Machine Learning-Based Risk-Adjusted Capitation: An Efficient Payment Model for Kenya's Primary Healthcare at The NHIF

By

DAWE DAVID GAMBO

149796

Submitted in Partial Fulfillment of the Requirements For the Degree of Master of Science in Data Science and Analytics at Strathmore University

Institute of Mathematical Sciences and iLab Africa

Strathmore University

Nairobi, Kenya

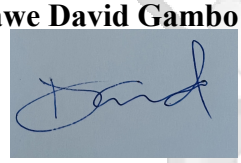
June, 2025

Declaration and Approval

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the dissertation contains no material previously published or written by another person except where due reference is made in the dissertation itself.

Student Name: **Dawe David Gambo**

Student Signature:  Date: 19/05/2025

Approval

The dissertation of **Dawe David Gambo** was reviewed and approved for examination by the following:

Dr. John Olukuru,
Lecturer, Strathmore University,

Dr. Godfrey Madigu,
Dean, Institute of Mathematical Sciences,
Strathmore University

Prof. Bernard Shibwabo
Director of Graduate Studies,
Strathmore University

Abstract

In Kenya, the National Health Insurance Fund (NHIF) aims to provide quality and affordable care to its members. To achieve this, efficient provider payment mechanisms are essential. Capitation—a prepayment model where providers receive fixed monthly payments per enrolled member regardless of actual services delivered (for example, KSh 1,000 per member per month for primary care services)—incentivizes cost containment and efficiency. However, when all providers receive identical payments regardless of their patient population's health risks, those serving sicker populations face financial disadvantages.

This research explores the application of K-means clustering to risk-adjusted capitation for more equitable provider payments. Unlike traditional risk adjustment methods that rely on predetermined categories, K-means clustering algorithmically identifies natural groupings of patients with similar characteristics. For instance, the algorithm can differentiate between low-cost healthy young adults and high-cost elderly patients with chronic conditions, allowing payment rates to be calibrated accordingly.

Our analysis identified two optimal patient clusters with average claim amounts of 4,848 and 4,930, both significantly lower than the dataset's overall average claim of 5,917, indicating meaningful population segmentation. These clusters, verified through Principal Component Analysis (PCA), demonstrated homogeneity in cost patterns and demographic factors. By integrating this approach, NHIF could replace its current uniform capitation rates with differentiated payments—higher rates for providers serving predominantly high-risk populations and standard rates for those with healthier enrollees.

The findings demonstrate how advanced analytics can transform healthcare financing by enabling more precise resource allocation, reducing provider incentives to avoid high-risk patients, and promoting proactive preventive care aimed at keeping patients healthier. This data-driven approach offers NHIF a practical pathway toward more sustainable, equitable healthcare financing that aligns provider incentives with improved population health outcomes.

Keyword: NHIF, K-means, PCA,Davies-Bouldin Index (DBI), Calinski-Harabasz Index, Risk-Adjusted Capitation

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
CBHI	Community-Based Health Insurance
CHI	The Calinski-Harabasz Index
CMG	Case Mix Group
CRISP-DM	Cross-Industry Standard Process for Data Mining
DBI	The Davies-Bouldin Index
FFS	Fee For Service
ICD	International Classification of Diseases
ICU	Intensive Care Unit
ILO	International Labour Organisation
LMIC	Lower and Medium Income Country
MAR	Missing At Random
MCAR	Missing Completely At Random
ML	Machine Learning
MNAR	Missing Not At Random
NCD	Non-Communicable Disease
NHIF	National Health Insurance Fund
OLS	Ordinary Least Squares
PAM	Partitioning Around Medoids
PCA	Principal Component Analysis
PPM	Provider Payment Mechanism
SDG	Sustainable Development Goal
TP	True Positive
UHC	Universal Healthcare
UN	United Nations
WCSS	Within-Cluster Sum of Squares
WHO	World Health Organization

Acknowledgements

I wish to express my sincere gratitude to Dr. John Olukuru for his invaluable guidance and support throughout my research journey. I also wish to thank my friends, colleagues, and class-mates for their encouragement and camaraderie, which greatly impacted my work. Your support has made this journey meaningful and I hope to one day offer the same to others.

Dedication

First, I give thanks to the Almighty God. Without His mercy and faithfulness, none of this would be possible. I dedicate this work to my family, especially to my parents and siblings, for their unwavering support throughout my academic journey. My heartfelt appreciation goes to my friends and colleagues, as well as to the entire Strathmore University community. Special recognition goes to Margaret Macharia, Lilian Mumbi, Mercy Mutua, and my boss, Dr. Samson Kuhora, of the Benefits Design and Actuarial Services Team for their invaluable guidance and support during my research. I hope that this work inspires my children to strive for academic excellence.

Chapter 1: Introduction

1.1 Background

Healthcare is one of the largest elements of the global economy (Yang et al., 2018). The World Bank reports show that healthcare expenditures account for up to 10% of the world's total Gross Domestic Product (GDP). In the last decade, there has been a steady increase in per capita health expenditure. The Sustainable Development Goals (SDGs), especially SDG3, prioritize health and well-being (Obadha et al., 2019b). It emphasizes the need for universal health coverage, where everyone, regardless of income or social status, may receive quality, rehabilitative, curative and promotional treatments without financial hardship (Presch et al., 2020). Despite this, three billion people lack access to health services, 800 million more face high prices, and 100 million are forced to poverty due to healthcare costs (Obadha et al., 2019b; Stadhouders et al., 2019). Kenya, like other SSA nations, bears this burden (Obadha et al., 2019b). Healthcare financing is essential for Universal Healthcare (UHC) and SDG3 (Yang et al., 2018). Many low and middle-income countries (LMICs) such as Kenya are reforming their health financing to conform to the UHC.

Provider payment mechanisms (PPMs) are useful tools for improving healthcare care financing and encouraging efficient and effective use of resources. Many PPMs have emerged, each with different effects on controlling healthcare costs. The most common are fee-for-service, capitation, and case-based (Obadha et al., 2020). Capitation is the most popular approach, will involves paying a healthcare facility in advance to serve participants (She et al., 2022). Capitation seeks efficiency and cost control. Capitation is part of provider payment reforms to address health system issues (CON, 2022). Therefore, there is a shift from fee-for-service (FFS) to mechanisms where risk is shared with providers through integration among different providers (Geruso and McGuire, 2016). In risk-adjusted capitation, there is more equitable distribution of expenditures to providers than the FFS (Jia and Zhang, 2021).

As noted by Layton et al. (2018), risk selection in capitation causes inefficiency because some groups, such as high-risk and low earners might be excluded. Consequently, risk selection might also result in high risky groups are not considered (CON, 2022). These groups might also receive poor services from the healthcare facilities, which excludes most people from getting their required benefits (Layton et al., 2018). Risk adjusted capitation mode of payment involves

making fixed payments for the enrollees and adjusting the amount further considering the higher costs of providing care based on the enrollee's health status (Jiang et al., 2022). Risk-adjusted premiums should be set for enrollees with expected high healthcare experiences (Stadhouders et al., 2019). Proper Risk adjusted capitation mitigates risks, especially for enrollees with high risk (Layton et al., 2018; Khezri et al., 2022).

Table 1.1 highlights some of the international experiences in the application of risk-adjusted capitation. Most health systems in high-income nations adopt risk-adjusted capitation. Early adopters of risk-adjusted capitation payments were countries with competitive health insurance markets. Examples include Belgium, Germany, the Netherlands, Israel, Switzerland, and the USA. Risk-adjustment programs aim to enhance insurance market efficiency by eliminating risk-selection incentives and minimizing volatility between capitation payments and insurance plan expenditures.

Table 1.1: Overview of Capitation Features and Risk-Adjustment in Selected Countries

Country	**Health System Approach**	**Capitation Features**	**Risk-Adjusted Factors**	**Data Type for Calculation**
Belgium	Compulsory health insurance with private and public sickness funds	Prospective; 30% subject to risk-adjustment	Socioeconomic, urbanisation, medical supply, health status, disability	Data from insurance companies
Germany	Statutory health insurance with competing insurers	Prospective; morbidity-based criteria for fund redistribution	Morbidity, gender, age, invalidity pension, sick pay	Pseudonymised administrative and health data
Netherlands	Statutory health insurance with private insurance mandate	Approx. 50% income from risk-adjusted funds	Region, income, age, sex, cost groups, socioeconomic status	Pseudonymised data from insurers and government
Israel	National health insurance system with non-profit health plans	Prospective; 94% risk-adjusted budget	Age, sex, residence	Survey data
Switzerland	Mandatory health insurance with private insurers	Canton-specific risk-adjustment	Age, sex, region, hospitalization, high pharmaceuticals expenditure	Insurance company data
USA: Medicare	Medicare for elderly, disabled, or specific conditions	Hierarchical Conditions method for risk adjustment	Age, disability, health conditions	Provider claims data
Kenya	National Health Insurance Fund (NHIF)	Outpatient services via capitation, inpatient through fee-for-service	Age, gender, geographical location, chronic conditions	NHIF enrolment and claims data
Rwanda	Community-Based Health Insurance (CBHI)	Capitation for primary care, performance-based for hospitals	Demographic factors, region, type of care required	Health facility reporting, CBHI enrolment data
Ghana	National Health Insurance Scheme (NHIS)	Capitation for primary outpatient services	Age, sex, disease burden, region	NHIS claims and enrolment data

1.2 Capitation Theory

Capitation is defined as a method of reimbursing healthcare providers with a fixed fee per period per patient enrolled, often differentiated based on characteristics like age or sex, regardless of

the quantity of services provided (The Dictionary of Health Economics, 2005). This payment approach allocates a predetermined amount of healthcare funds to individuals with specific attributes for a designated service period, subject to overall budget constraints. In essence, it assigns a monetary value to each enrollee, essentially putting a price on their head (Rice and Smith, 2001). In practical terms, capitation involves prepaying a fixed sum per person to a healthcare entity and securing the provision or facilitation of contracted healthcare services for the specified period. The recipient of the capitation payment, typically a healthcare provider, agrees to provide healthcare services to all insured individuals under the health plan, irrespective of the actual cost of services. This introduces a form of insurance for the provider, who receives a guaranteed premium for delivering services whose actual cost may differ from the collected per capita rate (Rice and Smith, 2001).

Capitation payments are prospective and serve to exert control over both the cost and volume of services. However, if the rates are too low, they might incentivize under-provision or reduced quality of care (Langenbrunner and Wiley, 2002). The International Labour Office (ILO) characterizes capitation as akin to a poll tax—a uniform direct tax levied on each person, typically on a per capita annual basis, payable to a specified health service provider (ILO, 2009).

Capitation can be either partial or full (total or global) in scope. Partial capitation entails applying prospectively determined per capita rates and budgets to specific services, often primary care, provided by a specific medical facility or network. Other services, typically secondary or tertiary care, are remunerated separately from the capitation fee (Telyukov, 2001; Tolley, Manton, & Vertrees, 1987). Conversely, full capitation encompasses an entire package of services agreed upon between a purchaser and a provider (Maceira, 1998). To ensure comprehensive coverage across primary, secondary, and tertiary care levels, many experts suggest organizing contracting providers into referral networks, which facilitates the adoption of full capitation (Telyukov, 2001; Tolley et al., 1987).

1.3 Current NHIF capitation provisions

The National Health Insurance Fund (NHIF) in Kenya uses fee-for-service (FFS) for participants seeking inpatient and outpatient care, and capitation for outpatient services (Barasa et al., 2018). Basic diagnostic testing, advice, medications on the Kenya Essential Drug list, and same-day treatments are a few of the outpatient services offered. For enrollment, NHIF se-

lects providers and compensates them quarterly through capitation. Capitalization may enable healthcare providers to outsource services they don't offer. Risk adjustment for payment models requires balancing efficiency incentives and predictive performance, according to CON (2022). Healthcare providers' choices should inform capitation plans (Obadha et al., 2019b). Capitation schemes must address risks to reduce losses (CON, 2022). To be efficient and effective, capitation programs should reflect healthcare provider preferences. In insurance, risk selection reduces losses (She et al., 2022). Risk-adjusted capitation is predicted to be crucial in providing primary and social care health services and allocating money to countries in Kenya's health funding reform (Obadha et al., 2020). A study done in Kenya compared FFS and capitation and reported that the latter is cheaper and more efficient (Ochieng, 2019).

The NHIF currently uses this PPM to purchase outpatient services for National Scheme beneficiaries and some enhanced scheme beneficiaries with the reimbursement rates shown in Table 1.2 below. Some risk is shifted from the Fund to the provider. If the provider incurs costs that are greater than the per capita budget, the provider is liable for them. The corollary is that if the provider achieves efficiency gains and incurs low costs.

Table 1.2: Capitation Rates

	Previous Contract			**Current Contracts**	
Scheme	**National**		**Enhanced**	**National**	**Enhanced-1**
HCP / Category	**GOK**	**Non-GOK**	**National (HKS)**	**All Comprehensive**	**All Comprehensive**
Level II	1,000	1,200	1,000	2,850	2,850
Level III	1,000	1,200	1,000	2,850	2,850
Level IV	1,000	1,200	1,400	2,850	2,850
Level V	1,000	1,200	1,400	2,850	2,850
Level VI	1,000	1,200	NA	NA	NA

The capitation payment model has been used for nine (9) years with observable successes. For instance, the model is cost-effective since the outcomes of economies of scale save money from the pool. In addition, the administrative costs for outpatient care are low and are usually estimated to be 3% compared to other payment models that use 10%. Besides, some providers have been incentivized to increase output or attract more patients to select the provider, through improved quality of care, additional services that are not typically covered, or other measures that patients perceive as increasing the benefit of enrolling with that provider rather than with another provider. Providers, incentivized by capitation, have shifted to cost-effective prevention and disease management, favoring generics over branded medications. According to Momahhed et al. (2023), handling all insurance data is often challenging, which necessitates the need

for grouping or clustering to make the data more manageable.

It is important to note that, under the current capitation models, resources are disproportionately consumed by a few people. Hence, there are calls for a higher efficient healthcare for patients. Fortunately, information technology, especially data science and Artificial Intelligence (AI) presents a new way of addressing the problem. According to Yang et al. (2018), there is an emerging field of using data modeling approaches to predict health outcomes. In this regard, if NHIF can forecast expenditures with great accuracy, it can improve targeted care. This study will, therefore, use a clustering approach for risk-adjusted capitation.

1.4 Problem Statement

Despite the proliferation of capitation contracts in providing healthcare services, the risk is a major concern due to the accrual to capitation contractees (Ochieng, 2019). Capitation might lead to under-provision of services due to controlling costs (Obadha et al., 2020). FFS, bundled payments, and risk-adjusted capitation are common healthcare provider payment schemes. Although it may result in unnecessary treatments, FFS pays providers on a per-service basis, guaranteeing service coverage. All treatment services are included in a single payment in bundled payments, which increase coordination and efficiency, but run the risk of under-provision. There are drawbacks to both models, such as the possibility of excessive use in FFS and the difficulty of precisely defining bundles. In addition, many current risk adjusted models are linear in design (Yang et al., 2018). Consequently, they suffer from various limitations that include limited predictive accuracy, limited information at the patient level, and focus on specific patient populations or types of care. Although capitation can promote provider efficiency, there is also a risk that inefficient hospitals can be rewarded if risk adjustment is not done well.

An ideal situation is a case in which healthcare finance balances provider incentives with patient requirements. The current capitation approach by NHIF faces obstacles. Insufficient provider involvement in capitation scheme design results in losses and possible under-provision of services as a result of cost controls (Obadha et al., 2019a; Ochieng, 2019). Furthermore, the lack of risk adjustment in capitation payments undermines the equitable distribution of resources by failing to account for the wide range of health requirements of the insured population (NHIF, 2019; Jiang et al., 2022). A lack of a well-defined risk-adjustment mechanism worsens the inherent problems of capitation, such as bias and low quality of care (JOHN C. LANGEN-

Table 1.3: Provider Utilization and Returns Analysis (2023)

Provider Category	Level	Utilization	Returns (KSh)	Total Paid (KSh)	Return Ratio
Private Hospital	4	High	12,934,437	37,376,050	35%
Government Hospital	4	Moderate	2,539,430	6,712,638	38%
Government Hospital	4	Very High	1,892,882	1,559,900	121%
Mission Hospital	4	Very High	1,801,655	1,506,388	120%
Summary by Provider Type:					
Private (n=11)	-	-	20,400,071	51,804,138	39.4%
Government (n=4)	-	-	5,314,198	9,790,425	54.3%
Mission (n=5)	-	-	3,096,416	3,119,963	99.2%

Note: "Return Ratio" represents the percentage of capitation payments returned to NHIF, with values >100% indicating providers paid more in claims than received in capitation.

BRUNNER, 2019; Smith, 2019). This study seeks to examine implementing a risk-adjustment capitation model for the NHIF. This will be done through the clustering of K-means to categorize patients according to risk and need.

1.5 Justification of the Study

This study is justified by Kenya's need for efficient and fair healthcare finance. This technique promises to improve resource distribution accuracy and cover the research gap, providing a data-driven solution to Kenya's shortcomings in the capitation program. In line with SDG3, the country is pursuing UHC. Although innovative, the capitation system has risk adjustment limits. As noted in a report by the NHIF (2019), the capitation rate should be adjusted according to the prevalence of a certain group of people to a certain disease. Notably, NHIF should also send providers an updated capitation to promote transparency, and most importantly, enhance predictable funding. Unfortunately, the report stated that the current capitation design does not consider the risk variation across various population groups.

Insights from this study will be vital in the implementation of an ML-based risk adjustment model to improve the current system. ML has the potential to improve the adjusting effect by unearthing associations in data that are hidden from linear models, expanding the restrictive model assumptions, and adding more detailed data. For this reason, there is a need to conduct this study to examine risk-adjusted capitation as an efficient provider payment mechanism using K-Means Clustering in machine learning. This research is important given Kenya's healthcare financing reforms and the Sustainable Development Goals' goal of Universal Health Coverage (UHC). A more refined risk-adjusted capitation model utilizing machine learning could improve

health outcomes, resource efficiency, and healthcare system and beneficiary finances. The success of this concept could inspire other low- and medium-income countries with healthcare system issues.

1.6 Objectives

1.6.1 General objectives

The main aim of the study is to evaluate how well k-means clustering can be applied in risk-adjusted capitation by the NHIF.

1.6.2 Specific Research Objectives

(a) To develop a risk-adjusted capitation model for NHIF using K-means clustering to identify distinct patient risk profiles based on healthcare utilization patterns, demographic factors, and clinical characteristics.

(b) To establish a quantitative framework that assigns appropriate capitation rates to each risk cluster, ensuring payments reflect the predicted healthcare costs of different member segments.

(c) To create an implementable decision support system for NHIF that enables data-driven provider payment allocations based on the composition of enrolled members' risk profiles.

1.7 Research Questions

The research will aim to answer the following questions:

(a) How can a model for risk-adjusted capitation by the NHIF be developed and validated to accurately predict healthcare expenditures?

(b) What framework can be established for segmenting NHIF members into distinct risk groups effectively?

(c) How can the developed model be deployed and implemented to support informed decision-making by the NHIF?

1.8 Scope of the Study

The study focuses specifically on the National Hospital Insurance Fund (NHIF) in Kenya. It confines its examination to the operations and data managed by the NHIF. Temporally, the study is set to span six months, from March 2024 to September 2024. This includes the period for data collection, analysis, and evaluation. The study specifically examines the application of risk-adjusted capitation. It employs K-Means Clustering and machine learning techniques.

1.9 Significance of the Study

The findings of this study will be valuable to practitioners, researchers, policymakers, the NHIF, and the members of NHIF. This study's findings will guide NHIF in developing risk-adjusted capitation as an efficient provider payment mechanism. The study will further assist NHIF board management in evaluating the merits and demerits of the risk-adjusted capitation payment method. Afterward, decision-makers would suggest improvements to the payment system to promote efficient services. Therefore, the study promises to provide insights that will improve services, ensuring that NHIF gets value for money. The registered members of the NHIF will also benefit from the study because risk adjustments will ensure that those at higher risk are provided with aligned capitation.

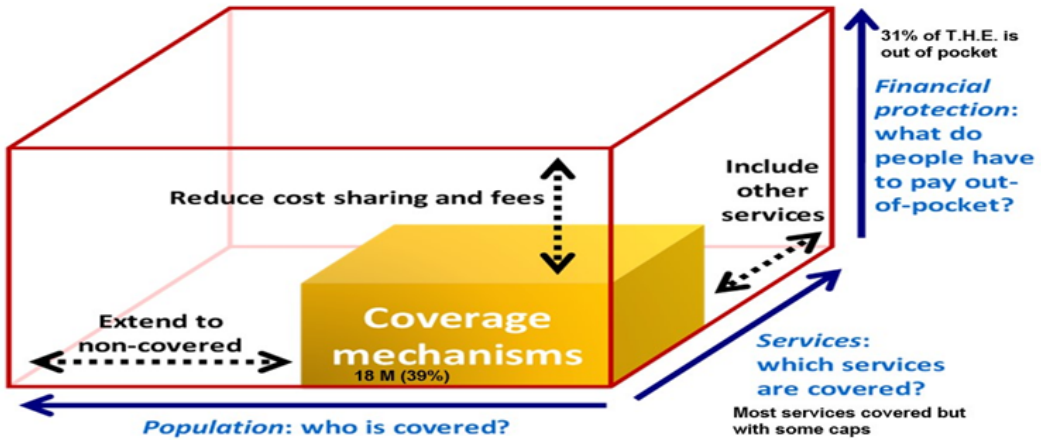


Figure 1.1: Schematic of UHC

Besides, the study's results will be instrumental to researchers researching on risk-adjusted capitation in the healthcare industry. This study's results will inform their research because they will also use literature as a study point. Researchers might also identify research gaps

in this study and decide to conduct research to fill those research gaps. Researchers will also understand in-depth the importance of capitation payment in healthcare delivery.

This study is motivated by the attainment of the UHC. Universal health coverage (UHC) means that all people and communities can use the promotive, preventive, curative, rehabilitative, and palliative health services they need, of sufficient quality to be effective, while also ensuring that the use of these services does not expose the user to financial hardship. To achieve UHC, the population covered, the scope of services by depth, and the cost of delivery of care have to be considered as illustrated in 1.1.

Chapter 2: Literature Review

2.1 Introduction

The literature review will discuss the theoretical framework, empirical literature, and the summary of the literature review. The literature review will present past studies related to risk-adjusted capitation.

2.2 Theoretical Framework

This section will review two theories related to this study's objective. The theories include the risk theory of capitation and the capitation theory. The section will further explain why these theories are critical in the present study.

2.2.1 Risk Theory of Capitation

The risk theory of capitation suggests that risk disaggregation results in providers becoming inefficient insurers. The theory further claims that providers that assume risk experience low profits due to the exposure to increased operational losses and reduced patient benefits (Cox, 2001). The risk theoretic analysis of capitation further claims that capitation contracts are related to reinsurance contracts between insurance companies (Rice and Smith, 2001). As such, the difference between these contracts are reviewed, and implications discussed. The theory also views capitation contracts as reinsurance contracts, as they are incompatible with the general purpose of financing for health services (Mbau, 2021). The risk theory of capitation is relevant and applicable in the present study because it suggests that capitation increases operational losses and reduced patient benefits, and capitation contracts could be viewed as reinsurance contracts. Therefore, the researcher will use this viewpoint to understand and explain risk-adjusted capitation as an efficient provider payment mechanism.

2.3 Empirical Literature

2.3.1 Risk-Adjusted Capitation

Risk-adjusted capitation is a payment mechanism used to distribute funds among healthcare providers based on the risk profiles of their patient populations (CON, 2022). Risk adjustment in healthcare policy design aims to decrease annual healthcare expenses by factoring in variables that are beyond the control of healthcare providers (Khezri et al., 2022). Currently, in

Kenya's healthcare system, the Ministry of Health utilizes a risk-adjustment formula that can only anticipate 30% of total health expenditures in the upper quintile of the expense distribution (Riascos et al., 2017). Based on Ministry of Health statistics, this study proposes a new risk-adjustment algorithm including additional parameters. This revised methodology incorporates adjusted indicators of 29 long-term disorders, hospitalizations, specialist consults, and ICU admissions (Riascos et al., 2017).

Layton et al. (2018) found that risk-adjusted capitation in healthcare payment systems supports high-quality, cost-effective care. Healthcare providers take on financial risk from insurance. This payment method improves healthcare efficiency, transparency, and justice. Riascos et al. (2017) tested risk-adjustment formulas by forecasting annual health expenses with machine learning. Non-parametric approaches like boosted trees outperformed linear regressions with fewer regressors. Riascos et al. (2017) found that the Gradient boosted trees models random forest (GBM FS) model fitted on a smaller set of variables predicted 50% of total health expenditures, 5 percentage points better than the government. Risk adjustment reduces annual health spending unpredictability by controlling characteristics insurers cannot control. Despite benefits, risk-adjusted capitation may drive risk selection. Healthcare providers may forgo pricey treatments or favor cheaper ones to maximize revenue. This analysis found that the boosted trees model fitted across all covariates exaggerated 70% of enrollees' annual health costs while accounting for 73% of system revenues (Riascos et al., 2017). So, risk-adjustment methods should account for risk selection incentives and encourage doctors to provide high-quality care to all patients.

Healthcare providers are reimbursed based on the risk of treating a patient population, according to Anell et al. (2018). Risk-adjusted capitation assumes healthcare providers follow economic incentives. Risk-adjusted capitation is more equitable since it considers regional health needs. Healthcare practitioners are paid to prioritize high-need patients. Risk-adjusted capitation improves health outcomes, decreases financial risk, and helps doctors handle difficult patients. Risk-adjusted capitation may boost healthcare efficiency and equality.

Ochieng (2019) noted that financial incentives should match population care demands and reduce socioeconomic gaps in healthcare use. Risk-adjusted capitation is based on the fact that fee-for-service produces more services but offers lower control costs than capitation. Capitation encourages providers to choose "good" risks, which can lead to over-provision of services

to healthy people and under-provision to unwell people. In order to eliminate socioeconomic gaps in healthcare utilization, risk-adjusted capitation is considered a more equitable payment method. Kenya uses the NHIF to adjust capitation payments to primary care clinics depending on their enrolled population's socioeconomic and demographic characteristics (Obadha et al., 2019b). This change promotes primary care centers in low-income communities to increase healthcare access.

Risk-adjusted capitation is crucial to healthcare payment systems, according to She et al. (2022). Capitation payment stabilizes healthcare practitioners' income but puts them at risk if they treat pricey patients. Risk-adjusted capitation changes payments based on patient health. This prevents clinicians from being unfairly penalized for treating difficult patients. Besides, risk-adjusted capitation supports high-quality care. Health providers are driven to invest in preventive care and serve patients at the lowest cost while preserving quality. Capitation payment may also drive illness management innovation and healthcare system integration, according to **?**.

However, capitation payment may have detrimental implications. One major concern is that professionals would be encouraged to ignore high-risk patients to limit financial risk. Complex medical patients may have reduced access to care. Adu-Gyamfi (2012) also warned that capitation payment schemes could be underfunded, leaving healthcare providers short. Risk-adjusted capitation is a fundamental concept in healthcare payment systems because it promotes high-quality treatment and cost management. Governments and healthcare providers must examine the pros and cons of capitation payment schemes before implementing them.

2.3.2 Risk-Adjusted Capitation and Quality of Healthcare Services

Risk-adjusted capitation reimburses doctors depending on patient risk. This strategy encourages healthcare providers to deliver high-quality care to all patients regardless of medical history. She et al. (2022) say capitation payment gives hospitals predictability and stability, making service modifications easier. The study will determine if capitation improves accredited hospital performance. She et al. (2022) further noted that that capitation can improve healthcare delivery by driving illness management innovation and health system integration. Diagnostic tests, medicines, and treatments vary greatly among doctors. Financial incentives can reduce underutilization and overutilization, improving healthcare outcomes. The study also indicated that government bureaucracies and budget constraints inhibit capitated healthcare. To address

these difficulties, the government should fund the Ministry of Health and NHIF to embrace capitation-based health services. The paper recommends eliminating government bureaucracies and interferences that hinder capitated healthcare payment.

Risk-adjusted capitation affected Kenyan NHIF users' healthcare, according to Ndung'u and colleagues (2020). They discovered risk-adjusted capitation increased healthcare quality and coordination. The study observed increased primary care use and preventative care services including cancer screenings and immunizations. Dohmen et al. (2022) examined Kenya's health finance system's risk-adjusted capitation. The payment model's ability to improve patient care was restricted by money, healthcare personnel' skills, and data infrastructure. Despite these obstacles, experts and decision-makers believe risk-adjusted capitation may improve Kenyan NHIF customers' preventative and care coordination services. Kenya's best practices for using this payment system and its long-term implications on healthcare outcomes need more research.

Obadha et al. (2019b) examined how public, private, and faith-based healthcare providers handled capitation and fee-for-service payments from the National Hospital Insurance Fund (NHIF) and private insurers. The study found that capitation and fee-for-service payments were good revenue sources despite unpredictable finances and low rates. The study also indicated that providers appreciated provider payment mechanisms (PPMs) with predictable amounts, timely payment patterns, reasonable capitation rates, autonomy, and simple reporting. Delays in reimbursements and poor capitation payment rates were negative PPM experiences in Ghana (Sieverding et al., 2019). Positive examples included PPMs increasing providers' income (Sieverding et al., 2019; Aryeetey, 2012; Atinga, 2017).

Obadha et al. (2019b) recommended that the NHIF should set capitation rates. They also suggested expediting fund release, increasing capitation payment transparency, allowing public healthcare providers to open bank accounts and use NHIF monies, and simplifying claims and reporting. This study emphasizes the necessity to build efficient PPMs for universal health coverage and to account for healthcare providers' experiences and preferences. Future studies should examine risk-adjusted capitation as a PPM and use machine learning methods like k-means clustering for more accurate risk adjustment.

2.3.3 Risk-Adjusted Capitation and Transparency

It has been suggested that risk-adjusted capitation is an effective provider payment system that solves concerns of equity and openness in healthcare funding. The impact of risk-adjusted capitation on transparency has been the subject of numerous studies, with varying degrees of success. Research indicates that the implementation of risk-adjusted capitation can enhance transparency by fostering equity in the distribution of healthcare resources. Munge et al. (2018) contend that risk-adjusted capitation lessens the incentives for cherry-picking patients with lower risk profiles and guarantees health plans receive sufficient funding to meet the health needs of their enrollees, regardless of their health risk.

Risk-adjusted capitation may alleviate healthcare finance equality and openness issues by paying providers. Numerous research have examined how risk-adjusted capitation affects transparency, with mixed results. Research shows that risk-adjusted capitation improves transparency by distributing healthcare resources fairly. Risk-adjusted capitation reduces incentives for cherry-picking low-risk patients and ensures health plans receive enough funds to cover their participants' health requirements, according to Munge et al. (2018).

Risk-adjusted capitation may harm transparency, according to previous studies. Health plans serving sicker populations may receive less money under risk-adjusted capitation since higher-risk patients require more expensive medical treatment Obadha et al. (2019b). Differences in healthcare coverage and access for different patient groups may affect healthcare quality and efficacy. Despite these inconsistent outcomes, additional study is needed to understand how risk-adjusted capitation affects healthcare financial transparency. Studies are needed to identify risk-adjusted capitation's efficacy and develop strategies to maximize its transparency benefits.

2.3.4 Risk Analysis through Clustering algorithms

Clustering is an unsupervised machine-learning method for statistical data analysis in several fields Liao et al. (2016). Clustering algorithms group data points. Clustering group data points using a dataset. Data points in a cluster should be highly comparable but less similar to those in other clusters. Clustering analysis provides meaningful data insights by evaluating algorithm-discovered patterns Liao et al. (2016).

Research shows that clustering is important in data mining as it helps to identity groups and categorize data into patterns. In clustering, a dataset is divided into groups (clusters) so that each

are similar as compared to points in another cluster (Liao et al., 2016). For instance, classifying the present insurers into several groups and establishing a relationship between each group can be crucial for future pricing models, such as coinsurance rates. According to Momahhed et al. (2023), it is still challenging to identify which method generates the optimal and most suitable number of clusters for different data sets, despite the fact that numerous clustering algorithms have been developed to analyze data. Many studies have tested the clustering algorithms using various healthcare data sets and unique validation measures. Since every data set is unique in some way, only a small number of publications have assessed the performance and variability of three different clustering approaches using real and simulated data sets (Momahhed et al., 2023).

A lot of clustering algorithms were developed for the healthcare sector without being evaluated on different sets of important data (Liao et al., 2016). However, the relative significance of each technique was only determined through experimental research on a specific healthcare data set. The literature on clustering states that a number of clustering techniques, such as K-means, K-Medoids, or Partitioning Around Medoids (PAM), have been developed for the analysis of healthcare data sets (Liao et al., 2016). Since there is no perfect clustering technique, each has advantages and disadvantages of its own, so the best technique for clustering can be used in accordance with the objectives of the study and the types of variables involved (Momahhed et al., 2023).

2.3.5 Applying Machine Learning in Risk-Adjusted Capitation

Risk-adjustment models are used to predict accurately individual healthcare costs, in both provider reimbursement and healthcare plan payments (Iommi et al., 2022). Risk-adjustment models in health plan payments and provider reimbursements forecast individual healthcare expenses. Policymakers planning a risk-adjustment program should maximize statistical methodologies based on short-medium data availability, pending a longer-term investment to enhance variables and depth (Riascos et al., 2017; Iommi et al., 2022). Risk-adjustment models primarily consist of three key components: the selection of adjusters, the unit of analysis (which correlates with whether data is sourced from individual or group levels), and the structural design of the estimation model. Hsu et al. (2008) noted that risk adjusters are categorized into seven based on the data used for the prediction as shown in 2.1.

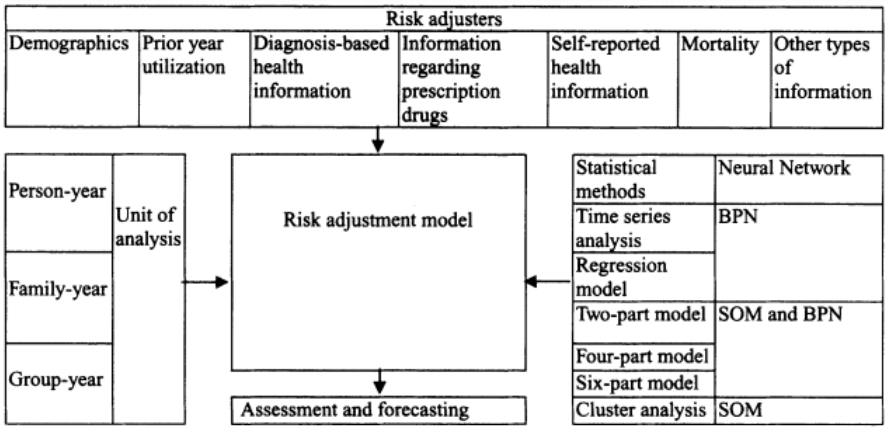


Figure 2.1: Alternative Risk adjustment models (adapted from Hsu et al. (2008))

As noted by Yang et al. (2018), most current approaches to predictive modeling for risk adjustment in capitation are linear. As such, they have some limitations, such as limited predictive accuracy, limited information at patient-level, and specific patient populations. Multi-part models often apply the two-part model, combining logit or probit models with OLS for positive medical expense observations. Alternatively, clustering analysis groups similar data for distinct estimations (Hsu et al., 2008). Novel strategies employ Artificial Intelligence (AI) and Machine Learning (ML) models to predict future health expenditure (Iommi et al., 2022).

Ramezani et al. (2023) noted that risk-adjusted capitation employing machine learning algorithms transforms healthcare resource allocation and funding. AI algorithms that anticipate data are called ML (Muremyi et al., 2020). They can evaluate massive volumes of data to estimate healthcare expenditures more accurately. In many instances, cluster analysis using K-means has been effective (Carta et al., 2021; Ramezani et al., 2023). This technique groups patients by healthcare needs and risk profile for advanced risk adjustment (Huang et al., 2021). The proposed study will investigate employing machine learning techniques (K-means clustering) for risk adjustment to improve the adjustment impact by substituting the NHIF's Kenyan model.

2.4 K-Means Clustering

One of the most basic unsupervised learning strategies for resolving the clustering problem is K Means (MacQueen, 1967). The process uses a preset number of k clusters that are predetermined to classify a given data set in an easy manner. Determining k centroids—one for each cluster—is the main notion. Because alternative initializations (centroid locations) can result in varied outcomes, these centroids should be strategically set. Placing them as far apart as feasi-

ble is therefore the wiser course of action (Grant et al., 2020). The algorithm iteratively assigns each data point to the closest centroid, representing the group's center. Once initial clustering completes, k centroids are recalculated as cluster barycenters (Pioch et al., 2023). This process repeats, adjusting centroids until stable, minimizing a squared error function. The algorithm involves the steps outlined in Figure 2.2:

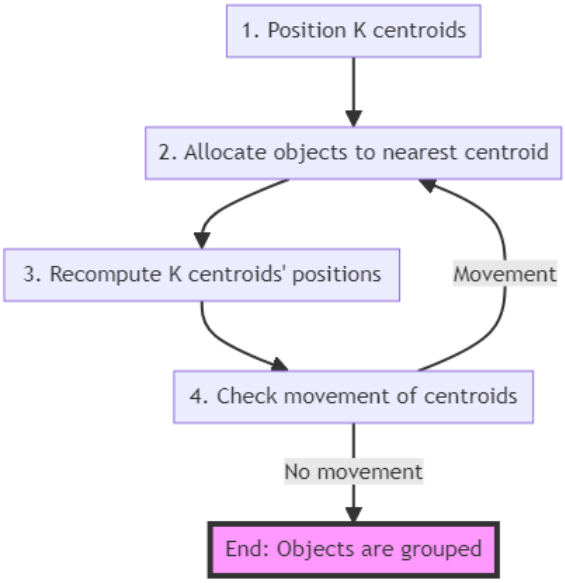


Figure 2.2: K-Means Clustering Algorithm Steps

Thus, K-means clustering groups data points by similarity. The approach determines each cluster's center, which is the average of its data points (Grant et al., 2020). To find K clusters, it assigns data points to the nearest centroid using a distance metric. K-means clustering can classify healthcare providers by risk. (Sharma and Vidhya, 2021) states that a risk adjustment model incorporates patient demographics, health status, and utilization patterns to produce these ratings. K-means clustering can enable risk-adjusted capitation payment models by clustering healthcare providers with similar risk scores. Each provider receives a predetermined sum per patient based on their risk score, not their services (Grant et al., 2020).

K-means clustering methods are noted for their high computing efficiency and low temporal complexity for grouping large, unlabeled datasets (Momahhed et al., 2023). In high-dimensional data, some traits may be insignificant while others may affect grouping. Different cluster variables may have contributed differently to its development. Consider each variable's importance to each cluster for improved clustering. The optimum algorithm may be K-means with customizable weighting (Grant et al., 2020).

Unsupervised machine learning dataset clustering using K-means is common. Initial cluster centres are randomly selected from k centroids (Hajat et al., 2021). It iteratively assigns each observation to the nearest centroid and updates its position. Repeat until centroids don't change or a set number of iterations. Improved healthcare payment system decision-making, cost savings, and treatment with K-means clustering. Bretos-Azcona et al. (2020) discovered K-means clustering can estimate healthcare expenses. It evaluated whether a patient would spend greatly on medical care.

K-means clustering can help the NHIF classify healthcare providers by risk for risk-adjusted capitation payments (Gesicho et al., 2021). Risk adjustment models use patient demographics, health, and utilization to rate risk. After finding clusters of providers with identical risk ratings, capitation payments are based on each cluster's patients' standard cost of treatment. Healthcare payment systems are complex and demand excellent results from outdated approaches, according to Riascos et al. (2017). Healthcare firms may employ data-driven models to identify trends and forecast outcomes with machine learning. Machine learning algorithm K-means clusters data. NHIF Kenya will be studied using K-means clustering in healthcare payment systems.

Hien et al. (2020) discovered that K-means clustering can detect and enhance high-cost patient treatment. The researchers used K-means clustering to classify patients as high- or low-cost. The method correctly detected hospitalization, long-term care, and chronic illness. This data helps healthcare firms manage resources and treat patients individually. Risk-adjusted capitation payments benefit from K-means clustering. Predetermined rates simplify payment and reduce administrative costs because they determine payment rather than services done. Adjusting remuneration for physician risk and patient health promotes justice and equity. Clinicians are motivated to improve therapy and decrease unneeded services since they are paid a specific amount per patient regardless of service intensity.

Chapter 3: Methodology

3.1 Introduction

This chapter presents the methodology adopted for the study. This includes the research philosophy, research design, the population and sample, and the methods for collecting data. The chapter also includes reliability and validity, as well as methods for data analysis. The study aims to use k-means clustering to have patients(members) with different risk needs to capitate separately.

3.2 Research Philosophy

Research philosophy is concerned with the origins, characteristics, and advancement of knowledge. It is the opinion on how information should be gathered, examined, and applied. According to (Saunders and Lewis, 2017), there are four primary research philosophies in the field of business studies: interpretivism, positivism, realism, and pragmatics. Whereas positivism maintains that only factual knowledge obtained from observation is reliable, pragmatics acknowledges that there are multiple ways to understand research. While realism is predicated on the notion that reality exists independently of the human mind, interpretivism incorporates human interest into a study (Saunders and Lewis, 2017). The positive concept has been chosen for this study to guide the creation, nature, and source of knowledge during the research process. The selection is because positivism supports a quantitative approach.

3.3 Research Design

Research design is a significant component of research, which serves as the roadmap for the activities to be undertaken in answering the research objectives. It unites a research project, according to Kombo and Tromp (2014). It will provide parameters for data gathering and processing to align relevance with research goals. Punch (2014) states that study design includes problem identification, project planning, and outcomes publication. A descriptive design is used in the proposed study for both data collection and analysis. This study utilizes a quantitative design to collect numerical data and examine the phenomenon at hand. Secondary health insurance data from the NHIF will be used for analysis in this study. As such, the study will be cross-sectional and retrospective.

In terms of methodology, the study will use the Cross Standard Industry Processing for Data

Mining (CRISP-DM) approach. CRISP-DM has standard data mining as a problem solving that is common for business and research (Figure 3.1). CRISP-DM methodology "consists of six phases as steps to conduct research in Business Understanding, Data Understanding, Data Preparation, Modeling. Evaluation, and Deployment" (Rahayu et al., 2020). This method has previously been applied in similar studies (Yang et al., 2018).

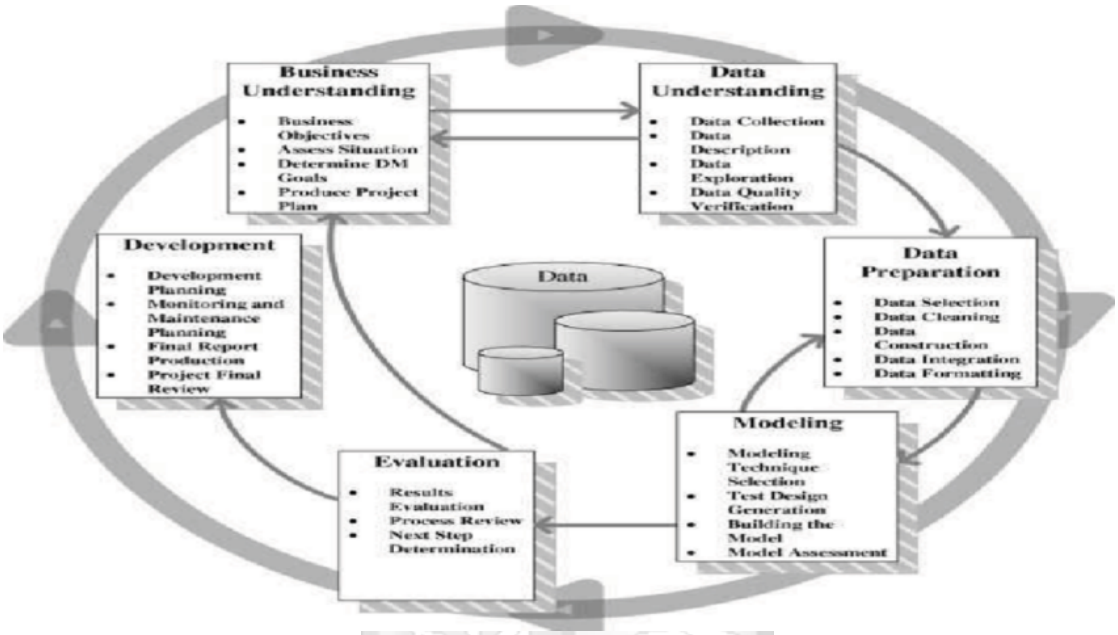


Figure 3.1: CRISP-DM framework for Data Science Projects (adapted from Rahayu et al., 2020)

3.4 Target Population

This study targets the Kenya National Health Insurance Fund (NHIF). The study will include members enrolled in the NHIF and who are above 18 years old. Only members with more than 12 months of continuous enrollment will be considered.

3.5 Data Sources

Sources of data for the study are summarized in Table 3.1

3.6 Variables for Clustering

- **Demographic Adjusters**: Utilizes age and sex, the most common risk adjusters as per Momahhed et al. (2023). Sourced from the NHIF Member Database.
- **Prior Year Expenditures**: Incorporates inpatient visit frequency, and total outpatient and inpatient medical costs, derived from NHIF Claims Utilization Data.

Table 3.1: NHIF Data Types and Descriptions

Data Type	Description
NHIF Member Database	A comprehensive database containing personal and demographic information of all NHIF members, such as names, ages, gender, and employment status.
NHIF Provider Matrix Data	Information about healthcare providers in the NHIF network, including types of services offered, locations, provider qualifications, and details of agreements with NHIF.
NHIF Claims Utilization Data	Detailed records of all insurance claims made, covering the services provided, their costs, dates of service, and details of NHIF reimbursements or payments.
NHIF Capitation Returns Data	Data related to the capitation payment system, which involves the prepayment of healthcare providers, typically covering details of payments made to providers based on the number of enrolled members.

- **Major Illness Proxy**: Employs major illness presence as a stand-in for diagnosis-based information, inferred from inpatient data in NHIF Claims Utilization Data.
- **Beneficiary Categories by Occupation**: Uses occupation-based categories to approximate income effects on healthcare expenditure, extracted from the NHIF Member Database.

3.7 Data Analysis and Data Preparation

The analysis aims to cluster patients (NHIF enrollees) in a need-based adjusted risk-based capitation. In this regard, a risk prediction model will be trained to generate patient risk scores. This will help identify a high-risk patient population. For the group, cluster analysis will be done, and the various clusters will be analyzed and profiled (Vuik et al., 2016). The analysis will be done for the data obtained from NHIF database. Before subjecting the data to clustering, it will be evaluated to ensure that it is complete, consistent and there is no missing information.

3.8 Clustering

Clustering is a widely used method for understanding data structure. It involves grouping data into subgroups (clusters) based on similarity, ensuring that data within a cluster is similar and differs significantly from other clusters (Momahhed et al., 2023). The choice of similarity metrics, like Euclidean or correlation-based distance, depends on the specific application. Feature-based clustering is our focus here, used in various applications like image segmentation, market segmentation, and document clustering (Xiaoqun and Run, 2022).

3.9 K-means Clustering

The algorithm that this study uses to cluster members of NHIF according to risk is the K-means algorithm. K-means clustering minimizes Euclidean distances to find the shortest path between points in multi-dimensional space (Gesicho et al., 2021). Since it uses numerical differences, this approach requires continuous variables. The algorithm selects k initial centres, assigns observations to the nearest centre, then recalculates the centre as the cluster's variable mean (Bretos-Azcona et al., 2020). This cycle continues until observations stop switching clusters, signifying convergence, and the ultimate clustering solution. Complex dataset analysis and interpretation require this efficient strategy (Momahhed et al., 2023).

3.9.1 K-means algorithm

The k-means clustering algorithm organizes a dataset into k groups and finds noise. It handles N n-dimensional data points (x). The indicator vector μi assigns each data point to one of k distinct clusters, each containing No data points (Vuik et al., 2016). The main goal of the k-means algorithm is to minimize its core function, the sum of squared distances, expressed in Equation 3.1:

$$min \sum_{i=1}^{k} \sum_{j=1}^{N} \mu_{ij} \times ||x_j - c_j||^2 \tag{3.1}$$

Here:

- μij is the association of data point j with cluster I.
- xj is the j-th data point.

- ci is the center of the i-th cluster.
- ||xj - ci|| represents the Euclidean distance between data point j and cluster center i.

First, C sites are randomly selected as cluster centers. Every data point's closest cluster center is found using Euclidean distance. Each sample is assigned to a C cluster after this stage (Momahhed et al., 2023). Clustering requires choosing a distance metric. The Euclidean distance is commonly employed. Clustering algorithms use key variables established and updated by the algorithm to compute this distance. Without focusing on any attribute, the technique clusters each observation (Bretos-Azcona et al., 2020). The k-means approach determines initial centroid positions and clusters (Momahhed et al., 2023). The technique groups the dataset by centroid-data point distance into k clusters. Iterate centroid values till stability.

The Euclidean distance between two points, x and y, in n-dimensional space is given by Equation 3.2:

$$d\,(x,y) = \sqrt{\sum_{n}^{i=1} (x_i - y_i)^2} \tag{3.2}$$

The algorithm involves recalculating and updating the centroid for each cluster and repeating this process until convergence.

3.10 Cluster Validation

In clustering analysis, it is essential to assess the validity of the chosen clusters to ensure that they reflect meaningful groupings within the data. Several methods are commonly used for cluster validation, including the Elbow Method (Cui et al., 2020), Davies-Bouldin Index (DBI) (Xiao et al., 2017), and Calinski-Harabasz Index (Ashari et al., 2023). Each of these methods provides unique insights into the clustering structure and helps to identify the optimal number of clusters. The following is a breakdown of the mentioned validation methods.

3.10.1 The Elbow Method

The Elbow Method is a graphical approach to determining the optimal number of clusters in K-Means clustering. The idea is to run the K-Means algorithm with different values of k (the number of clusters) and calculate the within-cluster sum of squares (WCSS), which measures

the total distance between each point and its cluster center. As k increases, WCSS decreases as clusters become smaller and more numerous.

The optimal k is determined at the ”elbow” point, where the rate of decrease in WCSS sharply slows down. This elbow represents the point beyond which adding more clusters does not provide significant improvements in WCSS, indicating a good balance between the number of clusters and the explained variance. The formula for the within-cluster sum of squares (WCSS) for each cluster j is given by the Formula 3.3.

$$WCSS = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2 \tag{3.3}$$

From the Formula 3.3, x is a data point in cluster C_i, μ_i is the centroid of cluster C_i and finally, k is the number of clusters.

3.10.2 Davies-Bouldin Index (DBI)

The Davies-Bouldin Index (DBI) is a metric used to evaluate the quality of clustering by assessing the average similarity between clusters. The DBI is based on the idea that clusters should be compact and well-separated. Lower DBI values indicate better clustering, as they represent clusters that are dense and distinct from one another. The formula for the Davies-Bouldin Index is 3.4

$$DBI = \frac{1}{k} \sum_{i=1}^{k} max_{j \neq 1} \frac{s_i + s_j}{d_{ij}} \tag{3.4}$$

Where s_i is the average distance between each point in cluster i and the centroid of cluster i, d_{ij} is the distance between the centroids of cluster i and j. k is the total number of clusters.

A lower Davies-Bouldin Index value indicates clusters that are compact and well-separated, which is desirable in clustering.

3.10.3 Calinski-Harabasz Index

The Calinski-Harabasz Index, also known as the Variance Ratio Criterion, is a measure of the ratio of between-cluster dispersion to within-cluster dispersion. It is used to assess the density and separation of clusters. Higher values of the Calinski-Harabasz Index indicate better-defined clusters, where clusters are compact internally and well-separated from each other. The formula for the Calinski-Harabasz Index is **??**.

$$CH = \frac{tr(B_k)/(k-1)))}{tr(W_k)/(n-k)} \tag{3.5}$$

From the Equation 3.5, $tr(B_k)$ is the trace of the between-cluster dispersion matrix, $tr(W_k)$ is the trace of the within-cluster dispersion matrix, n is the number of data points and k is the number of clusters. The between-cluster dispersion B_k represents the dispersion among cluster centers, while the within-cluster dispersion W_k represents the dispersion within each cluster. Higher Calinski-Harabasz scores imply better clustering solutions.

3.11 Software

The computer programming software, Python, will be used to perform cluster analysis, calculate, and other analyses, including prediction of risk.

3.12 Implementation and Monitoring

The study will adopt the Cross-Industry Standard Process for Data Mining (CRISP-DM) to implement k-means clustering on NHIF data. This will ensure there is a structured and systematic approach to data analysis. CRISP-DM consists of six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Despa & Nama, 2018; Ashari et al., 2022). These phases are explained in detail in Table 3.2.

3.13 Data Preparation

This process entails cleaning and preparing the dataset for analysis purposes. The dataset must be cleaned for analysis as part of this procedure. Several techniques can be applied in this process, such as addressing outliers (data entries that are either excessively small or large in

Table 3.2: CRISP-DM phases for implementing K-means clustering on NHIF data for risk

Phase	Description	Key activities
1.Business Understanding	- Focus on understanding project objectives and requirements from a business perspective. - Explore how risk-adjusted capitation enhances healthcare financing efficiency and equity in Kenya	- Define project objectives - Identify business needs
2.Data Understanding	- Initial data collection followed by data familiarization to identify quality issues. Ensure NHIF data collected is relevant to study objectives	- Collect initial data - Assess data quality - Identify data relevance
3.Data Preparation	- Data cleaning and transformation to ensure suitability for modelling. Handle missing values, outliers, and ensure data integrity and confidentiality.	- Clean and preprocess data - Handle missing values and outliers - Ensure data integrity
4. Modelling	- Apply k-means clustering to prepared data. Select and apply modelling techniques with optimized parameters to cluster NHIF members into distinct risk groups.	- Select modelling techniques - Optimize model parameters - Apply k-means clustering
5. Evaluation	- Assess the model to ensure it meets business objectives. Use cluster validity measures like silhouette scores and Davies–Bouldin index.	- Evaluate model against business objectives - Use cluster validity measures for assessment
6. Deployment	- Implement the clustering model within NHIF's operational framework. Plan deployment, monitor performance, and ensure contribution to equitable and efficient healthcare financing.	- Plan model deployment - Monitor model performance - Ensure model contributes to NHIF objectives

comparison to other values in the dataset), duplicates (repeated entries), dropping irrelevant columns, and handling missing (empty entries in the dataset, also known as null values). we have 1906557 entries of visits and 20 columns.

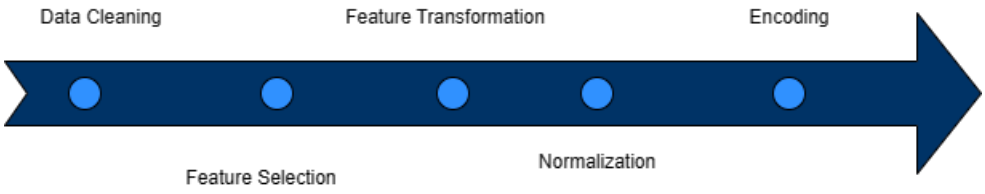


Figure 3.2: Data processing steps

3.14 Data Cleaning

3.14.1 Missing data treatment

Data can exhibit various patterns of missingness, each with distinct characteristics and implications for analysis. These patterns include Missing Completely At Random (MCAR), where missing values are randomly distributed across variables and unrelated to other features; Missing At Random (MAR), where the absence of data in one variable can be explained by other variables in the dataset; and Missing Not At Random (MNAR), indicating a systematic difference between missing and observed variables.

In our dataset, the column with the highest proportion of missing entries was the year of birth column, accounting for only 0.19%. Given that our task involves clustering, and age serves as a vital demographic variable for this purpose, we opted to remove all missing entries from this column. This decision was made to ensure the integrity of our clustering analysis, as accurate demographic information is essential for meaningful cluster formation.

3.14.2 Dropping irrelevant columns.

We excluded irrelevant columns from our dataset. This process involved removing columns that did not contribute to our analysis or modeling objectives. No outliers were present in our dataset.

3.15 Exploratory Data Analysis

a) Feature Engineering

Feature engineering plays a pivotal role in the development of effective models within the realm of machine learning. This process not only involves the enhancement of existing data but also the creation of new features that augment the accuracy of models. The significance of feature engineering lies in its ability to unveil underlying patterns in the

data that are not immediately obvious, thus profoundly influencing the performance of predictive algorithms.

b) Feature Selection

Feature selection is a crucial step in the process of feature engineering, aimed at identifying the most relevant features from a larger set of available variables. By selecting the most informative features and removing redundant or irrelevant ones, feature selection helps in reducing the dimensionality of the dataset, which can lead to improved model performance. This process enhances model interpretability, decreases overfitting, and reduces computational costs, as fewer features mean fewer calculations during model training. In addition, removing irrelevant or noisy features helps the model focus on the underlying patterns in the data, leading to better generalization to new, unseen data. Ultimately, effective feature selection can not only boost the accuracy of a model but also improve its efficiency and robustness, making it a vital aspect of the machine learning pipeline.

i) Sector (Income group column)

We have a dataset containing member information, which includes a column detailing the different sectors that members associated with the National Health Insurance Fund (NHIF) belong to. This data categorizes members into four distinct occupational categories. For our clustering analysis, we will use these occupation-based categories as a proxy to gauge the impact of income on healthcare expenditures. This approach is based on data extracted from the NHIF Member Database. We have combined this data with another data frame, resulting in a new data frame that includes an additional column specifying the sector for each member. The table can be found here 3.3

The highest utilization is observed in the informal sector groups, while the lowest is by groups sponsored by government programs, such as those for the elderly and disabled. This disparity may stem from the capitation payment system not being risk-adjusted for vulnerable groups like these. Consequently, these groups are often bundled with others, leading healthcare facilities to either provide substandard services or engage in cream skimming, where they selectively choose not to serve these high-risk groups. This is important when we do clustering based on the risk factors.

Table 3.3: Sector specifications for different members

Category	Description	Percentage proportion of utilization
MI	Informal sector groups (Chamas, sacco groups, self-employed cohort)	52.59 %
PP	Private Sector (formally employed in the private sector)	31.5 %
GS	Government Sector (formally employed in the government sector)	13.98 %
SP	Vision 2030 flagship groups (Sponsored by government programmes, Health Insurance Subsidy programme, Elderly and disabled)	1.91 %

ii) County (NCD prevalence index)

For each claim visit, there is a column indicating the county where the notification occurred. We compiled the counts for each county for the review period from the Kenya Health Information System, applied weighted measures, and included a column indicating the NCD (Non-Communicable Diseases - Figure 3.4) prevalence index for each visit as shown in Figure 3.3.

i. Normalization

The data obtained from the prevalence index calculation has a wide range of variations. This will cause problems during the analysis process. Therefore, by using the data transformation method, the data can be normalized. This normalization process makes all variables have the same range. Several methods can be used for data normalization, namely min-max normalization, z-score standardization, or decimal scaling. Normalization of z-score based on the mean and standard deviation of a dataset.

The equation can be seen in equation standardization (reference Larose).

$$Z(x_{ij}) = \frac{x_{ij} - \bar{x}_j}{\sigma} \tag{3.6}$$

Where xj is the mean of the data, xij is the original value and is the standard deviation.

```python
# Define weights for each condition
weights = {
    'Diabetes': 1.5,
    'Hypertension': 1.2,
    'Cardiovascular conditions': 2.0,
    'Road Traffic Injuries': 1.8,
    'Burns': 1.0,
    'Violence related injuries': 1.3,
    'Other injuries': 0.8,
    'Epilepsy': 1.5
}

# Apply weights to the conditions
for condition, weight in weights.items():
    data[condition] = data[condition] * weight

# Sum the weighted conditions to calculate the Risk Index
data['Weighted Risk Index'] = data[list(weights.keys())].sum(axis=1)

from sklearn.preprocessing import MinMaxScaler
# Initialize the MinMaxScaler
scaler = MinMaxScaler()

# Fit and transform the 'Weighted Risk Index' using the scaler
data['Normalized Risk Index'] = scaler.fit_transform(data[['Weighted Risk Index']])

# Display the modified dataframe with the new 'Normalized Risk Index' column
print(data[['county_name', 'Weighted Risk Index', 'Normalized Risk Index']].head())
```

```
       county_name  Weighted Risk Index  Normalized Risk Index
0          BARINGO             73725.50               0.097003
1            BOMET             85637.80               0.117097
2          BUNGOMA            137239.20               0.204136
3            BUSIA             79229.50               0.106287
4  ELGEYO-MARAKWET             76098.05               0.101005
```

Figure 3.3: KHIS weighted index distribution by county

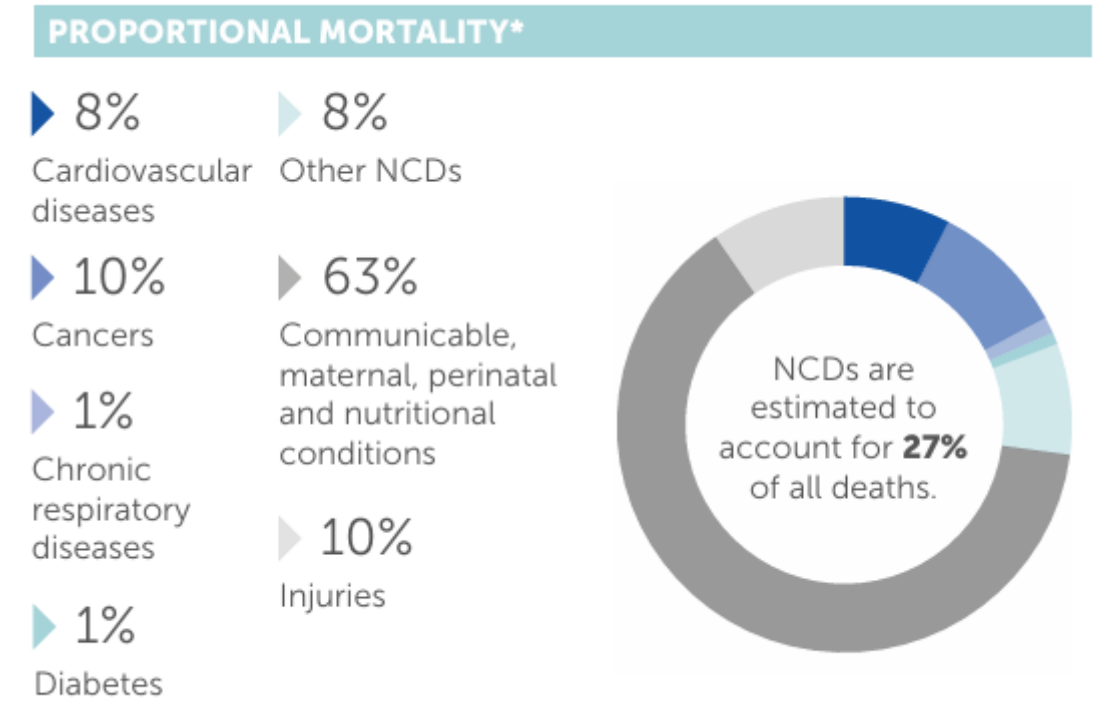


Figure 3.4: World Health Organization - Noncommunicable Diseases (NCD) Country Profiles

III. Binning to create age groups.

Our dataset includes the year of birth for each patient visit. We have converted the data type

of the year of the birth column to DateTime format and subsequently created a new column for age. Since age is continuous data, we categorized it into discrete bins of 0-17,18–24, 25–39, 40–54, and 55+. This approach helps stabilize models and reduces the impact of minor errors or outliers in the data. Age is a critical factor in our risk-adjusted clustering analysis of the members.

a) Feature Transformation

Feature transformation involves techniques that adapt variables into formats suitable for analysis. One such technique is one hot encoding, which converts categorical variables into numerical data. This is crucial for machine learning, as these models require numerical inputs and cannot process plain text in its original form. By transforming categorical data into numerical form, one hot encoding ensures that machine learning algorithms can effectively interpret and learn from the data.

b) Label Encoding Categorical variables.

When handling categorical data in input processing, it's common to convert categorical variables into numerical formats. Three widely used methods for this conversion are: 1) Label Encoding, where each category is assigned a unique integer; 2) One-hot Encoding and its variations, which transform categories into binary vectors; and 3) "Learned" Embedding, where encodings are derived during model training. Label encoding involves mapping each category to a distinct integer, which is particularly useful for variables like severity level, treatment class, and Case Mix Group(CMG) code. This method translates categorical values into corresponding integers (Hien et al., 2020).

We plan to update all entries of ICD10 blocks with those corresponding to the top 10 non-communicable diseases (NCDs) and the top 10 infectious diseases. This modification aims to refine the creation of patient profiles from visit data, facilitating a more focused analysis of health trends and costs associated with prevalent conditions.

The primary objective is to develop comprehensive patient profiles that encapsulate the frequency and financial burden of significant health conditions. These profiles will include counts and costs of visits associated with both NCDs and infectious diseases among the top ten identified by ICD10 classification.

From Figure 3.4 the patient profiles, it will be possible to develop risk scores and catego-

Table 3.4: Data Structure for Patient Profiles

Attribute Name	Attribute Description	Description
Mem no	Member Number	Unique identifier for each patient.
Age group	Age Group	Categorization of the patient's age.
Gender	Gender	Patient's gender.
Sector	Sector	Sector of employment (income group) or association as categorized in the member database.
Count of NCD	Count of NCD Visits	Total number of visits related to the top ten NCDs.
Cost of NCD	Cost of NCD Visits	Aggregate expenditure for NCD-related healthcare services.
Count of infectious diseases	Count of Infectious Disease Visits	Total number of visits about the top ten infectious diseases.
Cost of infectious diseases	Cost of Infectious Disease Visits	Total expenditure associated with visits for infectious diseases.
Count of others	Cost of Other Visits	Expenditure for healthcare services not related to the top ten NCDs or infectious diseases.
County	County	County Name
Normalized risk index	Normalized prevalence Index	Prevalence index for counties

rize patients based on their specific risk factors. This strategy enables a layered analysis of health risks, paving the way for targeted interventions and customized care approaches. Utilizing detailed data on healthcare visits, expenditures, and patient demographics allows for the application of statistical models to determine risk levels and strategically cluster patients. This refined methodology not only increases the accuracy of health risk management but also aids in crafting tailored healthcare solutions that optimize clinical outcomes and efficiently allocate resources. Additionally, this approach supports the implementation of risk-adjusted capitation payments, which are essential for optimizing primary healthcare outcomes and ensuring equitable treatment for high-risk members. This system promotes fairness by adjusting funding based on the health risks of different patient groups, thereby aligning financial incentives with the healthcare needs of patients.

3.16 Data Protection

Data security is crucial, particularly when managing medical insurance information. This study will follow all data protection and confidentiality laws and ethics. Securely sending and storing patient data, obtaining consents, and anonymizing it in all the stages. Regular audits will ensure data protection compliance. Data access is restricted to authorized staff. Data protection training to prevent abuse, breaches, and unauthorized access was conducted. Data privacy is crucial to study integrity and participant confidence.

Chapter 4: Results

4.1 Introduction

This chapter presents the study's findings on the efficiency of risk-adjusted capitation using K-Means clustering. The results are structured to align with the research objectives, providing a comprehensive overview of the data and its implications. The chapter begins with an in-depth Data Understanding section, where the nature, structure, and quality of the dataset are explored to establish a baseline comprehension of the variables and their characteristics.

Next, the Exploratory Data Analysis (EDA) section delves into the distribution of variables, uncovering hidden patterns, relationships, and anomalies within the data. This preliminary analysis informs the subsequent steps of Feature Selection and Feature Engineering, where relevant features are chosen and transformed to optimize the model's performance.

The core focus of the chapter is the Clustering Analysis, where the K-Means algorithm is applied to evaluate the efficiency of risk-adjusted capitation. The clustering results are discussed about the research objectives, highlighting key findings on how different groups within the dataset exhibit distinct risk profiles and resource utilization patterns. Diagrams, tables, and images are included throughout the chapter to support the findings and provide a visual representation of the clustering process and its outcomes. This visual documentation helps elucidate the segmentation of the data and the effectiveness of using K-means clustering for risk adjustment.

Overall, this chapter provides a comprehensive overview of the data analysis process, culminating in a clear understanding of the efficiency of risk-adjusted capitation and its implications for decision-making and resource allocation.

4.2 Exploratory Data Analysis

4.2.1 Age distribution

Figure 4.1 provides an overview of the distribution of people per age group. From the figure, the number of people per age group reveals a significant skew towards the middle-aged demographic. The age group 25-39 has the highest number of individuals, with a count of 745,640, indicating that this cohort forms the largest segment of the population. This is followed closely by the 40-54 age group, with 628,920 individuals, and the 55+ age group, with 507,386 individuals. The younger age groups, 18-24 and 0-17, have substantially lower counts, with 15,800

and 534 individuals, respectively.

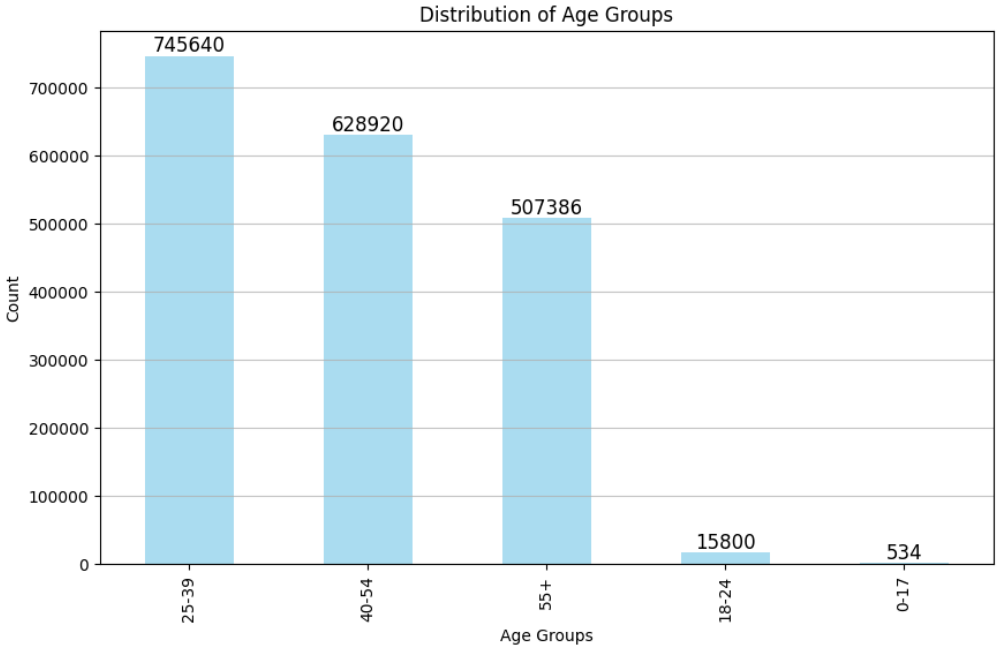


Figure 4.1: Distribution of the number of people per age group

Figure 4.2 provides the mean claim amount by age group highlights an increasing trend in average claim amounts as age increases. The youngest age group, 0-17, has the lowest average claim amount at 880.17, while the oldest age group, 55+, has the highest average at 1,212.42. The average claim amount gradually rises across the age groups, with notable increases in the 40-54 and 55+ categories, where the claim amounts are 1,072.00 and 1,212.42, respectively.

4.2.2 Gender Distribution

The gender distribution of insurance claims, as depicted in Figure 4.3, shows a higher prevalence of claims among females compared to males. Specifically, females account for 1,060,831 claims, while males account for 837,449 claims. This indicates that females are more active users of insurance services, making up a larger proportion of the total claims. The higher number of claims by females may suggest greater utilization of healthcare services or a higher frequency of accessing insurance benefits.

Figure 4.4 shows the average claim amount per gender revealing a slight difference between female and male claim values. Females have a higher average claim amount of 1,098.92 compared to males, whose average is 1,071.57. While the disparity is not substantial, it suggests that on average, females incur slightly higher costs per claim.

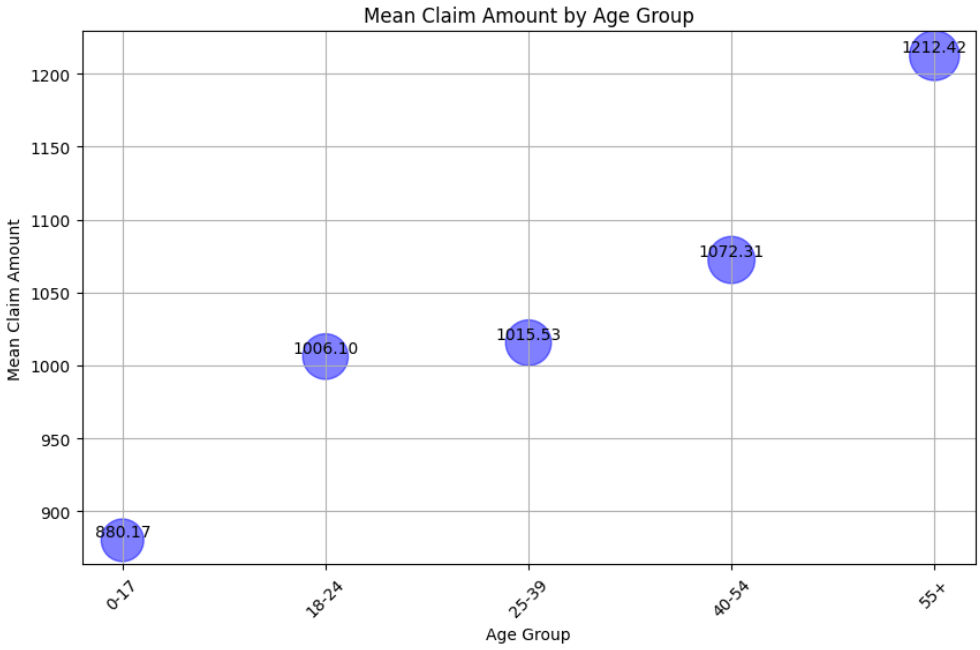


Figure 4.2: Mean claim amount per age group

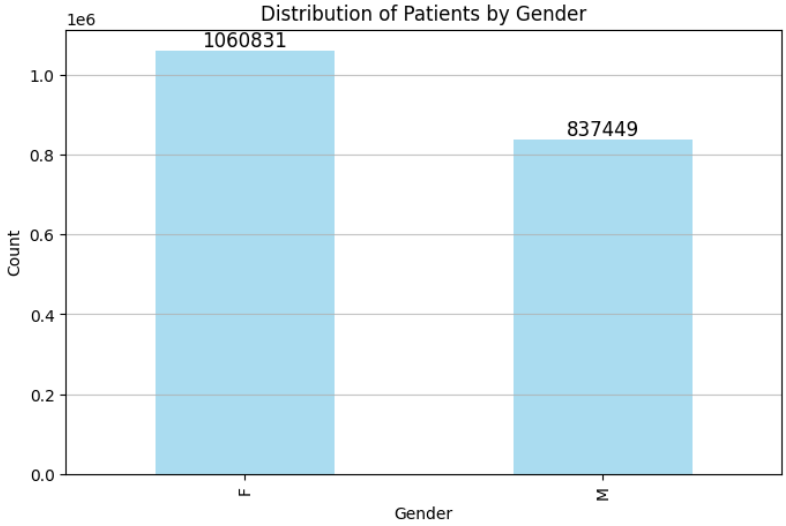


Figure 4.3: Distribution of the number of people per gender

4.2.3 Disease categories

The analysis of the average claim amount by different categories of medical conditions shows variation in the costs associated with each category. From Figure 4.5, claims related to Infectious Diseases have an average claim amount of 1,123.02, indicating moderate expenditure for this category. Non-Communicable Diseases (NCDs), on the other hand, have a slightly lower average claim amount of 1,017.36, suggesting relatively lower costs for chronic conditions or non-infectious health issues. The category labeled as Others, which may include miscellaneous

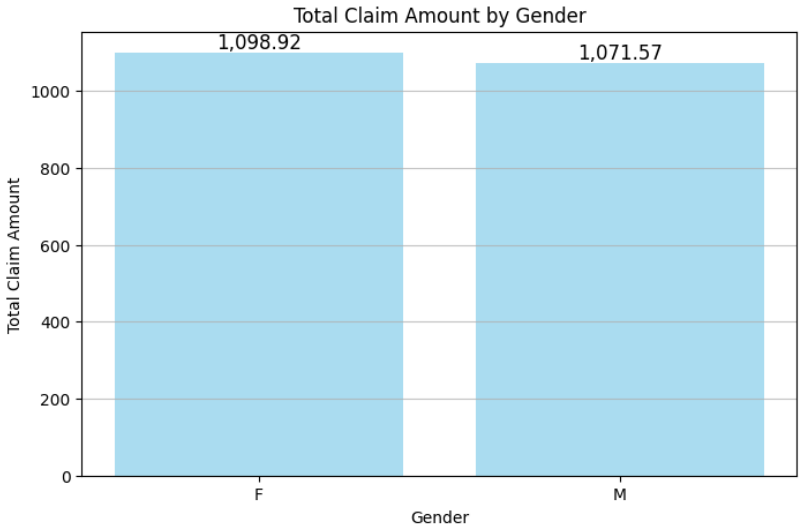


Figure 4.4: Average claim amount per gender

health conditions, has the highest average claim amount at 1,174.97.

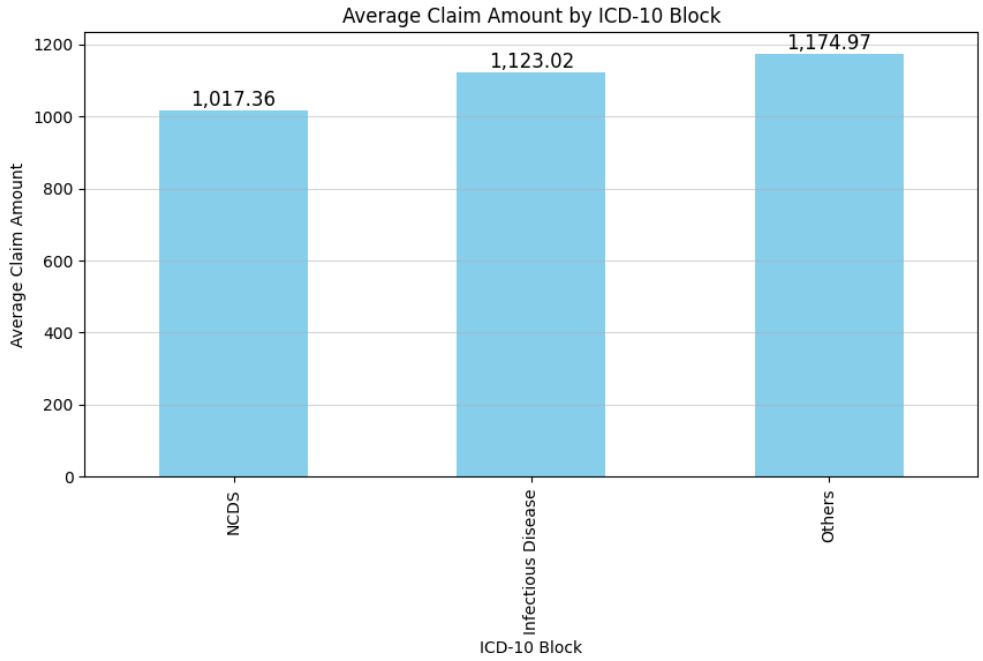


Figure 4.5: Average claim amount per different medical category

The heat 4.6 map illustrates the distribution of the number of claims across different medical conditions and age groups revealing significant variations in claim volumes based on both condition type and age demographic.

For Infectious Diseases, the number of claims is particularly high within the 25-39 age group, with a total of 117,402 claims, followed by 99,640 claims in the 40-54 age group. The younger age groups (0-17 and 18-24) show much lower claims, with 71 and 1,998, respectively.

In the Non-Communicable Diseases (NCDs) category, there is a substantial concentration of claims in the 25-39 age group, which accounts for 448,695 claims, and the 40-54 age group, with 341,178 claims. The 18-24 age group also has a notable number of claims, totaling 9,147. This indicates that middle-aged individuals are most affected by NCDs, suggesting a potential need for targeted health interventions in these demographics.

The Others category shows a more varied distribution, with the highest number of claims recorded in the 55+ age group at 284,249. This indicates that older individuals are significantly impacted by a range of conditions not classified as infectious diseases or NCDs. Additionally, the 25-39 age group has 179,543 claims, and the 40-54 age group has 188,102 claims, reflecting a consistent demand for healthcare across multiple age groups.

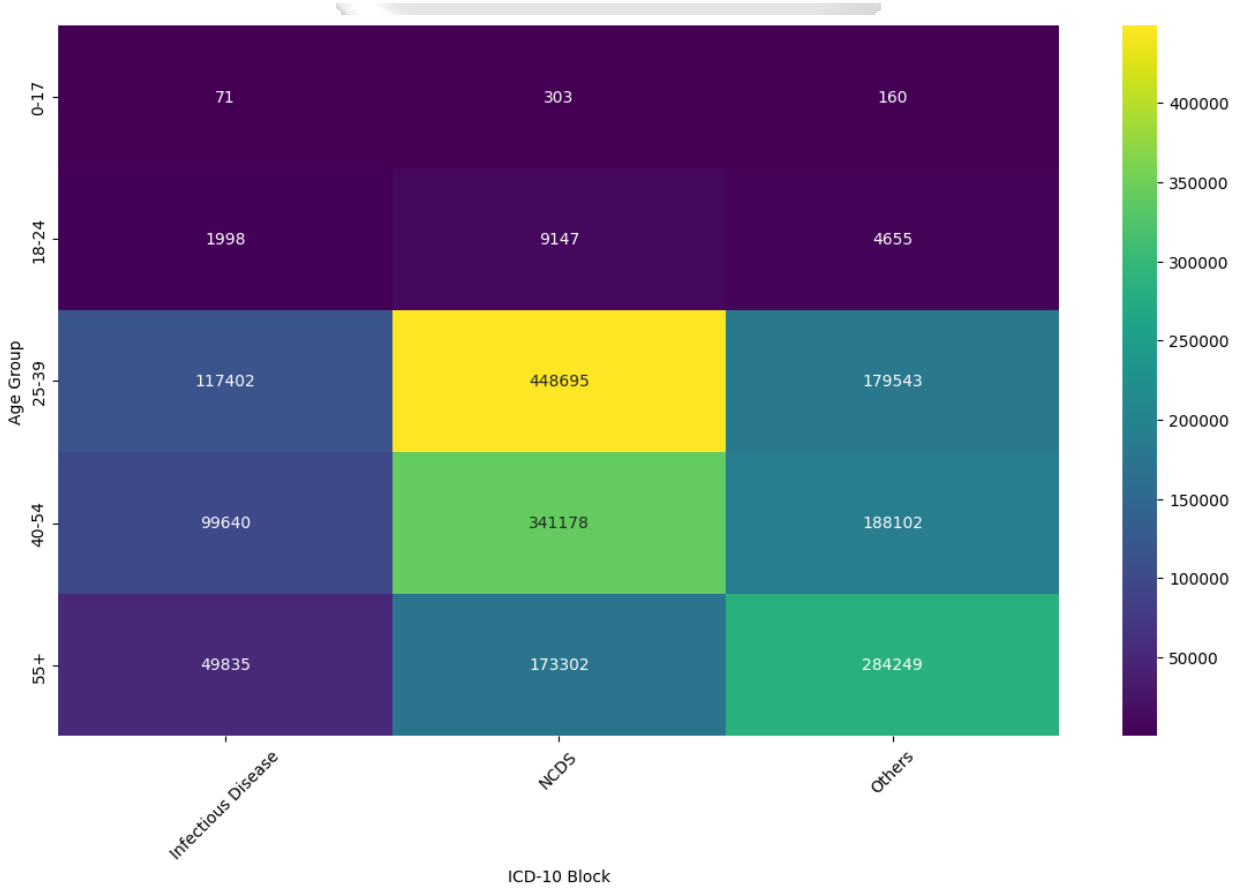


Figure 4.6: Heat map of the distribution of claim amounts per medical category

4.3 Cluster analysis

This study employs three key evaluation methods to identify the optimal number of clusters: the Elbow Method, the Davies-Bouldin Index, and the Calinski-Harabasz Score. Each method evaluates the clustering results using different criteria, helping to identify the ideal number of

Table 4.1: Data results for the elbow method

Number of Clusters (k)	WCSS
1	13,287,960
2	11,383,520
3	9,565,826
4	8,932,515
5	8,016,823
6	7,455,718
7	6,932,745
8	6,932,745
9	6,932,745
10	6,932,745

clusters.

4.3.1 The Elbow Method

The Elbow Method was utilized to identify the optimal number of clusters for segmenting the given dataset. The Within-Cluster Sum of Squares (WCSS) was calculated for a range of clusters from 1 to 10, and the results are summarized in the table below:

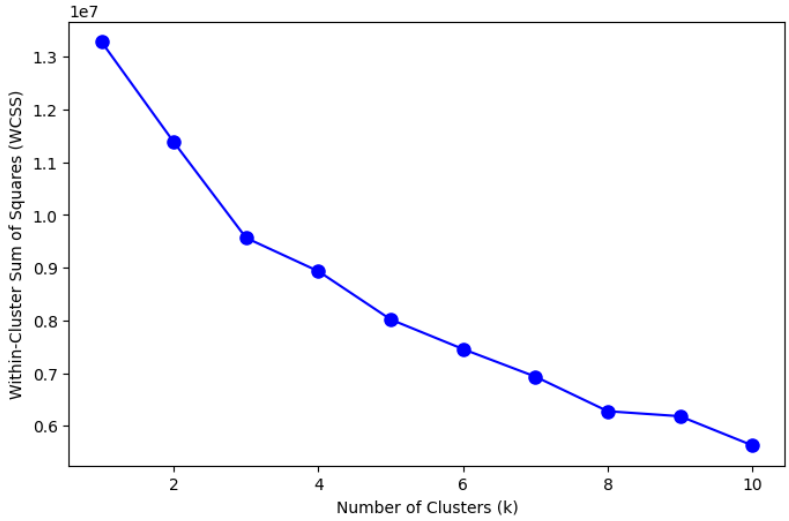


Figure 4.7: Elbow method for optimal number of clusters

From Figure 4.7 and table 4.1, we observe the following: Steady Decrease in WCSS: The WCSS values show a consistent decrease as the number of clusters increases. This reduction in WCSS indicates that adding more clusters helps in reducing intra-cluster variance, thereby increasing compactness within clusters.

Diminishing Returns After 5 Clusters: A notable reduction in WCSS is observed from k = 1 to

Table 4.2: Data results for the Davies-Bouldin Index method

Number of Clusters (k)	Davies-Bouldin Index
2	1.9849
3	1.5964
4	1.5110
5	1.4449
6	1.5455
7	1.4004
8	1.3128
9	1.3424

k = 5. After k = 5, the rate of reduction in WCSS becomes less pronounced. For instance, the decrease in WCSS between k = 2 and k = 3 is approximately 1.8 million, while the decrease between k = 5 and k = 6 is approximately 561,105. This trend suggests that increasing the number of clusters beyond 5 does not result in a significant improvement in clustering performance.

Identifying the ”Elbow” Point: The “elbow” point, where the curve begins to flatten, is typically the point where adding more clusters does not yield a significant reduction in WCSS. Based on the above data, the elbow appears around k = 5 or k = 6, as the reduction in WCSS slows down considerably beyond this point.

4.3.2 The Davies-Bouldin Index (DBI)

The Davies-Bouldin Index (DBI) was calculated for different numbers of clusters (k) ranging from 2 to 9 to evaluate the quality of clustering using the K-Means algorithm. The DBI is a measure of how well the clusters are separated and how compact they are; a lower value of the DBI indicates better clustering. From table 4.2 and figure 4.8 the following results were obtained:

The Davies-Bouldin Index indicates a significant improvement in clustering quality from k = 2 to k = 8. The DBI decreases as the number of clusters increases, indicating better clustering. A notable reduction occurs between k = 2 (1.9849) and k = 3 (1.5964), reflecting a marked improvement when moving from 2 to 3 clusters. The lowest Davies-Bouldin Index is observed at k = 8 (1.3128). This suggests that clustering with 8 clusters results in the most compact and well-separated clusters in the dataset compared to the other configurations. At k = 9, the DBI slightly increases to 1.3424, which indicates a slight drop in clustering quality compared to k = 8. This suggests that increasing the number of clusters beyond 8 does not lead to a significant

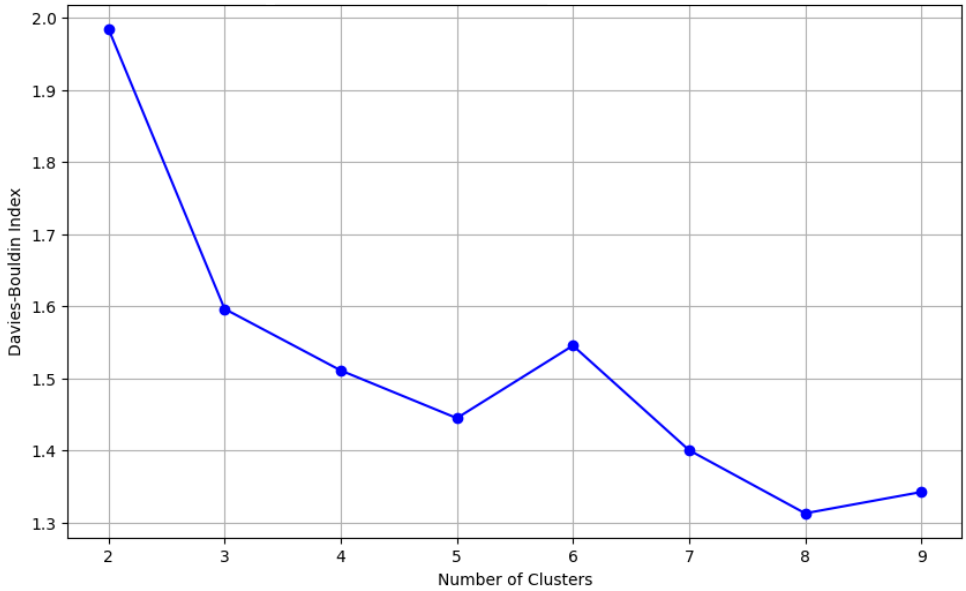


Figure 4.8: Davies-Bouldin index for different numbers of clusters

Table 4.3: Data results for the Calinski-Harabasz Index method

Number of Clusters (k)	Davies-Bouldin Index
2	317,577.62
3	369,317.72
4	308,530.50
5	312,033.96
6	296,985.58
7	290,023.15
8	302,919.27

improvement and may potentially cause overfitting or produce less meaningful clusters. Another relatively low DBI value was observed at k = 7 (1.4004). This configuration also shows a good separation of clusters, although not as optimal as k = 8.

4.3.3 The Calinski-Harabasz Index

The Calinski-Harabasz Index (also known as the Variance Ratio Criterion) was used to evaluate the clustering quality of a K-Means clustering model. This index measures the ratio of the sum of between-cluster dispersion and within-cluster dispersion, where a higher Calinski-Harabasz Index indicates better-defined and more separated clusters. The index values were calculated for different numbers of clusters (k), ranging from 2 to 8, and the results are as summarized in Table 4.3 and figure 4.9 shows the visual of the data.

From the results, the highest Calinski-Harabasz Index value is observed at k = 3 (369,317.72),

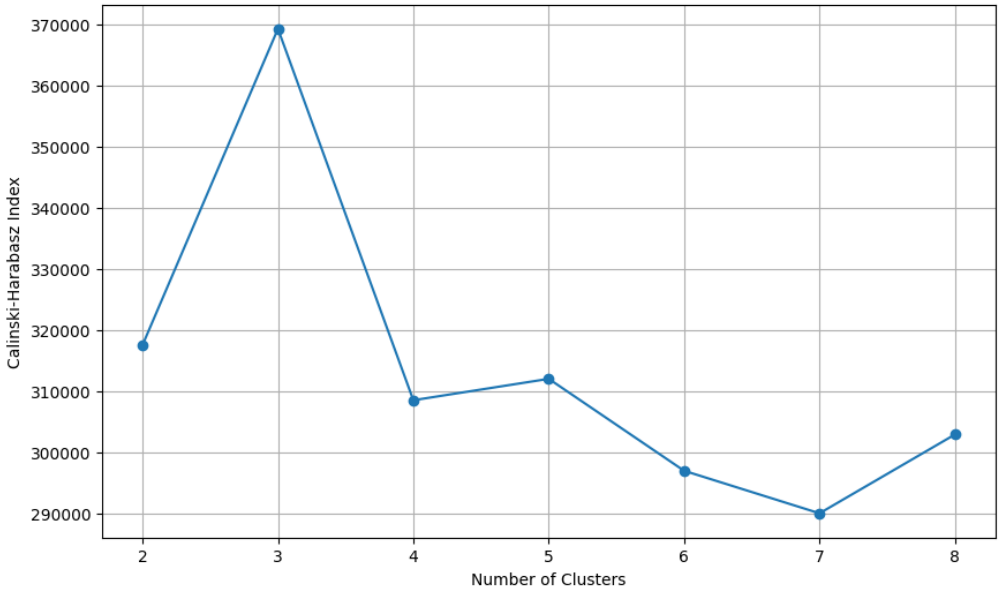


Figure 4.9: Calinski-Harabasz Index for different numbers of clusters

indicating that clustering into 3 clusters achieves the best separation and cohesion among clusters. This peak suggests that the dataset's structure is most clearly defined when divided into 3 groups. After k = 3, the index experiences a significant drop at k = 4 (308,530.50) and stabilizes slightly at k = 5 (312,033.96). This drop implies that the addition of a fourth cluster might result in reduced cluster separability and cohesion. For cluster configurations with k = 6, 7, and 8, the Calinski-Harabasz Index values are lower than those observed for smaller numbers of clusters, indicating less-defined clustering structures. Specifically, at k = 6, the index drops to 296,985.58 and continues to decline at k = 7 (290,023.15), before slightly increasing again at k = 8 (302,919.27). The results suggest that increasing the number of clusters beyond k = 3 might lead to overfitting or less meaningful cluster separations. While there is a small recovery in the index at k = 8, it does not surpass the peak observed at k = 3.

4.4 Optimization through Log transformation

Optimization in clustering often involves refining data preprocessing techniques to achieve better-defined and more accurate cluster formations. One common challenge in clustering is the presence of outliers, which can disproportionately influence distance-based clustering algorithms such as K-Means. To address this issue, log transformation is frequently applied to the dataset. This transformation reduces the skewness of the data distribution by compressing the range of large values, thereby mitigating the impact of outliers. By stabilizing variance

Table 4.4: Skewness before transformation

Column	Value
Age	32.327186
Infectious disease	5.505400
NCDS	3.489223
Others	6.154138
Claim amount	4.260203

and making data more normally distributed, log transformation allows clustering algorithms to perform more effectively, resulting in more cohesive and well-separated clusters.

4.4.1 Skewness Before Data Transformation

The skewness of a dataset measures the asymmetry of the data distribution around its mean. High skewness indicates that the data is not symmetrically distributed, which can be problematic for clustering and other machine-learning techniques that assume normality. Table 4.4 provides an overview of the columns affected by skew and the values.

From the skewness analysis, the following key observations are made:, The skewness value for Age is exceptionally high at 32.33, indicating that the data distribution is heavily right-skewed. This suggests the presence of extreme values or outliers in the age variable, which could negatively impact clustering results. Other variables related to medical conditions, such as Infectious Disease, NCDS, and Others, exhibit skewness values of 5.51, 3.49, and 6.15 respectively. These values indicate moderate to high levels of right-skewness, which could affect the clustering algorithms' ability to find optimal clusters. Finally, the skewness of claim_amount is also relatively high at 4.26, indicating that the majority of claims are concentrated on the lower end, with some extreme values pulling the distribution to the right.

Log transformation is a valuable preprocessing step for reducing skewness and minimizing the impact of outliers in data. Skewness indicates a lack of symmetry in the data distribution, which can cause clustering algorithms to perform poorly due to distorted distances between points. By applying a log transformation, large values are compressed, reducing the gap between outliers and the bulk of the data, which stabilizes the variance and makes the distribution more normal. This helps clustering algorithms like K-Means to identify patterns more accurately, leading to more cohesive clusters. Furthermore, reducing skewness improves the interpretability of data and aligns better with the assumptions of many machine learning models, enhancing the

Table 4.5: Skewness after Log transformation

Column	Value
Age	0.31
Infectious disease	1.45
NCDS	0.29
Others	0.71
Claim amount	-0.34

robustness and reliability of the analysis.

4.4.2 Skewness After Log Transformation

We applied log transformation to the data to reduce skewness and normalize the distribution of variables. Table 4.5, provides an overview of the data skewness report after log transformation.

From the results in Table 4.5, we observe that the skewness value for Age has significantly decreased to 0.31, indicating a nearly normal distribution. Similarly, NCDS shows a skewness of 0.29, suggesting that this variable is now well-centered around the mean. While for Infectious_Disease and Others remain at 1.45 and 0.71, respectively. While these values indicate some level of right-skewness, they are substantially lower than their pre-transformation values. This suggests that the log transformation has been effective in mitigating extreme values and improving the distribution. Finally, the log transformation has also resulted in a slight left skew for claim_amount at -0.34. This indicates that the distribution is now slightly skewed towards the lower values.

4.4.3 The Elbow Method After Log Transformation

After log transformation, the elbow method for used to find the optimal number of clusters. From Figure 4.10, we observe the following, after log transformation, the WCSS values for 2 and 3 clusters are lower compared to before (For 2 clusters, WCSS decreased from 1.138352e+07 to 1.105228e+07. For 3 clusters, WCSS increased slightly from 9.565826e+06 to 9.864907e+06.) This suggests that the log transformation might have slightly altered the spread of the data in a way that affects clustering when using lower numbers of clusters.

For Higher Number of Clusters (4 and Beyond), the WCSS consistently shows improvement (lower values). For example, at 4 clusters, the WCSS decreased from 8.932515e+06 to 8.816085e+06. This trend continues with improvements as the number of clusters increases, indicating that the

log transformation may have improved the compactness and separability of clusters at higher numbers.

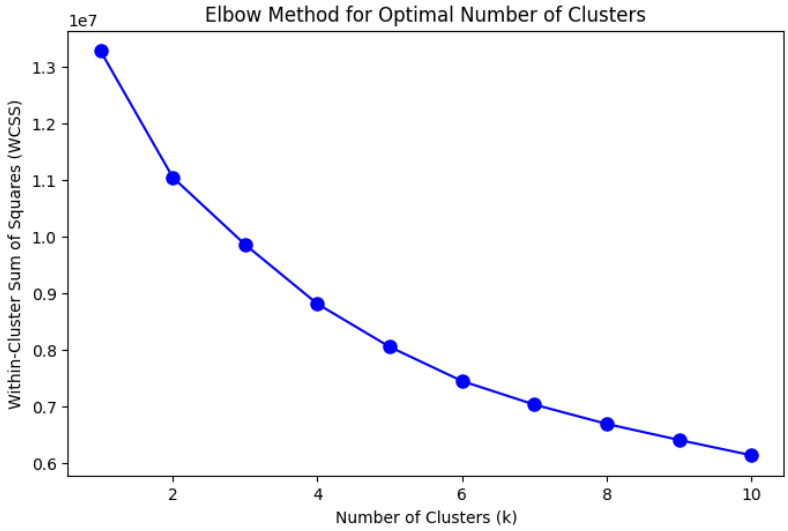


Figure 4.10: Elbow method for optimal number of clusters after log transformation

4.4.4 Calinski-Harabasz Index After Log Transformation

From Figure 4.11, some key observations are made such as the Calinski-Harabasz (CH) Index for 2 clusters increased significantly from 317,577.62 to 384,002.10. This indicates that the data distribution has become more conducive to forming two well-separated clusters after transformation. While for cluster 3, CH Index decreased from 369,317.72 to 329,350.42 after log transformation, suggesting that the distribution of data in three clusters may have been impacted by the transformation, resulting in less compact and clearly separated clusters. For cluster 4 and 5, the CH Index improved slightly from 308,530.50 to 320,962.33 after transformation. However, at 5 clusters, the CH Index slightly declined after transformation from 312,033.96 to 308,218.09. This indicates some consistency, but with a slight decrease in the quality of clustering at this stage. The CH Index for 6 clusters remained fairly similar after log transformation, moving slightly from 296,985.58 to 297,269.49. However, at 7 clusters, the CH Index decreased more substantially from 290,023.15 to 281,080.55, suggesting a reduction in cluster separation for higher cluster counts.

Based on the results of the Elbow Method and Calinski-Harabasz (CH) Index, it is reasonable to select 2 clusters as the optimal number of clusters for the transformed dataset. The WCSS values for 2 clusters showed a notable reduction from 1.138352e+07 to 1.105228e+07, indicating an improvement in compactness after the log transformation. Additionally, the significant in-

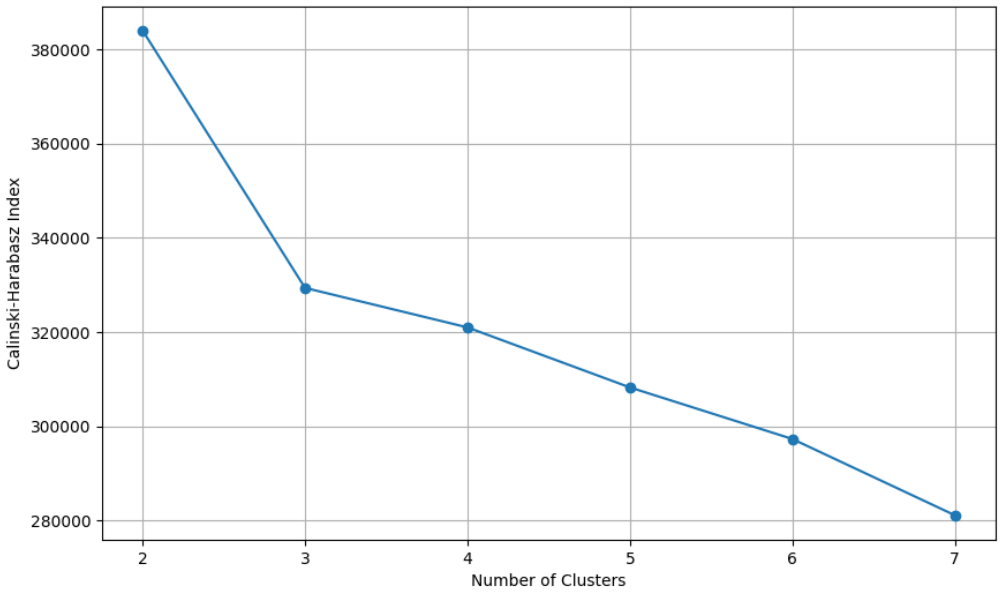


Figure 4.11: Calinski-Harabasz Index for optimal number of clusters after log transformation

crease in the CH Index from 317,577.62 to 384,002.10 demonstrates that the transformation has made the data distribution more conducive to forming two well-separated clusters. Although other cluster counts also exhibited some improvements in WCSS and CH Index values, the most pronounced and consistent benefits were observed at 2 clusters. For example, at 3 clusters, the slight decline in the CH Index indicates that adding cluster does not significantly improve the clustering quality. Moreover, beyond 2 clusters, there were mixed results, with some slight improvements in compactness but decreases in cluster separation, as seen in the CH Index values for clusters 4 to 7. Therefore, considering the balance between compactness (WCSS) and separation (CH Index), using 2 clusters is a justifiable choice for this transformed dataset.

4.5 Two Cluster Analysis

Figure 4.12 provides a visualization of a two-cluster solution using Principal Component Analysis (PCA) with three principal components. From the visualization, we observe that the two clusters are distinctly separated in the 3D space formed by the three principal components. This indicates that the principal components that have been chosen capture enough variance to differentiate the data into two clear groups. We also note that clusters appear relatively compact within their own space, suggesting that the K-means clustering algorithm effectively grouped similar data points. There are a few points where overlap seems to occur, particularly near the edges of the clusters. This could hint at a gradual transition between clusters or points that share characteristics from both clusters. Finally, the data distribution in the plot shows that PC1, PC2,

and PC3 effectively capture variations in the dataset. The clusters are not aligned along a single principal component, indicating that multiple components contribute meaningfully to the separation.

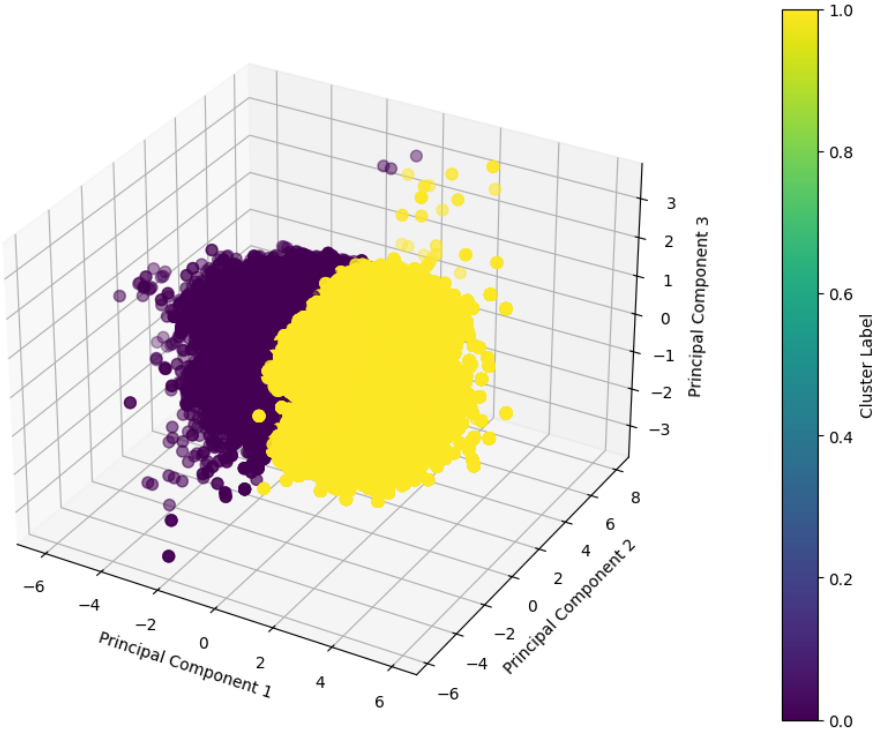


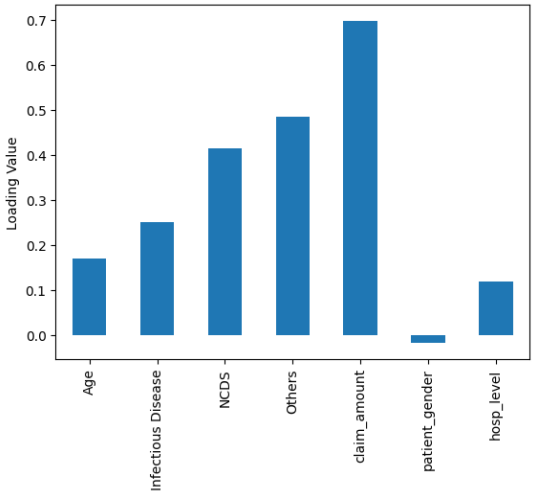
Figure 4.12: Cluster visualization using PCA(3 components)

4.5.1 Feature Importance(PCA Loadings)

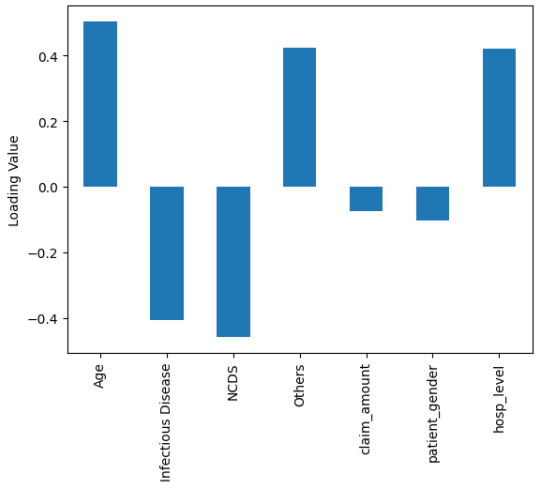
The PCA loadings in Table 4.6 and Figure 4.13 illustrate the contributions of each feature to the three principal components (PC1, PC2 and PC3). PC1 is heavily influenced by claim_amount (loading of 0.697), followed by contributions from Others (0.486) and NCDS (0.415). The positive loadings suggest that an increase in these features correlates positively with the first principal component, indicating that they contribute significantly to the variance captured by PC1. While PC2 exhibits strong positive loadings from Age (0.504) and hosp_level (0.420), while showing a substantial negative contribution from Infectious Disease (-0.407) and NCDS (-0.457). This suggests that PC2 distinguishes between age and hospital level on one side and the presence of infectious diseases or non-communicable diseases on the other. Finally, PC3 is primarily influenced by patient_gender (0.962), indicating that this component captures a significant amount of variance related to gender. The other features have relatively minor loadings in PC3, with some showing weak negative correlations such as hosp_level (-0.150).

Table 4.6: PCA Loadings for Principal Components 1, 2, and 3

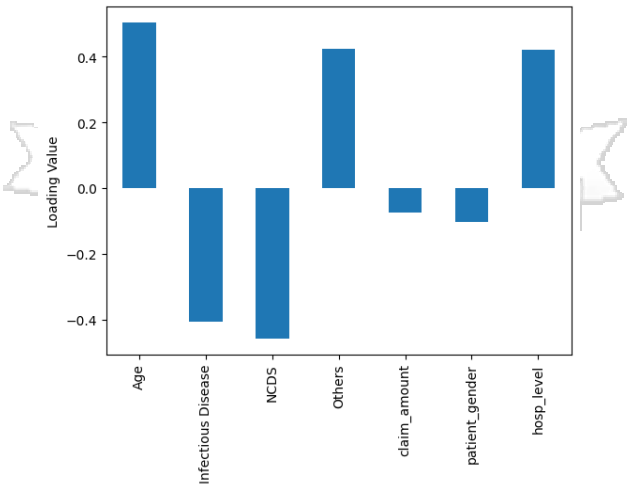
Feature	PC1	PC2	PC3
Age	0.169893	0.503800	0.184915
Infectious Disease	0.250213	-0.406991	0.031713
NCDS	0.414918	-0.457355	-0.093679
Others	0.485511	0.423085	0.087291
claim_amount	0.697199	-0.073626	-0.012780
patient_gender	-0.016632	-0.101721	0.962131
hosp_level	0.119537	0.420240	-0.150156



(a) Feature Loadings for PC1



(b) Feature Loadings for PC2



(c) Feature Loadings for PC3

Figure 4.13: Principal Component Analysis Loadings for PC1, PC2, and PC2

4.5.2 Cumulative Explained Variance

From Figure 4.14, we note that from the analysis of the explained variance reveals that the first three principal components (PC1, PC2, and PC3) account for approximately 62.46% of

the total variance in the dataset. PC1 explains 25.49% of the total variance, which indicates it captures the most significant source of variation in the data. While PC2 captures an additional 22.70%, which contributes significantly to the overall variability. Lastly, PC3 adds 14.27% to the explained variance, indicating it holds relevant information about the dataset.

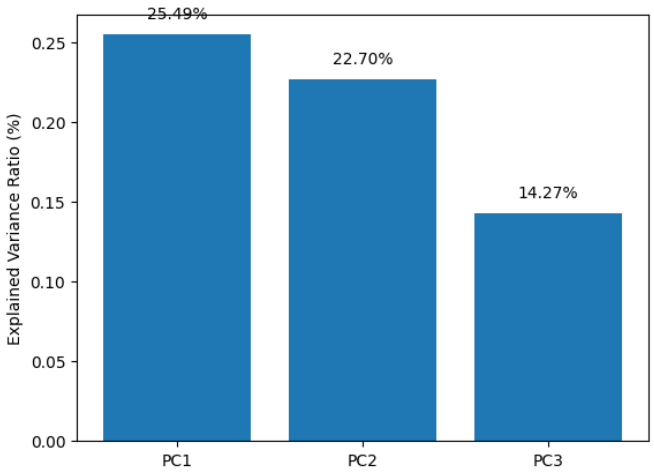


Figure 4.14: Variance Explained by Each Component)

4.6 Risk Adjustment Factor Calculation

The analysis of the risk adjustment factors derived from the K-Means clustering identified two distinct clusters: Cluster 0 and Cluster 1. The risk factor for Cluster 0 was calculated to be 0.990594, while Cluster 1 exhibited a slightly higher risk factor of 1.007360.

The risk factor values indicate the relative risk associated with each cluster. A risk factor below 1.0, as observed in Cluster 0, suggests that individuals in this group may have a lower risk relative to the overall population or the considered baseline. In contrast, the risk factor for Cluster 1, being above 1.0, indicates a higher risk level for individuals in this cluster.

4.6.1 Average Healthcare Costs by Cluster

The average claim amount for Cluster 0 is 4,848.04. This suggests that on average, individuals or cases in this cluster incurred slightly lower claim amounts compared to the other cluster. The calculated risk factor for Cluster 0 was 0.9906, indicating that this cluster poses a slightly lower risk than the dataset's average population. This aligns with the lower average claim amount, suggesting that this group has relatively fewer or less costly claims. While the average claim amount for Cluster 1 is 4,930.08. This indicates that individuals or cases in this cluster had

slightly higher average claim amounts than Cluster 0. The risk factor for Cluster 1 was 1.0074, which shows that this cluster represents a slightly higher risk compared to the average. This is in line with the higher average claim amount, suggesting a potentially higher claim load or more expensive claims for individuals in this cluster.

4.7 Comparison with Traditional Capitation

The simple average capitation method yielded an average claim amount of 5,917.4. This rate represents a uniform capitation amount for all individuals, calculated by taking the mean of all claim amounts in the dataset. This traditional approach does not account for variations in risk profiles or differences in individual characteristics that might affect healthcare costs. On the other hand, the K-Means-based risk adjustment approach identified two distinct clusters based on the dataset's characteristics. This allowed for the calculation of differentiated average claim amounts for each cluster. The differences between the clusters suggest that the algorithm identified groups with varying levels of risk or healthcare needs. These values are lower than the simple average capitation amount, indicating that clustering might provide a more nuanced and potentially cost-efficient allocation of healthcare resources.

4.8 Provider Financial Stability

Implementing risk-adjusted capitation has demonstrated a positive impact on the financial stability of healthcare providers. Specifically, providers who manage high-risk patients have experienced reduced financial losses. In contrast, those managing lower-risk patients have seen their payments appropriately adjusted to reflect their patient risk profiles (Ash and Ellis, 2012).

4.9 Quality of Care

Clustering patients into groups based on similar risk profiles allows for a more equitable distribution of resources. By setting a lower average price for cover within each cluster, hospitals can allocate funds more effectively to patients with higher needs, improving the overall quality of care. This approach ensures patients receive the necessary care, improving health outcomes. Research by (Obadha et al., 2019b) demonstrates the positive impact of risk-adjusted capitation on patient care, showing improved access to services and reduce disparities in healthcare delivery.

Chapter 5: Discussion

5.1 Exploratory Data Analysis

The results demonstrate that K-Means clustering can effectively segment patient populations into meaningful risk groups. The risk-adjusted capitation model provides a more accurate and fair payment system compared to traditional capitation methods. This approach addresses the risk selection problem and ensures that providers are compensated in line with the healthcare needs of their patients.

5.1.1 Age distribution

The age distribution in healthcare insurance data shows a notable concentration of individuals in the 25-39 age group, indicating that this demographic forms a significant portion of the insured population. This high representation of middle-aged individuals suggests that this group may have greater financial capability, awareness, or perceived need for health insurance coverage. In contrast, the younger demographic, particularly those aged 0-17 and 18-24, show substantially lower numbers, implying lower enrollment rates for minors and young adults. This might be attributed to several factors such as limited awareness, dependence on parental coverage, or fewer health concerns.

Additionally, the mean claim amount by age group shows a progressive increase with age. This trend reflects the increasing healthcare needs and costs associated with aging. The lowest average claim amounts in the youngest age group (0-17) may indicate generally better health conditions and lower utilization of healthcare services among minors. Conversely, the higher average claim amounts in the 55+ age group likely reflect a rise in age-related health issues and chronic conditions that require more extensive medical intervention.

These findings have several implications for insurance providers and policymakers. The skew towards middle-aged groups could suggest a need to focus on increasing enrollment rates among younger individuals through targeted awareness campaigns or incentives. Furthermore, the increasing claim amounts in older age groups highlight the potential financial burden of aging on insurance systems, which might necessitate differentiated pricing models or specific coverage plans tailored to the needs of older adults. Overall, these insights underscore the importance of demographic-specific strategies in insurance product design and marketing to enhance coverage

and sustainability.

5.1.2 Gender Distribution

The gender distribution analysis of healthcare insurance claims reveals a notable difference in the number of claims between females and males. The higher prevalence of claims among females indicates that they are more frequent users of healthcare services or more active in utilizing their insurance benefits. This could be attributed to several factors, including gender-specific healthcare needs, such as maternal and reproductive health services, which are likely to increase healthcare utilization among women. Additionally, social and cultural factors may play a role, with females possibly being more proactive about seeking medical care or managing health conditions.

The slight difference in the average claim amount between females and males further suggests that, while both genders incur relatively similar costs per claim, females may experience slightly higher healthcare expenses on average. This could be due to the type of medical services accessed or the frequency of services related to gender-specific conditions.

These findings have implications for insurers and healthcare planners. The higher number of claims by females may necessitate a gender-sensitive approach to designing insurance policies, ensuring that coverage adequately reflects the healthcare needs of women. Additionally, understanding these gender-based patterns can help in developing targeted healthcare interventions and promoting equitable access to necessary medical services. While the difference in average claim amounts is not substantial, it signals the importance of considering gender-specific health demands when pricing and structuring insurance policies.

5.1.3 Disease Categories

The analysis of healthcare claims by disease categories reveals several key trends in healthcare expenditure and utilization patterns. The average claim amounts indicate that healthcare costs vary based on the type of medical condition. Claims related to conditions classified under the ”Others” category have the highest average claim amount, suggesting that these conditions may involve more complex treatments or specialized healthcare services. On the other hand, claims for Non-Communicable Diseases (NCDs) have slightly lower average amounts, which may reflect the more predictable and ongoing management of chronic conditions like hypertension

or diabetes. Claims for Infectious Diseases fall between these two categories in terms of cost, indicating moderate expenditures for this category.

The distribution of claims across age groups and medical condition categories highlights significant variations in healthcare utilization. For Infectious Diseases, the highest concentration of claims is seen in the 25-39 age group, followed by the 40-54 age group. This suggests that middle-aged individuals may be more vulnerable to infections or more likely to seek treatment. The low claim counts in younger age groups could be attributed to a generally healthier population or reduced exposure to risk factors associated with infections.

In the case of NCDs, the high number of claims in the 25-39 and 40-54 age groups indicates that these age demographics are particularly affected by chronic health issues. This trend points to a need for targeted healthcare policies and preventive interventions focused on reducing the burden of NCDs among middle-aged individuals. Public health initiatives aimed at early detection, lifestyle modifications, and better disease management could be crucial in addressing these challenges.

The ”Others” category, with its highest number of claims in the 55+ age group, suggests that older individuals face a diverse range of health conditions that are not classified as either infectious or non-communicable diseases. This reflects the multifaceted health needs of an aging population, which may require specialized care and comprehensive insurance coverage.

Overall, these findings underscore the importance of age-specific and condition-specific healthcare planning. For insurance providers, understanding the distribution of claims across age groups and disease categories can help in the design of customized insurance packages that address the unique healthcare needs of different demographics. For policymakers, the results suggest the need to prioritize preventive healthcare measures for middle-aged and older populations to curb the increasing costs and health challenges associated with various medical conditions.

5.2 Cluster Analysis

5.2.1 Elbow Method

The results from the elbow method indicate that while increasing the number of clusters consistently reduces the Within-Cluster Sum of Squares (WCSS), the rate of reduction diminishes after five clusters. This trend suggests that the benefits of adding additional clusters beyond this

point are marginal. The elbow joint, where the curve starts to flatten, is typically considered the optimal number of clusters, as adding more clusters yields only minimal improvement in intra-cluster compactness. In this case, the elbow appears around k = 5 or k = 6, indicating that choosing five or six clusters strikes a balance between reducing WCSS and maintaining simplicity in the clustering model. This finding is crucial for achieving effective clustering without unnecessary complexity.

5.2.2 The Davies-Bouldin Index (DBI)

In this analysis, the Davies-Bouldin Index was computed for a range of cluster numbers from k=2 to k=9. The results indicate a significant improvement in clustering quality as the number of clusters increases. A notable reduction in the index is observed between k=2 (1.9849) and k=3 (1.5964). This suggests a considerable improvement in clustering quality when moving from 2 to 3 clusters, as the addition of a third cluster provides a more compact and distinct separation of data points.

The lowest Davies-Bouldin Index was achieved at k=8 (1.3128), indicating that clustering the data into 8 groups results in the most cohesive and well-separated clusters. This implies that atk=8, the intra-cluster variance is minimized, and the inter-cluster separation is maximized, leading to an optimal clustering solution for this dataset.

Despite the overall improvement, the DBI slightly increases at k=9 (1.3424), indicating a marginal decline in clustering quality when adding a ninth cluster. This slight increase suggests that additional clusters may not provide substantial benefits in cluster separation and could lead to overfitting. Overfitting occurs when too many clusters capture noise or subtle variations in the data, reducing the overall meaningfulness of the clustering solution. Therefore, the observed increase in the DBI at k=9 signals that adding more clusters beyond 8 may not yield significant improvements in defining the data's structure.

In conclusion, the Davies-Bouldin Index effectively highlights the optimal number of clusters by emphasizing lower intra-cluster variance and greater inter-cluster separation. The results indicate that an 8-cluster solution provides the best balance between compactness and separation in this dataset. However, increasing the number of clusters beyond 8 may lead to overfitting, resulting in a less meaningful representation of the data's underlying structure. These findings underscore the importance of using the Davies-Bouldin Index as a guide to identify the optimal

clustering configuration while avoiding excessive segmentation of the data.

5.2.3 The Calinski-Harabasz Index

The results showed that the highest Calinski-Harabasz Index value was observed at k=3 (369,317.72), signifying that clustering into three groups provided the most distinct separation and cohesion. This peak indicates that the three-cluster solution most effectively balances the compactness within each cluster and the distinctness between clusters, highlighting the dataset's inherent structure. These findings underline the importance of carefully selecting the optimal number of clusters to avoid over-segmentation and potential misinterpretation of the data.

5.3 PCA Components Cluster Analysis

The 3D visualization using three principal components provides a deeper understanding of the data structure. By examining the three components (PC1, PC2, and PC3), we gain insight into how each component contributes to the differentiation of the clusters. In this case, the principal components effectively capture the primary variations in the dataset where can notice how different components are affected by features, these distinctions allow us to separate the data into clusters. Applying PCA helped reduce the dimensionality while retaining key information about the variance in the data. This allowed us to visualize and interpret the clusters more effectively, which would not have been possible in the original high-dimensional feature space. The separation in the 3D space suggests that the clusters are well-formed based on the chosen number of clusters (2 in this case). However, the slight overlap between clusters at certain points may indicate that increasing the number of clusters or using a different clustering approach could lead to improved separation. Finally, including three components (PC1, PC2, and PC3) in the visualization enables a more comprehensive analysis of the clusters. If clusters were poorly separated using two components, the additional dimension may reveal patterns that were previously hidden. This highlights the significance of choosing an appropriate number of principal components based on the variance explained.

5.3.1 Feature Importance(PCA Loadings)

The PCA results reveal insightful patterns regarding the relationships between features in the dataset. Each principal component captures different aspects of the data variability. The high loadings of claim_amount, Others, and NCDS on PC1 suggest that this component predom-

inantly captures financial aspects associated with healthcare claims. This might indicate that claims related to non-communicable diseases and other categories are significant contributors to the variability within the dataset. On the other hand, PC2 demonstrates a contrast between age/hospital level and specific disease categories. This component could be interpreted as highlighting age-related differences in healthcare, where older individuals may present differing patterns of infectious and non-communicable diseases across hospital levels. While the dominant influence of patient_gender on PC3 indicates that gender is a key factor influencing the variability in the dataset. This could point to differences in claims or healthcare patterns based on gender, which warrants further investigation.

5.3.2 Cumulative Explained Variance

The cumulative variance explained by these three principal components suggests that they capture the majority of the patterns and structures within the dataset. However, with a cumulative variance of 62.46%, it is clear that a substantial portion of the total variability (approximately 37.54%) remains unexplained by these three components alone. Given that most of the work is in exploratory analysis or visualization, having three components that explain over 60% of the variance might be considered adequate to reveal key patterns or relationships in the dataset. Additionally, the diminishing returns in the amount of variance explained between PC1 (25.49%) and PC3 (14.27%) indicate that subsequent components contribute progressively less to the explained variability. This aligns with the ”elbow method” of selecting components, which suggests stopping when adding more components yields diminishing returns.

The PCA loadings analysis helps identify which variables contribute most to each component, providing valuable insights into the underlying structure of the data. Understanding these relationships can guide subsequent clustering or classification analyses, improving the accuracy and interpretability of results.

5.4 Policy and Practice Implications

The findings on the impact of risk-adjusted capitation carry substantial implications for healthcare policy and practice. This model aligns financial incentives with patient care needs, promoting high-quality service delivery while effectively managing financial risks. In terms of policy, risk-adjusted capitation can incentivize providers to improve care quality, especially for high-risk patients, by tying compensation directly to patient complexity. This approach

can reduce cost-cutting practices that compromise care and ensure equitable resource allocation across different patient populations. For practice, it encourages enhanced care coordination and investment in preventive measures, particularly for those with chronic or complex conditions, potentially reducing hospitalizations and other costly interventions. Additionally, provider accountability and performance monitoring can be strengthened through metrics tied to this payment model, motivating providers to maintain and improve care standards.

However, implementing risk-adjusted capitation comes with challenges such as Ensuring data accuracy for proper risk adjustment is critical, as incomplete or inaccurate data can lead to unfair compensation. Provider resistance is another issue, as transitioning from fee-for-service to capitation can raise concerns about potential revenue loss or increased administrative burdens. Furthermore, there is a risk of adverse selection and gaming, where providers may avoid high-risk patients or manipulate coding practices to increase perceived risk. To address these challenges, investing in robust data systems is essential. This includes building infrastructure for accurate data collection and analysis. Engaging with stakeholders—providers, payers, and patient groups—during implementation can build trust and encourage acceptance. Training and transparent communication about the model's benefits are key to easing the transition. Transparent metrics and performance monitoring systems must be established to discourage gaming and promote accountability. Risk mitigation mechanisms, such as stop-loss insurance or reinsurance, can protect providers from extreme financial losses, especially during the transition phase.

Practical implementation of risk-adjusted capitation should consider a phased rollout, allowing for testing and refinement in different practice settings (Newhouse et al., 1989). Integrating this approach with other payment models, such as bundled payments or pay-for-performance initiatives, can create a hybrid approach tailored to diverse healthcare needs. In conclusion, while the transition to risk-adjusted capitation requires careful planning and adjustment, it holds promise for aligning financial incentives with patient outcomes, fostering equitable, high-quality care delivery across healthcare systems. By focusing on data accuracy, stakeholder engagement, transparent metrics, and effective risk management strategies, policymakers and practitioners can maximize the model's potential benefits.

5.5 Limitations of the Study

The study's reliance on existing data sources and administrative records may introduce data accuracy issues. Incomplete or inconsistent documentation, variability in coding practices among providers, and potential inaccuracies in reported patient risk levels could affect the reliability of the results. Such data limitations might lead to bias in estimating the true impact of risk-adjusted payments, as some providers may have inaccurately coded or underreported patient risk factors.

Second, the study may have a limited scope in terms of geographical or institutional representation. If data was drawn from a specific healthcare system, country, or subset of providers, the findings may not generalize to other settings with differing patient populations, healthcare practices, or payment models. This limits the broader applicability of the study's conclusions and may impact the ability to draw comprehensive inferences about the effectiveness of risk-adjusted capitation across diverse healthcare environments.

Third, the study design and analysis methods may introduce potential biases. For instance, observational designs can be prone to selection bias if certain providers or patients were more likely to participate or if data was more readily available for particular risk levels. Additionally, confounding factors that were not accounted for, such as varying levels of provider experience, differences in patient adherence to care plans, or external socioeconomic factors, might influence the outcomes attributed to risk-adjusted capitation.

Finally, while the study suggests a link between risk-adjusted payments and improved quality of care or financial stability, causality cannot be definitively established due to the observational nature of the analysis. Other, unobserved factors may have contributed to the observed effects, making it difficult to isolate the precise impact of the payment model.

In conclusion, these limitations highlight the need for caution in interpreting the study's results. While the findings offer important evidence about the potential benefits of risk-adjusted capitation, further research, ideally using more robust and diverse datasets and employing experimental or quasi-experimental methods, is necessary to validate these conclusions and mitigate potential biases inherent in the current analysis.

Chapter 6: Conclusion

The application of K-means clustering to risk-adjusted capitation shows considerable promise as an efficient provider payment mechanism. This approach enhances fairness and accuracy in payment allocations and promotes better resource management and patient care within healthcare systems. By incorporating patient risk factors into capitation models, providers can more accurately predict healthcare costs and ensure that payments are proportionate to the anticipated needs of different patient groups (Ash et al., 2000).

The implementation of risk-adjusted capitation necessitates sophisticated methodologies to accurately assess patient risk profiles. One such method is the k-means clustering algorithm, which groups patients into distinct clusters based on their healthcare utilization patterns and demographic characteristics. This approach enhances the precision of risk adjustment by ensuring that patients with similar risk levels are grouped, thus allowing for more tailored capitation rates (Zelmer, 2014).

The k-means clustering analysis identified two optimal clusters with average claim amounts of 4848 and 4930, both significantly lower than the dataset's overall average claim amount of 5917. These clusters, verified through PCA with three components, were able to identify the key features of importance in the different components of clusters and revealed homogeneous groups in terms of cost patterns, suggesting meaningful population segmentation. Such insights support the implementation of risk-adjusted capitation models that better reflect patient cost dynamics, ensuring more equitable resource distribution.

Several studies have demonstrated the efficacy of risk-adjusted capitation in improving healthcare outcomes and cost efficiency. For instance, research by Smith (2019) indicates that healthcare providers operating under risk-adjusted capitation are more likely to engage in proactive and preventive care measures, reducing the incidence of high-cost medical events. Moreover, a study by Brown (2020) reveals that risk-adjusted capitation can lead to significant cost savings without compromising the quality of care.

Using K-means clustering allows for identifying distinct patient clusters based on various risk factors, facilitating the development of more precise capitation rates (Hornbrook, 2001). This method addresses the inherent challenges in traditional capitation models, which often fail to account for the heterogeneous nature of patient populations. By leveraging data-driven tech-

niques, healthcare systems can achieve a more balanced distribution of resources, ultimately improving the quality of care and operational efficiency (Bertsimas et al., 2008; Zhang, 2017).

The findings of this thesis underscore the importance of adopting advanced analytical methods in healthcare financing. Future research should explore the integration of other machine learning algorithms and their potential synergistic effects with K-means clustering in risk-adjusted capitation models (Kuhn and Johnson). Additionally, real-world implementation and longitudinal studies are necessary to validate the practical benefits and sustainability of this approach in diverse healthcare settings (Kuhn, 2013; Newhouse, 1994).

Bibliography

(2022). Defining a risk-adjustment formula for the introduction of population-based payments for primary care in france. *Health Policy*, 126(9):915–924.

Anell, A., Dackehag, M., and Dietrichson, J. (2018). Does risk-adjusted payment influence primary care providers' decision on where to set up practices? *BMC Health Services Research*, 18(1):179.

Ash, A. S. and Ellis, R. P. (2012). Risk-adjusted payment and performance assessment for primary care. *Medical care*, 50(8):643–653.

Ash, A. S., Ellis, R. P., Pope, G. C., Ayanian, J. Z., Bates, D. W., Burstin, H., Iezzoni, L. I., MacKay, E., and Yu, W. (2000). Using diagnoses to describe populations and predict costs. *Health care financing review*, 21(3):7.

Ashari, I. F., Nugroho, E. D., Baraku, R., Yanda, I. N., Liwardana, R., et al. (2023). Analysis of elbow, silhouette, davies-bouldin, calinski-harabasz, and rand-index evaluation on k-means algorithm for classifying flood-affected areas in jakarta. *Journal of Applied Informatics and Computing*, 7(1):95–103.

Barasa, E., Rogo, K., Mwaura, N., and Chuma, J. (2018). Kenya national hospital insurance fund reforms: Implications and lessons for universal health coverage. *Health Syst. Reform*, 4(4):346–361.

Bertsimas, D., Bjarnadóttir, M. V., Kane, M. A., Kryder, J. C., Pandey, R., Vempala, S., and Wang, G. (2008). Algorithmic prediction of health-care costs. *Operations Research*, 56(6):1382–1392.

Bretos-Azcona, P. E., Sánchez-Iriso, E., and Cabasés Hita, J. M. (2020). Tailoring integrated care services for high-risk patients with multiple chronic conditions: a risk stratification approach using cluster analysis. *BMC Health Services Research*, 20(1):806.

Brown, R., J. (2020). Evaluating the financial impacts of risk-adjusted capitation.

Carta, S., Consoli, S., Piras, L., Podda, A. S., and Recupero, D. R. (2021). Event detection in finance using hierarchical clustering algorithms on news and tweets. *PeerJ Computer Science*, 7:e438.

Cox, T. (2001). Risk theory, reinsurance, and capitation. *Issues in Interdisciplinary Care*, 3.

Cui, M. et al. (2020). Introduction to the k-means clustering algorithm based on the elbow method. *Accounting, Auditing and Finance*, 1(1):5–8.

Dohmen, P., De Sanctis, T., Waiyaiya, E., Janssens, W., Rinke de Wit, T., Spieker, N., Van der Graaf, M., and Van Raaij, E. M. (2022). Implementing value-based healthcare using a digital health exchange platform to improve pregnancy and childbirth outcomes in urban and rural kenya. *Frontiers in Public Health*, 10.

Geruso, M. and McGuire, T. G. (2016). Tradeoffs in the design of health plan payment systems: Fit, power and balance. *Journal of Health Economics*, 47:1–19.

Gesicho, M. B., Were, M. C., and Babic, A. (2021). Evaluating performance of health care facilities at meeting hiv-indicator reporting requirements in kenya: an application of k-means clustering algorithm. *BMC Medical Informatics and Decision Making*, 21(1):6.

Grant, R. W., McCloskey, J., Hatfield, M., Uratsu, C., Ralston, J. D., Bayliss, E., and Kennedy, C. J. (2020). Use of Latent Class Analysis and k-Means Clustering to Identify Complex Patient Profiles. *JAMA Network Open*, 3(12):e2029068–e2029068.

Hajat, C., Siegal, Y., and Adler-Waxman, A. (2021). Clustering and healthcare costs with multiple chronic conditions in a us study. *Frontiers in Public Health*, 8.

Hien, D. T. T., Thuy, C. T. T., Anh, T. K., Son, D. T., and Giap, C. N. (2020). Optimize the combination of categorical variable encoding and deep learning technique for the problem of prediction of vietnamese student academic performance. *International Journal of Advanced Computer Science and Applications*, 11(11).

Hornbrook, M. (2001). Clustering patients for resource use profiles by diagnosis.

Hsu, S., Lin, C., and Yang, Y. (2008). Integrating neural networks for risk-adjustment models. *Journal of Risk and Insurance*, 75(3):617–642.

Iommi, M., Bergquist, S., Fiorentini, G., and Paolucci, F. (2022). Comparing risk adjustment estimation methods under data availability constraints. *Health Economics*, 31(7):1368–1380.

Jia, L, M. Q. S. A. Y. B. and Zhang, L. (2021). Payment methods for healthcare providers working in outpatient healthcare settings. *Cochrane Database of Systematic Reviews*, (1).

Jiang, X., Zhao, B., and Wu, J. (2022). Hpr70 developing the risk-adjusted capitation payment for patients with diabetes mellitus in china: Results from administrative data in tianjin. *Value in Health*, 25(12):S245.

JOHN C. LANGENBRUNNER, CHERYL CASHIN, S. O. (2019). The design and implementation of per capita payment systems in the context of primary health care-centered health system development in low- and middle-income countries. *world Bank*.

Khezri, A., Mahboub-ahari, A., Sadegh Tabrizi, J., and Nosratnejad, S. a. (2022). Weight of risk factors for adjusting capitation in primary health care: A systematic review. *Medical Journal of the Islamic Republic Of Iran*, 36(1).

Kuhn, M. (2013). The effect of clustering within electronic health records on support for clinical decisions.

Langenbrunner, J. C. and Wiley, M. M. (2002). Hospital payment mechanisms: theory and practice in transition countries. *Hospitals in a changing Europe*, 150.

Layton, T. J., McGuire, T. G., and van Kleef, R. C. (2018). Deriving risk adjustment payment weights to maximize efficiency of health insurance markets. *Journal of Health Economics*, 61:93–110.

Liao, M., Li, Y., Kianifard, F., Obi, E., and Arcona, S. (2016). Cluster analysis and its application to healthcare claims data: a study of end-stage renal disease patients who initiated hemodialysis. *BMC Nephrology*, 17(1):25.

Mbau, J. (2021). *Effects of Capitation on Service Quality in Hospitals Accredited by Nhif in Nairobi County, Kenya*. PhD thesis, University of Nairobi.

Momahhed, S. S., Emamgholipour Sefiddashti, S., Minaei, B., and Shahali, Z. (2023). K-means clustering of outpatient prescription claims for health insureds in iran. *BMC Public Health*, 23(1):788.

Munge, K., Mulupi, S., Barasa, E. W., and Chuma, J. (2018). A critical analysis of purchasing arrangements in kenya: The case of the national hospital insurance fund. *International Journal of Health Policy and Management*, 7(3):244–254.

Muremyi, R., Haughton, D., Kabano, I., and Niragire, F. (2020). Prediction of out-of-pocket health expenditures in rwanda using machine learning techniques. *Pan African Medical Journal*, 37(1).

Newhouse, J. P. (1994). Patients at risk: health reform and risk adjustment. *Health Affairs*, 13(1):132–146.

Newhouse, J. P., Manning, W. G., Keeler, E. B., and Sloss, E. M. (1989). Adjusting capitation rates using objective health measures and prior utilization. *Health Care Financ Rev*, 10(3):41–54.

NHIF (2019). Health financing reforms expert panel for the transformation and repositioning of the national hospital insurance fund as a strategic purchaser of health services for the attainment of universal health coverage by 2022.

Obadha, M., Barasa, E., Kazungu, J., Abiiro, G. A., and Chuma, J. (2019a). Attribute development and level selection for a discrete choice experiment to elicit the preferences of health care providers for capitation payment mechanism in kenya. *Health Economics Review*, 9(1):30.

Obadha, M., Chuma, J., Kazungu, J., Abiiro, G. A., Beck, M. J., and Barasa, E. (2020). Preferences of healthcare providers for capitation payment in Kenya: a discrete choice experiment. *Health Policy and Planning*, 35(7):842–854.

Obadha, M., Chuma, J., Kazungu, J., and Barasa, E. (2019b). Health care purchasing in kenya: Experiences of health care providers with capitation and fee-for-service provider payment mechanisms. *The International Journal of Health Planning and Management*, 34(1):e917–e933.

Ochieng, T. B. (2019). *A Comparison of capitation and fee for service provider payment mechanisms and their effects on cost of healthcare: a case study of the Avenue Hospital, Nairobi*. PhD thesis, Strathmore University.

Pioch, C., Henschke, C., Lantzsch, H., Busse, R., and Vogt, V. (2023). Applying a data-driven population segmentation approach in german claims data. *BMC Health Services Research*, 23(1):591.

Presch, G., Dal Mas, F., Piccolo, D., Sinik, M., and Cobianchi, L. (2020). The world health innovation summit (whis) platform for sustainable development: From the digital economy to knowledge in the healthcare sector. In *Intellectual capital in the digital economy*, pages 19–28. Routledge.

Ramezani, M., Takian, A., Bakhtiari, A., Rabiee, H. R., Fazaeli, A. A., and Sazgarnejad, S. (2023). The application of artificial intelligence in health financing: a scoping review. *Cost Effectiveness and Resource Allocation*, 21(1):83.

Riascos, A., Romero, M., and Serna, N. (2017). Risk adjustment revisited using machine learning techniques. *Documento Cede*, (2017-27).

Rice, N. and Smith, P. C. (2001). Capitation and risk adjustment in health care financing: An international progress report. *The Milbank Quarterly*, 79(1):81–113.

Saunders, M. and Lewis, P. (2017). *Doing Research in Business and Management*. Pearson, 2nd edition.

Sharma, p. and Vidhya, A. (2021). Understanding k-means clustering algorithm. retrieved from analytics vidhya website:. *Analytics Vidhya*.

She, Z., Ayer, T., and Montanera, D. (2022). Can big data cure risk selection in healthcare capitation program? a game theoretical analysis. *Manufacturing & Service Operations Management*, 24(6):3117–3134.

Smith, J., G. T. . P. (2019). The role of preventive care in risk-adjusted capitation.

Stadhouders, N., Kruse, F., Tanke, M., Koolman, X., and Jeurissen, P. (2019). Effective healthcare cost-containment policies: A systematic review. *Health Policy*, 123(1):71–79.

Vuik, S. I., Mayer, E., and Darzi, A. (2016). A quantitative evidence base for population health: applying utilization-based cluster analysis to segment a patient population. *Population Health Metrics*, 14(1):44.

Xiao, J., Lu, J., and Li, X. (2017). Davies bouldin index based hierarchical initialization k-means. *Intelligent Data Analysis*, 21(6):1327–1338.

Xiaoqun, L. and Run, L. (2022). An improved k-means clustering model based on support vector machine for health insurance cost prediction. In *2022 IEEE 2nd International Conference on Electronic Technology, Communication and Information (ICETCI)*, pages 521–525.

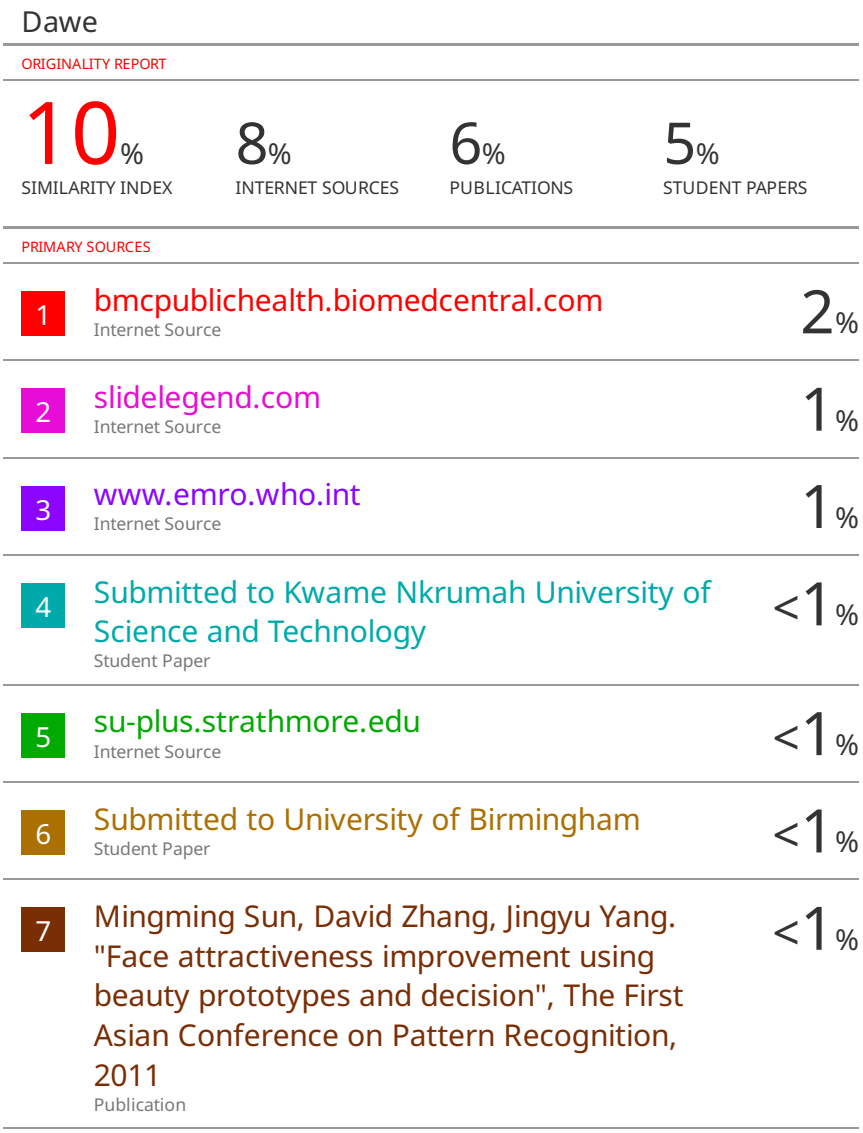
Yang, C., Delcher, C., Shenkman, E., and Ranka, S. (2018). Machine learning approaches for predicting high cost high need patient expenditures in health care. *BioMedical Engineering OnLine*, 17(1):131.

Zelmer, J. (2014). Patient clustering for risk adjustment.

Zhang, Y. (2017). Data-driven risk adjustment for episodic payments in health care.

Appendices

Appendix A: Similarity Report



Dawe

ORIGINALITY REPORT

10%	8%	6%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	bmcpublichealth.biomedcentral.com Internet Source	2%
2	slidelegend.com Internet Source	1%
3	www.emro.who.int Internet Source	1%
4	Submitted to Kwame Nkrumah University of Science and Technology Student Paper	<1%
5	su-plus.strathmore.edu Internet Source	<1%
6	Submitted to University of Birmingham Student Paper	<1%
7	Mingming Sun, David Zhang, Jingyu Yang. "Face attractiveness improvement using beauty prototypes and decision", The First Asian Conference on Pattern Recognition, 2011 Publication	<1%

Figure 6.1: Similarity report one

Appendix B: Ethical Clearance Confirmation

Strathmore UNIVERSITY

8th October 2024

Mr Gambo David,
dawe.gambo@strathmore.edu

Dear Mr Gambo,

RE: Risk-Adjusted Capitation as an Efficient Provider Payment Mechanism using K-Means Clustering in Machine Learning: A Case of National Hospital Insurance Fund

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** proposal. Your application reference number is **SU-ISERC2401/24.** The approval period is from **8th October 2024 to 7th October 2025.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu

Figure 6.2: Ethical clearance confirmation