banks can identify the reasons behind customer attrition and determine the last transaction conducted before switching banks. Armed with this insight, banks can enhance their operations and implement strategies to better retain their customer base [24].

2.5. Clustering Techniques for Customer Segmentation

Utilizing machine learning methods to uncover insights and patterns within customer data proves highly effective. Decision-makers can harness artificial intelligence models as powerful tools, allowing them to accurately delineate client segments—a task considerably more arduous to accomplish manually or through traditional analytical methods. A range of machine learning algorithms exists, each tailored to specific types of problems. Among these, the k-means clustering approach stands out as widely employed for customer segmentation tasks. Additionally, other clustering techniques such as Density-Based

Spatial Clustering of Applications with Noise (DBSCAN), agglomerative clustering, mean-shift, among others, are available for similar purposes [25].

Clustering is a machine learning method where data points are grouped together based on similarities. Using a clustering algorithm, each data point within a dataset can be assigned to a specific group. The fundamental idea is that data points within the same group share common properties or traits, while those in different groups exhibit distinct features. Clustering operates under unsupervised learning, representing a statistical approach to data analysis utilized across various fields [26].

In the process of clustering, the dataset undergoes segmentation, resulting in several groups where data points within each segment exhibit greater similarity to one another compared to data points in other clusters. These segments are formed by identifying relationships based on the variables present in the raw data.

The objective of clustering is to identify an appropriate number of clusters that offer meaningful insights for analysis and evaluation. This is an iterative process involving the examination of extensive raw data to identify similarities, relationships, and patterns. Through this process, uncategorized data is scrutinized to extract relevant knowledge, and data points are subsequently assigned to appropriate clusters.

Various clustering algorithms employ distinct methods to cluster data points based on their characteristics [10].

- Hierarchical Clustering encompasses two methods: agglomerative (bottom-up) and divisive (top-down) approaches.
- Grid-based clustering algorithms partition data points into grid structures composed of multiple cells. This approach utilizes subspace and hierarchical clustering methods. Sting and Clique are examples of grid-based clustering algorithms.
- Partitioned-Based Clustering initially treats all data points as a single cluster. Subsequently, these points are iteratively grouped into clusters by aligning objects between clusters. Examples of partitioning algorithms include K-means, K-medoids, and K-modes.

- Density-based clustering identifies clusters as regions with higher density compared to other parts of the dataset. Core, noise, and border points serve as differentiators for objects within this approach.

In unsupervised machine learning algorithms, various types of clustering methods cater to diverse datasets. Among these algorithms, we have utilized four: DBSCAN, agglomerative algorithm, K-means, and mean-shift algorithms.

2.5.1. Partition Clustering Approach

For a database comprising n objects, the partitioning technique generates k partitions of data, ensuring that each object belongs to precisely one group and that each group contains at least one object. Partitioning techniques enhance iterative movement by transferring objects from one group to another. The primary aim of the partition clustering algorithm is to divide data points into K partitions. This method enhances partitioning by establishing initial partitions and utilizing iterative movement to relocate objects among groups. Unlike multi-step clustering methods, partitioning clustering produces clusters in a single step, culminating in the formation of one set of clusters, though multiple sets may be generated internally [27]. The choice of partitioning method is guided by specific objective functions, such as minimizing the square error criterion. However, a weakness of such algorithms arises when the distance between two points from the center closely aligns with another cluster, potentially resulting in poor or misleading outcomes due to data point overlap. Among the most renowned and frequently used partitioning methods are K-means and K-medoids.

K-Means Clustering

K-means stands as one of the most widely used centroid-based clustering algorithms [10]. While effective, centroid-based clustering algorithms like K-means are highly sensitive to initial conditions and outliers. In the case of K-means, it is necessary to initially select several classes or groups and randomly initialize their respective centroid points. Identifying the initial classes typically involves a quick examination of the dataset to identify any distinct groupings. Each data point is then clustered by computing the distance

between that point and each centroid, assigning the point to the cluster whose centroid is closest to it. Following this classification, the centroids are recalculated by averaging all the vectors within each cluster. K-means offers advantages such as fast computation and effectiveness, making it suitable for handling large datasets.

Let's consider a dataset, denoted as S, comprising n objects within Euclidean space. Partitioning methods aim to distribute the objects in S into k clusters, denoted as C1,..., Ck, where each Ci is a subset of S, and Ci ∩ Cj = ∅ for ($1 \leq i, j \leq k$). An objective function is employed to evaluate the quality of partitioning, ensuring that objects within a cluster exhibit similarity to one another while being dissimilar to objects in other clusters. This objective function seeks to maximize intra-cluster similarity and minimize inter-cluster similarity. In centroid-based partitioning techniques, each cluster Ci is represented by its centroid, ci. The disparity between an object p ∈ Ci and ci, the centroid of the cluster, is measured using the Euclidean distance, denoted as dist(p, ci), where dist(x, y) represents the Euclidean distance between two points x and y.

The effectiveness of cluster Ci can be assessed by its within-cluster variation, calculated as the sum of squared errors between all objects in Ci and its centroid ci [28]:

$$E = \sum_{i=1}^{k} \sum_{p \in Ci} dist(p, ci)\ 2 \qquad (1)$$

Here, E represents the sum of squared errors for all objects in the dataset; p denotes the point in space representing a specific object; Ci signifies the centroid of cluster Ci. Both p and ci are multidimensional.

The algorithm can be summarized in the following four steps:

Step 1: Select k objects randomly. These objects represent the initial group of centroids.

Step 2: Assign each object to the group that has the closest centroid.

Step 3: When all objects have been assigned, recalculate the positions of the k centroids.

Step 4: Repeat Steps 2 and 3 until the centroids no longer move.

Mean shift Clustering

Mean shift seeks to identify clusters within the intricate density of samples. It operates as a centroid-based algorithm, iteratively updating candidate centroids such that they represent the average of points within a specified region. Subsequently, these candidates undergo filtering in a post-processing phase to eliminate nearly identical duplicates, resulting in the formation of the final set of centroids.

When considering a candidate centroid xi for iteration ti, it is updated based on the following equation [29]:

$$x_i^{t+1} = m(x_i^t) \qquad (2)$$

Here, N(xi) represents the neighbourhood of samples within a specified distance around xi, and m denotes the mean shift vector calculated for each centroid. This vector points towards an area where there is the maximum increase in the density of points.

This calculation involves updating a centre to compute the mean of the samples within its neighbourhood [29]:

$$m(x_i) = \frac{\sum xj \epsilon N(xi) K(xj - xi) xj}{\sum xj \epsilon N(xi)\, K(xj - xi)} \qquad (3)$$

The algorithm autonomously determines the number of clusters instead of depending on a parameter known as bandwidth, which determines the size of the region to explore. While this parameter can be manually configured, an estimate bandwidth function is available to estimate it if not explicitly set.

Although the algorithm is guaranteed to converge, it is not highly scalable because it necessitates multiple nearest neighbour searches during execution. Iterations cease when there is minimal change in centroids.

Hierarchical Clustering Approach

Hierarchical clustering, also referred to as connectivity-based clustering, aims to construct a hierarchy of clusters. It operates on the principle that objects are more closely related to nearby objects than those farther away. These algorithms form clusters by linking "objects" based on their distances. A cluster is primarily defined by the maximum distance required to connect its components. Different clusters emerge at various distances, depicted through a dendrogram. Hence, the term "hierarchical clustering" originates from the fact that these algorithms don't yield a single partition of the dataset but rather offer a broad hierarchy of clusters that merge at specific distances. In a dendrogram, the y-axis indicates the merging distance of clusters, while objects are positioned along the x-axis to prevent cluster overlap [30]. Strategies for hierarchical clustering generally fall into two categories:

Divisive method

This method, also referred to as the top-down approach, begins with all objects in a single cluster. Through successive iterations, clusters are divided into smaller clusters until each object is in its own cluster or until a termination condition is met. This approach is inflexible, meaning that once a merging or splitting operation is performed, it cannot be reversed.

Agglomerative Clustering method

Agglomerative clustering, also referred to as the bottom-up approach, begins with each object forming its own individual cluster. The method then proceeds to merge neighbouring objects until all clusters are merged into one or until a termination condition is met. This technique supports three distinct approaches: single-link, complete-link, and average-link. In the single-link approach, clusters are merged based on the smallest distance between their closest members in each step. On the other hand, the complete-link approach merges clusters with the smallest diameter or maximum pairwise distance in each step. Meanwhile, the average-link approach represents a compromise between the sensitivity of complete-link clustering to outliers and the tendency of single-link clustering to create elongated

chains that don't align with the conventional notion of clusters as compact, spherical entities [30].

Hierarchical Clustering is widely employed for managing complex systems, wherein elements of such systems can initially be grouped into N clusters. These clusters can then be further subdivided into M sub-clusters (where M > N) until the desired number of clusters is achieved [31].

It's a method that merges smaller clusters to form larger clusters. In the subsequent section, we'll introduce hierarchical agglomerative clustering, as recommended by [32].

Let's suppose X represents a dataset of a complex system, comprising n elements {X1, X2,…, Xn}, and the objective is to group these elements into K clusters. The clustering algorithm proceeds as follows:

> *Step 1:* Define the distance between clusters, distance between elements, and distance between element and cluster.
>
> *Step 2:* Compute the distance between elements pair-wisely, and the pair which have the smallest distance and merge them into a new cluster.
>
> *Step 3:* Compute the distance between clusters and the distance between elements and clusters, then merge the pair with the smallest distance. Keep merging until the number of clusters reduces to K.

It's important to recognize that the distance mentioned above can be defined in three ways: the distance between clusters, the distance between elements, and the distance between an element and a cluster. When considering elements as singleton clusters, all distances become distances between clusters. Theoretically, the result of agglomerative clustering may be indeterminate if different pairs of elements or clusters have identical distances. However, such a scenario rarely occurs in practice.

According to [33], various methods exist for defining inter-cluster distance or similarity. Here are some of them:

a. Single-linkage, utilized in hierarchical clustering, combines groups based on the shortest distance across all potential pairs, as depicted in figure 2.1. It determines the shortest distance between two points, i and j, belonging to clusters R and S, respectively. This is accomplished by identifying the shortest distance between any point in cluster R and any point in cluster S.

$$L(R,S) = min\big(D(i,j)\big), i \in R, j \in S \qquad (4)$$

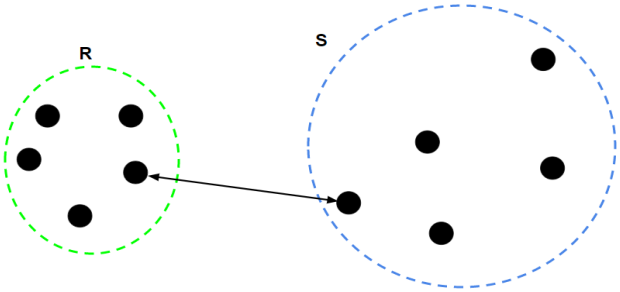

Figure 2. 1*: Single linkage looks at the minimum distance between all inter-group pairs.*

b. In complete-linkage clustering, instead of selecting the shortest distance, the distance between two groups is determined by the largest distance across all potential pairs, as illustrated in figure 2.2. It computes the maximum distance between two points, i and j, where i belongs to cluster R and j belongs to cluster S, for two clusters R and S. This approach prioritizes the maximum distance between any point in cluster R and any point in cluster S.

$$L(R,S) = max\big(D(i,j)\big), i \in R, j \in \qquad (5)$$

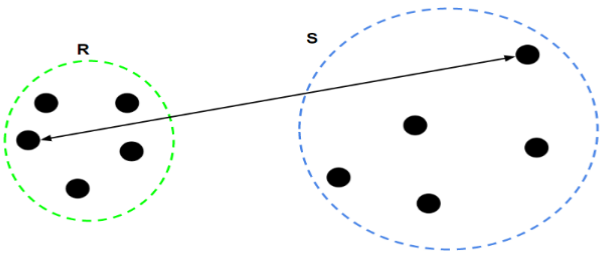

Figure 2. 2*: Complete linkage looks at the maximum distance between all inter-group pair*

c. The average linkage method can be considered a middle ground between the single and complete linkage criteria. It generates clusters that are relatively compact yet may exhibit some elongated shapes, as depicted in figure 2.3. The method computes the arithmetic mean of distances between all pairs of points, one from cluster R and the other from cluster S, for two clusters R and S. This mean distance value is returned as the result of the average linkage process.

$$L(R,S) = \frac{1}{n_R + n_S} \sum_{i=1}^{n_R} \sum_{j=1}^{n_S} D(i.j), i \in R, j \in S \quad (6)$$

Where, n_R Number of data points in R and n_S Number of data points in S

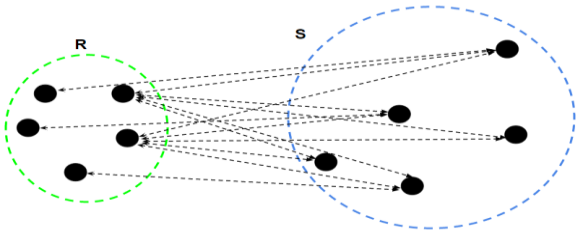

Figure 2. 3*: Average linkage uses the average distance between all inter-group pairs.*

Density-Based Clustering Approach

This approach relies on the concept of density. The fundamental concept is to expand the designated cluster as long as the density within its vicinity surpasses a certain threshold. In other words, for every data point within a specific cluster, the radius of that cluster must encompass at least a minimum number of points. Density-based regions are well-suited for capturing arbitrary-shaped clusters, although attribute selection and cluster selection with these algorithms tend to be more intricate. One notable feature is the ability to merge two clusters that are sufficiently proximate to each other.

DBSCAN

Density-based clustering is an unsupervised learning technique used to identify clusters or groups within data. It operates on the principle that a cluster represents a contiguous region

of high point density in data space, separated from other clusters by regions of low point density. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) serves as a fundamental algorithm for density-based clustering. It has the capability to uncover clusters of varying shapes and sizes within large datasets, even in the presence of noise and outliers. Two parameters are essential for applying the DBSCAN algorithm. The first parameter, minPts, specifies the minimum number of points required to be clustered together for a region to be considered dense. The second parameter, eps (ε), determines the distance measure used to identify points within the neighbourhood of any given point [34]. Epsilon and minPts can be defined through the epsilon and min point parameters, respectively. DBSCAN commences by selecting an arbitrary starting point that has not yet been visited. It then retrieves the epsilon-neighbourhood of this point, and if it contains a sufficient number of points, a cluster is initiated. Otherwise, the point is labelled as noise.

It's important to note that a point initially classified as noise might later be incorporated into a cluster if it is discovered to be within the sufficiently sized epsilon-neighbourhood of another point. When a point is identified as a dense part of a cluster, its epsilon-neighbourhood also becomes part of that cluster. Consequently, all points within the epsilon-neighbourhood are included, along with their own epsilon-neighbourhoods if they are also deemed dense. This process persists until the density-connected cluster is fully delineated. Subsequently, an unvisited point is retrieved and processed, leading to the identification of another cluster or noise. In cases where there is no "id" attribute present, this procedure will generate one. The label "Cluster 0" assigned by the DBSCAN operator signifies points categorized as noise, indicating those with fewer than the minimum required points in their epsilon-neighbourhood [30].

According to researcher [35], this algorithm categorizes data points into three types.

- **Core Point:** A point is classified as a core point if it possesses more than MinPts points within the specified eps distance.
- **Border Point:** A point with fewer than MinPts points within eps, yet it resides within the neighbourhood of a core point.

- **Noise or Outlier:** A point that does not meet the criteria to be labelled as a core point or border point.

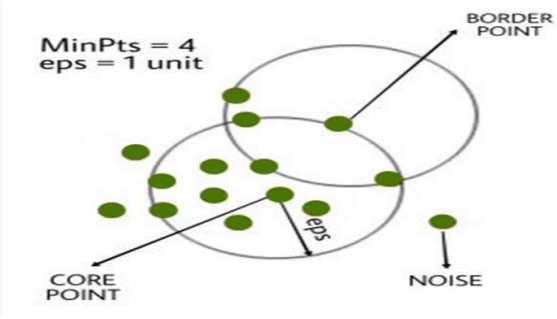

Figure 2. 4*: Four (4) Density-based clustering*

2.6. Model Evaluation Metrics

Assessing the outcomes of a clustering algorithm is a crucial aspect of the data clustering procedure. When scrutinizing clustering results, various attributes need to be considered to validate the efficacy of the clustering algorithm results [36]:

- Identifying the clustering tendency within the dataset
- Establishing the appropriate number of clusters
- Evaluating the quality of the clustering outcomes independent of external information

To assess the effectiveness of algorithms in CBEBirr customer segmentation, the following two internal cluster validation metrics will be employed.

2.6.1. Silhouette Coefficient

The silhouette coefficient evaluates both cluster cohesion and separation. It measures how closely customer data points are grouped together within their assigned cluster and how distinct they are from data points in other clusters. This metric, ranging from -1 to 1, assesses the effectiveness of a clustering technique [37].

1: This means clusters are well apart from each other and distinguished.

0: This means clusters are indifferent, or we can say that the distance between clusters is not significant.

-1: This means clusters are assigned in the wrong way.

Silhouette Score is calculated as [37]:

$$\bar{S} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{b(i)-a(i)}{max\{a(i),b(i)\}}\right) \quad (7)$$

In this context:

- $a(i)$ signifies the average distance from sample i to other samples within the same cluster,
- $b(i)$ denotes the minimum distance from sample i to any sample in a different cluster.

Internal cluster validation aims to enhance the similarity among data points identified within the same cluster while maximizing the dissimilarity between data points belonging to different clusters. It strives for minimal average distance among objects within a cluster and maximal average distance among objects across distinct clusters. The coherence, stability, and distinctness of each cluster partition are typically considered during cluster evaluation.

2.6.2. Davies-Bouldin Index

The Davies-Bouldin index (DBI) is a commonly employed metric for evaluating clustering algorithms. It primarily gauges the effectiveness of a clustering algorithm in dividing data into distinct clusters for a given number of clusters [38]. In essence, the DBI quantifies the average similarity of each cluster to its most similar cluster, thus determining its score.

When the average similarity is reduced, it signifies improved differentiation among clusters and consequently better outcomes from the clustering process. The calculation of the Davies-Bouldin Index is as follows:

$$\bar{R} = \frac{1}{N}\sum_{i=1}^{N} R_i \quad (8)$$

In essence, the Davies-Bouldin Index involves summing up the similarity scores between each cluster and its closest counterpart in terms of similarity. A smaller value of R' signifies that the optimal cluster selection minimizes the average similarity, indicating clusters with clearer definition.

2.7. Related works

Research on customer segmentation is gaining traction within the financial industry. This includes both global and local studies, as discussed in Sections 2.7.1 and 2.7.2, respectively.

2.7.1. Global Studies on Customer Segmentation

Asha Panyako. M. [39] conducted a study titled "Customer Segmentation on Mobile Money Users in Kenya". The aim of the research was to examine mobile money data and identify homogeneous customer groups by combining behavioral, demographic, and psychographic data. The researcher employed Hierarchical clustering, K-means, and affinity propagation algorithms to segment customers, comparing them using internal validation measures. The dataset included various demographic and behavioral features obtained from a telecommunications company's data warehouse. The correlation among these features was tested, leading to the selection of age, network revenue, mobile money transaction amounts, frequency of loan uptake, customer and organization transfers, goods and service payments, and deposits and withdrawals as modelling features. The dataset was then applied to their algorithms. Agglomerative clustering resulted in seven clusters with a normalized mutual score of 0.5526, an adjusted rand score of 0.5436, and a Silhouette coefficient of 0.4523. K-means generated 11 clusters with an NMI score of 0.5168, an adjusted rand score of 0.3315, and a Silhouette coefficient of 0.2369.

In addition, I.Smeureanu, et al. [40] authored a study titled "Customer segmentation in the private banking sector using machine learning techniques." The objective of this research is to discern customers within private banks based on the value they bring to the bank through specific behaviours, enabling targeted offerings such as deposits and credit cards. The researchers sought to differentiate between "affluent" customers and those with "mass"

characteristics based on certain behaviors. Two prominent machine learning techniques, namely Neural Networks and Support Vector Machines (SVM), were employed for the analysis. The dataset comprised 2,783 observations representing active cardholders at the Commercial Bank of Romania. For the neural network model, the data was divided into training (60%), validation (20%), and out-of-sample testing (20%) sets. Conversely, for Support Vector Machine (SVM), the initial data split involved 80% for training and 20% for testing, without requiring a validation sample. The researchers noted that both machine learning techniques performed effectively in the segmentation process. The SVM model, particularly when using the RBF kernel function, exhibited superior performance in detecting "affluent" customers compared to the MLP using a gradient descent algorithm. Despite minor variations in the overall detection rate on the test sample between the two approaches, there is a recognized necessity for employing more advanced algorithms.

The research conducted by Tushar et al. [28] is titled "Customer Segmentation using K-means Clustering." Clustering has demonstrated its effectiveness in customer segmentation and falls within the realm of unsupervised learning, capable of identifying clusters within unlabelled datasets [28]. In this study, three distinct clustering algorithms (K-means, Agglomerative, and Mean Shift) are applied to segment customers using a dataset containing two features and 200 records. Subsequently, the results of the clusters obtained from these algorithms are compared. The researchers utilized two internal clustering metrics, namely the Silhouette score and the Calinski-Harabasz index, to evaluate and compare the outcomes of the clusters derived from the algorithms.

Furthermore, Eric Umuhoza, et al. [41] conducted a study titled "Using Unsupervised Machine Learning Techniques for Behavioral-based Credit Card User Segmentation in Africa." The objective of the research was to revolutionize the existing credit card business model and campaign strategy at CIB by shifting from traditional value-based campaigns to targeted campaigns aligned with customers' needs, lifestyles, and usage preferences. The datasets utilized for this investigation comprised credit card transaction data from the first quarter to the last quarter of 2016 to 2017, encompassing 143,975 observations per quarter and a total of 60 variables collected from the Commercial International Bank of Egypt.

Due to the predominantly unlabelled nature of credit card usage data, the researchers employed an unsupervised machine learning approach. To evaluate the model's performance, an optimal number of clusters was determined using various methods including elbow, Silhouette, and Calinski-Harbaz. The researchers ultimately recommend that CIB integrate the findings of the project into their marketing strategies and update this model quarterly to capture new customer behaviours.

Another research endeavour titled "Customer segmentation model based on two-step optimization in the big data era" was undertaken by W. Gao, et al. [42]. In this study, the authors devised a two-step optimization model utilizing genetic algorithm (GA) and cluster ensemble (CE) for customer segmentation. The data utilized in this investigation originates from the primary data of a Chinese city business bank's credit card system, encompassing fundamental customer information and detailed transaction records. Basic customer information includes age, gender, educational background, marital status, etc., while transaction records provide details such as transaction date, amount, and type. The experiment involved 3,544 customers and 157,756 trading records in total. Since preserving information integrity is crucial in customer segmentation, the FCEN model significantly reduces input data at the onset while eliminating the impact of irrelevant variables on segmentation outcomes. Additionally, FCEN, as a favourable customer segmentation model, enhances validity and robustness and can effectively subdivide simplified sample data. The findings indicate that compared to other single clustering algorithms, the FSGA-FCEN approach in this study yielded the most favorable outcomes, making it an optimal choice in terms of both efficiency and practicality.

In their study, R. Baradaran, et al. [43] conducted research titled "Profiling bank customer behavior using cluster analysis for profitability," focusing on Saman bank cardholders. The CRISP-DM Methodology was adopted for the research process. The dataset comprised customer information such as age, gender, marital status, and transactional data from approximately 900,000 customers, involving nearly 26 million transactions. The research involved two primary operations: segmentation and classification for customer profiling. Segmentation was performed on three channels—ATM, Terminal, and web—utilizing the

K-means clustering algorithm. The resulting clusters were labeled as long-term potential customers, disloyal customers, runaway customers, hostage customers, consumer customers, official customers, paying bill customers, future customers, and good customers. Based on the analysis of the ruleset, the researchers recommended that the bank formulate its future strategy based on three suggestions derived from the study concerning the three channels of segmentation. Firstly, recognizing the high installation costs of bank terminals by analyzing merchant behavior and transactions. Secondly, considering the importance of using RFM models for customer ranking. Lastly, employing time series analysis to predict transaction times.

V. Mihovaa, et al. [44] conducted a study aimed at identifying three clusters (segments) of loyal borrowers: "platinum," "gold," and "silver." The study involved analyzing initial segmentation variables and standardized variables from a database of 100 borrowers obtained from a commercial bank branch that offered secured consumer loans. The objective was to categorize the 100 subjects into three groups based on loan amount, duration with the bank, and the worst status, utilizing these three variables. K-means clustering was employed for the analysis, and a comparative examination of the results from both methods was conducted. The findings suggest that if the bank seeks to incentivize long-term clients with a positive credit history, it would be suitable to employ two-step clustering or K-means cluster analysis with standardized segmentation variables. Conversely, if the aim is to target customers with substantial loan amounts (who are either long-term clients or possess a good credit history), it would be preferable to utilize K-means cluster analysis with initial segmentation variables.

2.7.2. Local Studies on Customer Segmentation

Local research studies were conducted to explore the utilization of data mining techniques for customer segmentation. Henock W. [45] delved into this topic in his work titled "Application of data mining techniques to support customer relationship management at Ethiopian Airlines." The study focused on applying data mining techniques to the database of frequent flyer customers to identify strategic customer segments that could enhance CRM activities at Ethiopian Airlines. The dataset included demographic information and

the current status of each program member, totaling 90,833 records across 10 fields. The researcher employed the Cross Industry Standard Process for Data Mining (CRISP-DM) process model for the study. K-means clustering algorithm was utilized to segment individual customer records into clusters exhibiting similar behaviors. Decision tree classification was then employed to establish rules that could be applied to assign new customer records to these segments. The cluster findings affirmed existing business knowledge that frequent travelers typically generate higher revenue. Additionally, the clusters revealed discrepancies in revenue contribution among customers with similar travel frequencies, with one segment generating more revenue than the other. Overall, the results of this study were promising, demonstrating the feasibility of segmenting customer data using data mining techniques that align with business objectives. The researcher posits that a more thorough investigation utilizing data mining techniques could enhance business advantages derived from customer insights and support CRM activities. Finally, the researcher recommends further segmentation leveraging comprehensive customer demographic data, employing predictive models such as association rule algorithms for target marketing and opportunity identification.

The study titled "Mining customs data for customer segmentation: the case of Ethiopian Revenues and Customs Authority" was conducted by [46]. This research investigates the potential of utilizing clustering data mining techniques to support CRM activities for the Ethiopian Customs and Revenue Authority (ERCA), employing the CRISP-DM process model approach. Following necessary data preparation steps, a dataset comprising 65,535 records/instances was utilized to construct a clustering model. For customer segmentation, the K-means clustering algorithm was applied, experimenting with various values for k (4, 5, and 6) and seed sizes (10, 100, and 1000). The model with k=4 and a seed size of 100 exhibited superior clustering performance. This model facilitated the segmentation of the authority's customers into distinct clusters representing high, medium, and low-value customer groups with minimal algorithm iterations and a minimal sum of square error (SSE) between records within a cluster, totaling 6 and 2142.82, respectively.

N. Kannaiya Raja, et al. [47] present a study titled "Data Mining Model to Analyze CRM in Banking Sector," proposing an effective data mining model for customer relationship management (CRM) to forecast customer relationships within banking applications. The focus lies on profiling loyal and lost banking customers based on RFM (Recency, Frequency, and Monetary) scores. The study encompasses customer transactions and demographic data from Ethiopian commercial banks, involving over 70,000 customers. In their developed model, customer groups play a pivotal role in CRM, prompting an exploration and assessment of various customer segments based on shared characteristics. Customer classification and grouping are accomplished by integrating SVM-Ensemble-Classifier-Random Forest and K-means methodologies with applications in the RFM model (Recency, Frequency, Monetary). Additionally, customer predictions are conducted using a logistic regression model. The authors emphasize the critical importance for organizations to possess a comprehensive understanding of their customers' attributes, behaviors, demographics, etc. They suggest that a more comprehensive examination utilizing data mining techniques can enhance business advantages derived from customer insights and support CRM activities in Ethiopia.

Gutema D. M [48] conducted a study titled "Application of data mining techniques for customer segmentation in the insurance business: the case of Ethiopian Insurance Corporation." The aim of the research was to utilize data mining techniques within the insurance sector to construct models capable of segmenting customers according to their value. To achieve this objective, the researcher employed the CRISP-DM methodology. In order to develop the customer segmentation models, both the K-means clustering algorithm and the J48 decision tree algorithms were tested using WEKA to unveil meaningful patterns and analyze the data. The findings of the study highlighted the necessity of employing customer segmentation models that combine classification and clustering data mining techniques for the Life Addis District (LAD) and marketing department of Ethiopian Insurance Corporation (EIC). These models aid in identifying valuable customer segments and other underlying factors contributing to variations in customer value.

Yeshitla T. [49], in his study titled "Assessing opportunities and challenges of CBEBirr mobile money service: a case study on Commercial Bank of Ethiopia," sought to evaluate the obstacles, potential advantages, and factors affecting customers' inclination to utilize the CBEBirr Mobile money service. Employing quantitative methods, the researcher analysed survey data to scrutinize the challenges, opportunities, and determinants influencing customers' behavioural intention to adopt the CBEBirr Mobile money service for enhancing financial inclusion in Ethiopia. The study targeted customers registered for the CBEBirr Mobile money service, specifically within the eastern district of the Commercial Bank of Ethiopia. From this district, the researcher selected a sample of 8,375 customers, constituting the focus population for the research.

Moreover, a number of recent local investigations [13], [14] , [15] have endeavoured to evaluate the difficulties encountered with the CBEBirr service, the practices surrounding CBEBirr services, and the operational obstacles linked to CBEBirr services. These studies have explored various aspects including challenges, prospects, and operational barriers concerning E-Banking (CBEBirr).

2.8. Gaps in literature review

All the research works cited in section 2.7.1 and section 2.7.2 delve into the exploration of data mining and machine learning techniques for customer segmentation, focusing on foreign banks and other institutions. However, there has been a notable absence of research conducted on customer segmentation specifically pertaining to CBEBirr users of the Commercial Bank of Ethiopia using machine learning techniques. Therefore, this study endeavours to address this gap created by the dearth of local investigations specifically concerning CBEBirr customers of the Commercial Bank of Ethiopia. Unlike banks in foreign countries, Ethiopian banks exhibit differences in terms of the banking market, services offered, customer demographics, and banking regulations. Consequently, it is inappropriate to directly apply the findings from the aforementioned foreign studies to the context of Ethiopian banks.

Local research has been undertaken concerning customer segmentation, customer relationship management, and churn prediction. However, there exists a notable dearth of substantial studies focusing on CBEBirr customer segmentation within the Commercial Bank of Ethiopia. Kassahun G. [50] conducted an innovative study on customer churn prediction at the CBE, employing a predictive model to identify the characteristics exhibited by customers before switching to another competitive bank. The study advocates for further research to ascertain customer demographics or other characteristics. Furthermore, investigations have been conducted on the application of customer segmentation in various sectors such as the Ethiopian Customs Authority, Ethiopian Airlines [45] , and insurance companies. This highlights a deficiency in studies addressing customer segmentation within the banking industry, particularly among mobile money customers in Ethiopia. Additionally, local research by [13], [14], [15] aimed to assess challenges, practices, and operational barriers associated with CBEBirr services. These studies utilized both qualitative and quantitative approaches, basing their analyses on primary data from CBEBirr customers collected through questionnaires from diverse sources.

The banking sector, being a service industry, holds the vital responsibility of upholding robust and efficient customer relationship management practices. Achieving this necessitates banks to allocate resources towards gaining a deeper understanding of both current and potential customers. By concentrating on segmentation, banks can enhance relationships with existing clientele, draw in new customers, and bolster the bank's share of the market. Thus, this study aims to explore unsupervised machine learning methods for customer segmentation specifically for CBEBirr users of the Commercial Bank of Ethiopia, utilizing secondary data derived from CBEBirr customers.

Table 2. 1:*Summary table of related works*

Reference	Objective	Approach Used	Algorithm Used	Evaluation metrics	Number of Datasets	Results achieved
[46]	Explore the applicability of clustering data mining techniques to support Customer Relationship Management (CRM)	Cross-Industry Standard Process for Data Mining (CRISP-DM)	Expected Maximization, Filtered cluster, Hierarchical Cluster, DBSCAN, and K-means algorithms	WEKA tool to measure the model building Performanc e.	65,535	The K-means algorithm was identified as the most efficient and the researcher got the value of K=4 and clustered the given dataset.
[28]	Apply clustering algorithms to segment customers based on their behaviours and patterns.	Machine learning	k-Means, Agglomerative , and mean shifts	silhouette score and Calinski-Harabasz index.	200	• Based on the Silhouette score value both k-Means and Agglomerative clustering were selected as the best model. • By applying both clustering, 5 segments of a cluster have been formed and labelled as Careless, Careful, Standard, Target, and Sensible customers.

Reference	Objective	Approach Used	Algorithm Used	Evaluation metrics	Number of Datasets	Results achieved
[45]	Testing the application of data mining techniques to support CRM activities at Ethiopian Airlines.	Cross-Industry Standard Process for Data Mining (CRISP-DM)	K-means clustering and Decision tree	The researcher did not use model evaluation metrics to evaluate the algorithms.	90,833	**Using K-means(K=5)** • Three of the clusters contained 21% of customer records that generated the highest revenue and differed in the total frequency of trips and tenure of customers. • The cluster analysis included medium and low revenue-generating customers, with 27% and 52% customer records respectively. **Using Decision Tree** • The decision tree model that was generated from the cluster results correctly assigned 92.18% of new records to the five clusters, with 'Total Revenue' as the splitting variable. This result was found to be lower than the results obtained with 'Total Trips' and 'Tenure in

Reference	Objective	Approach Used	Algorithm Used	Evaluation metrics	Number of Datasets	Results achieved
						Months' as splitting variables, but the difference was quite negligible, and the decision tree model with 'Total Revenue' making the initial split was chosen as a working model.
[48]	To apply data mining techniques in the insurance business to build models that can segment customers based on their value	Cross-Industry Standard Process for Data Mining (CRISP-DM)	K-means clustering and J48 decision tree	Sum of Squared Error (SSE) and average accuracy rate revealed in the correctly classified instances and the confusion matrix.	27,845	**In the case of K-means:** • 55% of customers clustered under high value. In this group: o Most policyholders are female, whose age is greater than 28, and who paid a sum insured value above the average. o Cover types such as death, Complementary accident insurance – death, total, partial, and permanent disability medical (CAI), and pre-needed funeral expense insurance (PNFE) were mostly purchased by female customers.

Reference	Objective	Approach Used	Algorithm Used	Evaluation metrics	Number of Datasets	Results achieved
						○ The youth and adult policyholders who paid a sum insured value less than average contribute low value for the firm. **In the case J48 algorithm of the decision tree:** • Higher accuracy rates, with the first experiment showing a 99.9815% accuracy rate in predicting customer value.
[39]	Analyse mobile money data	Unsupervised Machine learning	Hierarchical clustering, K-means & affinity propagation	Adjusted Rand-score, silhouette coefficient & normalized mutual score	12,350	Out of three algorithms, K-means is the best one and the researcher achieved 11 clusters for the given data set.
[40]	Perform customer segmentation in the private banking sector using machine learning techniques.	Machine learning	Neural Networks and Support Vector Machines	Radial basis function (RBF) for SVM and	2,783	**In the case Neural Network model:** The data was split as follows: 60% for the training sample; 20% for the validation

Reference	Objective	Approach Used	Algorithm Used	Evaluation metrics	Number of Datasets	Results achieved
				Multilayer perceptions (MLP) for ANN		set; 20% for an out of sample testing process. • After 191 cycles with 8 hidden neurons, with a total detection rate on the test sample of 93.885%. • However, the detection rate for "affluent" customers is rather low, with only 41% of the "affluent" customers being detected by the model **In the case SVM Model:** • The data sample was randomly split as 80% for training and 20% for the testing process. In the case of SVM, there is no need for a validation sample. • The researcher observed that the accuracy of the "affluent" customers is somewhat

Reference	Objective	Approach Used	Algorithm Used	Evaluation metrics	Number of Datasets	Results achieved
						smaller (83%) than the overall detection rate.
[41]	Build a behavioural-based segmentation model for African credit cardholders based on their purchase data.	Unsupervised machine learning	K-Means clustering algorithm	Silhouette, and Calinski-Harbaz	143,945	• Customers are grouped into four distinct segments. • Tailored marketing strategies are suggested based on customer purchasing behaviours.
[42]	Implement customer segmentation	Unsupervised Machine learning	• Two-step optimization model (FSGA-FCEN) • Genetic algorithm (GA) and cluster ensemble (CE)	Clustering Variance and overall clustering quality(ocq)	157,756	• The two-step model is efficient and practical for customer segmentation. • Outperforms K-means, FCM, and MAJ models.
[43]	Profile bank customers' behavior using cluster analysis for profitability.	Machine Learning	K-means, Classification and Regression Trees (CRT),	Not used	71,000	• **ATM channel:** Most single men customers have little

Reference	Objective	Approach Used	Algorithm Used	Evaluation metrics	Number of Datasets	Results achieved
			Chi-squared Automatic Interaction Detection (CHAID), and C5.0			loyalty to the use of Saman Bank's ATM. • **Web Channel:** The Bank can invest in paying bill services for customers who have been born in Tehran province and offer them incentive programs. • **Terminal channel:** The Bank should invest in female customers for using terminals and increase the number of these customers because they have the most profit for the bank.
[44]	Improve customer segmentation in commercial banks.	Unsupervised Machine learning	K-means clustering	Not used	100	• Three clusters of loyal borrowers, namely "platinum," "gold," and "silver," are identified based on credit history and three segmentation variables: loan amount, time with the bank, and worst status in the last 12 months.

Reference	Objective	Approach Used	Algorithm Used	Evaluation metrics	Number of Datasets	Results achieved
						• The loan amount is found to have the biggest influence on cluster formation, while the time with the bank has the least influence.

2.9. Summary

This chapter elaborated on clustering, which involves grouping similar data into distinct clusters, as well as partitioning collected data into subsets where the data within each subset shares similarity according to predefined criteria. Clustering algorithms serve as valuable tools for data exploration, with K-means, DBSCAN, HAC, and mean shift being particularly noteworthy and widely utilized. The chapter provides an overview of the most common clustering methods and underscores the significance of clustering across various sectors, notably within the banking industry, for tasks such as classification and customer segmentation. Moreover, the chapter underscores the pivotal role of the banking sector in Ethiopia, with the Commercial Bank of Ethiopia (CBE) standing out as the largest state-owned bank. It highlights the substantial profits and growth witnessed by CBE over the past decade. Various types of customer segmentation, including demographic, geographic, and psychological segmentation, are discussed. The chapter also reviews numerous related researches works conducted locally and globally on customer segmentation, revealing diverse approaches to clustering customers based on different feature or attribute values, leading to distinct clusters and segments. Furthermore, the chapter suggests that the study could significantly enhance by incorporating multiple criteria to evaluate user behaviour across various CBEBirr services. Additionally, it reviews and discusses common model evaluation metrics, emphasizing the importance of assessing clustering results to gauge performance. Finally, the chapter identifies gaps in the literature review pertinent to the study at hand.

CHAPTER THREE

DESIGN OF UNSUPERVISED CLUSTERING MODEL FOR CBEBIRR CUSTOMERS

3.1. Introduction

In this chapter, we explore the suggested system architecture for CBEBirr customer segmentation using machine learning algorithms. Additionally, we delve into the processes of data gathering, data pre-processing, dimensionality reduction, and feature selection. Moreover, this chapter examines clustering methodologies and the algorithms slated for utilization in this research. Furthermore, we discuss the model evaluation metrics intended for measuring the effectiveness of the algorithms.

3.2. Design Considerations

During the model design process, technical considerations involve minimizing the dimensions of large-scale CBEBirr data. Dimensionality reduction refers to a method of representation within the data that retains relevant information for clustering while discarding extraneous variance. Dimension reduction for extensive datasets has become increasingly significant due to the challenges posed by high dimensionality, which can hinder the efficiency of many algorithms. Feature extraction serves to mitigate computational complexity, facilitating segmentation.

Understanding the CBEBirr data entails grasping the actual usage behaviour of customers. The attributes utilized as input for the clustering algorithm include Age, Gender, recruitment source, Number of Buy Air Time Transactions, Number of Send Money Transactions, Number of Pay Bill Transactions, Number of Buy Goods Transactions, Number of Cash-in Transactions, and Number of Cash-Out Transactions.

3.3. Proposed Architecture for CBEBirr Customer Segmentation

The proposed architecture for the CBEBirr customer segmentation is shown in figure 3.1. Figure 3.1 demonstrates that the proposed architecture consists of four major components, namely data collection, data pre-processing (containing data cleaning, transformation, and dimensionality reduction); feature selection, and finally clustering algorithms (K-means, DBSCAN, HAC, and mean shift) is applied to create a group of related customers.

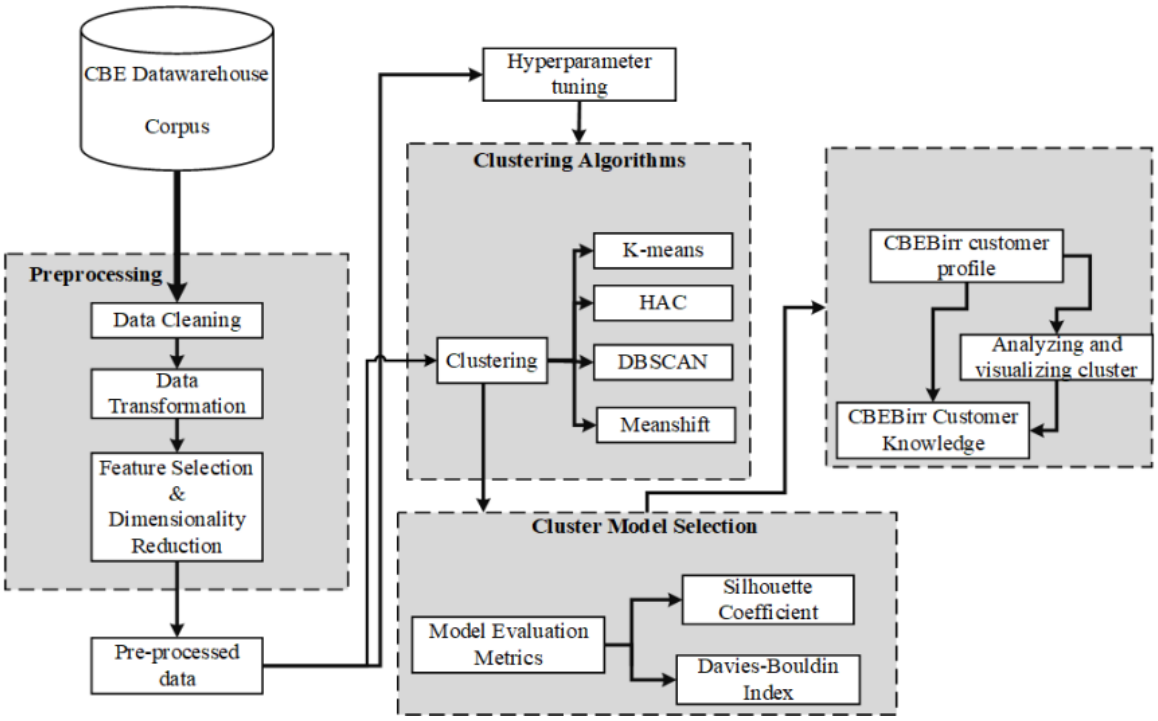

***Figure 3. 1**: Proposed architecture for CBEBirr customer segmentation*

Furthermore, figure 3.1 depicts the overall architecture of the CBEBirr customer segmentation process. It outlines the sequential steps undertaken in this study, beginning with the collection of CBEBirr customer data, followed by pre-processing to eliminate redundant elements and transforming the data into the CBEBirr customer dataset. Various clustering algorithms were then applied to the dataset to obtain meaningful clusters or groups of CBEBirr customers. The subsequent section provides a detailed description of each component and module of the architecture.

3.4. Data Collection

For our research, we obtained data from the data warehouse, comprising 250,000 unique CBEBirr customer records spanning from February 2023 to July 2023 from the Commercial Bank of Ethiopia. The data collection process involved collaborating with the management information system administrator, who utilized Oracle structured query language (SQL) queries to extract the data. Subsequently, we organized the data in Excel format to facilitate further processing.

3.5. Data Pre-Processing

Data pre-processing is a crucial stage in preparing the data for the construction of a CBEBirr customer segmentation model. Extracting pertinent insights from these extensive CBEBirr customer datasets and segmenting them for analysis requires pre-processing of the data. This process encompasses tasks such as data cleaning, which involves rectifying incomplete, inconsistent, and noisy data, as well as data transformation using various techniques. The data pre-processing component addresses various issues related to attribute values inherent in the stored data to render it suitable for analysis. The effectiveness of clustering customers is heavily influenced by the pre-processing steps undertaken.

3.5.1. Data Cleaning

The CBEBirr dataset may contain instances of missing values. Before proceeding with data clustering tasks, it's essential to address these missing values. Various techniques can be employed to handle missing values, including disregarding the tuple, manually filling in missing values, using a global constant to fill in missing values, substituting with attribute mean, or employing the most probable value. The CBEBirr dataset exhibits several irrelevant and missing sections, necessitating data cleaning procedures to address missing data, noisy data, and so forth. Some tuples within the CBEBirr dataset contain multiple missing values, underscoring the importance of addressing these gaps before proceeding with data analysis to ensure that clustering outcomes align with business objectives.

Among the aforementioned techniques, we opted to employ the "ignoring the tuple" approach to eliminate missing values from our collected dataset. By implementing these techniques, missing values for features such as CustomerID, Kebele, Wereda, Nationality, City, and Region were removed, as they were deemed irrelevant to our study. As depicted in figure 3.2, our CBEBirr customer dataset now contains no missing values.

```
Merchant                              0
Branch                                0
Agent                                 0
Gender                                0
Number of Buy Air Time Transaction    0
Number of Send Money Transaction      0
Number of Pay Bill Transaction        0
Number of Buy Goods Transaction       0
Number of Cash in Transaction         0
Number of Cash Out Transaction        0
Amount of Transaction performed       0
Age__15-25                            0
Age__26-35                            0
Age__36-45                            0
Age__45-55                            0
Age__Above 55                         0
dtype: int64
```

Figure 3. 2:*Missing value count*

Noisy data refer to values that contain errors or outliers that deviate from the expected pattern. In our dataset, certain records exhibit discrepancies between age and the number of transactions performed. However, these discrepancies may not necessarily be attributed to noise but rather to outliers, which are values that significantly deviate from the typical pattern. To address this issue, we identified and eliminated outliers from the "amount of transactions performed" attribute in our dataset. This process involves two steps: firstly, we detect outliers using the unsupervised attribute Interquartile Range filter, and subsequently, we remove them using the unsupervised instance RemoveWithValues filter.

Figure 3. 3: *Outlier Detection*

As depicted in figure 3.3, outliers were detected in the CBEBirr customer dataset. However, instead of deleting records with outliers, they were addressed through log transformation.

3.5.2. Data Transformation

Data Transformation involves modifying raw data into a format or structure that is more suitable for model development and data analysis. This stage is crucial in feature engineering, enabling insights to be gained [51].

In this study, to prepare the dataset for model building, all categorical data such as gender, age (based on age range value) and customer recruitment source (recruited_by attribute) were encoded and converted to float type to facilitate easier understanding by machine algorithms. Encoding of these attributes was achieved using the one-hot encoding method.

Following data transformation and normalization, we selected 11 primary attributes that are highly pertinent for the segmentation model.

Normalization is employed to standardize dataset features with varying value ranges to a unified scale, typically ranging from 0 to 1 or, in some instances, from -1 to 1. This scaling process is crucial because disparate ranges can significantly impact the learning process. Various normalization methods are available, including the standard scaler and the min-max scaler. With the standard scaler, scaling is conducted independently on each feature by computing relevant statistics from the dataset. Conversely, the min-max scaling approach scales each feature individually using the minimum and maximum values within the dataset. The choice between these methods depends on the intended use for training the models. Normalization and standardization were implemented on the CBEBirr customer data. figure 3.4 illustrates box plots of features within the CBEBirr customer data following this process.

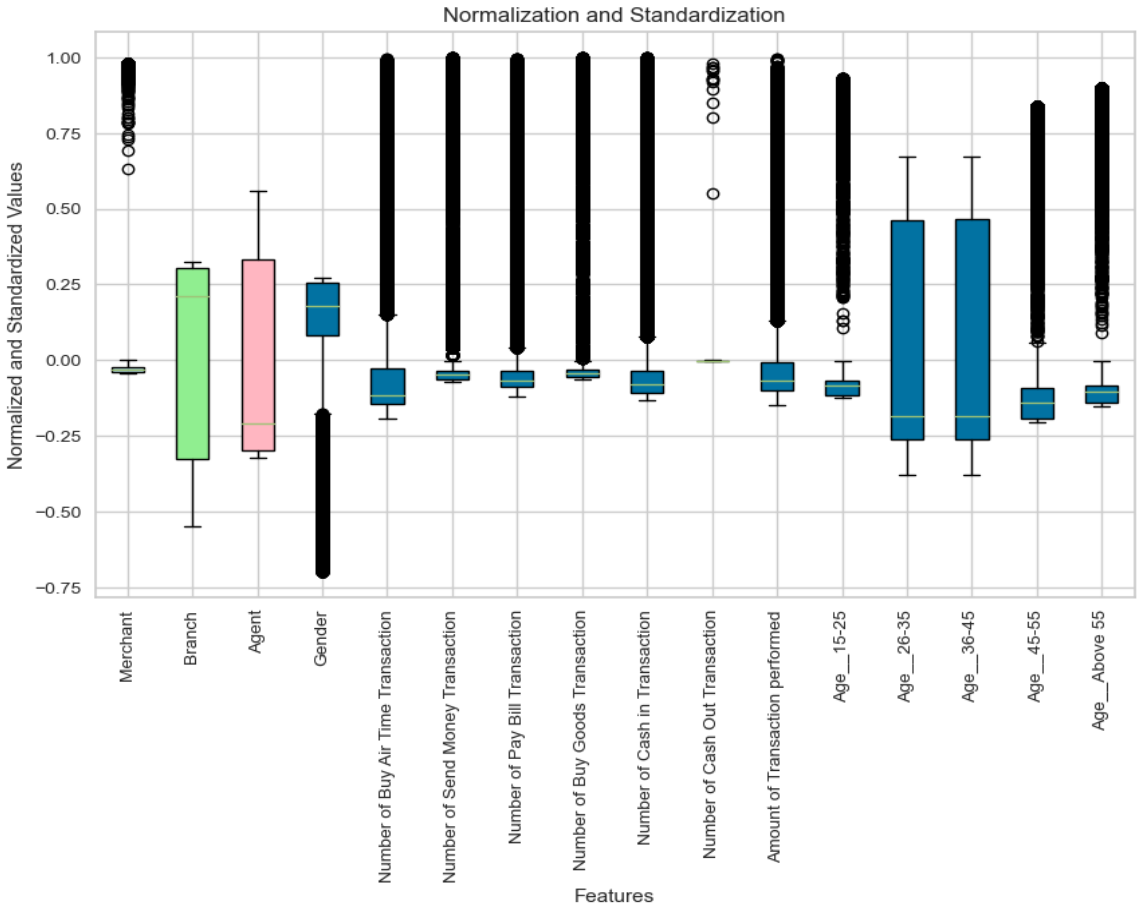

Figure 3. 4*: Normalization and Standardization*

3.6. Feature Selection and Dimensionality Reduction

In the realm of machine learning, features encompass a set of input variables that represent specific characteristics or properties of the entity under observation. The objective of this stage is to identify features or attributes with significant potential for the intended purpose. However, the clustering effectiveness of a feature can only be assessed within the context of a particular clustering problem and dataset. As a result of this stage, a list of features is generated, which may or may not be utilized depending on their efficacy in addressing the specific clustering problem. Arguably, this represents one of the most crucial steps in the data pre-processing phase, as it directly influences model performance. Selecting appropriate data for features can be challenging, as it is often difficult to anticipate which data will yield features with strong clustering capabilities [52]. In this study, features or attributes were primarily selected and defined based on the available data obtained during data collection and sampling. Taking into account their importance in model development, a list of potentially valuable features or attributes was selected and detailed in table 3.1.

Table 3. 1:*Selected feature for our study*

Feature category	Features		Data type
Demographical data	Age		Float
	Gender		Float
	Recruited by	Recruited by Branch	Float
		Recruited by Merchant	
		Recruited by Agent	
Behavioural data	Number of Buy Air Time Transaction		Float
	Number of Send Money Transactions		Float
	Number of Pay Bill Transaction		Float
	Number of Buy Goods Transaction		Float
	Number of Cash-in Transactions		Float
	Number of Cash-Out Transactions		Float

Utilizing a method called dimensionality reduction proves beneficial in the domain of machine learning. This technique aids in reducing the number of features, each representing a dimension that only partially describes the objects under consideration. With the addition of more features, the data becomes increasingly sparse, leading to challenges

associated with the curse of dimensionality. Furthermore, working with smaller datasets becomes more manageable. As additional characteristics are introduced, the data tends to become sparser, impeding analysis due to increased dimensionality. Hence, dimensionality reduction is crucial for obtaining smaller datasets, facilitating easier processing [53].

Dimensionality reduction can be carried out in two ways:

- Choosing from the available features (feature selection)
- Combining existing features to extract new features (feature extraction)

Principal Component Analysis (PCA) serves as the main technique for feature extraction [53]. PCA aims to identify the optimal linear transformation to minimize the number of dimensions while retaining the maximum amount of information. Occasionally, the discarded information is regarded as noise—information that stems from other, typically unobserved processes and does not accurately represent the phenomena being modelled.

In this study, our dataset comprises 11 features or attributes. We employ Principal Component Analysis (PCA) to condense these features or attributes into two-dimensional data, as depicted in the figure below.

	P1	P2
0	0.244687	0.086395
1	0.244687	0.086395
2	0.244687	0.086395
3	0.244687	0.086395
4	0.244687	0.086395

Figure 3. 5:*Dimensionality Reduction Using Principal Component Analysis*

3.7. Clustering Model

3.7.1. K-means Clustering

K-means clustering is an algorithm designed to partition data into k groups, where k represents a predefined number. The algorithm operates by initially assigning each data

point to the nearest cluster center and subsequently updating these centers based on the average of the points assigned to them. This process iterates until either the cluster assignments remain unchanged or a maximum number of iterations is reached.

For the implementation of K-means clustering in Python, we utilized the sklearn.cluster module available within the scikit-learn library. This module offers the K-means class, which encompasses methods for fitting the data, predicting cluster labels, and assessing clustering performance. Additionally, we employed visualization techniques such as scatter plots or elbow plots to visualize the clustering outcomes.

When utilizing the K-means class, we specified several parameters, including the number of clusters (k), initialization method, maximum number of iterations, tolerance, and random state. The optimal value of k for our data was determined based on domain knowledge, heuristics, or validation measures.

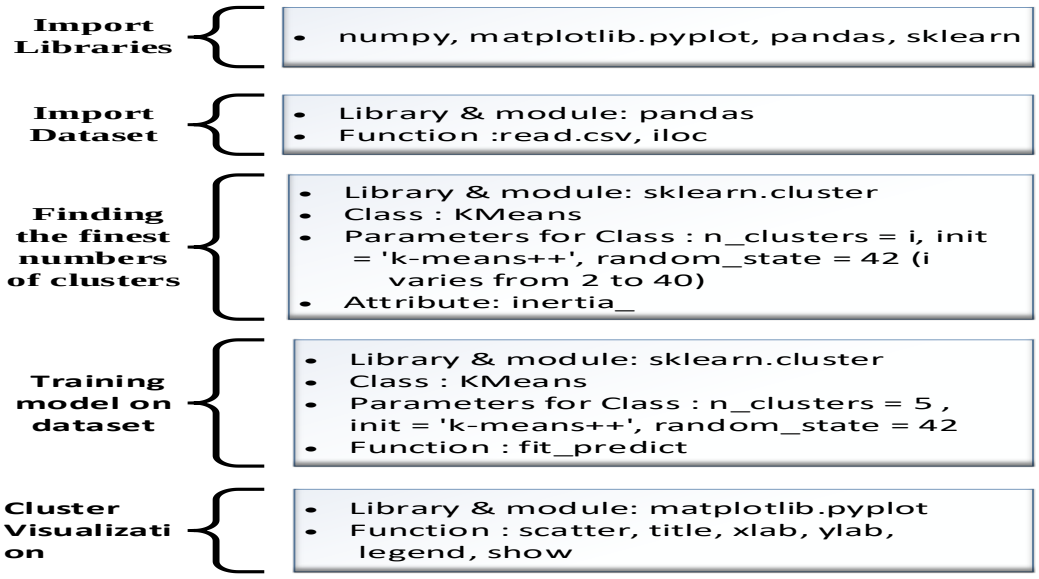

Figure 3. 6: *Python libraries for implementing K-means clustering*

3.7.2. Agglomerative Clustering Algorithm

Agglomerative clustering, a form of hierarchical clustering, operates by merging clusters based on their proximity. It commences with each data point constituting a separate cluster and progressively combines the closest clusters until only one remains. The outcome of

agglomerative clustering can be depicted through a dendrogram, which is a tree-like diagram illustrating the sequence of merges or splits.

For the implementation of agglomerative clustering in Python, we employed the sklearn.cluster module from the scikit-learn library. Within this module, the AgglomerativeClustering class facilitates fitting the data, predicting cluster labels, and estimating the number of clusters. Additionally, we utilized visualization techniques such as scatter plots to visualize the clustering outcomes.

When utilizing the AgglomerativeClustering class, several parameters need to be specified, including the linkage criterion, distance metric, and the number of clusters for the algorithm. The linkage criterion dictates how the distance between clusters is evaluated and can be selected from options like ward, complete, average, or single. In this context, we opted for the average method as our linkage criterion.

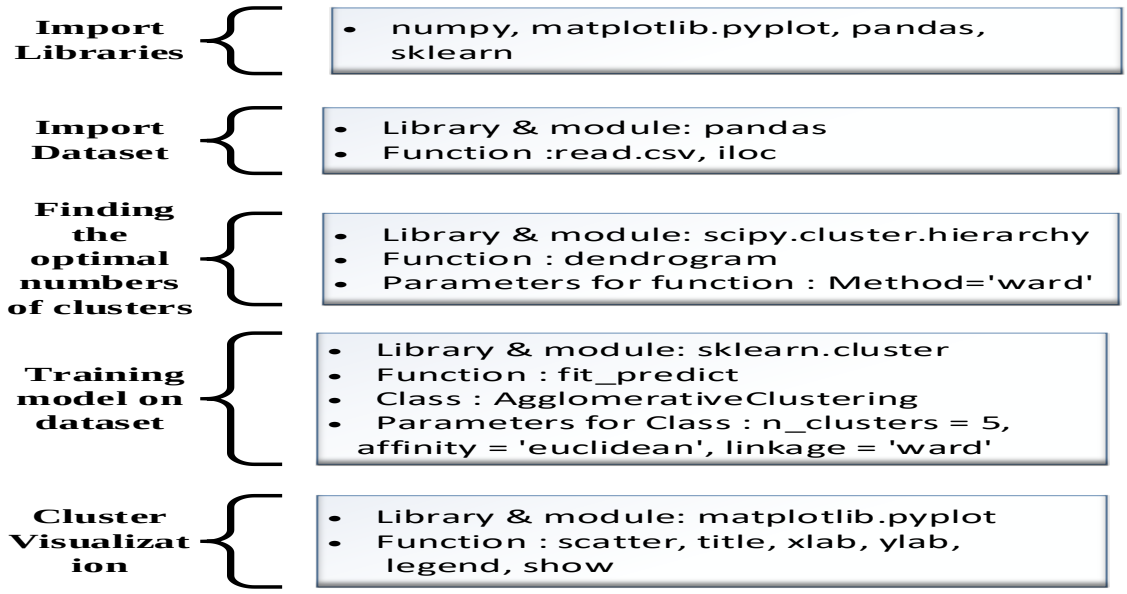

Figure 3. 7*: Python libraries for implementing Agglomerative clustering*

3.7.3. DBSCAN

DBSCAN, a clustering algorithm, groups data points based on their density, specifically the number of points within a designated radius. Notably, DBSCAN can identify clusters of various shapes and sizes and is proficient at detecting outliers or noise points.

For implementing DBSCAN clustering in Python, we utilized the sklearn.cluster module from the scikit-learn library. Within this module, the DBSCAN class facilitates fitting the data, predicting cluster labels, and estimating the bandwidth parameter. Additionally, we employed visualization methods like scatter plots to portray the clustering outcomes.

When utilizing the DBSCAN class, two essential parameters must be specified: eps and min_samples. Eps denotes the radius of the neighborhood surrounding each point, while min_samples represent the minimum number of points required to form a dense region. In this context, a core point is identified if it possesses at least min_samples point within its eps-neighborhood. A border point, on the other hand, lies within the eps-neighborhood of a core point but contains fewer than min_samples point within its eps-neighborhood. Any point not classified as a core or border point is deemed a noise point.

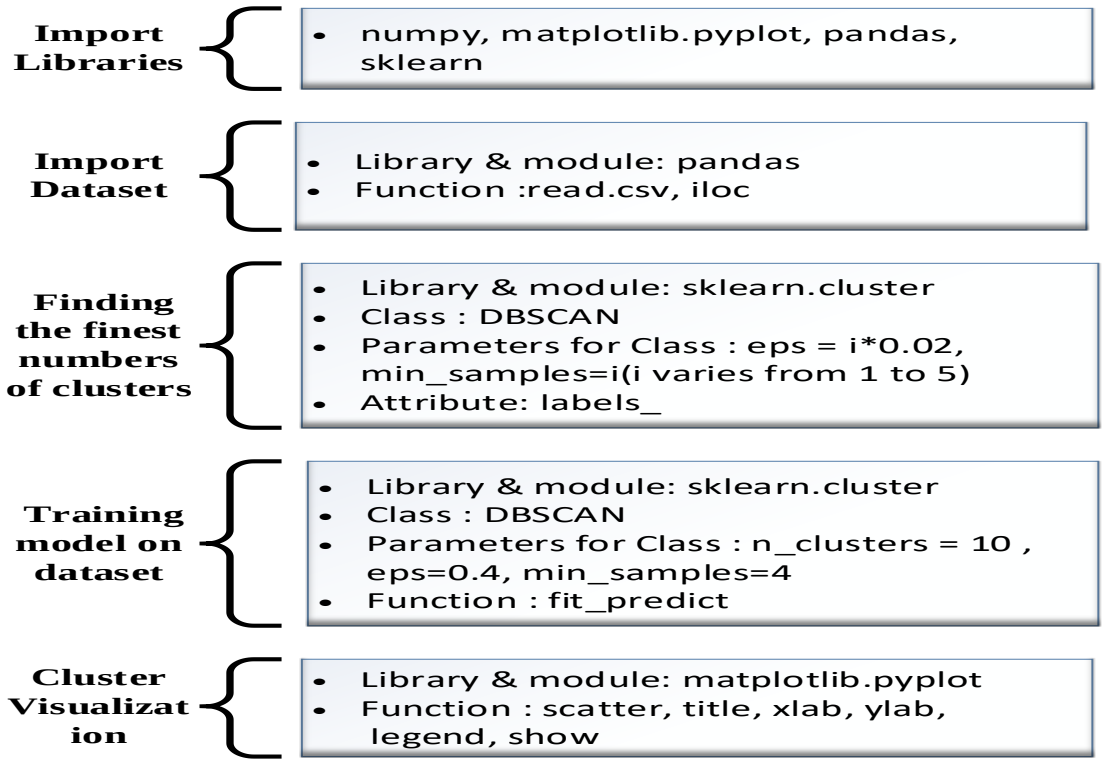

Figure 3. 8: *Python libraries for implementing DBSCAN clustering*

3.7.4. Mean Shift Algorithm

Mean-shift clustering is an unsupervised learning algorithm utilized to identify clusters within a dataset by considering the density of data points. The algorithm operates by iteratively shifting each data point towards the mean of its neighboring points within a specified radius, ultimately converging to local maxima of the density function, which represent the cluster centers. One notable advantage of mean-shift clustering is its ability to automatically determine the number of clusters without the need for prior specification. However, the algorithm's performance can be influenced by the selection of the radius parameter, impacting the size and shape of the resulting clusters.

For implementing mean-shift clustering in Python, we utilized the MeanShift class available in the sklearn cluster module. This class offers methods for fitting the data, predicting cluster labels, and estimating the bandwidth parameter.

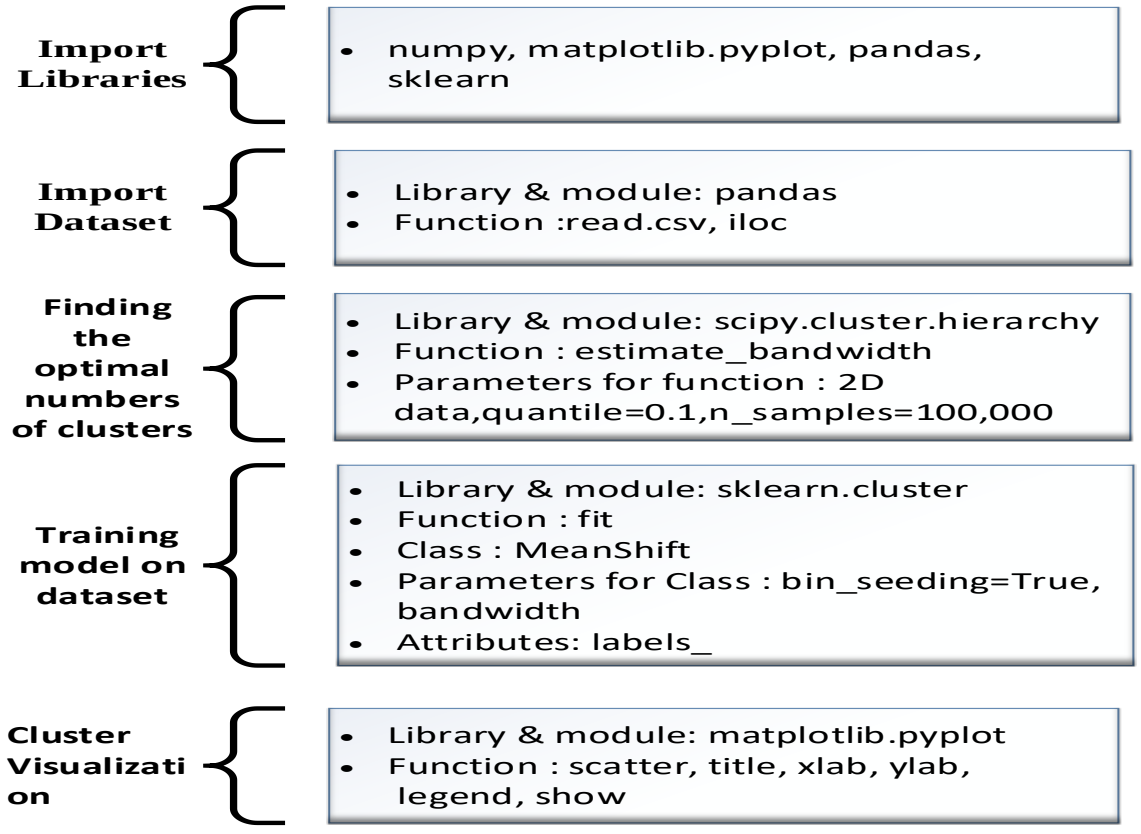

Figure 3. 9: *Python libraries for implementing mean shift clustering*

3.8. Model Evaluation Metrics

To evaluate the effectiveness of the algorithms applied to segment CBEBirr customers, several evaluation metrics were employed. Among the most widely utilized evaluation metrics for clustering algorithms are the Silhouette coefficient and the Davies-Bouldin Index. In this study, both the Silhouette coefficient and Davies-Bouldin Index were utilized to gauge the model's performance. These metrics are prevalent in unsupervised machine learning tasks and provide valuable insights into the clustering performance.

3.9. Hyperparameter Tuning

Hyperparameter tuning is a critical aspect of machine learning, and selecting the right values for hyperparameters is pivotal for achieving desirable outcomes. In our process of determining the optimal number of clusters, we applied hyperparameter tuning to refine our selection.

3.10. Summary

This chapter detailed the proposed architecture tailored for CBEBirr customer segmentation. We elucidated the foundational design elements and modules comprising the clustering methodology for customer data records. The CBEBirr data pre-processing module, encompassing cleaning and data transformation, emerged as a pivotal component of the model, essential for data preparation prior to clustering. Additionally, dimensionality reduction, a technique aimed at reducing the number of features or variables in a high-dimensional dataset while retaining crucial information, was explored. This method serves as a pre-processing step to enhance the performance of machine learning algorithms by reducing data dimensionality. Among the various options available, we delved into the chosen methodology for this study. The outlined methods shed light on how data is pre-processed for clustering and utilized to cluster customers based on their usage patterns. Furthermore, we provided algorithmic steps for the model we developed, along with illustrative examples in relevant sections. Lastly, the chapter delineated the model evaluation metrics employed to assess the performance of the algorithms.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1. Introduction

This chapter presents the findings of the research, which aimed to investigate the characteristics of CBEBirr customers and segment them based on their behavioural and demographic attributes using an unsupervised machine learning approach, as outlined in Chapter One. Four clustering algorithms, namely K-means, DBSCAN, Agglomerative clustering, and mean shift, were systematically tested to assess their performance with different sets of variables, aiming to determine the most effective clustering results. The effectiveness of these algorithms is gauged by their ability to effectively partition the dataset into distinct groups [54]. Therefore, we conducted a comparative analysis to evaluate the efficacy of these clustering algorithms in segmenting CBEBirr customers based on their demographic and behavioural profiles. This chapter thus focuses on experimentally validating whether the intended objective has been achieved and provides a detailed account of the experimental evaluation of our proposed approach, along with a discussion of the obtained results.

4.2. Exploratory data analysis

As outlined in Chapter 3, the dataset chosen for analysis is aimed at facilitating the modelling process. It encompasses usage data spanning six months for approximately 170,012 CBEBirr customers. The dataset is structured at the customer level and comprises 11 behavioural variables. The primary objective of utilizing this dataset is to derive customer segments based on their behavioural patterns, thereby enabling the bank to tailor its marketing strategies towards specific segments.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 170012 entries, 0 to 170011
Data columns (total 16 columns):
 #   Column                                Non-Null Count   Dtype
---  ------                                --------------   -----
 0   Merchant                              170012 non-null  float64
 1   Branch                                170012 non-null  float64
 2   Agent                                 170012 non-null  float64
 3   Gender                                170012 non-null  float64
 4   Number of Buy Air Time Transaction    170012 non-null  float64
 5   Number of Send Money Transaction      170012 non-null  float64
 6   Number of Pay Bill Transaction        170012 non-null  float64
 7   Number of Buy Goods Transaction       170012 non-null  float64
 8   Number of Cash in Transaction         170012 non-null  float64
 9   Number of Cash Out Transaction        170012 non-null  float64
 10  Amount of Transaction performed       170012 non-null  float64
 11  Age__15-25                            170012 non-null  float64
 12  Age__26-35                            170012 non-null  float64
 13  Age__36-45                            170012 non-null  float64
 14  Age__45-55                            170012 non-null  float64
 15  Age__Above 55                         170012 non-null  float64
dtypes: float64(16)
memory usage: 20.8 MB
```

Figure 4. 1:*Information of the data columns of the CBEBirr customers dataset*

As illustrated in figure 4.1, the info function is employed to display the count of records and their respective data types. Subsequently, the Numpy library in Python is utilized for fundamental quantitative analysis of the dataset. Descriptive statistics such as central tendency, range, standard deviation, mean, maximum, and minimum values are computed. Additionally, visualization tools including matplotlib, plotly, and Seaborn are utilized in conjunction with Python machine learning libraries for cluster analysis, model training, and evaluation. Table 4.1 provides an overview of the initial descriptive analysis of the raw data.

Table 4. 1: *Descriptive analysis of the data columns of the CBEBirr customers dataset*

	count	mean	std	min	25%	50%	75%	max
Merchant	170012.0	0.007488	0.086207	0.0	0.0	0.0	0.0	1.0
Branch	170012.0	0.706103	0.455547	0.0	0.0	1.0	1.0	1.0
Agent	170012.0	0.286027	0.451903	0.0	0.0	0.0	1.0	1.0
Gender	170012.0	0.764793	0.424129	0.0	1.0	1.0	1.0	1.0
Number of Buy Air Time Transaction	170012.0	9.408836	24.381790	0.0	0.0	0.0	7.0	859.0
Number of Send Money Transaction	170012.0	0.455556	4.205059	0.0	0.0	0.0	0.0	256.0
Number of Pay Bill Transaction	170012.0	2.353310	35.366766	0.0	0.0	0.0	1.0	9999.0
Number of Buy Goods Transaction	170012.0	0.588629	4.839424	0.0	0.0	0.0	0.0	340.0
Number of Cash in Transaction	170012.0	0.554484	2.334660	0.0	0.0	0.0	0.0	115.0
Number of Cash Out Transaction	170012.0	0.006135	0.742866	0.0	0.0	0.0	0.0	117.0
Amount of Transaction performed	170012.0	1296.797867	4271.294067	41.0	130.0	400.0	1180.0	675665.0
Age__15-25	170012.0	0.055837	0.229608	0.0	0.0	0.0	0.0	1.0
Age__26-35	170012.0	0.363068	0.480886	0.0	0.0	0.0	1.0	1.0
Age__36-45	170012.0	0.357292	0.479203	0.0	0.0	0.0	1.0	1.0
Age__45-55	170012.0	0.140796	0.347812	0.0	0.0	0.0	0.0	1.0
Age__Above 55	170012.0	0.083006	0.275892	0.0	0.0	0.0	0.0	1.0

Next, Exploratory Data Analysis (EDA) is conducted on the loaded dataset to enhance comprehension of its characteristics. Python libraries, specifically Matplotlib and Seaborn, are employed for data visualization. Histograms are frequently utilized to illustrate the distribution of numerical data. It's crucial to grasp the distribution of specific numerical variables within the dataset during exploration. Figures 4.2 to 4.14 showcase the distribution of various features across the dataset, with histograms employed for visualization purposes.

1. Age

This attribute indicates the age of customers that use CBEBirr service in the commercial bank of Ethiopia.

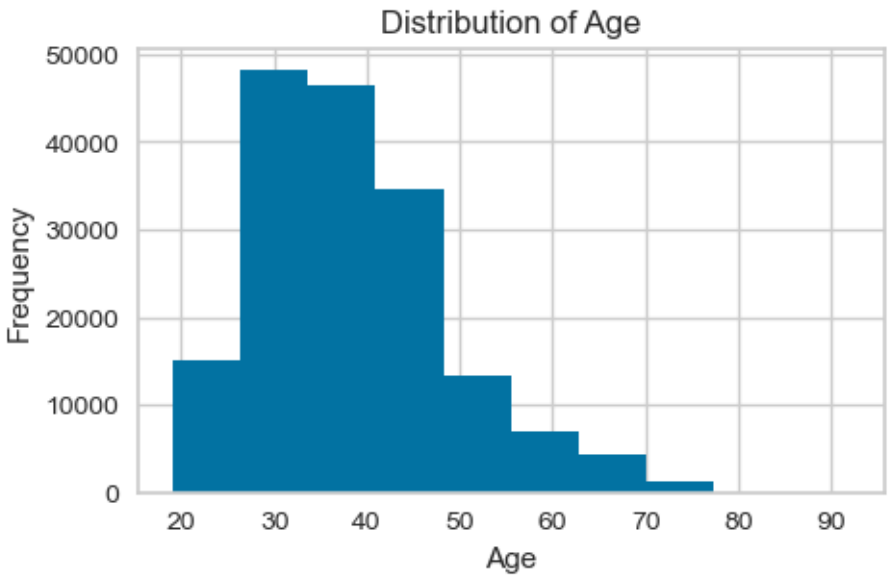

Figure 4. 2: *Age distribution*

2. Merchant

This is the seller or retailer who receives the payment via CBEBirr service in exchange for goods or services. Records between 0 and 1(1=indicates merchant, 0= not merchant)

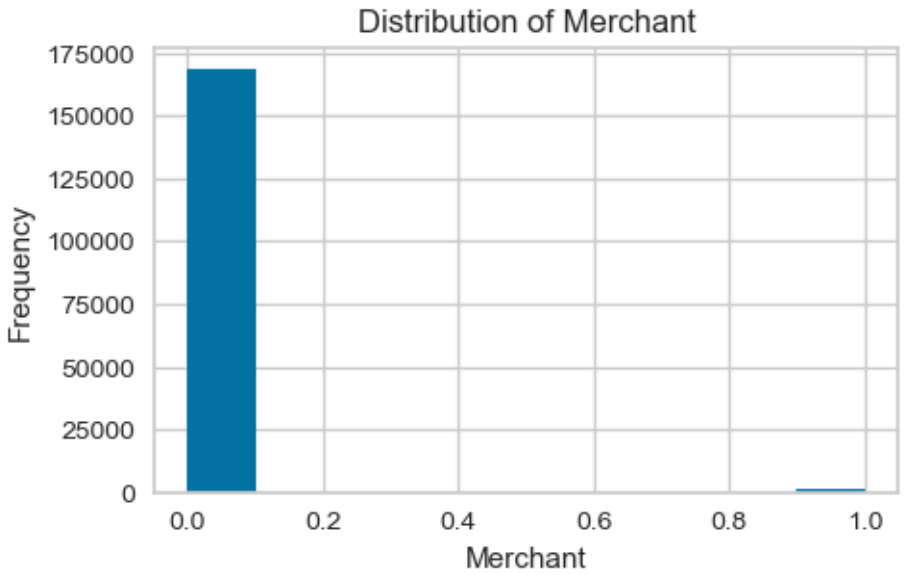

Figure 4. 3:*Merchant distribution*

3. Agent

This provides the merchant with ways to accept those payments by connecting them to an acquirer. Records between 0 and 1(1= indicates agent, 0= indicates not agent)

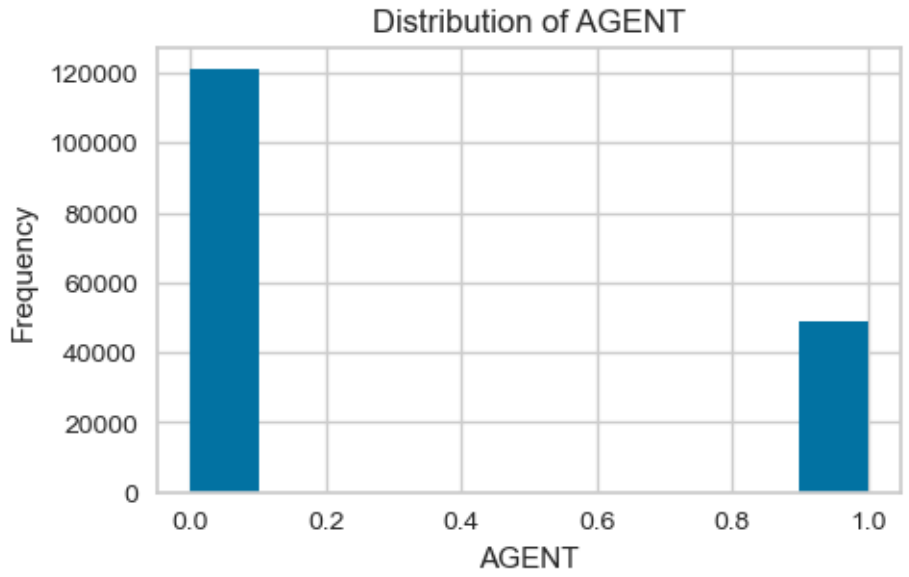

Figure 4. 4: *Agent distribution*

4. Branch

This is a retail location where CBE, credit union, or other financial institution offers a wide array of face-to-face and automated services to its customers. Record between 0 and 1 (1 = indicates branch, 0= indicates not branch).

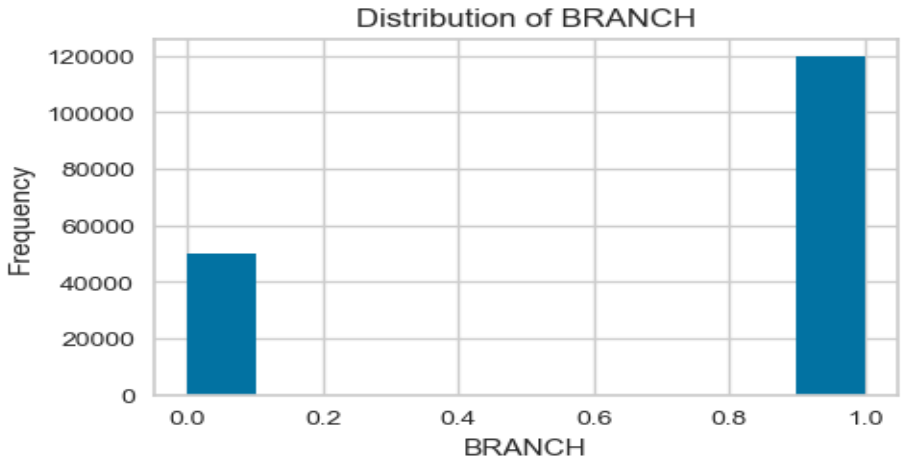

Figure 4. 5: *Branch distribution*

5. Gender

This indicates the gender of the CBEBirr customer. Records between 0 and 1(1= indicates male, 0= indicates female).

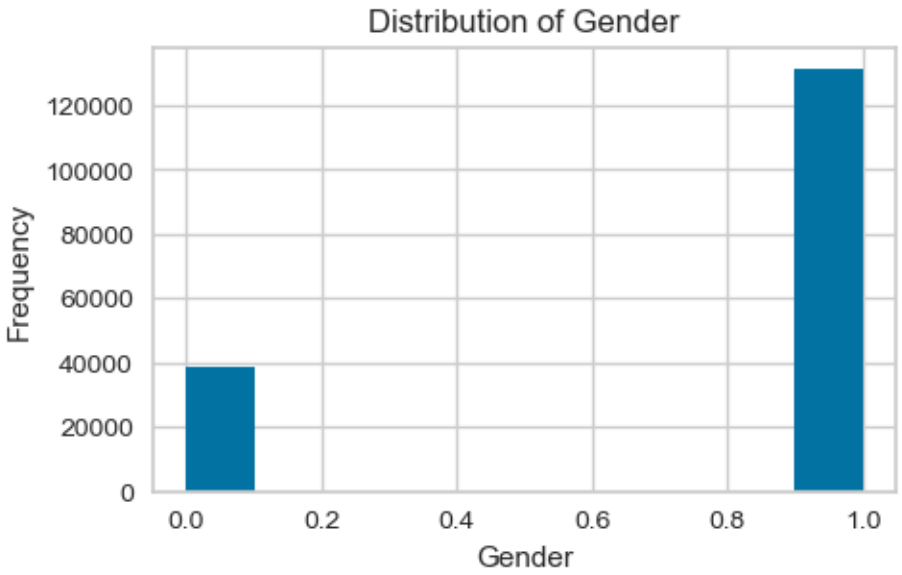

Figure 4. 6: *Gender distribution*

6. Number of buy air time transactions

This is CBEBirr customer purchases airtime for their mobile phone using their mobile number. CBE has different methods and codes for allowing customers to buy airtime from their accounts, such as Unstructured Supplementary Service Data (USSD) codes and CBEBirr apps.

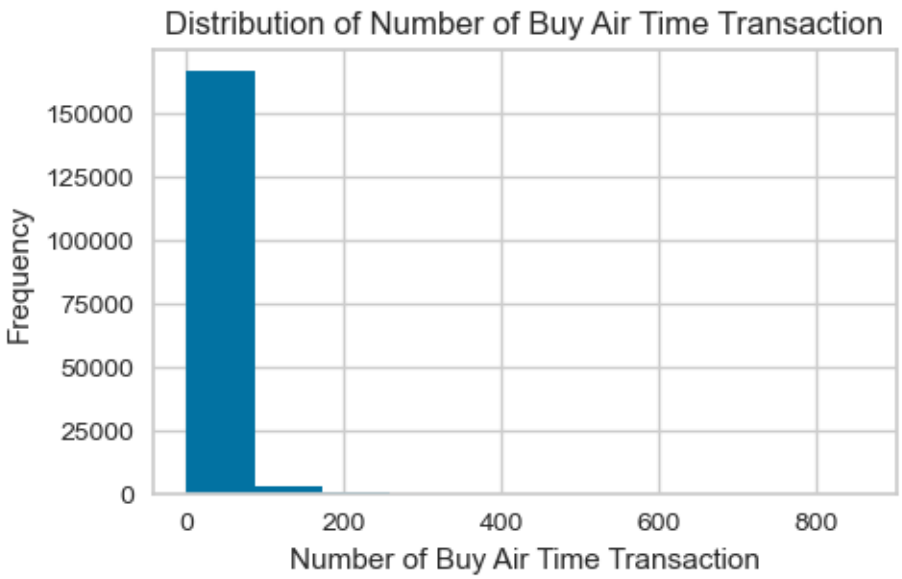

Figure 4. 7: *Number of buy air time transactions distribution*

7. Number of send money transactions

This is when CBEBirr customer transfers money from their CBEBirr account to another person's bank account, either within the same bank or across different banks using USSD and CBEBirr apps.

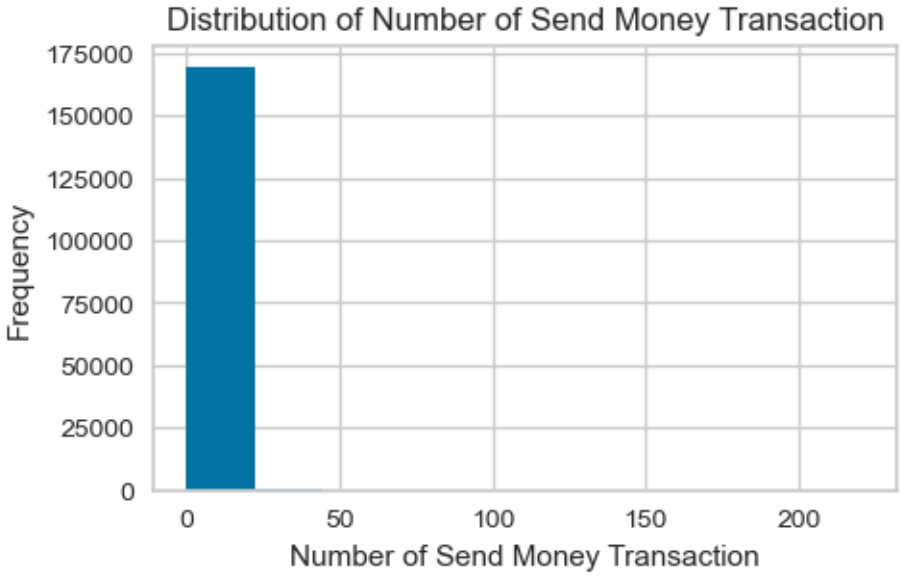

Figure 4. 8: *Number of send money transactions distribution*

8. Number of pay bill transactions

This is the number of times a customer pays a bill, such as utility, fuel, rent, immigration, airline ticket, and others using their CBEBirr service such as USSD or CBEBirr apps.

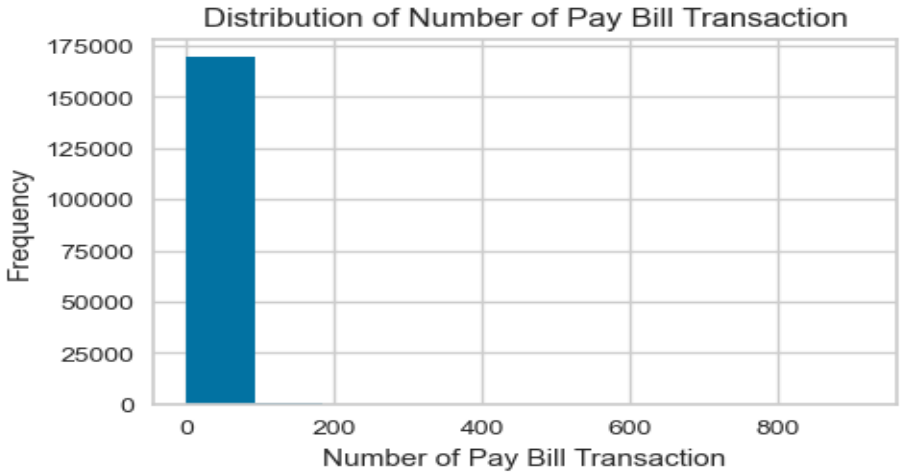

Figure 4. 9: *Number of pay bill transaction distribution*

9. Number of buy goods transaction

This is the number of times a customer purchases goods or services from a merchant using their CBEBirr account. This can be done through USSD or CBEBirr apps.

Figure 4. 10:*Number of buy goods transaction distribution*

10. Number of cash-in transactions

This is the number of times a customer deposits cash into their CBEBirr account.

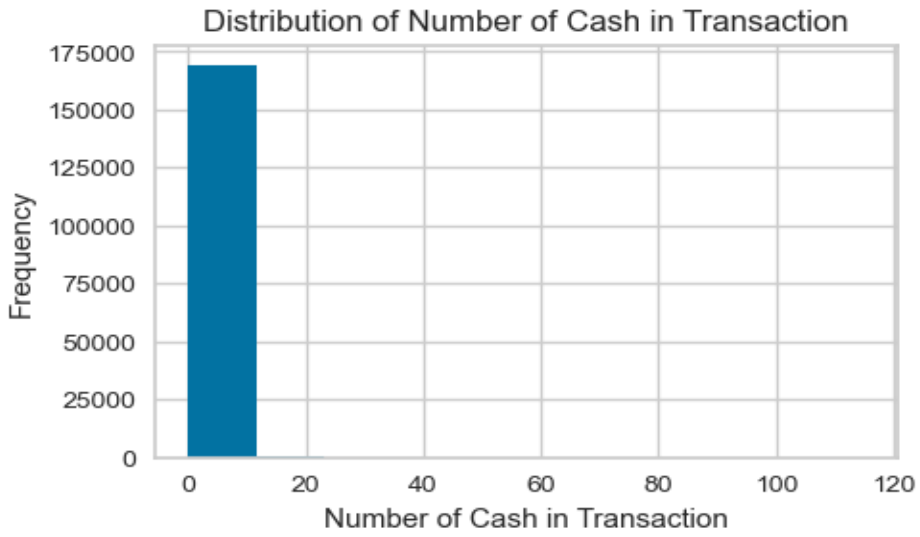

Figure 4. 11: *Number of cash-in transactions distribution*

11. Number of cash-out transactions

This is the number of times a customer withdraws cash from their CBEBirr account using CBEBirr apps or USSD.

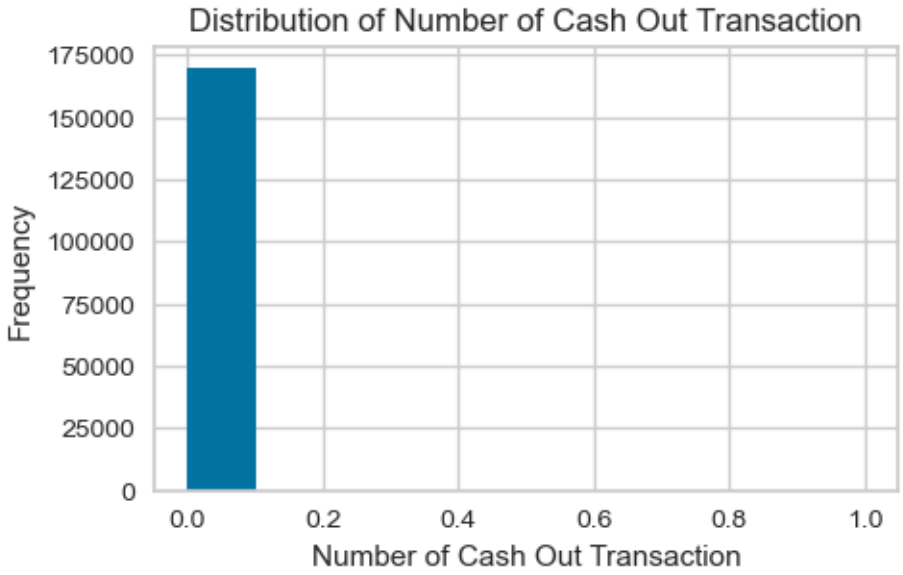

Figure 4. 12: *Number of cash-out transactions distribution*

12. Amount of transactions performed

This is the total number or value of transactions that are processed by a CBEBirr service within a given period of time.

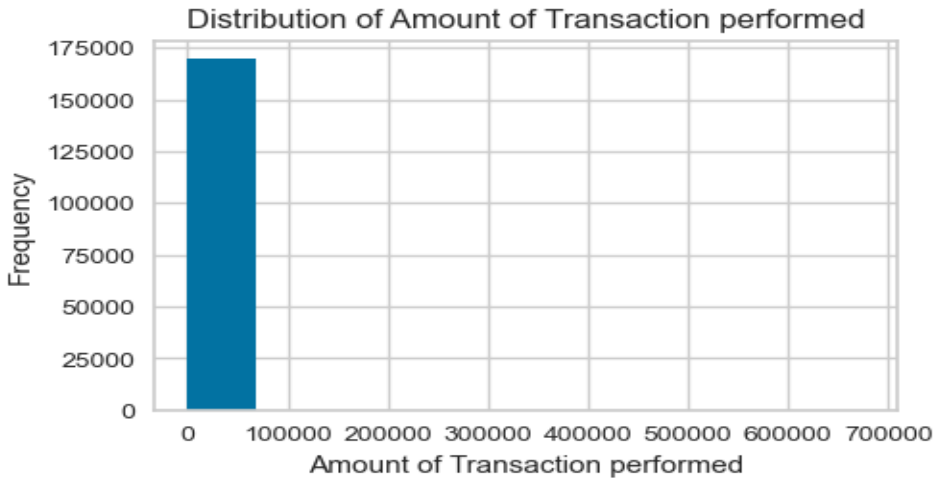

Figure 4. 13: *Amount of transactions performed distribution*

4.3. Hyperparameter tuning

Hyperparameter tuning is executed on the pre-processed dataset to determine the most suitable number of clusters for the K-means, DBSCAN, mean shift, and agglomerative clustering algorithms.

4.3.1. Hyperparameter tuning for K-means

The elbow method serves as a heuristic for determining the ideal number of clusters within a dataset. This technique involves plotting the explained variance against the number of clusters and selecting the "elbow" point on the curve as the optimal number of clusters. In our study, we initially aimed to identify the optimal number of clusters, denoted as 'k'. We conducted a series of experiments ranging from 2 to 40 clusters, recording the inertia at each iteration. Inertia represents the sum of squared distances of samples to their closest cluster center. Subsequently, we visualized the inertias for each 'k' value, as depicted in Figure 4.15. Additionally, we employed a visualizer tool, namely KElbowVisualizer from the yellow brick library in Python. Following our experimentation with the elbow method,

we determined that the optimal number of clusters ('k') is 6, corresponding to a point where the distortion is minimal. In Figure 4.15, the optimal number of clusters ('K') is graphed against the distortion, representing the total within-cluster sum of squares for a given number of clusters ('K'). The optimal 'K' value is identified at the point where the curve exhibits a bend or elbow.

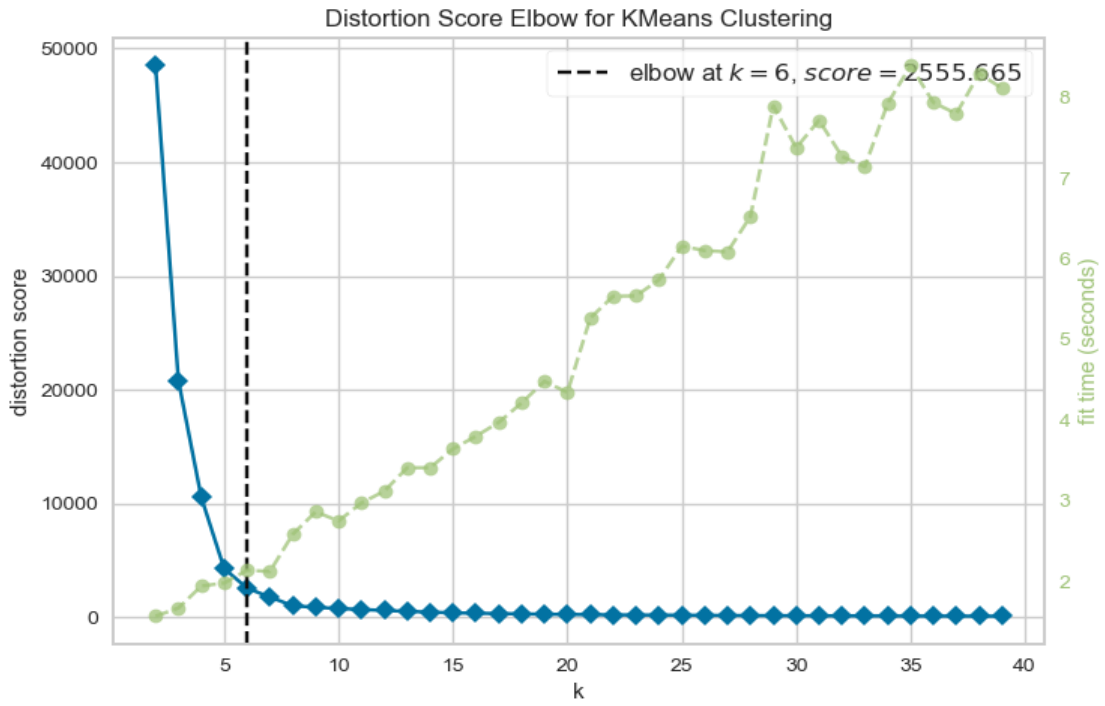

***Figure 4. 14**:K-value from elbow methods*

After conducting clustering on the data across a range of cluster numbers from 2 to 40, we collected the Silhouette coefficient and Davies-Bouldin index for each clustering outcome, which are presented in Table 4.2. These evaluation metrics aid in determining the most suitable number of clusters by evaluating the clustering quality and selecting the number of clusters that yield the best scores for these metrics. Ultimately, we identified six (6) clusters (K-value) as the optimal choice, maximizing the Silhouette coefficient at 0.792 and minimizing the Davies-Bouldin index at 0.291 for our dataset. Table 4.2 provides details on the inertia (distortion) value, the number of clusters, and the outcomes of the two-evaluation metrics.

Table 4. 2:*K-means hyperparameter tuning clustering results and evaluation Metrics*

Number of Clusters	Inertia(distortion)	Silhouette Score	Davies-Bouldin Index
2	48488.28	0.473	0.907
3	20817.62	0.612	0.661
4	10541	0.963	0.484
5	4276.9	0.756	0.369
6	2555.66	0.792	0.291
7	1767.48	0.794	0.301
8	1008.46	0.805	0.305
9	866.76	0.785	0.416
10	763.5	0.788	0.436
11	680.78	0.729	0.5
12	591.41	0.732	0.505
13	507.73	0.738	0.483
14	446.94	0.742	0.468
15	382.98	0.74	0.477
16	346.67	0.727	0.49
17	322.38	0.716	0.496
18	287.52	0.701	0.507
19	263.36	0.69	0.512
20	241.60	0.69	0.525
21	225.52	0.691	0.549
22	205.04	0.689	0.556
23	191.56	0.69	0.545
24	186.51	0.676	0.564
25	175.13	0.678	0.557
26	162.86	0.69	0.583
27	151.93	0.666	0.573
28	143.31	0.682	0.57
29	136.24	0.66	0.587
30	130.31	0.669	0.604
31	123.28	0.664	0.599
32	119.31	0.66	0.609
33	114.06	0.674	0.59
34	110.01	0.675	0.59
35	103.02	0.671	0.571

Number of Clusters	Inertia(distortion)	Silhouette Score	Davies-Bouldin Index
36	101.76	0.669	0.59
37	97.73	0.671	0.614
38	90.92	0.671	0.601
39	90.09	0.668	0.608

For additional examination, Figure 4.16 illustrates an Elbow plot featuring both the silhouette score and Davies Bouldin score for the K-means clustering algorithm. From this visualization, we can ascertain that the optimal value for k is 6 for clustering the customer dataset.

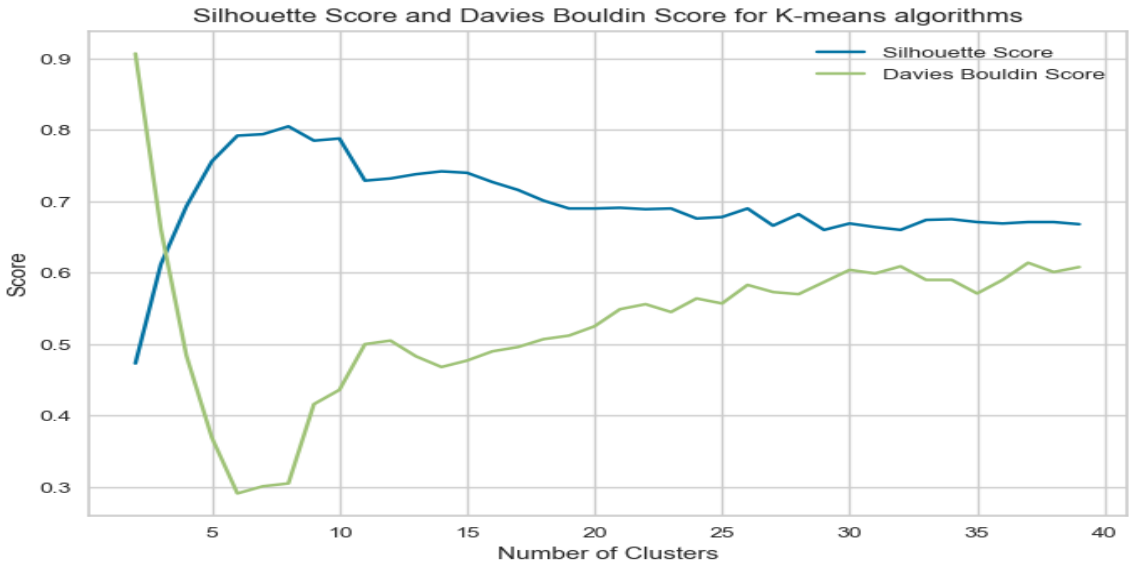

Figure 4. 15: *Silhouette score and Davies-Bouldin index plot*

4.3.2. Hyperparameter tuning for DBSCAN

To determine the most suitable cluster value, DBSCAN clustering relies on two key parameters: eps and min_samples, which respectively dictate the size of the neighbourhood and the minimum density required for clusters. The following steps outline the process for conducting DBSCAN clustering:

Step 1: Choose two parameters: eps and min_samples. Therefore, we defined the two parameters value eps= 0.01 initial value and min_samples range from 1 to 10.

Step 2: For each point in the data, find its neighbours within eps distance and count how many there are. If the number is greater than or equal to min_samples, mark the point as a core point. Otherwise, mark it as a noise point.

Step 3: For each core point, find all the points that are directly or indirectly reachable from it, meaning they are within eps distance from any core point in the same cluster. Assign them the same cluster label. Noise points are not assigned to any cluster.

Step 4: Repeat steps 2 and 3 until all points are processed.

Table 4.3 presents the epsilon, minimum sample size, and two evaluation metrics aimed at identifying the optimal cluster value by maximizing the Silhouette score and minimizing the Davies-Bouldin index. Consequently, we have derived a Silhouette score of 0.304626, Davies-Bouldin index of 0.106958, epsilon of 0.03, and a minimum sample size of 9. Based on these metrics, the ideal cluster count for the dataset is determined to be 3. This decision is informed by the evaluation of both the silhouette score and Davies Bouldin score, along with the epsilon and minimum sample size parameters, which collectively indicate that 4 clusters offer the best fit for our dataset.

Table 4. 3:*DBSCAN hyperparameter tuning clustering results and evaluation metrics*

Eps	Min_samples	Number of clusters	Silhouette Score	Davies-Bouldin Index
0.01	1	132	0.004947	0.85423
0.02	1	14	0.215894	0.913047
0.03	1	7	-0.046245	1.458476
0.04	1	3	-0.051617	1.494528
0.01	2	77	0.049193	1.77881
0.02	2	11	0.217325	2.149887
0.03	2	5	-0.011932	1.824663
0.04	2	3	-0.051617	1.494528
0.01	3	49	0.05739	2.924832
0.02	3	11	0.217325	2.149887
0.03	3	5	-0.011932	1.824663
0.04	3	3	-0.051617	1.494528
0.01	4	35	0.173843	2.18033
0.02	4	10	0.217716	1.367136

Eps	Min_samples	Number of clusters	Silhouette Score	Davies-Bouldin Index
0.03	4	5	-0.011932	1.824663
0.04	4	3	-0.051617	1.494528
0.01	5	35	0.137688	1.213479
0.02	5	10	0.233723	1.858973
0.03	5	4	-0.003106	1.399516
0.01	6	33	0.179761	1.227928
0.02	6	8	0.238448	1.606722
0.03	6	4	-0.003106	1.399516
0.01	7	29	0.195063	1.25519
0.02	7	8	0.238855	3.597232
0.03	7	5	0.269646	0.827738
0.01	8	27	0.19588	1.262492
0.02	8	8	0.238228	1.460852
0.03	8	5	0.273465	0.887562
0.01	9	26	0.203514	1.302924
0.02	9	8	0.237868	1.516592
0.03	9	3	0.304626	0.106958

In addition, the value for parameter epsilon can be determined by plotting the K-Distance graph. The maximum point curve is the graph is selected as the value for the parameter epsilon.

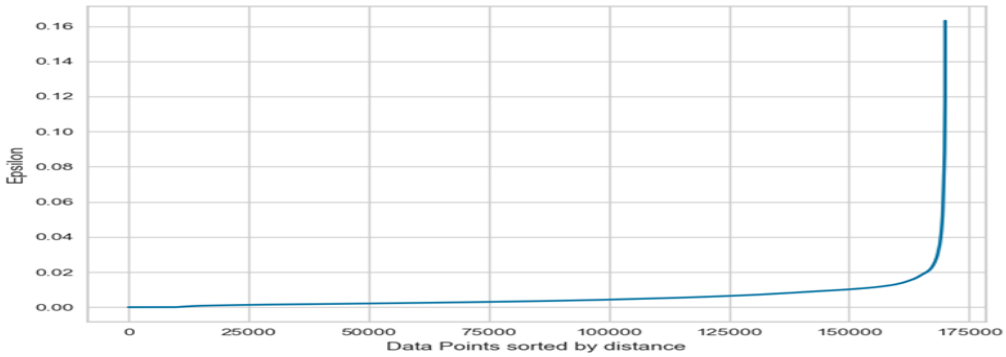

Figure 4. 16: *K-Distance Graph*

Figure 4.17 shows the plotted K-Distance graph for the processed dataset. According to this graph, the maximum curve point is at 0.03 value hence, this value is selected as the optimum value for parameter epsilon.

4.3.3. Hyperparameter tuning for mean shift

Table 4.4 summarizes the outcomes of the hyperparameter tuning analysis conducted for the Mean Shift clustering algorithm. This experimentation involved exploring various bandwidth parameter values to assess their impact on clustering performance. The results reveal that specific bandwidth values within a defined range notably enhance both clustering accuracy and convergence speed.

Furthermore, Table 4.4 comprises columns detailing the bandwidth value, number of clusters, Silhouette score, and Davies-Bouldin index. These metrics serve as benchmarks for determining the optimal cluster value based on evaluation criteria. By examining the relationship between bandwidth values and the associated clustering metrics, the experiment offers valuable insights into selecting the optimal bandwidth hyperparameter for the mean shift clustering algorithm.

Table 4. 4: *Mean shift hyperparameter tuning clustering results and evaluation metrics*

Bandwidth	Number of clusters	Silhouette Score	Davies-Bouldin Index
0.01	255	0.653	0.686
0.02	116	0.660	0.674
0.03	83	0.654	0.716
0.04	53	0.671	0.739
0.05	37	0.676	0.644
0.06	22	0.753	0.564
0.07	15	0.756	0.502
0.08	11	0.776	0.551
0.09	10	0.771	0.532
0.10	10	0.772	0.512
0.11	7	0.775	0.541
0.12	7	0.775	0.512
0.13	7	0.776	0.541
0.14	7	0.776	0.465
0.15	7	0.779	0.462
0.16	7	0.779	0.458
0.17	6	0.792	0.290
0.18	5	0.756	0.369

Bandwidth	Number of clusters	Silhouette Score	Davies-Bouldin Index
0.19	5	0.729	0.417
0.20	5	0.728	0.417

Moreover, prior to training the mean shift clustering model on the customer dataset, the process of hyperparameter tuning is conducted to determine the optimal number of clusters.

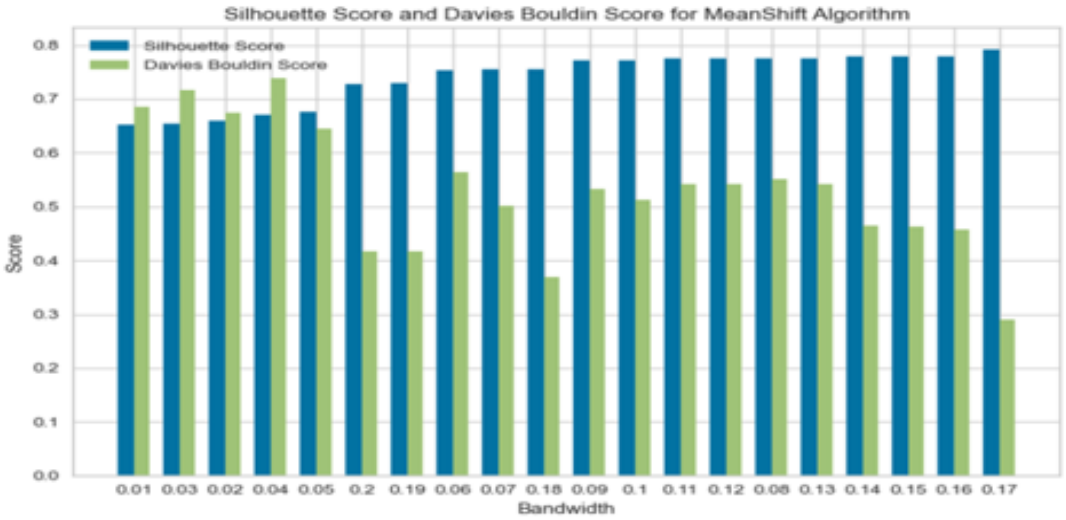

Figure 4. 17*:Mean-shift Silhouette score and Davies-Bouldin index*

The outcomes of the Silhouette and Davies-Bouldin index for different numbers of clusters are depicted in Figure 4.18. The Silhouette score reaches its peak at 0.792, while the Davies-Bouldin score achieves its lowest value of 0.290 with a bandwidth of 0.17. From the analysis presented in Table 4.4, it is observed that the optimal number of clusters is 4. Consequently, we have determined 6 to be the ideal number for training the mean shift clustering model.

4.3.4. Hyperparameter tuning for agglomerative clustering algorithm

The optimal number of clusters and the most suitable linkage method for the agglomerative clustering algorithm are determined through hyperparameter tuning. This process entails examining various numbers of clusters and linkage methods to determine the combination that yields the most effective clustering performance.

To accomplish this, we utilized metrics such as the silhouette score and the Davies-Bouldin score to gauge the quality of the clustering outcomes for different parameter configurations. By systematically exploring the parameter space and evaluating clustering performance using these metrics, we identified the ideal number of clusters and the optimal linkage method for our CBEBirr customer dataset. Table 4.5 displays the silhouette score, Davies-Bouldin score, number of clusters, and linkage method.

Table 4. 5*:Agglomerative clustering algorithm hyperparameter tuning clustering results and evaluation metrics*

Iteration	Number of clusters	Silhouette score	Davies-Bouldin Index	Linkage method
1	2	0.556	0.568	average
2	2	0.450	0.847	complete
3	3	0.385	0.596	Single
4	3	0.706	0.449	Average
5	3	0.696	0.459	Complete
6	4	0.276	0.711	Single
7	4	0.714	0.300	Average
8	4	0.575	0.598	Complete
9	5	0.273	0.716	Single
10	5	0.646	0.448	Average
11	5	0.569	0.683	Complete
12	6	0.259	0.707	Single
13	6	0.614	0.526	Average
14	6	0.626	0.565	Complete
15	7	0.230	0.751	Single
16	7	0.589	0.583	Average
17	7	0.591	0.608	Complete
18	8	0.184	0.793	Single
19	8	0.535	0.681	Average
20	8	0.582	0.664	Complete
21	9	0.172	0.767	Single
22	9	0.541	0.664	Average
23	9	0.540	0.618	Complete
24	10	0.158	0.780	Single
25	10	0.542	0.700	Average

Iteration	Number of clusters	Silhouette score	Davies-Bouldin Index	Linkage method
26	10	0.539	0.659	Complete
27	11	0.154	0.764	Single
28	11	0.530	0.758	Average
29	11	0.534	0.638	Complete
30	12	0.103	0.854	Single
31	12	0.513	0.732	Average
32	12	0.528	0.688	Complete
33	13	0.101	0.834	Single
34	13	0.498	0.703	Average
35	13	0.502	0.712	Complete
36	14	0.095	0.838	Single
37	14	0.503	0.701	Average
38	14	0.492	0.788	Complete

After 38 iterations, we obtained the optimum value of 4 clusters with the best linkage method being **"average"**, resulting in a maximum silhouette score of 0.714 and a Davie-Bouldin score of 0.300 at 7 iterations, as observed from Table 4.5. These results indicate the effectiveness of the clustering approach in capturing meaningful patterns in our dataset.

Further analysis, of the optimal number of clusters obtained through hyperparameter tuning before performing the agglomerative clustering algorithm on the processed CBEBirr customer dataset.

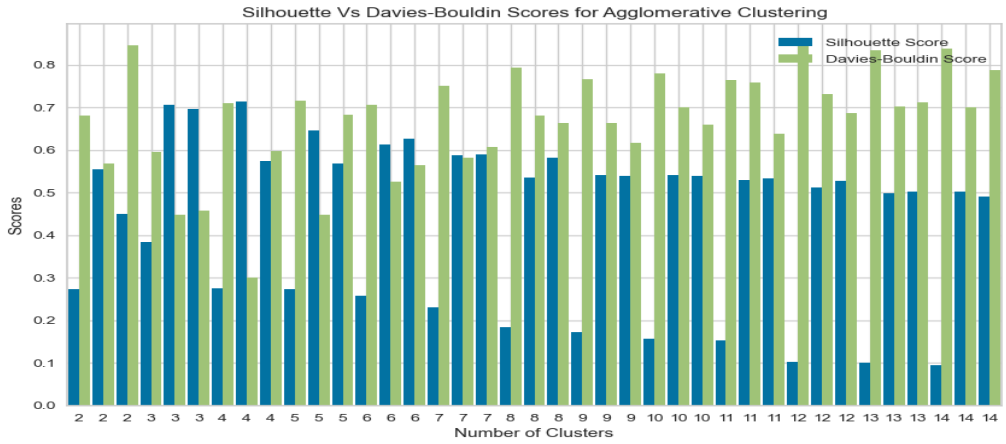

Figure 4. 18:: *Silhouette score and Davies-Bouldin index of Agglomerative clustering*

Figure 4.19 illustrates the Silhouette score and Davies Bouldin score achieved for Agglomerative clustering, aiding in the identification of the optimal number of clusters prior to clustering. Based on the depicted bar graph, we determined 4 to be the optimal number of clusters. As outlined in the Chapter 2 literature review, a higher Silhouette score and a lower Davies Bouldin score indicate superior clustering performance.

4.4. Clustering Model Analysis

This section presents the evaluation of cluster analysis, including the distribution and outcomes obtained from training four distinct clustering algorithms.

4.4.1. K-means Clustering

Following the determination of the optimal value for the clustering algorithm, the CBEBirr customer dataset undergoes clustering using the K-means algorithm. The clustering outcome is visualized in Figure 4.20, with PC1 and PC2 utilized for the 2D plot representation of clusters.

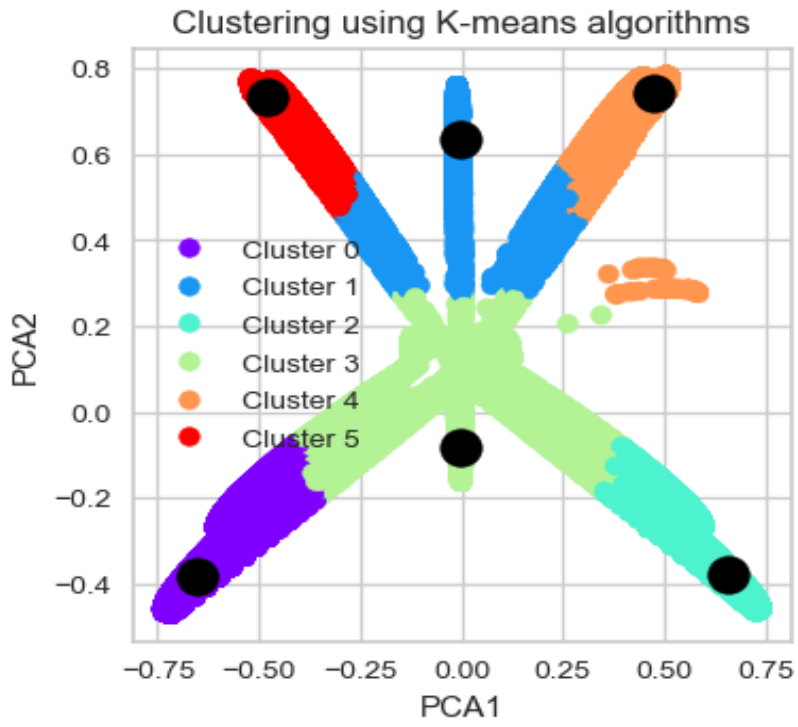

Figure 4. 19:*Cluster result obtained through K-means Algorithm*

Based on the results, the distribution of data across 6 clusters is depicted in Figure 4.21. In Cluster 0, there are 40,668 customers, while Cluster 1 comprises 13,843 customers. Cluster 2 encompasses 40,855 customers, Cluster 3 consists of 40,178 customers, Cluster 4 includes a total of 17,330 customers and cluster 5 includes a total of 17,138 customers.

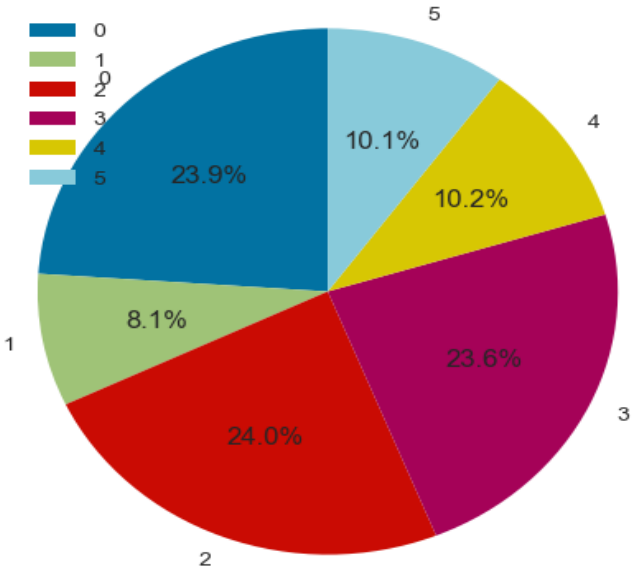

Figure 4. 20*:Customer data distribution among clusters - K-means clustering*

4.4.2. DBSCAN Clustering

Using the determined optimal values for epsilon and minimum samples, we established 3 clusters. Subsequently, we applied the DBSCAN clustering technique to the CBEBirr customer dataset based on these cluster numbers.

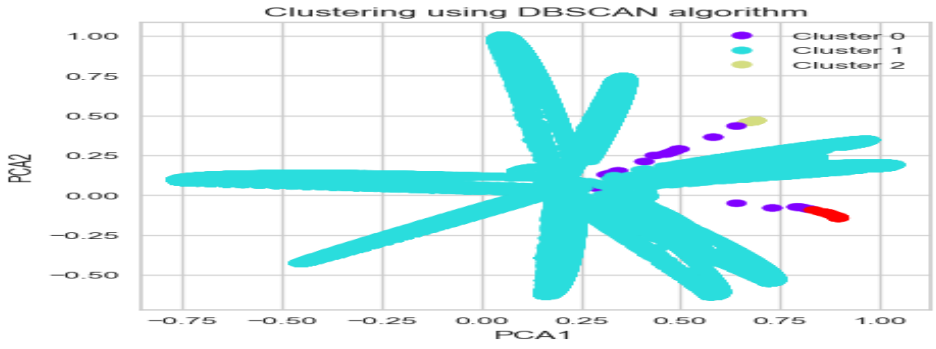

Figure 4. 21: *Cluster result obtained through DBSCAN clustering*

The result of applying the DBSCAN algorithm are illustrated in Figure 4.22. It demonstrates the segmentation of the customer base into four distinct clusters. The visual representation of this clustering is depicted using PC1 and PC2 components. However, upon analysing the distribution of clusters shown in Figure 4.23, it's apparent that the clustering distribution is not particularly uniform. Almost 100% of the customer dataset is categorized under cluster 0, comprising 169,947 customers. Notably, cluster -1 and 1 contains no customers, accounting for 0% of the dataset, with only 9 and 56 customers classified within this group respectively.

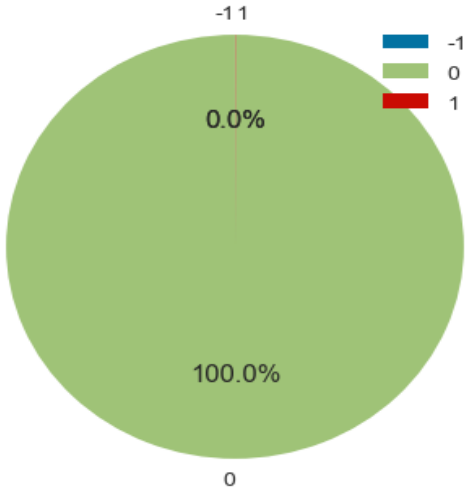

Figure 4. 22:*Customer data distribution among clusters - DBSCAN clustering*

4.4.3. Agglomerative clustering

Once the optimal value for the clustering algorithm was determined, the agglomerative clustering algorithm was employed on the CBEBirr customer dataset, and the resultant clustering outcome is depicted in figure 4.24. The 2D plot of clusters utilizes PC1 and PC2 for visualization.

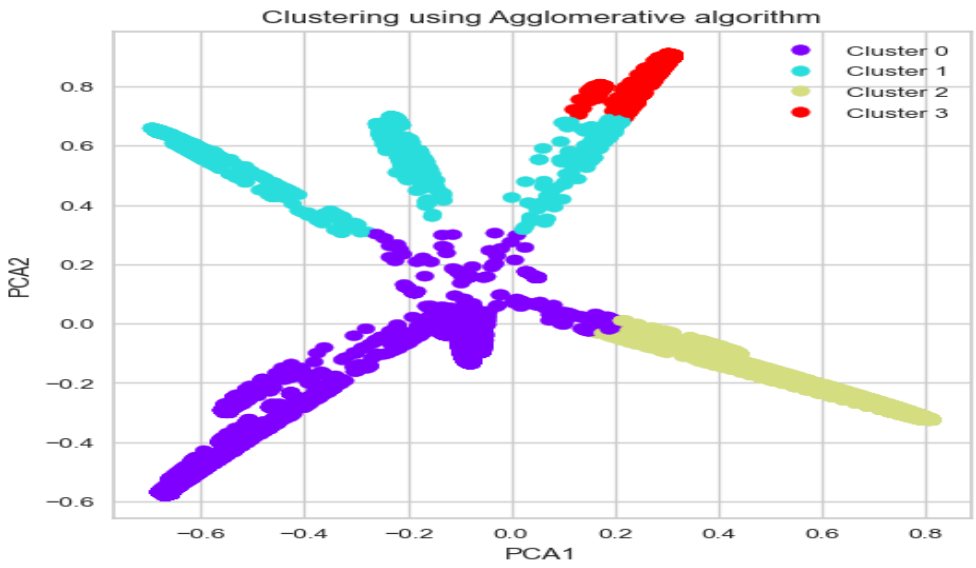

***Figure 4. 23**: Cluster result obtained through Agglomerative clustering*

As illustrated in figure 4.25, Cluster 0 encompasses approximately 43.4% of the CBEBirr customer dataset of 73,785 customers. Cluster 1 represents only 16.4% of the distribution, consisting of 27,882 customers. Cluster 2 comprises 28.7% of the dataset, with 48,793 customers classified within this cluster, while Cluster 3 accounts for 11.5% distribution, encompassing 19,551 customers. It is noticeable that the distributions of the remaining clusters are comparatively smaller in comparison to Cluster 0.

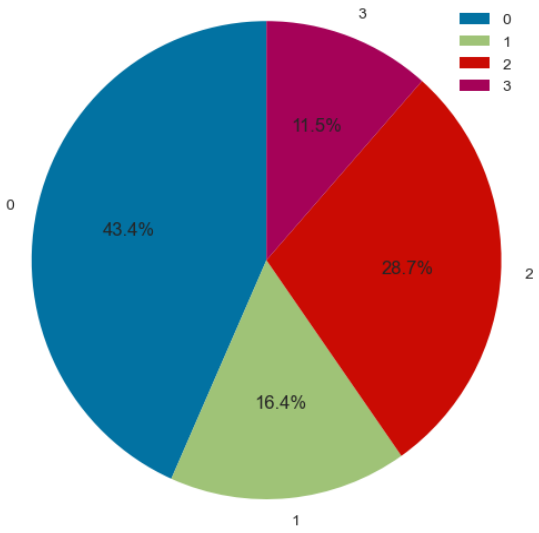

***Figure 4. 24**:Customer data distribution among clusters - Agglomerative clustering*

4.4.4. Mean Shift clustering

Upon determining the optimal cluster value, we implemented the mean shift algorithm on the CBEBirr customer dataset, resulting in the clustering outcome displayed in Figure 4.26. The visualization of the clusters utilizes the PC1 and PC2 components to represent the data in a 2D space.

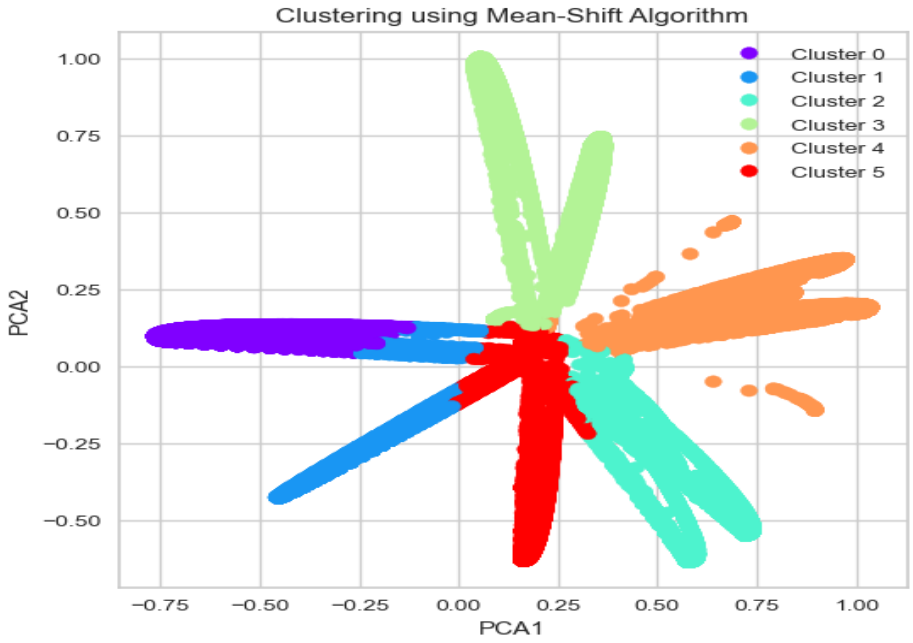

Figure 4. 25:*Cluster result obtained through Mean Shift clustering*

A pie chart, as depicted in Figure 4.27, was generated to illustrate the distribution of the CBEBirr client base across six distinct clusters. Cluster 0 comprises the largest portion, with 25.9% of customers, which are 40,794. Following this, cluster 1 constitutes 19.6% of customers, accounting for 40,621 individuals. Cluster 2 accounts for 18.5% of customers, with a total of 40,303, while cluster 3 represents 18% of customers, amounting to 17,339 individuals, cluster 4 represents 10.8% of customer, which contains 17,154 and cluster 5 represents 7.1% of customer, which contains 13,801. This chart provides insight into the customer demographics within each cluster.

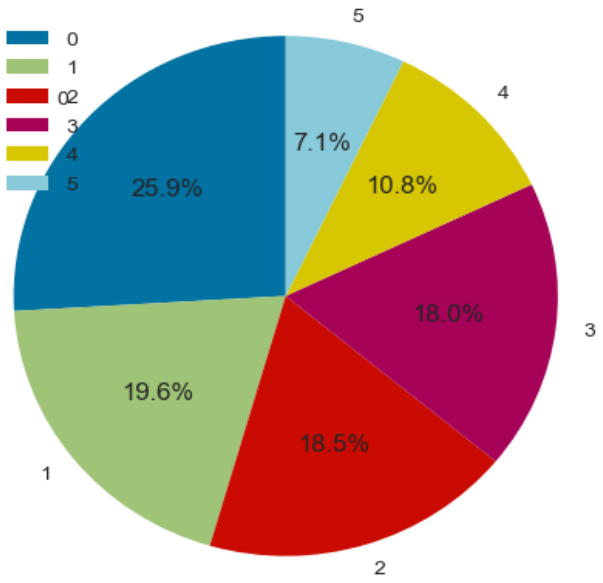

***Figure 4. 26**:Customer data distribution among clusters – Mean Shift clustering*

4.5. Cluster Results Evaluation and Interpretation

Multiple clustering algorithms exist for customer segmentation, each presenting distinct advantages and limitations based on factors such as complexity, performance, and adaptability to diverse datasets. In our CBEBirr customer segmentation study, we assessed four algorithms: K-means clustering, Agglomerative clustering, DBSCAN, and the mean shift algorithm, utilizing metrics such as the Silhouette coefficient and Davies Bouldin Index for evaluation.

Figure 4.28 depicts the Silhouette score and Davies-Bouldin index for the aforementioned algorithms employed in our investigation: K-means clustering, Agglomerative clustering, DBSCAN, and the mean shift algorithm.

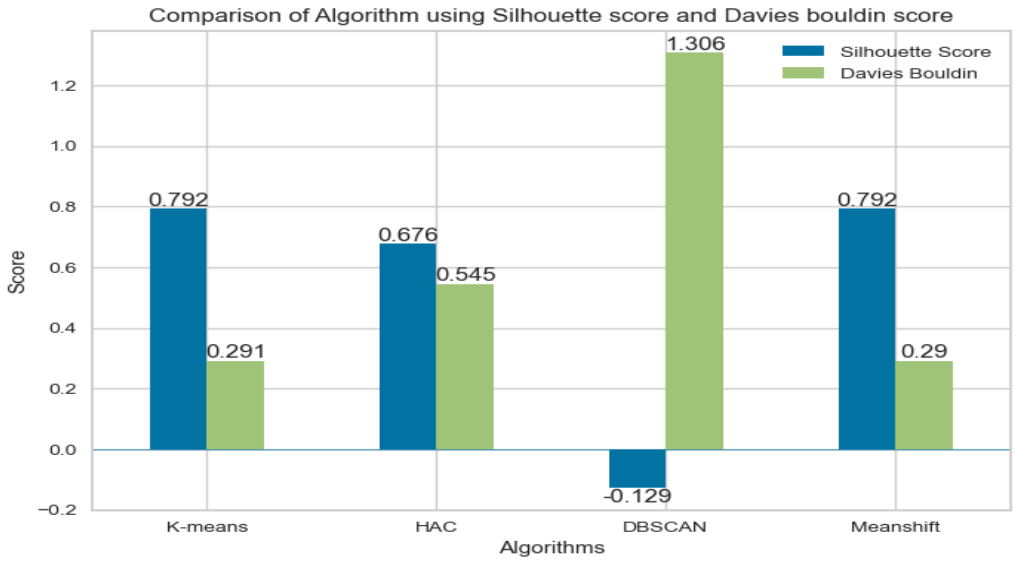

Figure 4. 27: *Summary of clustering algorithms performance*

To clarify, our experiment across K-means clustering, agglomerative clustering, DBSCAN, and the mean shift algorithm yielded higher Silhouette scores and lower Davies-Bouldin index values, as illustrated in Figure 4.28.

Considering these evaluation metrics across the four algorithms applied to cluster our CBEBirr customer data, the mean shift algorithm outperforms agglomerative clustering, DBSCAN, and the K-means algorithm, evident from its higher Silhouette score and lower Davies-Bouldin index.

Analysing and interpreting clusters is crucial for comparing them based on segmentation criteria. In our study, segmentation criteria encompass demographic data such as age, gender, and recruitment method (Agent, Merchant, and Branch), as well as behavioural features including the number of transactions (Buy Air Time, Send Money, Pay Bill, Buy Goods, Cash-in, and Cash-out) and transaction amounts.

Examining the outcomes of our clustering model for CBEBirr customers provides insights into the extent of usage among customers of the Commercial Bank of Ethiopia's CBEBirr services. This understanding serves as a foundation for crafting marketing strategies tailored to segments where the service is either underutilized or moderately utilized, guided

by the analysis results of each cluster. To gauge the significance of attributes within each cluster, we computed the mean (average) value of each record in every cluster using Python, as depicted in Table 4.6.

Table 4. 6:*Mean values on records of each cluster*

Cluster	Age Range					Merchant	Branch	Agent	Gender	Number of Buy Air time transaction	Number of Send money Transaction	Number of Pay bill Transaction	Number of Buy good transaction	Number of Cash in Transaction	Number of Cash out Transaction	Amount of Transaction Performed
	15-25	26-35	36-45	46-55	Above 55											
0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	11.0	0.0	3.0	1.0	0.0	0.0	1330.0
1	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	8.0	0.0	3.0	0.0	1.0	0.0	1472.0
2	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	9.0	1.0	2.0	1.0	1.0	0.0	1155.0
3	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	11.0	1.0	1.0	0.0	1.0	0.0	1173.0
4	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	10.0	0.0	2.0	0.0	0.0	0.0	1391.0
5	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	9.0	0.0	2.0	0.0	1.0	0.0	1136.0

Based on the findings depicted in Table 4.6, distinct customer groups have been discerned within the designated customer dataset. Furthermore, the distinctive characteristics of these six customer segments have been delineated.

Cluster 0

CBEBirr customers in this category are more male with all average age range which are 15-25,25-36,36-45,45-55, and above 55 years old, slightly used Buy Air Time, pay bill transactions, and buy good among others CBEBirr services, and customers in this cluster are also recruited through a branch. Most likely Customers in this cluster have performed an average amount of transactions of 1330.

Cluster 1

CBEBirr customers in this category are male with an average age range 26-35 years old, use CBEBirr service: Buy Air Time, Pay Bill, and cash in at some extent among others, and customers in this cluster are mainly recruited through branch. Most likely customers in this cluster have performed an average amount of transactions of 1472.

Cluster 2

CBEBirr customers in this category are more female with an average age range of 36-45 years old, used CBEBirr service: Buy Air Time, send money, pay bill and cash in at some extent, and customers in this cluster are also recruited through a branch. Most likely Customers in this cluster have performed an average amount of transactions of 1155.

Cluster 3

CBEBirr customers in this category are also more male with all an average age range, used CBEBirr service: Buy Air Time at high extent among others, and customers in this cluster are also recruited through branch. Most likely Customers in this cluster have performed an average amount of transactions of 1173.

Cluster 4

CBEBirr customers in this category are also more male with an average age range of 26-35, used CBEBirr service: Buy Air Time and pay bill at some extent among others, and customers in this cluster are also recruited through branch. Most likely Customers in this cluster have performed an average amount of transactions of 1391.

Cluster 5

CBEBirr customers in this category are also more female with an average age range of 36-45, used CBEBirr service: Buy Air Time at high extent among others, and customers in this cluster are also recruited through branch. Most likely Customers in this cluster have performed an average amount of transactions of 1136 which are considered to be the low transaction amount among others.

4.5.1. Comparison of our result with other researchers

Our analysis results were compared with those of other researchers to ensure the robustness and validity of our findings. The comparison revealed a high degree of alignment between our conclusions and those of our peers, providing further confidence in the accuracy of our analysis. This consistency across different research efforts lends further support to the

reliability of the conclusions drawn from our study and serves to validate the soundness of our methodology and approach.

Furthermore, the comparison highlighted areas of convergence as well as divergence between our analysis results and those of other researchers. These divergences provided valuable insights into potential areas for further investigation and exploration, as well as opportunities for nuanced interpretation and discussion of the findings. Overall, the comparison with other researchers' analysis results enriched our understanding of the subject matter and enhanced the depth and rigor of our research. Table 4.7 shows the comparison of our analysis results with others.

Table 4. 7: *Our result comparison with other researchers*

Reference	Algorithm used	Evaluation metrics	Number of datasets	Results achieved
[55]	K-means, agglomerative, DBSCAN	Silhouette score and Davies-Bouldin's index	216,721	• In agglomerative researcher got K=4 clusters. • In the case of K-means, the researcher got K=4 • In the case of DBSCAN K=6 However, the researcher identified the modelling algorithms based on the two-evaluation metrics i.e., K-means is the best one.
[41]	K-Means clustering algorithm	Silhouette, and Calinski-Harbaz	143,945	The researcher got the optimum value of cluster is 4 in K-means algorithms and identified K-means is the best algorithm for the given dataset.

Reference	Algorithm used	Evaluation metrics	Number of datasets	Results achieved
Our Analysis Result				
	K-means, agglomerative, DBSCAN, and means shift	Silhouette score and Davies-Bouldin's index	170,012	• In the case K-means algorithms K=6 • In the case of agglomerative K=4 • In the case of DBSCAN K=3 • In the case of mean shift K=6 Here, in our analysis we have identified mean shift as the best algorithm based on a maximum value Silhouette score and a minimum value of Davie-Bouldin index

Table 4.7 describes how different clustering techniques and evaluation criteria yield varied results when comparing our analysis with other researchers. In the study by [55], the researchers utilized K-means, agglomerative, and DBSCAN algorithms, with the Silhouette score and Davies-Bouldin's index serving as evaluation metrics across 216,721 datasets. They reported achieving K=4 clusters with the agglomerative algorithm, K=4 clusters with K-means, and K=6 clusters with DBSCAN. However, they identified K-means as the best algorithm based on the evaluation metrics. Similarly, the study by [41] focused solely on the K-means clustering algorithm, utilizing the Silhouette and Calinski-Harbaz metrics across 143,945 datasets, and concluded that K-means was the best algorithm, with an optimum cluster value of 4.

In our analysis, which involved the use of K-means, agglomerative, DBSCAN, and mean shift algorithms, along with the Silhouette score and Davies-Bouldin's index across 170,012 datasets, we obtained different results. Specifically, we identified K=6 clusters

with K-means, K=4 clusters with agglomerative, K=3 clusters with DBSCAN, and K=6 clusters with mean shift. Based on our analysis, mean shift emerged as the best algorithm, as indicated by a higher Silhouette score and a lower Davies-Bouldin index. These variations may be attributed to the diverse nature of the datasets and the unique characteristics of the algorithms applied.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1. Conclusion

These days, customer segmentation is essential for the business to be able to adapt to customer needs, innovate, and develop target specific offerings given the widespread use of mobile money services. Peoples here and there uses mobile money to buy goods and receive money for selling goods. Mobile money products or technologies have emerged mainly to create cashless societies and several banks are working by targeting those products, to expand the use of mobile money among their customers including Commercial Bank of Ethiopia.

CBEBirr is one of the mobile money services provided by the Commercial Bank of Ethiopia. Knowing the current customer behaviour and information is essential in expanding and advertising CBEBirr to gain its function of creating a cashless society. Machine learning techniques are an important approach to it. The machine learning techniques for CBEBirr customer segmentation based on demographic and behavioural data of Commercial Bank of Ethiopia's CBEBirr customers is the central theme of this entire master's thesis. With the aid of machine learning techniques, the study aimed to identify customer categories and pinpoint the crucial information regarding CBEBirr customers of the Commercial Bank of Ethiopia.

In our study, we collected CBEBirr customer data from the Commercial Bank of Ethiopia for experimentation. For this experiment, we collected 250,000 CBEBirr data records from the Commercial Bank of Ethiopia and we made it clean, pre-processed, and reduced to 170,012 data records to fit into our model.

Finally, to verify the objective, the CBEBirr customer segmentation model is done using an unsupervised machine learning approach, K-means clustering, agglomerative clustering, DBSCAN, and mean shift algorithm.

using data gathered from the Commercial Bank of Ethiopia. Among the four clustering algorithms, evaluation metrics, the Silhouette coefficient, and the Davies-Bouldin index were conducted to demonstrate the effectiveness of the K-means algorithm. We get silhouette score 0.792, 0.676, -0.129 and 0.792, and Davies-Bouldin Score 0.291,0.290,1.306, and 0.290 for K-means clustering, agglomerative clustering, DBSCAN, and mean shift algorithm respectively. Hence, this concludes that considering those evaluation metrics among the four algorithms we used to cluster our CBEBirr customer's data, the mean shift algorithm is better than among the three algorithms as displayed by the higher value of silhouette score and lower value of Davies Bouldin Score.

5.2. Contributions of the Study

As stated in chapter one, the major goal of our study was to examine the CBEBirr customers data from Commercial Bank of Ethiopia and create homogeneous customer groups using a mix of behavioural and demographic data to have an understanding of essential information about Commercial Bank of Ethiopia's CBEBirr customers. This has been accomplished by both clustering methods, except for a small difference in the number of clusters, where mean shift divides the dataset into four clusters, agglomerative clustering divides the dataset into four, DBSCAN divides the dataset into four clusters, and K-means clustered it into five clusters.

The study represents a significant contribution to the field, as we successfully prepared a comprehensive dataset for CBEBirr customer clustering and developed novel machine learning models to enable the clustering of CBEBirr customers for the first time. This advancement in CBEBirr customer clustering has the potential to greatly impact the Commercial Bank of Ethiopia and provide valuable insights for businesses to better understand and cater to their CBEBirr customer base. The development of these machine learning models represents a significant step forward in CBEBirr customer segmentation and targeting, offering valuable opportunities for improved marketing strategies and CBEBirr customer engagement.

Generally, through this analysis obtained from our study, one can dive into the essential information about CBEBirr customers of the Commercial Bank of Ethiopia. Through examining the features in each cluster, it is easily possible to recommend any idea related to CBEBirr services and marketing strategy to optimize the use of CBEBirr services.

5.3. Recommendations and Future Direction

The primary purpose of this study was to develop a model that can efficiently segment or cluster CBEBirr customers of Commercial Bank of Ethiopia using an unsupervised machine learning approach specifically using K-means clustering, agglomerative clustering, DBSCAN, and mean shift algorithm. Even if the study's result is capable of segmenting or clustering Commercial Bank of Ethiopia's CBEBirr customers, it has a limitation. Thus, the following recommendation is forwarded.

- It is best if demographic and behavioural data that are not considered in this study will be considered for future work.
- Future work on segmenting CBEBirr customers using geographic and psychographic data or features in addition to demographic and behavioural data which were used in this study. Adding those features, and geographic and psychographic data will enable obtaining a 360-degree view of the consumer.
- Further work experimentations on CBEBirr customer data with extended dimensionality and dataset using another machine learning algorithm.
- Additional future work considerations on performing churn prediction using a machine learning approach to identify or predict which segment has a high or low churn probability.

REFERENCES

[1] Gosavi, "'Can Mobile Money Help Firms Mitigate the Problem of Access to Finance in Eastern subSaharan Africa ? Can Mobile Money Help Firms Mitigate the Problem of Access to Finance in Eastern sub-Saharan. Journal of African Business.," vol. 19, no. 3, p. 343–360, 2018.

[2] Omigie, et al., "Customer pre-adoption choice behavior for M-PESA mobile financial services: extending the theory of consumption values.," vol. 117 Industrial Management & Data Systems, no. 5, p. 910–926, 2017.

[3] Kulondwa & Lukogo, "Financial inclusion in Africa through Mobile Money Services: A Swot Analysis of Mobile Money Services: Evidence from Bukavu," *International Journal of Commerce and Finance,* vol. 5, no. 2, pp. 22-30, 2019.

[4] Bhasin, M. L, "Data Mining: A Competitive Tool in the Banking and Retail Industries," *The Chartered Accountant,* pp. 588-594, 2006.

[5] Bhambri, "Implementation of Data Mining Techniques for Strategic CRM Issues.," *International journal of Computer. Technology and Application,* pp. 879-883, 2011.

[6] Choudhry, M., "An Introduction to Banking Principles, Strategy and Risk Management.," *United Kingdom: John Wiley & Sons, Ltd and Moorad Choudhry.,* 2011.

[7] Semonegna.com., "Semonegna.com.," 21 March 2018. [Online]. Available: https://semonegna.com/cbe-launches-mobile-money-platform-cbe-birr/. [Accessed 12 June 2023].

[8] Mark K.Y. M. et al., "A Financial Data Mining Model for Extracting Customer Behavior.," *The Hong Kong Polytechnic University, China.,* 2011.

[9] Shukla, D., "Mobile Money: The Age of Digital Payments.," 2018. [Online]. Available: www.electronicsforu.com: www.electronicsforu.com/technology-trends/tech-focus/mobile-money-digital-payments. [Accessed 15 May 2023].

[10] Tripathi, S. et al., "Approaches to Clustering in Customer Segmentation.," *International Journal of Engineering & Technology,* pp. 802-807, 2018.

[11] Sinja., "Predictive Modelling and Segmentation for Market Sizing and Product Design.," 2019.

[12] Hassan., "Customer profiling and segmentation in retail banks using data mining techniques.," 2020.

[13] Eshetu S., "Assessing opportunities and challenges of operating CBE-Birr mobile money.," 2020. [Online]. Available: https://www.coursehero.com/file/72162218/Hana-Tilahunpdf/. [Accessed 15 May 2023].

[14] Moges G., "Assessment of Practices and Operational Barriers of Mobile money Service in the Commercial Bank of Ethiopia: The Case of CBEBirr in East Addis District.," *Addis Ababa: AAU.,* 2018.

[15] Hana T., “The Challenges and Prospects of E-Banking System in Commercial Bank of Ethiopia: A Case Study of CBE-Birr Agent Banking.,” 2020.

[16] Pennsylvania State University, “PennState Eberly Collage of Science Online courses,” 7 August 2019. [Online]. Available: https://onlinecourses.science.psu.edu/stat505. [Accessed 19 August 2023].

[17] NBE., “National Bank of Ethiopia.,” [Online]. Available: https://nbebank.com/banks/. [Accessed 13 May 2023].

[18] Cepheus., “Ethiopia’s Banking Sector. Cepheus Research & Analytic.,” 2020.

[19] CBE, “Commercial Bank Of Ethiopia,” [Online]. Available: https://combanketh.et/commercial-bank-of-ethiopia/about-cbe/. [Accessed 13 May 2023].

[20] Commercial Bank of Ethiopia, “Annual Report of 2022/2023 Fiscal Year,” CBE, Addis Ababa, Ethiopia, June 30, 2023.

[21] Kolter, et al., “Optimove Solution,” Optimove Solution, 2009.. [Online]. Available: https://www.optimove.com.. [Accessed 13 May 2023].

[22] Fullerton, R. A., “Segmentation Strategies and Practices in the 19Th-Century German Book Trade:,” in *a Case Study in the Development of a Major Marketing Technique. Association for Consumer Research*, 1985, pp. 135-139.

[23] Camilleri el al., “Market Segmentation, Targeting and Positioning In Travel Marketing, Tourism Economics and the Airline Product.,” Department of Corporate Communication, Faculty of Media and Knowledge Sciences, University of Malta, Msida, 2018, pp. 69-83.

[24] Moin, K. I., “Use of Data Mining in Banking.,” *International Journal of Engineering Research and,* pp. 738-742, 2012.

[25] Manish Gupta, “Customer Segmentation,” [Online]. Available: https://manishgupta-ind.github.io/retail_pgp.html. [Accessed 20 August 2023].

[26] Anon., 2019. [Online]. Available: https://developers.google.com/machine-learning/clustering/clustering-algorithms.. [Accessed 23 May 2023].

[27] Swarndeep S., J., & Sharnil P., “Implementation of Extended K-Medoids Algorithm to Increase Efficiency and Scalability using Large Datasets.,” *Int J Comput Appl,* 2016.

[28] Tushar, et al., “Customer segmentation using K-means clustering.,” *2018 international conference on computational techniques, electronics and mechanical systems (CTEMS),* 2018.

[29] “Scikit-learn developers.,” 2007-2022. [Online]. Available: https://scikit-learn.org/stable/modules/clustering.html#mean-shift. [Accessed 16 May 2023].

[30] RapidMiner Documentation, “RapidMiner. Retrieved from rapidminer.,” 2021.

[31] Lu, Yilang, “Application of Clustering Methods to Trading Strategies in the US Equity Market.,” 2018.

[32] Kakushadze, Z., and Willie Yu, "Statistical industry classification.," *arXiv preprint arXiv:1607.04883,* 2016.

[33] Ryan P. Adams, "Hierarchical clustering.," *COS 324 – Elements of Machine Learning,* 2015.

[34] Chauhan., "DBSCAN Clustering Algorithm in Machine Learning," kdnuggets, 4 April 2022. [Online]. Available: https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html. [Accessed 03 June 2023].

[35] Monil et al., "Customer segmentation using machine learning.," *International Journal for Research in Applied Science and Engineering Technology (IJRASET),* vol. 8, no. 6, pp. 2104-2108, 2020.

[36] Palacio-Niño et al., "Evaluation Metrics for Unsupervised Learning Algorithms.," *arXiv preprint arXiv:1905.05667,* 2019.

[37] Wang, Xu, and Yusheng Xu., "An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index," *IOP Conference Series: Materials Science and Engineering.,* vol. 569, no. 5, p. 052024, 2019.

[38] PyShark, "Davies-Bouldin Index for K-Means Clustering Evaluation in Python," 2 June 2021. [Online]. Available: https://python-bloggers.com/2021/06/davies-bouldin-index-for-k-means-clustering-evaluation-in-python/.

[39] Asha Panyako. M., "Customer Segmentation on Mobile money users in Kenya," 2020.

[40] I.Smeureanu, et al., "Customer Segmentation in Private Banking Sector Using Machine Learning Techniques.," *Journal of Business Economics and Management,* vol. 14, no. 5, pp. 923-939, 2013.

[41] Eric Umuhoza, et al., "Using Unsupervised Machine Learning Techniques for Behavioral-based Credit Card Users Segmentation in Africa.," *SAIEE Africa Research Journal,* vol. 111, no. 3, pp. 95-101, 2020.

[42] W. Gao, et al., "Customer segmentation model based on two-step optimization in big data era.," *In 4th International Conference on Information Technology and Management Innovation,* pp. 800-803, 2015.

[43] R. Baradaran, et al., "Profiling bank customers behaviour using cluster analysis for profitability. International Conference on Industrial Engineering and Operations Management (pp. 458-467). Kuala Lumpur:," in *Kuala Lumpur: International Conference on Industrial Engineering and Operations Management.,* 2021.

[44] V. Mihovaa, et al., "A Customer Segmentation Approach in Commercial Banks.," *American Institute of Physics,* pp. 1-8, 2018.

[45] Henock W., "Application of data mining techniques to support customer relationship management at ethiopian airlines.," *Addis Ababa, Ethiopia: AAU.,* 2002.

[46] Sewagegn T., "Mining customs data for customer segmentation: the case of Ethiopian Revenues and Customs Authority.," *Addis Ababa: AAU.,* 2018.

[47] N.Kannaiya Raja, et al., “Data Mining Model to Analyse Crm in Banking Sector.,” *PalArch's Journal of Archaeology of Egypt/Egyptology*, vol. 17, no. 7, pp. 7170-7181, 2020.

[48] Gutema D. M, “ Application of data mining techniques for customer segmentation in insurance: siness: the case of Ethiopian Insurance Corporation,” *Addis Ababa: AAU,* 2016.

[49] Yeshitla T., “Assessing opportunities and challenges of CBE–Birr mobile money service: case study on Commercial Bank of Ethiopia.,” 2019.

[50] Kassahun G., “Application of Data Mining Techniques to Predict Customers churn At Commercial Bank Of Ethiopia.,” *Addis Ababa,Ethiopia: AAU.,* 2013.

[51] destingong.medium.com., “destingong.medium.com.,” 24 December 2022. [Online]. Available: https://destingong.medium.com/list/eda-and-feature-engineering-techniques-e0696974ed54. [Accessed 1 July 2023].

[52] Aggarwal, D. et al., “An insight into machine learning techniques for predictive analysis and feature selection.,” *International Journal of Innovative Technology and Exploring Engineering,* vol. 8, no. 9, pp. 342-349, 2019.

[53] Margel & Shtar , “Clustering and Dimensionality Reduction: Understanding the “Magic” Behind Machine Learning.,” 31 July 2017. [Online]. Available: https://www.imperva.com/blog/clustering-and-dimensionality-reduction-understanding-the-magic-behind-machine-learning/. [Accessed 09 July 2023].

[54] Kamande, S. W. et al, “Consumer Segmentation and Profiling using Demographic Data and Spending Habits Obtained through Daily Mobile Conversations,” 2018.

[55] Yinebeb K., “Application of Data Mining Techniques for Customer segmentation in Commercian Bank of Ethiopia,” Addis Ababa University, Addis Ababa, Ethiopia, 2021.