MODELING SELF-HELP GROUPS' IMPACT ON MEMBERS' LIVELIHOODS IN MURANG'A EAST SUB-COUNTY USING MACHINE LEARNING TECHNIQUES

Jane Wangui Runo

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Statistics of Murang'a University of Technology

July, 2024

DECLARATION

Student Declaration

I hereby declare that this Thesis is my original work and to the best of my knowledge has not been presented for a degree award in this or any other university. I understand what plagiarism is and I am aware of the University's policy in this regard. I understand that any false claim in respect of this work shall result in disciplinary action in accordance with the University Anti-Plagiarism Policy.

_____ _____

Jane Wangui Runo Date

AS400/5690/2021

Supervisors' Approval

The undersigned certify that they have read and hereby recommend for acceptance of Murang'a University of Technology a thesis entitled **"Modeling Self-Help Groups' Impact on Members' Livelihoods in Murang'a East Sub-County Using Machine Learning Techniques".**

_____ _____

Dr. Ayubu Anapapa, PhD Date

Department of Mathematics and Actuarial Science

Murang'a University of Technology

_____ _____

Dr. Euna Nyarige, PhD Date

Department of Mathematics and Statistics

Machakos University

DEDICATION

To my beloved family. My husband Jesse and our children Claudia and Ethan for the wonderful support and encouragement through this journey, you made it easier and bearable.

My mother Ann, my brothers and sisters especially Systol and Elizabeth for their great support in my academic journey. God truly bless you.

ACKNOWLEDGEMENT

Gratitude to God for His Grace this far. To my supervisors, Dr. Ayubu Anapapa and Dr. Euna Nyarige who have provided great insights and assistance throughout this research, I am forever grateful. Dr. John Mutuguta, thank you for all the support. To my wonderful family, I am grateful for your support and faith in me, especially to my husband Jesse and our children Claudia and Ethan thank you for the support and encouragement.

ABSTRACT

Approximately 8.9 million people, or 17% of Kenya's population, live below the poverty line of 1.9 USD on a daily basis, majority of them in the rural areas. This research aimed to model the impact of self-help groups on livelihoods in rural areas of Kenya using machine learning techniques. To achieve this, the study employed Logistic regression, Naïve bayes algorithm and Support vector machine and compared their performance using various metrics to find out whether or not one of the three models outperformed the other two models. The study used principal component analysis to evaluate the most significant factors that affected group performance with an aim to ensure that the self-help groups achieved their original set objective. The study used primary data obtained through the issuance of structured questionnaires to self-help group members, on their wealth status since joining the self-help groups on areas such as ability to save, access to credit services and acquiring assets, both income generating and household assets. The target population was 2250 and the sample size was obtained using the Yamane's formula. The study's findings helped identify the key predictors of members' livelihoods and provided insights into how self-help groups influence them. The research also explored the factors that contributed to the success or failure of self-help groups. The results of the principal component analysis indicated that the frequency of the meetings and the quality of discussions held was the factor that contributed the most to group performance, as well as access to credit facilities. The results of logistic regression indicated that 90.9% of the members had seen a significant improvement on their wealth status since joining self-help groups and the significant predictor variables were income generating assets, access to basic commodities and access to loans. The model's accuracy was 88.04%. The SVM model classified 93.98% of the members as having an improved wealth status since joining SHGs with an accuracy of 84.62%. Finally, the naïve bayes algorithm classified 87% of the members as having a positive improvement on their wealth status with an accuracy of 92.34%. The results of this research are valuable to NGOs and practitioners in determining the effectiveness of self-help groups in promoting sustainable livelihoods and shaping future interventions on SHGs as well as on group members on how best they can make their groups more effective. Furthermore, the study's results provided more knowledge into the use of Machine learning methods in performing classification tasks and in identifying which method performed best in this specific task. This research contributed to the existing body of knowledge on the impact of self-help groups on rural livelihoods and the use of SHGs as a tool that can be used to reduce or completely eradicate poverty in the lives of people living in rural areas of Kenya. The comparison of the three models to find out which among them performed the best in the classification task was also a significant contribution to the existing body of knowledge. Naïve bayes algorithm outperformed the other two models.

TABLE OF CONTENTS

DECLARATION ..... **ii**
DEDICATION ..... **iii**
ACKNOWLEDGEMENT ..... **iv**
ABSTRACT ..... **v**
LIST OF TABLES ..... **viii**
LIST OF FIGURES ..... **ix**
ACRONYMS AND ABBREVIATIONS ..... **x**
DEFINITION OF TERMS ..... **xi**
CHAPTER ONE: INTRODUCTION **1**
1.1 Background Information **1**
1.1.1 Self Help Groups (SHGs) 1
1.1.2 Logistic Regression Model 1
1.1.3 Naïve Bayes Algorithm 2
1.1.4 Support Vector Machine Model 2
1.2 Problem Statement. **2**
1.3 Objectives **3**
1.3.1 General Objective 3
1.3.2 Specific Objectives 3
1.4 Hypotheses **3**
1.5 Justification of the Study **4**
1.6 Scope and Limitations of the Study **4**
1.7 Thesis Contribution **5**
1.8 Thesis Organization **6**
CHAPTER TWO: LITERATURE REVIEW **7**
2.1 Introduction **7**
2.2 Theoretical Framework **7**
2.3 Empirical Literature Review **8**
2.3.1 Self Help Groups 8
2.3.2 Logistic Regression Model 9
2.3.3 Naïve Bayes Algorithm 10
2.3.4 Support Vector Machine (SVM) Model 11
2.4 Conceptual Framework **14**
2.5 Chapter Summary **13**
CHAPTER THREE: METHODOLOGY **16**
3.1 Introduction **16**

3.2 Design of the Research **16**
3.3 Population of Interest **16**
3.4 Sampling Techniques **16**
3.5 Determination of the Sample Size **17**
3.6 Procedure for Data Collection **17**
3.6.1 Pilot Testing 18
3.6.2 Research Instruments Validity 18
3.6.3 Research Instruments Reliability 18
3.7 Procedure of the Research **19**
3.8 Analysis of Data **19**
3.8.1 Principal Component Analysis 19
3.8.2 Logistic Regression Model Fitting 20
3.8.3 Fitting a Naïve Bayes Algorithm 23
3.8.4 Fitting the SVM Model 24
3.9 Model Selection **25**
3.10 Model Validation **27**
3.11 Ethical Considerations **27**
3.12 Chapter Summary **27**
CHAPTER FOUR: RESULTS AND DISCUSSIONS 29
4.1 Introduction 29
4.2 Principal Component Analysis 29
4.3 Logistic Regression Model Results 31
4.4 Naïve Bayes Algorithm. 37
4.5 Support Vector Machine (SVM) Model 41
4.6 Model Selection 43
4.7 Test of Hypotheses 44
4.8 Chapter Summary 45
CHAPTER FIVE: CONCLUSION AND RCOMMENDATIONS 47
5.1 Conclusion 47
5.2 Recommendations 48
5.3 Recommendations for Future Work 48
REFERENCES 50
APPENDICES 52

LIST OF TABLES

Table 3.1: Confusion Matrix 26
Table 4.1: PCA Results 30
Table 4.2: Logistic Regression Model Results 32
Table 4.3: Logistic Regression Results for social demographic variables 34
Table 4.4: A-priori probabilities of Y 38
Table 4.5: Conditional probabilities for categorical variables 39
Table 4.6: Conditional probabilities for continuous predictor variables 41
Table 4.7: SVM Model Testing Parameters 42
Table 4.8: SHGs Training and Testing Evaluation Results for SVM Model 42
Table 4.9: Confusion Matrix metrics for the three models 43

LIST OF FIGURES

Figure 2:1Conceptual Framework.....13
Figure 3:1 Sigmoid functions.....21

ACRONYMS AND ABBREVIATIONS

CI-Confidence Interval

KNN- K-Nearest Neighbor.

LR-Logistic Regression.

ML- Machine Learning

MESC-Murang'a East Sub- County

NBA- Naïve Bayes Algorithm

NGO-Non-governmental Organization

PC-Principal Components

PCA-Principal Component Analysis

RBF-Radial Basis Function

RMSE- Root Mean Squared Error

SVM- Support Vector Machine.

SHG- Self-Help Group.

SRS-Simple Random Sampling.

DEFINITION OF TERMS

Livelihood- Livelihood refers to the resources, capabilities and actions that allow individuals to generate income and sustain their lives (Serrat, 2017). It encompasses the capacity to attain essential necessities such as food, shelter and clothing as well as access to education, credit facilities and investment opportunities.

SHGs- A collective of individuals who come together with a shared objective, which may be difficult to accomplish alone. These members work together to attain the common goal, supporting each other throughout the process. The functions of the group are determined by the purpose that the members have agreed upon and the group is managed and governed by the members themselves.

Machine Learning- Machine learning utilizes data and algorithms to simulate human learning processes. This involves machines replicating human-like intelligence by learning from their environment (El Naqa & Murphy, 2015). ML can be applied for tasks such as speech recognition, pattern recognition and more. The algorithms used in machine learning make predictions based on historical data that is fed as input.

Support Vector Machine – Vapnik introduced SVM (Vapnik et al., 1995), which is a ML model that includes learning algorithms for both data classification and regression (Horak et al., 2020).

Naïve Bayes Algorithm- Naive Bayes algorithm employs Bayes' rule and has the assumption of a high level of independence among variables (Webb et al., 2010). It is commonly used for classification tasks.

Logistic Regression- Logistic regression is a statistical method used to model the relationship between a binary outcome variable (i.e., one that takes on only two possible values, such as 0 or 1) and one or more predictor variables (LaValley, 2008).

Model- It is a mathematical representation of a system or process that is used to describe or explain the relationship between variables or to make predictions about future observations. Statistical models are typically formulated using mathematical equations or functions that describe the relationship between predictor variables and the outcome variable.

CHAPTER ONE
INTRODUCTION

1.1 Background Information

1.1.1 Self Help Groups (SHGs)

SHGs are informal groups of individuals who work together towards a common goal. These groups are formed voluntarily by individuals who have a good relationship with one another and a desire to support each other in achieving their goals. The nature of the groups' objectives may vary depending on the interests of its members. SHGs are self-regulated, with members selecting their leaders from within the group.

There is a significant portion of the population, amounting to 16.1%, who struggle to access basic needs and are considered to be living in poverty. In response, the government has implemented several initiatives, including the Supplemental Nutrition Assistance Program (SNAP), which aims to assist low-income families in obtaining food and ensuring food security.

This demonstrates the need to find ways and measures to assist this group of people living below the poverty line to achieve financial independence and sustainability in their way of living. Livelihood refers to the approach through which people acquire the essentials of life. It encompasses the means and activities undertaken to obtain basic and supplementary needs.

1.1.2 Logistic Regression Model

According to (Raghavan et al., 2016), logistic regression is utilized to identify a boundary line that can differentiate between the different categories within a dataset. The model comprises of a response variable and predictor variables, which jointly determine the outcome of the dependent variable. Logistic regression helped identify the significant predictor variables in predicting the outcome or response variable using the p-values. The output variable in logistic regression is in binary form, yes or no.

1.1.3 Naïve Bayes Algorithm

Naive Bayes is a straightforward and efficient method of machine learning that predicts data category and is suitable for both binary and multi-class classification problems, particularly when the dependent variable is categorical. It assigns the greatest likelihood to the correct class and can effectively handle small training sets, irrelevant features and still produce precise predictions.

1.1.4 Support Vector Machine Model

SVM was initially proposed by Vladimir Vapnik in 1995 (Vapnik et al., 1995), for classifying data. Its primary purpose is to classify data into different categories. It examines input and recognizes patterns within the input. The main objective of SVM is to produce a decision line, which is called a hyperplane, that best separates the two classes by maximizing the distance between them.

1.2 Problem Statement.

Every individual, regardless of their race, gender, background, or skin color, has the fundamental human right to access basic necessities such as food, clothing, shelter and education. It is crucial that people can satisfy their daily needs and live with dignity and respect.

Unfortunately, the reality is that many people around the world face significant challenges in affording these basic needs. In particular, access to education can be a major obstacle for children whose families are struggling financially. This can create a cycle of poverty and disadvantage that is difficult to break out of.

The challenges faced by individuals in meeting their basic needs and accessing resources and opportunities demand continuous efforts to establish sustainable livelihoods. One such solution is the formation of SHGs, which facilitate people's collaboration and collective efforts towards enhancing their livelihoods.

This research intended to analyze the impact SHGs have had in the lives and livelihoods of their members. A comparison of the three models using self-help groups data had not been done in any research. Also, SVM and naïve bayes algorithm had

been used for classification in other studies but not using self-help groups' data. Logistic regression was also used in analysis of self-help groups before but on women self-help groups. The success of some groups and failure of others necessitated the need to study the most significant factors that affected the performance of SHGs, which was done using the PCA analysis.

1.3 Objectives

1.3.1 General Objective

This study aimed to use Machine learning methods to develop a model that assessed how SHGs affected their members' livelihoods in Murang'a East Sub- County.

1.3.2 Specific Objectives

i. To establish the key factors that influence the performance of self-help groups in Murang'a East Sub- County.
ii. To evaluate the effectiveness of logistic regression in predicting self-help groups' impact on members' livelihoods in Murang'a East Sub- County.
iii. To model the impact of self-help groups on members' livelihoods in Murang'a East Sub- County using naïve bayes algorithm.
iv. To evaluate the effectiveness of SVM in predicting self-help groups' impact on members livelihoods in Murang'a East Sub- County.
v. To establish the best performing model among logistic regression model, naïve bayes algorithm and the SVM model.

1.4 Hypotheses

i. **H_0:** There are no factors that are more significant than others that affect how SHGs perform in Murang'a East sub-county.
ii. **H_0:** Logistic regression as a machine learning method is not suitable for predicting SHGs' impact on members' livelihoods in Murang'a East sub-county.
iii. **H_0:** Naïve bayes algorithm is not a suitable machine learning method for predicting SHGs' impact on members' livelihoods in Murang'a East sub-county.

iv. **H_0:** Support vector machine is not a suitable machine learning method for predicting SHGs' impact on members' livelihoods in Murang'a East sub-county.

v. **H_0:** There is no best performing model among the three models.

1.5 Justification of the Study

The use of SHGs is largely recognized as an important strategy for reducing poverty and promoting sustainable livelihoods. Rural areas of Kenya, including Murang'a East sub-county face various challenges related to sustainable livelihoods that could potentially be addressed through the use of SHGs. This study employed Machine learning methods such as the Naïve bayes, logistic regression and SVM to investigate the impact of SHGs on the livelihoods of their members. By comparing the results of these models, the study sought to identify the most effective method for modeling SHGs' impact on members' livelihoods. The findings of this research will be beneficial for members of the society, non-governmental organizations and other parties looking to promote sustainable livelihoods and reducing poverty in rural areas of Kenya, beyond just Murang'a East sub-county. Moreover, the study presented an updated approach to understanding SHGs' impact on members' livelihoods since the analysis made use of machine learning techniques.

1.6 Scope and Limitations of the Study

This study aimed to analyze the impact of self-help groups on the livelihoods of their members in Murang'a East Sub-County, Kenya, using machine learning techniques. The study also focused on exploring the factors that contributed to the success of these groups in promoting sustainable livelihoods including access to credit and training opportunities. To achieve its objectives, the study used data from self-help groups and their members to build a predictive model using machine learning algorithms. Machine learning algorithms were preferable to traditional methods since they are flexible, scalable and can handle large datasets.

The limitations of the study included bias from the respondents in their responses and personal perceptions they individually have concerning self-help groups and their impact. This was however tackled using the test-retest method which involved

administering the same test twice. This was achieved by conducting the pilot study and then the actual study and evaluating the accuracy of the information provided. In both the pilot study and the actual study, the questionnaire captured the information it was intended to capture consistently. In addition, members were not left to decide whether the impact was positive or not, but their responses to various questions were analyzed and this decision arrived at using the machine learning techniques, this played a major role in doing away with bias of the respondents.

1.7 Thesis Contribution

The contribution of this study was in the explanation of how SHGs affect the livelihoods of members in Murang'a East Sub-County. Through findings obtained in this study, valuable insights were gained on how these groups have impacted the economic and personal well-being of members in the region.

The study's findings provide valuable insights for members of the society, leaders in the community, policymakers and development organizations to create more effective interventions that support and strengthen SHGs in Murang'a East sub-county. It also helped identify the factors that should be emphasized when carrying out group operations to ensure the success of SHGs.

The utilization of machine learning methods in this research was significant as it demonstrated a new approach to examining the influence of SHGs. The application of machine learning techniques enabled a more comprehensive understanding of the connections between SHGs and their members, providing a more precise representation of these groups' effects on their livelihoods.

The findings offered significant contributions to understanding SHGs' impact on promoting community development in MESC. The results obtained from this study will guide the design and implementation of future efforts for strengthening and supporting SHGs. In addition, using machine learning algorithms allowed for the identification of patterns and relationships that may be difficult for humans to detect, allowing the research to have a more accurate position on the impact of the SHGs on livelihoods. Machine learning helped get rid of the biasness aspect as it looked at

various patterns which allowed for the drawing of the conclusion as to the impact of SHGs as opposed to when people get to decide what impact the groups have had. Based on the data the members gave, machine learning helped identify patterns and hence the conclusions were not fully based on human perception.

1.8 Thesis Organization

This thesis was divided into five sections. The first section included the introduction, which gave a brief description of the study and the tools used for analysis, the objectives, the hypotheses, scope and limitations of the study, significance of the study and thesis organization.

The second chapter detailed the literature review, both theoretical and empirical. The theoretical literature review detailed theories in SHGs, in logistic regression, NBA and the SVM models. The theories included social capital theory and empowerment theory, probability theory and the concept of odds.

The third chapter entailed the methodology and the steps that were followed in data collection and analysis. It explained the population of interest, size of the sample, data analysis, model selection and model fitting.

The fourth chapter included the results and discussion of the results after data analysis and chapter five which was the last chapter included the conclusion, recommendations and recommendations for future work.

CHAPTER TWO
LITERATURE REVIEW

2.1 Introduction

In this chapter, an in-depth examination of relevant literature pertaining to the research topic is presented. It is categorized into two primary parts. The initial section involves an empirical review of the literature on subjects such as SHGs, the logistic regression, the naive bayes and SVM. The second part provides a comprehensive conclusion of the literature by summarizing the findings obtained from prior research in this area.

2.2 Theoretical Framework

SHGs are often studied through the lens of community development and poverty alleviation strategies. Research emphasizes their role in enhancing access to financial services, promoting savings habits, and facilitating small-scale entrepreneurship among marginalized communities. Studies highlight how SHGs contribute to social capital formation, women's empowerment, and community resilience through collective action and mutual support networks. Theoretical frameworks commonly used include social capital theory and empowerment theory, which explore how participation in SHGs can lead to improved socio-economic outcomes and empowerment, especially among women and marginalized groups.

Logistic regression is a statistical model widely used in research to predict binary outcomes based on one or more predictor variables. In literature, logistic regression is applied across diverse fields including medicine, social sciences, and business, due to its ability to model the probability of categorical outcomes. Studies often focus on its advantages in handling categorical data and interpreting odds ratios to understand the relationship between predictors and outcomes. Theoretical basis includes probability theory and the concept of odds, as logistic regression estimates the likelihood of an event occurring given the predictor variables.

Support Vector Machines (SVMs) have gained prominence in literature for their robustness in handling both linear and non-linear classification tasks. SVMs are widely applied in fields such as pattern recognition, bioinformatics and finance, where data separability is crucial. Theoretical reviews of SVMs focus on the mathematical

principles behind kernel methods and margin maximization, which form the basis of their classification ability.

Naive Bayes algorithms are probabilistic classifiers that assume independence among predictors. In literature, Naive Bayes is appreciated for its simplicity and efficiency in text classification, spam filtering and medical diagnosis, where computational speed and predictive accuracy are paramount. Theoretical discussions focus on Bayesian probability theory and conditional independence assumptions, critiquing the "naive" assumption of predictor independence against real-world data complexities.

2.3 Empirical Literature Review

2.3.1 Self Help Groups

According to (Brody et al. 2013), a SHG is a group of individuals who voluntarily unite with a shared purpose and work together to achieve that purpose. Self Help Groups (SHGs) are created to cater to the needs of underprivileged and marginalized individuals who can work together to achieve common goals. According to (Deininger and Liu, 2009), Self-Help Groups (SHGs) have become increasingly popular in India, particularly at the grassroots level and have had a beneficial effect on the economic empowerment and access to fundamental needs of their members.

A research study conducted in Bhandara District, India by Sukhdeve (2011) found that rural SHGs had a positive impact on the acquisition of household assets, with all participants reporting an increase. The study found that 76% of SHG members experienced an improvement in their homes' structures and access to electricity, resulting in greater comfort. Additionally, 66% of respondents were able to accumulate savings and 96% reported an increase in their business incomes. According to these findings, the study suggested that SHGs should consist of members of the same age group who share a common goal and receive regular training to stay informed of available services. The study employed logistic regression, which correctly classified 67% of women, indicating that the model was well-fitted to the data.

(Alrefaei et al., 2022) conducted a study on SHGs in India that aimed to reduce poverty and promote empowerment. The study identified key features of SHGs, including savings, democratic leadership and offering loans at lower interest rates than other financial institutions. The study found that SHGs contributed to economic development, created job opportunities and improved the quality of life.

The study conducted by (Jose et al., 2019) aimed to identify the problems encountered by women SHGs in Ernakulam District, India. The data was collected through questionnaires using the SRS technique and the challenges identified included illiteracy, inadequate training and a need for government and voluntary groups to work together to address these issues.

2.3.2 Logistic Regression Model

According to (LaValley 2008), logistic regression involves analyzing binary outcomes, where the dependent variable is limited to either 0 or 1. Independent variables can be either categorical or continuous. The logistic model predicts the likelihood of an event happening, taking into account multiple predictor variables.

Sukhdeve (2011) utilized logistic regression to investigate how SHGs impacted women empowerment in Bhandara District of Maharashtra. The form of the logistic regression model used in the study was:

$$P = \frac{1}{1 + \exp(-X)} \qquad (2.1)$$

The linear regression model below is used to explain the relationship between the predictor variables X and the dependent variable P.

$$P = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \qquad (2.2)$$

An accurate classification of 67% of rural women was made by the model, suggesting that it fit the data well.

Othman et al. (2012) employed logistic regression as one of the models in their study on Malaysian cooperatives, to identify factors that influenced cooperative membership

and share increment. The LR model was used to assess independent variables including age, gender and income level on the dependent variable, which was the decision of an individual to join a cooperative or increase their shares. The study found that age and occupation were significant predictor variables in the logistic model and that the model was a suitable fit for the data. The conclusions of the study were that the cooperative performance and progress were dependent on trust and member commitment and a recommendation to reach more young people was made for more growth of the Malaysian cooperatives.

(Magali 2013) employed logistic regression to examine the influence of Savings and Credit Cooperative Organizations or SACCOs and credit cooperatives on various factors in Tanzania. The study used health, education, new assets, income, business capital and renovation or building cost as independent variables. The logistic regression analysis revealed that SACCOs had a positive impact on health, education, purchasing new assets, crop yields and business capital. 73.5% of SACCO members had households' improvements, 50% experienced an improvement in health and 80% saw an improvement in physical assets. The study concluded that logistic regression fit the data well.

2.3.3 Naïve Bayes Algorithm

According to (Webb and Keogh 2010), it is an algorithm that is mainly used for classification and that utilizes Bayes rule with the assumption that the predictors are not related to one other. Although this independence assumption may not always be true in practice, the algorithm generally produces highly accurate results in its classification task.

Naïve Bayes Classifiers assume that features are independent of each other and they are built on Bayes Theorem (Webb et al., 2010). They are supervised learning algorithms that use labeled data to train, classify and predict outcomes. These classifiers are scalable and work well when there are only two possible outcomes, but their performance decreases as the number of outcomes increases (Rish, 2001).

(Krichene 2017) employed the Naïve Bayes Classifier to assess the risk of default. The predictor variables used in this analysis included working capital, cash flow indicators

and solvency. The dependent variable was the probability of default, which was classified as either healthy or unhealthy based on the output variable. The study achieved 59.63% accuracy in correctly classifying healthy firms and 40.37% accuracy in classifying unhealthy ones. Type I error was identified as a significant issue, where bad customers were falsely classified as good, while Type II error was the classification of reliable clients as unreliable. To improve the model's performance, the study introduced cash flow variables, resulting in an improvement in accuracy from 59.63% to 63.85%. However, the study only utilized quantitative variables and it was recommended that further research should be conducted to collect qualitative data.

(Okesola et al. 2017) employed the Naïve bayes algorithm to develop a bank scoring model that was improved, which was used to categorize the payment status of loans over a 3-month period as either good or bad. The dependent variable used was a binary variable that was coded 0 for bad payment status and 1 for good payment status. The sample size used for the study was 680 loans and the naïve bayes model demonstrated higher accuracy compared to the KNN model, achieving 83.3% accuracy while KNN only achieved 82.3%. In this study, the naïve bayes algorithm was found to effectively handle categorical variables.

Naïve Bayes is an appropriate algorithm for classification tasks, making it a relevant approach for this study. However, its practical use in the area of SHGs is restricted, making it a suitable choice for this research, as there is insufficient knowledge regarding its use in this specific field.

2.3.4 Support Vector Machine (SVM) Model

Vapnik is credited with creating the SVM in 1995 (Vapnik, 1995). The SVM is a machine learning method that is utilized for classification and regression analysis (Horak & Vrbka, 2020).

(Amzile et al. 2022) analyzed the performance of SVM in forecasting defaults and determining credit scores. The SVM algorithm was used to classify customers into two categories: creditworthy and not-creditworthy, based on a dataset of 3,000

customers. An SVM model was used to analyze the factors that contribute to default in revolving loans. The variables considered in the analysis included the client type (private or public), number of children, age, income and education level. The model achieved a 71.7% accuracy rate in correctly predicting loan defaults, which was verified through the creation of a confusion matrix.

The study trained the model on 2,250 members, assuming that the data was both linearly separable and non-linearly separable. The SVM-RBF kernel performed best with a 98% accuracy; support vector machine-linear achieved a level of accuracy of 62%. The researcher had a hard time choosing the best SVM model and had to test all of them to determine which among them was optimal.

(Harris 2015) conducted a study on credit scoring using SVM model. Prior to the analysis, pre-processing of the data was done to convert categorical variables into numerical variables and then normalization was done to enhance the performance of the model. Subsequently, partitioning of the data into two sets, 20% was the testing set while the training and cross-validation set making up the remaining 80%. Age, loan amount, loan purpose, employment status, monthly income and expenditure per month were among the variables utilized in the study. SVM was deemed suitable for analyzing the data, but it was observed that as the dataset grew, nonlinear methods became computationally costly.

Although SVM has been applied in various classification and regression problems, its application in the analysis of SHGs is limited. This research aims to utilize SVM to analyze and study SHGs, thus contributing new insights to the field.

2.4 Conceptual Framework

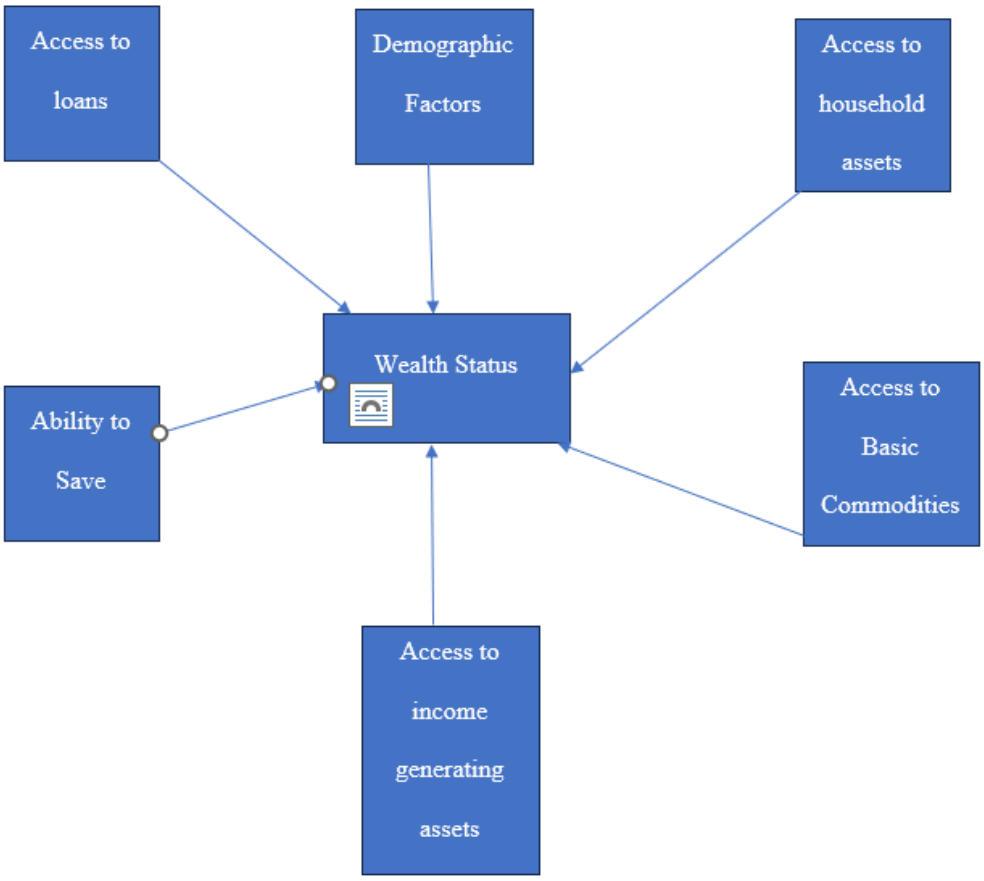

Figure 2:1Conceptual Framework

Wealth status - It refers to the financial condition or net worth of individuals or households. It reflects the economic well-being and financial stability of a person or family. Wealth status is influenced by various factors including income levels, savings habits, investments, property ownership, access to credit and economic opportunities.

Ability to save- It refers to an individual's capacity to set aside income for future use rather than immediate consumption.

Household assets- They are tangible and intangible resources owned by individuals or families, including property, vehicles, savings, and personal belongings. These assets played a crucial role in determining household wealth.

Income generating assets- The resources or investments that generated revenue or income for individuals or households. The ownership and management of income generating assets contributed significantly to financial security and wealth accumulation.

Basic commodities- They encompassed essential goods and services necessary for daily living, including food, clothing, shelter, and utilities. Access to these commodities was fundamental for maintaining a decent standard of living and overall well-being.

Access to loans- It refers to an individual or entity's ability to obtain financial credit from formal or informal sources. It was crucial for economic activities such as investment, consumption and entrepreneurship.

Demographic factors- They encompassed characteristics such as age, gender, marital status, and educational attainment within the population. These factors provided insights into the composition and diversity of individuals within the given context.

2.5 Chapter Summary

In conclusion, after conducting the analysis on literature, it was clear that there had been effective attempts made to use machine learning in various analysis involving SHGs, particularly logistic regression. SVM and NBA have been used in various classification tasks successfully. However, rather than using data from SHGs, SVM and NBA have been used to analyze data in banks and cooperatives successfully in

areas such as categorizing customers as credit worthy or not. Logistic regression was used to analyze SHGs data in India, but the analysis was purely on the impact SHGs have had on women SHGs. The machine learning methods described in the literature review for modeling SHGs impact on members' livelihoods have not been compared in any research involving classification using SHGs data.

The support vector machine and the naïve bayes models have been found to be suitable for classifying data, particularly when the output is binary. It is evident that SHGs impact has been positive in many aspects, but not all groups have experienced the same level of success and some were started and they failed to achieve the purpose for which they were created. This study aimed to identify the most significant factors leading to the success of SHGs in Murang'a East sub-county. This was achieved by issuing a questionnaire on factors that affect SHGs and then analyzing the responses to determine which factors most significantly affected the performance of SHGs. This research contributed to improving the effectiveness of SHGs in MESC and beyond as well as providing a better understanding on ML methods that were recommended in this study.

CHAPTER THREE
METHODOLOGY

3.1 Introduction

This part described the necessary steps to carry out the research, which involved identifying the intended group of participants, choosing suitable methods for selecting samples, determining the appropriate sample size and specifying the tools and approaches for collecting and analyzing data.

3.2 Design of the Research

This study utilized the quantitative research design, the design involved gathering data from the population using questionnaires. The analysis of the collected data allowed conclusions to be drawn.

3.3 Population of Interest

The scope of this research was to survey the SHG members located in Murang'a East sub-county. The study utilized questionnaires to gather relevant data from the individual members, which when analyzed provided insights on the impact of SHGs on the livelihoods of their members. The population size was 2250.

3.4 Sampling Techniques

A report obtained from Murang'a East sub-county social services indicated that there were 150 registered and active SHGs in the area, with each group consisting of at least 10-15 members.

To select SHGs and members for this research, the study used a sampling technique known as stratified sampling to obtain sub-groups (strata) from the population and then selected samples from each stratum. There were 11 areas that were considered i.e., Thangathi, Thaini, Kiangatia, Githuthi, Kambirwa, Township, Mbiri, Gakunja, Kimathi, Mirira and Gikindu. Each area was assigned 13 groups except Township ward which was assigned 20 groups due to its high population. 6 members, above half

the minimum number of group members, were selected per group. The study used SRS to select individual members within each stratum.

3.5 Determination of the Sample Size

The findings obtained from a sample are used to make inferences about the whole population. In this study, the population (N) was composed of 2250 individuals, divided into 150 groups, with at least 15 members per group.

Yamane's formula (1967) allowed for the calculation of a representative sample size. The formula was found most effective since it did not assume normality of data and also because the population size was known. Other formulas such as Mugenda and Mugenda was not effective because of its assumption of normality and is also not effective with population sizes less than 10,000. The Slovin's formula works best when the population size is unknown, hence the conclusion to use the Yamane's formula.

$$n = \frac{N}{1+N(e)^2} \qquad (3.1)$$

Where:

$n =$ size of the sample

$N =$Target population

$e =$ Margin of error. To calculate the size of the sample for this study, an error margin of 0.025 was applied. Using a smaller margin of error resulted in a larger sample size.

$$n = \frac{2250}{1 + 2250(0.025)^2} = 935 \qquad (3.2)$$

By this, the size of the sample for this study was determined to be 935 individuals.

3.6 Procedure for Data Collection

To achieve the objectives of this research, collection of data involved obtaining and analyzing responses from the target population. Primary data, collected through the use of questionnaires was used. The questions were in the form of a Likert scale and were used to investigate the impact of groups on the individual aspects of the

respondents' lives. The anonymity provided by the questionnaires gave respondents the freedom to answer questions with reduced bias.

3.6.1 Pilot Testing

Pilot testing involves testing the research approach with a smaller sample than the one to be used in the main study Bager et al., (2018). It involved doing what was to be done in the main study but with a smaller sample size to test how effective the research procedures were. 10% of the self-help groups in Murang'a East Sub-County were used in the pilot study. The results in the pilot study indicated that the research instrument captured what it was intended to capture.

3.6.2 Research Instruments Validity

Validity refers to the degree to which a research instrument measures what it is intended to measure. The validity was measured using face-validity. Face validity is based on subjective judgement and it formed a basis for further investigations on validity such conduction of the pilot study. Cross-validation was also used and it ensured that the machine learning models accurately classified the data. This involved splitting the data into training and testing sets and evaluating the performance of the model on the testing set.

3.6.3 Research Instruments Reliability

Ensuring the reliability of a research instrument is crucial to obtaining accurate and valid results. Reliability refers to the consistency and stability of the instrument's measurements over time and across different samples. Several methods to assess the reliability of research instruments exist and include test-retest reliability, which involves administering the same instrument to the same group of participants at two different times and comparing the results. This was achieved by carrying out a pilot test before carrying out the main data collection. The questionnaire was found to be consistent in its measurement.

3.7 Procedure of the Research

The research procedure involves the several steps taken by the researcher to gather the necessary data for a study. In this study, primary data was obtained by distributing questionnaires to the targeted population. Once the data was collected, it was analyzed using machine learning methods i.e., SVM, Naïve Bayes and Logistic Regression.

The questionnaire used in this study implemented a rating scale ranging from 1 to 5 to assess the participant's responses. The responses collected from the questionnaires were analyzed using the above-mentioned machine learning methods. The results obtained from the sample analysis were utilized to evaluate the impact of SHGs on their members.

3.8 Analysis of Data

Analysis of data involves refining and organizing unprocessed data to derive meaningful insights that aids in making informed decisions and drawing conclusions. This study employed the inferential and descriptive statistics. Descriptive statistics consist of measures of central tendency, frequency and dispersion and their purpose is to summarize and present the characteristics of the study sample, as stated by (Fisher et al. 2009). In contrast, inferential statistics utilize descriptive statistics to draw inferences and make conclusions that are applicable to the entire population.

Once the data had undergone the process of eliminating errors and redundant information, it was subjected to analysis using the R programming language to generate outputs that were relevant for drawing conclusions.

3.8.1 Principal Component Analysis

To determine which was the most significant factor that affected how SHGs performed in MESC, the factors affecting performance of SHGs were subjected to a PCA analysis. PCA is a reduction technique used in machine learning to transform data to fewer dimensions but when retaining the most important information. The original data is transformed to principal components. The analysis of the factors affecting SHGs was done using the PCA because the study wanted to find out which among the factors was the most significant and thus there was no response variable, just a comparison of the factors.

After the PCA was fit in R, the output was standard deviation, loadings, proportion of variance and cumulative proportion. The standard deviation of each principal component (PC) is an indication of the spread of the data along that component. Each principal component is a linear combination of the original variables. The proportion of variance represents the proportion of the total variance in the data that is captured by each principal component while the cumulative proportion represents the proportion of the total variance in the data that is captured by the principal components.

Loadings represent the relationship between the PC and the original variables. They represent the contribution of the original variables to the PCs. A high positive or negative loading for a variable indicate a high positive or negative relationship between the PC and the variable.

3.8.2 Logistic Regression Model Fitting

Logistic regression is a technique that is utilized when the outcomes are binary, where the anticipated value is either 0 or 1. In this method, each independent variable in the model is assigned a coefficient that gauges its effect on the output variable. These coefficients aid in comprehending the association between the independent variables and the binary outcome.

In this study, the output variable Y (wealth status) had two possible values: 1 meaning "improved" and 0 meaning "not improved". The Y variable represented the wealth status and the study aimed at evaluating whether the wealth status had improved or it had not. The coefficients assigned were denoted by β. The first coefficient, β_0, was the intercept, while each variable from $1 \ldots n$ were each assigned the coefficients, $\beta_1, \beta_2 \ldots \beta_n$. They determined the association between the independent and dependent variables. The result obtained from the analysis were in the form of odds ratio;

$$\frac{P}{1-P} = exp(\beta_0 + \sum_{i=1}^{n} \beta_i X_i) \qquad (3.3)$$

The probability of the occurrence of the event was denoted by P.

In this research, multiple logistic regression was employed since the prediction of a single binary output was done using multiple variables.

The equation for the multiple logistic regression function is:

$$P(x) = \frac{exp(\boldsymbol{X'\beta})}{1 + exp(\boldsymbol{X'\beta})} \qquad (3.4)$$

In machine learning, the logistic sigmoid function is widely used to transform any input value to a probability value that ranges between 0 and 1. This is different from the hyperbolic sigmoid function which transforms data values to a range between -1 and 1 and the arctangent function which transforms data values to a range between $\frac{-\pi}{2}$ and $\frac{\pi}{2}$. Below is a graphical representation of the common sigmoid curves.

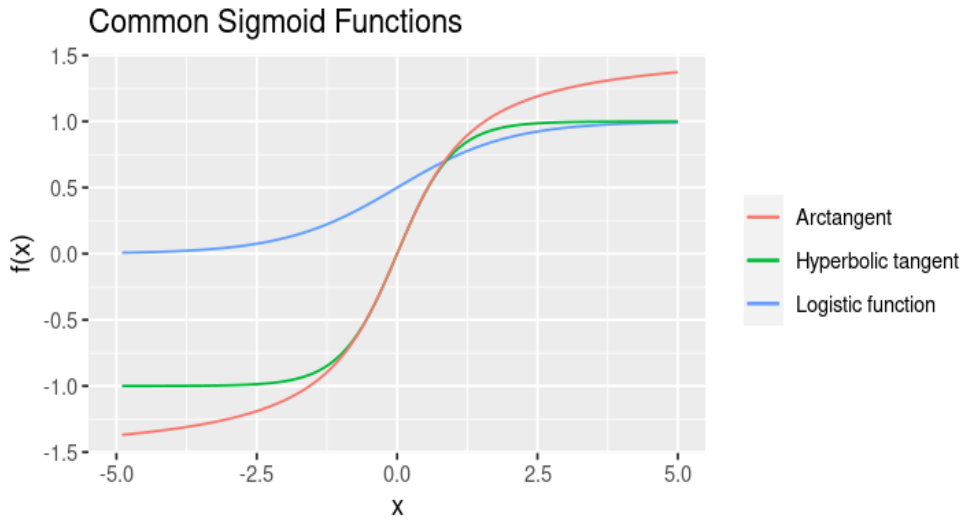

Figure 3:1 Sigmoid functions (Sigmoid Function Definition | DeepAI, n.d.)

The formula below represents the logistic sigmoid function:

$$P(x) = \frac{1}{1 + \exp(-x)} \qquad (3.5)$$

The LR model is a powerful tool for data analysis that is simple to understand and interpret, computationally efficient and can handle datasets that are large.

Additionally, updating new data is easy and it can handle outliers and data that is missing without requiring data preprocessing.

LR has assumptions such as the requirement for a binary or dichotomous outcome variable and a linear relationship between the predictor and outcome variables.

3.8.2.1 Logistic Regression Parameters Estimation

The maximum likelihood estimator is a statistical approach that is employed to estimate the parameters of a statistical model. It strives to identify the parameter values that are most likely to account for the observed data, given a specific set of observations. When it comes to LR, the goal is to develop a model that describes the relationship between a binary dependent variable and a set of independent variables. The form of the logistic regression is;

$$P(x) = \frac{exp(\boldsymbol{X'\beta})}{1 + ex\,p(\boldsymbol{X'\beta})} \tag{3.6}$$

Below is a representation of the likelihood function;

$$l(\beta, Y = 1/X = x_i) \tag{3.7}$$

$$= \pi \left[\frac{exp(\boldsymbol{X'\beta})}{1 + exp(\boldsymbol{X'\beta})}\right]^{\sum_i^n y_i} \left(1 - \left[\frac{exp(\boldsymbol{X'\beta})}{1 + exp(\boldsymbol{X'\beta}))}\right]^{n - \sum_i^n y_i}\right)$$

Where;

$l(\beta, Y = 1/X = x_i)$ =The likelihood function

π_i =Probability of observing the i^{th} observation

X=Predictor variables

β=coefficients of the logistic regression model

y_i=Outcome for the i^{th} observation which is either 0 or 1

n=Total number of observations

3.8.3 Fitting a Naïve Bayes Algorithm

According to Webb et al. (2010), Naïve Bayes Classifiers are based on Bayes Rule and assume that the independent variables are independent of one another. It is a supervised machine learning algorithm that requires a labeled dataset to train, classify and predict outcomes. This theorem enables the calculation of the likelihood of a particular class based on a several independent variables by multiplying the probability of those variables given the class with the prior likelihood of the class and then dividing by the marginal probability of the features. It is represented as follows;

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \quad (3.9)$$

Where;

$Y =$ wealth status

$X =$ The group of independent variables used in the analysis include access to household assets, access to basic commodities, access to income generating assets and ability to save.

P(Y|X) = the posterior probability of the class Y given the input features X.

P(X|Y) = the likelihood of the input features X given the class Y.

P(X) = the marginal likelihood of the input features X.

P(Y) = the prior probability of the class Y

Types of Naive Bayes algorithm (NBA) include: Gaussian Naïve Bayes, Multinomial Naïve bayes and Bernoulli Naïve bayes. Gaussian Naïve bayes is used when the variables are continuous and normally distributed. Multinomial NBA is used when the features are discrete and follow a multinomial distribution. Bernoulli NBA is used when the features are binary, meaning they can only take two values. The algorithm calculates the likelihood of each variable being present or absent in each class. The NBA is based on Bayes' theorem, which uses prior knowledge and conditional probability to determine the probability of an event occurring.

It is computationally efficient because of its simplicity and requires minimal computational resources. Naive Bayes makes the assumption that features within a

class are independent and unrelated. Although the independence assumption may not always hold true in real-world applications, the naïve bayes classifier is still effective in solving real-life problems.

3.8.4 Fitting the SVM Model

SVM is a machine learning model which takes input data points and produces a decision boundary that separates the points into different categories. The boundary is created by finding the best line or hyperplane that maximizes the margin between the different categories, ensuring that the data points are well-separated. Any new data points can then be classified based on which side of the decision boundary they fall on. This allows for clear and accurate classification of the input data.

Support Vector Machines (SVMs) can be classified into two types: Linear SVMs, which can separate data into two groups using a straight line and non-linear SVMs, which are used when data cannot be separated by a single straight line. When a straight line is insufficient, non-linear SVMs employ a technique known as the "kernel trick" to address this issue. The kernel trick allows SVMs to establish a non-linear decision boundary.

SVM searches for multiple hyperplanes that classify the inputs in the most optimal way and then selects the hyperplane with the maximum distance from the datapoints. However, SVM may not be efficient for large datasets since it requires high training time. Additionally, it makes an assumption that the margin between the hyperplane and the nearest datapoints should be maximized.

SVM utilizes the mathematical function known as the kernel function to change input data into the appropriate format for analysis. The kernel function varies from one SVM function to another. This research employed SVM with the radial basis function (RBF) kernel. The input variables for the SVM model were access to household assets, access to income-generating assets, access to basic commodities and the ability to save, while the output variable was wealth status. The SVM model was trained using the radial basis function (RBF). The RBF kernel is particularly useful for capturing complex, non-linear decision boundaries. The formula for the RBF kernel is as given below.

$$K(x, y) = exp(-\gamma \parallel x - y \parallel^2) \quad (3.11)$$

Where;

$\gamma = Positive\ Parameter$

$\parallel x - y \parallel = Distance\ between\ the\ points\ x\ and\ y$

The objective of this research was to use SVM (Support Vector Machine) to establish a decision line for evaluating the impact of SHGs on wealth status. The analysis involved classifying individuals into two categories based on their association with SHGs: those who have benefited from the self-help groups and those who have not. The analysis identified a hyperplane that best separated the data points, with the optimal hyperplane being the one that maximized the separation between the two categories.

3.9 Model Selection

In order to select the best m among the three, the research utilized a confusion matrix. Several metrics are typically used to assess the performance of a classification model, including sensitivity, precision and specificity. Sensitivity, representing true positives (TP) to the total actual positives (TP+FN), precision, representing true positives to the total predicted positives. Specificity is a metric measuring true negatives to the total actual negatives, whereas accuracy refers to the proportion of correct predictions out of the total number of predictions made by a classification model. To evaluate the performance of the models, the study summarized the actual class labels versus the predicted class labels in a confusion matrix consisting of four cells: TP (True positive), FP (False positive), TN (True negative) and FN (False negative).

Binary classification confusion matrix is shown below:

Table 3.1: Confusion Matrix

	Actual Positives	**Actual Negatives**
Positives Predicted	True Positives	False Positives
Negatives Predicted	False Negatives	True Negatives

The true positives (TP) indicate the cases correctly predicted as positive outcomes; the false positives (FP) indicate the cases incorrectly predicted as positive outcomes. The true negatives (TN) indicate the cases correctly predicted as negative outcomes while the false negative (FN) indicates the cases incorrectly predicted as negative outcomes.

$$Accuracy = \frac{TP + TN}{TN + TP + FN + FP} \quad (3.12)$$

$$Precision = \frac{TP}{FP + TP} \quad (3.13)$$

$$Sensitivity = \frac{TP}{FN + TP} \quad (3.14)$$

$$F1\ Score = \frac{2 \times (sensitivity \times precision)}{(sensitivity + precision)} \quad (3.15)$$

Precision is a metric used to measure the percentage of correctly predicted positive instances in relation to the total number of instances predicted as positive. Sensitivity, which is also called recall, is the ratio of the accurately predicted positive instances to the actual number of positive instances in the dataset. Accuracy is the percentage of instances classified correctly, whether positive or negative, compared to the total number of instances in the dataset. The F1 score is a measure of performance that provides a balanced evaluation of a model's precision and sensitivity by taking the harmonic mean of both metrics.

3.10 Model Validation

Model validation ensures that the model is accurate, reliable and can be used to make accurate predictions in real-world applications. The process involves splitting the available data into training and testing datasets, training the model on the training data and evaluating its performance on the testing data. Performance metrics such as accuracy, precision, recall, or F1-score were used to compare the predicted values to the actual values. Model validation is a critical step in building reliable predictive models and it ensures that they are robust and can be used to make accurate predictions in real-world applications.

3.11 Ethical Considerations

Ethical considerations are an essential aspect of any research project that involves human subjects.

This research ensured that the welfare and privacy of the participants was protected. The ethical considerations surrounding this research included consent from respondents, confidentiality, privacy and avoiding harm to participants. Informed consent from the participants was obtained before conducting the research. This means that participants were informed of the research's purpose and benefits before agreeing to participate.

Confidentiality and privacy were also crucial ethical considerations in this research and the steps to protect the participants' identities and personal information were taken ensuring that this information was not disclosed. Additionally, this research ensured that it did not cause any harm to participants, whether it be physical, psychological, or emotional.

The study also ensured there was no coercion or pressure to participate in the study but that it was absolutely by their own choice.

3.12 Chapter Summary

This chapter provided a comprehensive analysis of the research process, explaining the steps involved in conducting the study. It began by identifying the target population, SHG members in Murang'a East sub-county and applied the quantitative research design using questionnaires to gather data. Stratified sampling was employed

to select SHGs from various areas, followed by simple random sampling of individual members, resulting in a sample size of 935 determined using Yamane's formula.

Data collection involved anonymous Likert-scale questionnaires to ensure unbiased responses. Validity and reliability of the research instruments were established through face validity, cross-validation and pilot testing. Data analysis was conducted using R programming, employing methods such as PCA, logistic regression, Naïve Bayes and SVM to make conclusions and evaluate SHG impacts.

Model performance was assessed using a confusion matrix and validation ensured accuracy and reliability. Ethical considerations included informed consent, confidentiality, privacy and voluntary participation without coercion.

CHAPTER FOUR
RESULTS AND DISCUSSION

4.1 Introduction

Data analysis in this study was done using the R environment because of its wide library of libraries and ease of implementation. The version of R used was 4.3.1. The fitting of the models, the outcomes of the techniques used as well as the best performing model among the three are explained in this chapter. Every model's result is displayed.

4.2 Principal Component Analysis

To determine which was the most significant factor that affected how SHGs performed in MESC, the factors stated above, B1-B9 were subjected to a PCA analysis. This output is tabulated below.

Table 4.1: PCA Results

	Standard Deviation	Proportion of Variance	Cumulative Proportion	Variable	Loadings
PC1	1.3378	0.1988	0.1988	B1	0.2645
PC2	1.0184	0.1152	0.3141	B2	0.2857
PC3	1.0047	0.1121	0.4262	B3	0.2323
PC4	0.9692	0.1044	0.5306	B4	0.2985
PC5	0.9557	0.1015	0.6321	B5	0.2656
PC6	0.94066	0.09832	0.73040	B6	0.2864
PC7	0.93166	0.09644	0.82684	B7	0.2802
PC8	0.86100	0.08237	1.00000	B9	0.2992

A high standard deviation in PCA implies that the data points are more dispersed. Components with high standard deviations are considered significant because they capture a large portion of the variability in the data. Components with high proportions of variance are also considered significant because they explain a large percentage of the total variability in the dataset.

The loadings provided above represented the average contribution of each variable in every principal component. The variable with the highest loading was the most significant variable because it had the highest relationship with the principal components.

From the loadings above, the order in which they appeared from highest to lowest was B8 (0.2992), B4 (0.2985), B6 (0.2864), B2 (0.2857), B7 (0.2802), B5 (0.2656), B1 (0.2645), B3 (0.2323). The variable with the highest loading is the variable with the highest significance, meaning that it is the factor that has the greatest impact on the performance of SHGs in MESC.

B8= The frequency of the meetings and the quality of the discussions held during the meetings is a significant factor in the performance of SHGs in Murang'a East sub-county.

B4= Access to financial resources and credit facilities is a significant factor in the performance of SHGs in impacting lives in Murang'a East sub-county.

B6= Collaboration and partnerships with other organizations and stakeholders is a significant factor in the performance of SHGs.
B2= The level of participation and engagement of self-help group members is a significant factor in impacting lives in Murang'a East sub- County.
B7= The level of community support and involvement is a significant factor in the performance of SHGs in impacting lives in Murang'a East sub-county.
B5= The presence of supportive policies and regulations is a significant factor in the performance of SHGs in Murang'a East sub-county.
B1= The availability of training and capacity-building programs is a significant factor in the performance of SHGs in Murang'a East sub-county.
B3= The presence of effective leadership and management structures is a significant factor in the performance of SHGs in Murang'a East sub-county.
From the PCA results, the frequency of the meetings and the quality of the discussions held was the most significant factor that affected how SHGs performed in MESC. Access to financial resources and credit facilities was the second most significant factor that affected the groups' performance. The order of the factors with their significance is stated above.

4.3 Logistic Regression Model Results

Logistic regression model fits data with predictor variables and a dependent variable which is in dichotomous form. This section gives the results of the predictor variables when fit in the logistic regression model. Predictor variables such as access to household assets, access to income generating assets, access to loans, access to basic commodities and ability to save were fit to the model to investigate their relationship with the dependent variable, wealth status. The R software was used in the analysis. Standard errors, z-values and p-values which determine the significance of the predictor variables are shown below. Logistic regression in R was fit using the glm function.

The table below shows the results when the predictor variables were fit in the logistic regression model using the R environment. The table shows the estimated coefficients, the std error, the z-value, the p-value and the odds ratio and the 95% CI.

Table 4.2: Logistic Regression Model Results

Variable	Estimate	Std Error	z-value	p-value	OR
Intercept	-5.9973	1.8858	-3.170	0.001526	0.0025
Household Asset	-0.0607	0.2196	-0.276	0.782463	0.9412
Income Generating Assets	0.6633	0.1853	3.580	0.000344	1.9411
Basic commodities	0.5154	0.1921	2.684	0.007284	1.6743
Loans	0.8611	0.2522	3.415	0.000638	2.3658
Savings	0.2494	0.2631	0.948	0.343176	1.2832

Out of the fit predictor variables, access to income generating assets, access to basic commodities and access to loans were found to have significant impact on the dependent variable, wealth status. Savings and access to household assets did not have a significant impact on the dependent variable.

The intercept was -5.9773 ($OR = 0.0025, 95\%\ CI: -9.72\ -2.31$). It represented the log-odds of the response variable (wealth status) when all other predictor variables were set to 0. The coefficient for access to household assets was -0.0607 ($OR = 0.9412, 95\%\ CI: -0.497\ 0.366$). A one-level increase in the variable access to household assets was associated with a decrease of 0.0607 in the log-odds (probability of the event occurring to the probability of the event not occurring) of the response variable. However, it was not statistically significant (p-value = 0.782463) the 95% CI also included zero indicating that the coefficient was not statistically significant at 0.05 level of significance and hence was removed from this study. The coefficient for access to income generating assets was 0.6633 ($OR = 1.9411, 95\%\ CI: 0.299\ 1.027$).The 95% CI also did not include zero indicating that it was statistically significant and the fact that it was positive showed that the relationship between this variable and the outcome variable was positive, an increase

in income generating assets lead to an increase in wealth status. A one-level increase in the variable access to income generating assets was associated with an increase of 0.6633 in the log-odds of the response variable. It was statistically significant (p-value = 0.000344) and thus had a significant impact on the response variable. The coefficient for access to basic commodities was 0.5154 ($OR = 1.6743, 95\%\ CI$: 0.1321 0.888). A one-level increase in the variable access to basic commodities was associated with an increase of 0.5154 in the log-odds of the response variable. It was statistically significant (p-value = 0.007284). The coefficient for access to loans was 0.8611 ($OR = 2.3658, 95\%\ CI$: 0.369 1.361). A one-level increase in the variable access to loans was associated with an increase of 0.8611 in the log-odds of the response variable. It was statistically significant since its p-value = 0.000638. The coefficient for ability to save was 0.2494 ($OR = 1.2832, 95\%\ CI$: $-$ 0.267 0.766). A one-level increase in the variable ability to save was associated with an increase of 0.2494 in the log-odds of the response variable. However, it was not statistically significant (p-value = 0.343176) meaning it did not affect the response variable and hence it was removed from this study.

The table below shows the results when the demographic factors were fit in the logistic regression model. The factors included age, marital status, education level, gender and source of income. The table shows the estimated coefficients, the std error, the z-value, the p-value and the odds ratio and the 95% CI.

Table 4.3: Logistic Regression Results for social demographic variables

Variable	Estimate	Std Error	z-value	p-value	OR
Intercept	2.8699	0.6112	4.695	2.66e-06	17.636
Age (years)					
18-30 (ref)	-	-	-	-	-
31-45	-0.1484	0.3729	-0.398	0.6907	0.8621
46-60	-0.2600	0.3899	-0.667	0.5048	0.7710
Over 60	-0.4649	0.4154	-1.119	0.2631	0.6282
Marital status					
Divorced	-	-	-	-	-
Married	0.0906	0.3388	-2.316	0.0206	1.0949
Single	-0.7847	0.3778	0.240	0.8105	0.4562
Widowed	-0.0009	0.3631	-0.002	0.9981	0.9991
Education					
College	-	-	-	-	-
No Education	-0.4367	0.3824	-1.142	0.2534	0.6462
Primary	-0.4542	0.3658	1.242	0.2143	0.6349
Secondary	-0.601	0.3730	-1.611	0.1072	0.5484
Gender					
Female (ref)	-	-	-	-	-
Male	0.5479	0.2577	2.126	0.0335	1.7295
Income					
Retired	1.1503	1.0981	1.048	0.2949	3.1590
Self Employed	-0.3620	0.425	-0.852	0.3943	0.6963
Unemployed	-1.1465	0.5805	-1.975	0.0483	0.3177
Employed (ref)	-	-	-	-	-

Demographic factors such as gender, source of income and marital status were found to be significant predictors of wealth status. Factors such as education level and age were not found to be significant predictors to wealth status.

The intercept, which was the log-odds when all other predictor variables were zero had an estimated coefficient of 2.87 and a p-value 2.66e-06 indicating that the intercept was significantly different from zero ($OR = 17.636, 95\%\ CI: -9.7188 - 2.3092$).

Age was categorized into four groups i.e., 18-30, 31-45, 46-60 and over 60 years. The category 18-30years was used as the reference category or the baseline category against which the other age categories were compared. The estimated log-odds for this category were assumed to be zero. The estimated coefficient for the age category 31-45 were -0.1484. This represented the change in the log-odds of the response variable for individuals aged 31-45 compared to those aged 18-30. The standard error for this estimate was 0.3729. The z-value was -0.398 and the associated p-value was 0.6907, indicating that the coefficient was not statistically significant. The odds ratio (OR) for this category was 0.8612 while the 95% confidence interval for the odds ratio was between 0.7710 (lower) and 0.9582 (upper). The estimated coefficient for the age bracket 46-60 years was -0.2600, representing the change in the log-odds of the response variable for individuals aged 46-60 compared to those in the reference category. The associated p-value is 0.5048, indicating that the coefficient was not statistically significant. The odds ratio (OR) for this category was 0.7708 while the 95% confidence interval for the odds ratio was between 0.6619 (lower) and 0.8970 (upper).

The estimated coefficient for Over 60 is -0.4649, representing the change in the log-odds of the response variable for individuals aged over 60 compared to those aged 18-30.
The p-value was 0.2631, indicating that the coefficient was not statistically significant. The odds ratio (OR) for this category was 0.6282 and the 95% confidence interval for the odds ratio was between 0.5618 (lower) and 0.7033 (upper). None of the age categories (31-45, 46-60, Over 60) had coefficients that were statistically significant at the 0.05 significance level.

Gender of the respondents were categorized as either male or female. The female category was used as the reference category and hence its coefficient was assumed to

be zero. The coefficient for the male gender was 0.5479. This shows that compared to the reference category, being male was associated with an estimated increase in the log-odds (OR=1.7295) of the response variable by 0.5479. The p-value is 0.0335 which is less than 0.05, indicating that this effect was statistically significant at the 0.05 significance level. The 95% CI was 0.0498 (lower) and 1.0633(upper) did not include zero indicating the variable was statistically significant and the relationship between this predictor variable and the response variable is positive.

The marital status category was divided into single, married, widowed and divorced. Divorced was used as the reference category. The coefficient for marital status married was 0.0906, being married was associated with an estimated increase in the log-odds of the response variable by 0.0906 (OR 1.0949). The p-value of 0.0206 was less than 0.05, indicating that this effect was statistically significant at the 0.05 significance level. The 95% CI was 0.1305 (lower) and 1.4660 (upper). The coefficient for marital status single was estimated to be -0.7847. The negative coefficient suggested a decrease in the log-odds for individuals who were single. The high p-value (0.8105) indicated that this category was not statistically significant. The OR were 0.4562 and the 95% CI was -0.6516 (lower) and 0.8413 (upper). The coefficient for marital status widowed was -0.0009, the p-value was (0.9981) suggesting that this category was not statistically significant ($OR = 0.9991, 95\% CI: -0.7194\ \ 0.7140$).

The level of education was categorized into no formal education, primary education, secondary education and college or tertiary level. The college category was used as the reference category. The coefficient for no formal education was -0.4367 implying a negative impact on the log-odds for individuals with no education. The p-value was 0.2534 ($OR = 0.6462, 95\% CI: -1.2026\ \ 0.3075$). This category was not found to be statistically significant. Primary education had a coefficient -0.4542 having only primary education appears to have a negative impact on the log-odds compared to the reference category, but the effect is not statistically significant (p-value = 0.2143) ($OR = 0.6349, 95\% CI: -1.1921\ 0.2522$). The coefficient for secondary education (-0.601) suggested a negative impact of having secondary education on the log-odds, but the effect was not statistically significant (p-value = 0.1072) ($OR = 0.5484\ CI: -1.3521\ 0.1201$).

The last category was source of income which was divided into unemployed, self-employed, employed and retired. The employed category was used as the reference category. Individuals with a source of income as retired had a positive impact on the log-odds, coefficient 1.1503, but the effect was not statistically significant (p-value = 0.2949) $(OR = 3.1590\ 95\%\ CI\colon -0.6549\ 4.1083)$. The coefficient (-0.3620) for the self-employed category suggested a negative impact on the log-odds for individuals who were self-employed, but again, the effect was not statistically significant (p-value = 0.3943)$(OR = 0.6963\ 95\%\ CI\colon -1.2793\ 0.4105)$. Individuals with a source of income as unemployed had a negative impact on the log-odds, coefficient -1.1465 and the effect was statistically significant at a 5% significance level (p-value = 0.0483) $(OR = 0.3177\ 95\%CI\colon -2.3125 - 0.0044)$.

The logistic regression model was then fit using only the significant predictor variables and the coefficients of this model were extracted. The logistic regression equation with these coefficients is expressed as shown below.

$$ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = -4.4359 + 0.5915\ x_1 + 0.4994x_2 + 0.7142x_3 \qquad (4.1)$$

Where;

$x_1 =$ Income generating assets

$x_{2=}$Basic Commodities

$x_{2=}$ Access to loans

The model classified 19 members under class 0 which represented a total percentage of 9.1 and 190 under class 1 which was a percentage of 90.9. This was done using the test set. The accuracy of the logistic regression model in classifying members of SHG in this study as either having an improved wealth status or not since joining a SHG was 88.04%.

4.4 Naïve Bayes Algorithm.

The predictor variables were fit into the naïve bayes algorithm to determine their effect on the response variable wealth status. The model was fit in the R environment. Data preprocessing was done to convert the variables into binary form so as to fit the Bernoulli naïve bayes algorithm since the output was binary.

Table 4.4: A-priori probabilities of Y

Improved	Frequency	Percent (%)
No	98	10.11%
Yes	871	89.89%

A-priori probabilities of Y are predictions made on the output variable (wealth status) before considering any features. The results shown above suggest that class 0 has a prior probability of 10.11% while class 1 has a prior probability of 89.89%.

Table 4.5: Conditional probabilities for categorical variables

Variable	0	1
Age (years)		
18-30	0.1688	0.1625
31-45	0.5455	0.5124
46-60	0.1948	0.2578
Over 60	0.0909	0.0673
Marital status		
Divorced	0.052	0.0922
Married	0.909	0.8389
Single	0.026	0.0294
Widowed	0.013	0.0395
Education		
College	0.0779	0.0732
No Education	0.0909	0.082
Primary	0.2857	0.2972
Secondary	0.5455	0.5476
Gender		
Female	0.5844	0.6237
Male	0.4156	0.3763
Income		
Retired	0.013	0.0659
Self Employed	0.7013	0.7584
Unemployed	0.0909	0.0586
Employed	0.1948	0.1171

Table 4.5 shows the probabilities of belonging to class 0 or to class 1 given categorical variables such as age, source of income, level of education, marital status and gender. 16.88% was the probability of belonging to age category 18-30 years and being classified under class 0 while 16.25% was the probability of being in the age group and being categorized under class 1. The probability of being in the age category 31-45 years and being categorized under class 0 was 54.55% while that of being in this age category and being classified under class 1 was 51.24%. 19.48% was the probability of being in the age group 46-60 and being in class 0 while 25.77% was the

probability of being in the age group and be classified under class 1. The probability of being in the category over 60 years and in class 0 was 9.09%, while 6.73% was the probability of being over 60 and in class 1.

Gender was categorized as male and female. There was a 58.44% probability of being female and being classified in class 0 while the probability was 41.56% for men. The probability of being female and being in class 1 was 62.37% while that of being male and being in class 1 was 37.63%.

The other category was marital status, categorized as single, married, divorced and widowed. The probability of being divorced and in class 0 was 5.19%, being married and in class 0 was 90.91% being single and in class 0 was 2.60% and 1.30% was the probability of being widowed and in class 0. The probability of being in class 1 was 9.22% when divorced, 83.89% when married, 2.93% when single and 3.95% when widowed.

Source of income was categorized as unemployed, self-employed, employed and retired. The probability of being in class 0 was 19.48% when employed, 1.30% when retired, 70.13% when self-employed and 9.09% when unemployed. For class 1, there was an 11.71% probability of being categorized here when employed, 6.59% when retired, 75.84% when self-employed and 5.86% when unemployed.

For the category education level, the probability of being categorized in class 0 was 7.79% for members with college education, 9.09% for those with no education, 28.57% for primary education and 54.55% for those with secondary education. The probability of being categorized in class 1 was 7.32% for members with college/tertiary education, 8.2% for those with no education, 29.72% for the ones with primary education and 54.76% for members with secondary education.

The predictor variables such as access to loans, access to household assets, access to income generating assets and ability to save had data in numeric form. Naïve bayes algorithm generated the mean and the standard deviation for each class. The results are tabulated below.

Table 4.6: Conditional probabilities for continuous predictor variables

Variables	0		1	
	Mean	Standard Deviation	Mean	Standard Deviation
Household Asset	3.44	0.6	3.316	0.62
Income Generating Assets	3.28	0.76	3.68	0.67
Basic commodities	3.56	0.66	3.88	0.55
Loans	3.61	0.63	3.89	0.45
Savings	3.68	0.55	3.7	0.52

Table 4.6 represents the conditional probabilities for the continuous predictor variables. Naïve bayes algorithm uses mean and standard deviation for continuous variables for each class, 0 and 1 to calculate the likelihood of observed data being in a given class. The likelihood calculated is combined with prior probabilities to compute posterior probabilities. The assumption of the naïve bayes algorithm is that features within a class are independent and unrelated, this feature however does not always hold but the naïve bayes algorithm is still able to make predictions in real world applications. When the naïve bayes algorithm was put into test in the classification of members of SHGs as either having improved wealth status or not, it achieved an accuracy of 92.34%.

Naïve bayes algorithm classified 12 members in class zero which was a percentage of 5.74 and 197 members as class 1 which was a percentage of 94.26 of the test set.

4.5 Support Vector Machine (SVM) Model

The SVM model was also fit using the predictor variables, access to loans, ability to save, access to income generating assets, age, marital status, access to household assets, access to basic commodities and the outcome variable, wealth status, to assess the impact of the predictor variables on the dependent variable. The analysis used SVM with the RBF function. The RBF kernel function outperformed the linear

function when they were put to the test. The linear SVM classified all the data points under one class meaning it was not able to capture complexities in the data and hence the choice of the RBF kernel.

Table 4.7: SVM Model Testing Parameters

Parameters	Values
SVM-Kernel	RBF
SVM-Type	Classification
Support Vectors	149(class 0 =122, class 1=27)
Cost	10
Gamma	0.111

The model was trained using 80% of the data and tested using 20% of the entire data set (Uçar et al 2020). A higher percentage of the data was used for training to enable the model to learn patterns and relationships present in the data effectively. 80-20 percentage is the most acceptable ratio from literature since it allows enough data for training without overfitting the model by using a higher percentage as well as having sufficient data for testing. Support vectors are the data points that are crucial for the SVM algorithm. These data points support or define the structure of the decision boundary and the SVM algorithm relies on them to make accurate predictions on new, unseen data. The cost parameter in SVM helps balance the desire for a smooth decision boundary (to prevent overfitting) and the goal of correctly classifying training data. A small cost produces a simple model that is less prone to overfitting. The gamma parameter controls the shape and flexibility of the decision boundary. The results obtained for both the training and the testing sets are represented in the table below.

Table 4.8: SHGs Training and Testing Evaluation Results for SVM Model

Training		**Testing**	
Parameter	**Value**	**Parameter**	**Value**
Number of Observations	748	Number of Observations	209
Predicted class (0)	45	Predicted class (0)	9
Predicted class (1)	703	Predicted class (1)	200

From the results, the SVM model made a prediction on the training set that 93.98% of the observations or individuals had had a positive impact on their wealth status since joining a SHG. A 6.02% of the predictions made using the training set were classified as class 0 meaning they had not seen any significant improvement on their wealth status since joining SHGs. The predictions done on the test set indicated that 4.31% of members had not seen an improvement on their wealth status since joining SHGs while 95.69% of the members were classified as class 1 meaning they had seen a significant impact on their wealth status since joining SHGs. The evaluation of the model performance was done by constructing a confusion matrix. The results obtained from the confusion matrix were discussed under the model selection section.

The accuracy of the SVM model in classifying the wealth status of the members as improved or not improved since joining a SHG was 84.62%.

4.6 Model Selection

To determine which model among the three models used in this study was the best, the study utilized a confusion matrix. The metrics of a confusion matrix such as sensitivity, precision, F1 score and specificity were used and the results are as shown below. The metrics of the confusion matrix are calculated using the test set or the validation set. The test set tests the model's performance on new, unseen data that it has

The results of the confusion matrix of the three models are tabulated below.

Table 4.9: Confusion Matrix metrics for the three models

	Logistic Regression	Naïve Bayes	SVM
Accuracy (%)	88.04	92.34	84.62
Precision (%)	94.21	94.92	87.74
Recall (%)	92.75	96.89	95.88
F1 Score (%)	93.47	95.89	91.63

From the results, the naïve bayes algorithm had the highest level of accuracy which was 92.34%, followed by the logistic regression model which had an accuracy of 88.04% and finally the SVM model with an accuracy of 84.62%. The naïve bayes algorithm also had a higher precision (94.92%) compared to the logistic regression model whose precision was 94.21% and the SVM model whose precision was 87.74%. Naïve bayes algorithm also had a higher sensitivity or recall which was at 96.89%, followed by the SVM model with 95.88% and the logistic regression model with 92.75%. The F1 score was highest in NBA (95.89%), followed by the logistic regression model (93.47%) and finally the SVM model with 91.63%. From the results tabulated above, it is clear that the naïve bayes algorithm was the best performing model since it outperformed the other two models in most of the metrics. The second-best performing model was the logistic regression model while the SVM model had the worst performance compared to the other two models.

4.7 Test of Hypotheses

Hypotheses on whether SVM, naïve bayes and logistic regression were suitable machine learning models were tested using the chi-square test and the results were as shown.

	Logistic Regression	Naïve Bayes	SVM
χ^2	76.039	3.7215	14.528
DF	1	1	1
P-value	2.2e-16	0.04	0.0001381

To test the first hypothesis, hoteling's T^2 which tests the significant differences represented by the principal components (PC1, PC2, ..., PC8) was performed and the p-value obtained was 0.008815166 which was less than 0.05 and hence H_0 was rejected and it was concluded that there were more significant factors affecting SHGs performance than others.

To test whether there was any model that outperformed the other, the two-sample z test proportion was carried out and a Z calculated value of 0.01578 was observed which was below 0.05 and hence the null hypotheses was rejected.

All the null hypotheses were rejected and it was concluded that;

i. There were factors that were more significant than others that affected how groups performed.
ii. Logistic regression was a suitable machine learning technique for predicting SHGs impact within MESC because it achieved an accuracy of 88.04%.
iii. Naïve bayes algorithm was a suitable machine learning technique for predicting SHGs impact within MESC because it achieved an accuracy of 92.34% accuracy.
iv. SVM was a suitable machine learning technique for predicting SHGs impact within MESC because it achieved an accuracy of 84.62%.
v. The best performing model was the Naïve bayes algorithm.

4.8 Chapter Summary

4.8.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was performed to identify the key factors influencing Self-Help Groups (SHGs) in Murang'a East sub-county (MESC). The analysis identified several critical variables influencing SHG performance. Key factors included the frequency of meetings (B8, loading = 0.2992), access to financial resources (B4, loading = 0.2985), and collaboration with stakeholders (B6, loading = 0.2864). These findings indicate the importance of organizational practices and resource availability in determining SHG effectiveness in the region.

4.8.2 Logistic Regression Model

The Logistic Regression model examined predictors affecting the wealth status of SHG members. Significant predictors included access to income-generating assets (coefficient = 0.6633, $p < 0.001$), access to basic commodities (coefficient = 0.5154, $p = 0.007$), and access to loans (coefficient = 0.8611, $p < 0.001$). Demographic variables such as gender and marital status also influenced wealth status, with males and married individuals showing higher wealth status. The LR model achieved an 88.04% accuracy in predicting wealth status.

4.8.3 Naïve Bayes Algorithm

A-priori probabilities in Naive Bayes Algorithm indicated an 89.89% likelihood of improved wealth status. Conditional probabilities revealed demographic factors such as age, education level, gender, marital status, and income sources play significant

roles. For instance, being married and having secondary education positively impacted wealth status, while unemployment negatively affected it.

4.8.4 SVM Model

Based on the analysis conducted using SVM with an RBF kernel function, the model was effectively trained and tested to assess the impact of various predictor variables on the wealth status of members in Self-Help Groups (SHGs). The RBF kernel, chosen for its ability to handle non-linear relationships, outperformed the linear kernel by capturing a higher number of support vectors, which are crucial for defining the decision boundary.

The SVM model achieved an accuracy of 84.62% in distinguishing between members whose wealth status had improved versus those who had not.

CHAPTER FIVE
CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

The PCA identified significant factors affecting SHG performance, with the frequency of meetings and the quality of discussions being the most influential factor. Other significant factors included access to financial resources, collaboration with stakeholders, member participation, community support, policy support, training availability, and effective leadership.

In the logistic regression model, access to income-generating assets, access to basic commodities, and access to loans were found to significantly impact the wealth status of SHG members. Factors such as savings and access to household assets did not show significant effects on wealth status. Demographic factors like gender (male) and being married were associated with higher wealth status.

The Naïve Bayes model highlighting probabilities of wealth status across different demographic and socioeconomic factors with an accuracy of 92.34%.
The study highlighted the importance of frequent and high-quality meetings in SHGs, as well as access to financial resources, which are crucial for enhancing their performance.
The logistic regression model achieved an accuracy of 88.04% in classifying SHG members based on their improved wealth status.

Policymakers and SHG facilitators can use these findings to prioritize resource allocation and support mechanisms that enhance meeting frequency and financial access.
The study provides valuable insights into the factors influencing SHG performance in Murang'a East, emphasizing practical implications for stakeholders aiming to support and strengthen these community-based organizations.

5.2 Recommendations

The logistic regression model further analyzed the relationship between predictor variables and wealth status. Access to income-generating assets, basic commodities, and loans were significant predictors. Specifically, access to income-generating assets had the highest positive impact on wealth status, followed by access to loans and basic commodities.

Based on the analysis and results obtained from the Naïve Bayes, Support Vector Machine (SVM), and Logistic Regression models, it is recommended to utilize the Naïve Bayes algorithm for predicting the wealth status of members in self-help groups (SHGs). The Naïve Bayes algorithm demonstrated the highest accuracy (92.34%), precision (94.92%), recall (96.89%), and F1 score (95.89%) among the models tested, indicating its superior performance in classifying the wealth status as either improved or not improved.

The most significant factor affecting SHG performance in MESC is the frequency and quality of meetings, highlighting the importance of regular and effective discussions. Additionally, financial resources, collaboration, and member engagement play crucial roles in improving SHG performance. The logistic regression analysis supports these findings, emphasizing the positive impact of financial access and resource availability on wealth status.

5.3 Recommendations for Future Work

A recommendation on further studies to help understand poverty and how it can be eliminated or reduced in Kenya and even in Africa should be done and other tools that can be used to achieve this discussed. This will help place Africa in the global stage and improve lives and livelihoods within the continent. The goal to make people live decent lives within the continent should be a topic that should be widely researched on and explored to ensure that lives are made better.

A future study on groups that were started but collapsed should be made to further understand what members of those groups have to say on factors affecting group performance.

More studies and analysis should be conducted using the SVM and the naïve bayes algorithm with SHGs data since there is limited application of these models in analyzing this type of data.

REFERENCES

Alrefaei, N., Aquinas, P. G., & Al-Maamari, O. A. (2022). Self-help group (SHG) in India: A path toward empowerment and poverty reduction. *Social Work with Groups*, 1-15.

Amzile, K., & Habachi, M. (2022). Assessment of Support Vector Machine performance for default prediction and credit rating. *Banks and Bank Systems*, 17(1), 161-175.

Bager, L., Elsbernd, A., Nissen, A., Daugaard, G., & Pappot, H. (2018). Danish translation and pilot testing of the European Organization for Research and Treatment of Cancer QLQ-TC26 (EORTC QLQ-TC26) questionnaire to assess health-related quality of life in patients with testicular cancer. *Health and Quality of Life Outcomes*, 16(1), 1-6.

Brody, C., Dworkin, S., Dunbar, M., Murthy, P., & Pascoe, L. (2013). Protocol: The effects of economic self-help group programs on women's empowerment: A systematic review. *Campbell Systematic Reviews*, 9(1), 1-43.

Deininger, K., & Liu, Y. (2009). Economic and social impacts of SHGs in India. *World Bank Policy Research Working Paper*, (4884).

El Naqa, I., & Murphy, M. J. (2015). What is machine learning? In *Machine Learning in Radiation Oncology* (pp. 3-11). Springer, Cham.

Fisher, M. J., & Marshall, A. P. (2009). Understanding descriptive statistics. *Australian Critical Care*, 22(2), 93-97.

Harris, T. (2015). Credit scoring using the clustered support vector machine. *Expert Systems with Applications*, 42(2), 741-750.

Hoffmann, V., Rao, V., Surendra, V., & Datta, U. (2021). Relief from usury: Impact of a self-help group lending program in rural India. *Journal of Development Economics*, 148, 102567.

Horak, J., Vrbka, J., & Suler, P. (2020). Support vector machine methods and artificial neural networks used for the development of bankruptcy prediction models and their comparison. *Journal of Risk and Financial Management*, 13(3), 60.

Jose, S., Chockalingam, S. M., & Velmurugan, R. (2019). Problems of women self-help group members in Ernakulam District. *Journal of Critical Reviews*, 7(1), 141-143.

Kenya: Extreme poverty rate 2016-2025. (n.d.). *Statista*.

Krichene, A. (2017). Using a naive Bayesian classifier methodology for loan risk assessment: Evidence from a Tunisian commercial bank. *Journal of Economics, Finance and Administrative Science*, 22(42), 3-24.

LaValley, M. P. (2008). Logistic regression. *Circulation*, 117(18), 2395-2399.

Magali, J. (2013). Impacts of rural savings and credits cooperative societies (SACCOS') loans on borrowers in Tanzania. *International Journal of Management Sciences and Business Research*, 2(12).

Mwai, J. (2017). Impact of Saccos in poverty eradication among farmers in Murang'a County, Kenya.

Okesola, O. J., Okokpujie, K. O., Adewale, A. A., John, S. N., & Omoruyi, O. (2017, December). An improved bank credit scoring model: A naïve Bayesian approach. In *2017 International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 228-233). IEEE.

Othman, A., Kari, F., Jani, R., & Hamdan, R. (2012). Factors influencing cooperative membership and share increment: An application of the logistic regression analysis in the Malaysian cooperatives. *World Review of Business Research*, 2(5), 24-35.

Raghavan, V. V., Gudivada, V. N., Govindaraju, V., & Rao, C. R. (2016). *Cognitive computing: Theory and applications*.

Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence* (Vol. 3, No. 22, pp. 41-46). Chicago.

Serrat, O. (2017). The sustainable livelihoods approach. In *Knowledge Solutions* (pp. 21-26). Springer, Singapore.

Sukhdeve, M. I. (2011). Factors influencing empowerment of rural SHG women in Bhandara District of Maharashtra: Logistic regression model. *International Journal of Microfinance*, 1(1), 102-108.

Uçar, M. K., Nour, M., Sindi, H., & Polat, K. (2020). The effect of training and testing process on machine learning in biomedical datasets. *Mathematical Problems in Engineering*, 2020.

Vapnik, V., Golowich, S. E., & Smola, A. (1995). Support vector method for function approximation, regression estimation, and signal processing. *Advances in Neural Information Processing Systems*, 281-287.

Watson, R. (2015). Quantitative research. *Nursing Standard*, 29(31), 44.

Webb, G. I., Keogh, E., & Miikkulainen, R. (2010). Naïve Bayes. In *Encyclopedia of Machine Learning* (pp. 713-714).

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques* (4th ed.). Morgan Kaufmann.

Yamane, T. (1967). *Statistics: An Introductory Analysis* (2nd ed.). Harper and Row.

APPENDICES

Appendix 1: Introductory Letter

Hello,

I hope this letter finds you well. I am writing to invite you to participate in a survey that is part of my research project. Your involvement in this survey is completely voluntary and confidential. Your participation is important to help me gain a better understanding of the topic, which will contribute to making informed decisions. It is important that you respond to each question honestly and to the best of your ability. Your cooperation is highly appreciated and I would like to express my sincere gratitude for taking the time to complete the questionnaire. If you have any questions or concerns about the survey, please do not hesitate to contact me. Thank you for your cooperation.

Best regards,

Runo Jane

Appendix 2: Questionnaire

I am conducting a study on the impact of SHGs on members livelihoods in Murang'a East sub-county. Your participation in this study is highly appreciated as your responses will contribute to a better understanding of the topic. Please answer the following questions truthfully and to the best of your ability.

SECTION A: DEMOGRAPHIC FACTORS.

Age:

a. Under 18 years [ ]
b. 18-30 years [ ]
c. 31-45 years [ ]
d. 46-60 years [ ]
e. Over 60 years [ ]

Gender:

a. Male [ ]
b. Female [ ]

Marital Status:

a. Single [ ]
b. Married [ ]
c. Divorced [ ]
d. Widowed [ ]

Education Level:

a. No formal education [ ]
b. Primary education [ ]
c. Secondary education [ ]
d. College/university education [ ]

Occupation:

a. Unemployed [ ]
b. Employed in the formal sector [ ]
c. Employed in the informal sector [ ]
d. Self-employed [ ]

e. Retired [ ]

What is your primary source of livelihood?

a. Agriculture [ ]

b. Business [ ]

c. Salaried employment [ ]

d. Other [ ]

Are you a member of any other SHGs or community organizations in Murang'a East sub-county?

a. Yes [ ]

b. No [ ]

SECTION B: FACTORS IMPACTING PERFORMANCE OF SHGS.

1. The availability of training and capacity-building programs is a significant factor in the performance of SHGs in Murang'a East sub-county.

   Strongly agree [ ]

   Agree [ ]

   Neutral [ ]

   Disagree [ ]

   Strongly disagree [ ]

2. The level of participation and engagement of self-help group members is a significant factor in impacting lives in Murang'a East sub- County.

   Strongly agree [ ]

   Agree [ ]

   Neutral [ ]

   Disagree [ ]

   Strongly disagree [ ]

3. The presence of effective leadership and management structures is a significant factor in the performance of SHGs in Murang'a East sub-county.

   Strongly agree [ ]

   Agree [ ]

Neutral [ ]

Disagree [ ]

Strongly disagree [ ]

4. Access to financial resources and credit facilities is a significant factor in the performance of SHGs in impacting lives in Murang'a East sub-county.

   Strongly agree [ ]

   Agree [ ]

   Neutral [ ]

   Disagree [ ]

   Strongly disagree [ ]

5. The presence of supportive policies and regulations is a significant factor in the performance of SHGs in Murang'a East sub-county.

   Strongly agree [ ]

   Agree [ ]

   Neutral [ ]

   Disagree [ ]

   Strongly disagree [ ]

6. Collaboration and partnerships with other organizations and stakeholders is a significant factor in the performance of SHGs in impacting lives in Murang'a East sub-county.

   Strongly agree [ ]

   Agree [ ]

   Neutral [ ]

   Disagree [ ]

   Strongly disagree [ ]

7. The level of community support and involvement is a significant factor in the performance of SHGs in impacting lives in Murang'a East sub-county.

   Strongly agree [ ]

   Agree [ ]

   Neutral [ ]

   Disagree [ ]

   Strongly disagree [ ]

8. The frequency of the meetings and the quality of the discussions held during the meetings is a significant factor in the performance of SHGs in Murang'a East sub-county.

Strongly agree [ ]
Agree [ ]
Neutral [ ]
Disagree [ ]
Strongly disagree [ ]

SECTION C: Access to household assets

1. I have seen an improvement in household assets since joining the self-help group.

Strongly agree [ ]
Agree [ ]
Neutral [ ]
Disagree [ ]
Strongly disagree [ ]

2. My family and I now have access to assets that we did not have before joining the self-help group.

Strongly agree [ ]
Agree [ ]
Neutral [ ]
Disagree [ ]
Strongly disagree [ ]

3. The self-help group has helped me acquire household assets like furniture and household appliances.

Strongly agree [ ]
Agree [ ]
Neutral [ ]
Disagree [ ]
Strongly disagree [ ]

4. I am able to maintain and take care of my household assets more effectively since joining the self-help group.

   Strongly agree [ ]
   Agree [ ]
   Neutral [ ]
   Disagree [ ]
   Strongly disagree [ ]

5. I feel more secure and comfortable in my home since acquiring these assets through the self-help group.

   Strongly agree [ ]
   Agree [ ]
   Neutral [ ]
   Disagree [ ]
   Strongly disagree [ ]

SECTION D: Access to income generating assets

1. I have seen an increase in my income since joining the self-help group.

   Strongly agree [ ]
   Agree [ ]
   Neutral [ ]
   Disagree [ ]
   Strongly disagree [ ]

2. I have access to income generating assets like livestock or equipment that I did not have before joining the self-help group.

   Strongly agree [ ]
   Agree [ ]
   Neutral [ ]
   Disagree [ ]
   Strongly disagree [ ]

3. The self-help group has helped me start or grow my business.

Strongly agree	[ ]
Agree	[ ]
Neutral	[ ]
Disagree	[ ]
Strongly disagree	[ ]

4. I feel more confident in my ability to generate income since joining the self-help group.

Strongly agree	[ ]
Agree	[ ]
Neutral	[ ]
Disagree	[ ]
Strongly disagree	[ ]

5. I am able to provide for my family more effectively since having access to these income generating assets.

Strongly agree	[ ]
Agree	[ ]
Neutral	[ ]
Disagree	[ ]
Strongly disagree	[ ]

SECTION E: Access to basic commodities

1. I am able to access basic commodities like food and medicine more easily since joining the self-help group.

Strongly agree	[ ]
Agree	[ ]
Neutral	[ ]
Disagree	[ ]
Strongly disagree	[ ]

2. The self-help group has helped me purchase basic commodities at a lower cost.

   Strongly agree [ ]
   Agree [ ]
   Neutral [ ]
   Disagree [ ]
   Strongly disagree [ ]

3. My family and I now have better access to basic necessities like clean water and adequate shelter.

   Strongly agree [ ]
   Agree [ ]
   Neutral [ ]
   Disagree [ ]
   Strongly disagree [ ]

4. I feel more secure in meeting my family's basic needs since joining the self-help group.

   Strongly agree [ ]
   Agree [ ]
   Neutral [ ]
   Disagree [ ]
   Strongly disagree [ ]

5. I am able to allocate more resources to other needs since having access to these basic commodities at a lower cost.

   Strongly agree [ ]
   Agree [ ]
   Neutral [ ]
   Disagree [ ]
   Strongly disagree [ ]

SECTION F: Access to loans

1. I have been able to access loans through the self-help group that I would not have been able to obtain otherwise.

   Strongly agree [ ]
   Agree [ ]
   Neutral [ ]
   Disagree [ ]
   Strongly disagree [ ]

2. The interest rates on loans obtained through the self-help group are lower than those offered by other financial institutions.

   Strongly agree [ ]
   Agree [ ]
   Neutral [ ]
   Disagree [ ]
   Strongly disagree [ ]

3. I feel more financially secure since having access to these loans.

   Strongly agree [ ]
   Agree [ ]
   Neutral [ ]
   Disagree [ ]
   Strongly disagree [ ]

4. I am able to invest in my business or household more effectively since having access to these loans.

   Strongly agree [ ]
   Agree [ ]
   Neutral [ ]
   Disagree [ ]
   Strongly disagree [ ]

5. I am able to repay the loans obtained through the self-help group on time and without difficulty.

   Strongly agree [ ]
   Agree [ ]
   Neutral [ ]
   Disagree [ ]
   Strongly disagree [ ]

SECTION G: Ability to save

1. I am able to save more effectively since joining the self-help group.

   Strongly agree [ ]
   Agree [ ]
   Neutral [ ]
   Disagree [ ]
   Strongly disagree [ ]

2. I feel more secure knowing that I have savings set aside for emergencies or future needs.

   Strongly agree [ ]
   Agree [ ]
   Neutral [ ]
   Disagree [ ]
   Strongly disagree [ ]

3. I am able to set aside money for my children's education or future needs since having access to these savings.

   Strongly agree [ ]
   Agree [ ]
   Neutral [ ]
   Disagree [ ]
   Strongly disagree [ ]

4. I am able to make larger purchases or investments since having access to these savings.

   Strongly agree [ ]
   Agree [ ]
   Neutral [ ]
   Disagree [ ]
   Strongly disagree [ ]

5. I feel more in control of my financial situation since having access to these savings.

   Strongly agree [ ]
   Agree [ ]
   Neutral [ ]
   Disagree [ ]
   Strongly disagree [ ]

SECTION H: WEALTH STATUS

1. I have seen an improvement in my overall wealth status since joining the self-help group.

   Strongly agree [ ]
   Agree [ ]
   Neutral [ ]
   Disagree [ ]
   Strongly disagree [ ]

2. The self-help group has helped me access financial resources and opportunities that I did not have before.

   Strongly agree [ ]
   Agree [ ]
   Neutral [ ]
   Disagree [ ]
   Strongly disagree [ ]

3. I feel more financially secure since becoming a member of the self-help group.

   Strongly agree [ ]

Agree [ ]

Neutral [ ]

Disagree [ ]

Strongly disagree [ ]

4. I am able to invest more effectively in my business or household since becoming a member of the self-help group.

Strongly agree [ ]

Agree [ ]

Neutral [ ]

Disagree [ ]

Strongly disagree [ ]

5. I am able to provide a better standard of living for my family since becoming a member of the self-help group.

Strongly agree [ ]

Agree [ ]

Neutral [ ]

Disagree [ ]

Strongly disagree [ ]

Appendix 3: Proposal Approval

MURANG'A UNIVERSITY OF TECHNOLOGY

Office of the Director, Postgraduate Studies

E-mail: bps@mut.ac.ke
Website: www.mut.ac.ke

P.O. Box 75-10200
Murang'a, Kenya

Ref: MUT/ARP/PGS/19/2020 /VOL.1

Date: 17th July, 2023

Ms. Jane Wangui Runo (AS400/5690/2021)
School of Pure, Applied and Health Sciences,
Murang'a University of Technology,
P.O. Box 75-10,200
Murang'a.

Dear Ms. Runo,

RE: APPROVAL OF RESEARCH PROPOSAL AND SUPERVISORS.

I am pleased to inform you that the Directorate of Postgraduate Studies on 3rd July, 2023 considered and approved your Masters research proposal entitled ***"Modeling Self-Help Groups' Impact on Members' Livelihood in Murang'a East Sub-County Using Machine Learning"*** and appointed the following as supervisors:

1. Dr. Ayubu Anapapa - Murang'a University of Technology
2. Dr. Euna Nyarige - Machakos University

You may now proceed with your data collection subject to obtaining research permit from NACOSTI, if required. You should also begin consulting your supervisors and submit through them quarterly progress reports to the Director Postgraduate Studies through your CoD and School Dean. Progress Reports forms can be accessed in the Student Portal.

It is the policy and regulations of the University that you observe deadlines. The guidelines on Postgraduate supervision can be accessed in the post graduate Handbook.

Your responsibilities as a student will include, among others;

i. Maintain regular consultation with your supervisor(s), at least once a month
ii. Submit quarterly reports on time, through your supervisors, CoD, Dean and to the Director of Postgraduate Studies;
iii. Ensure quality work all through;

iv. Present your research findings at 2-3 seminars/conferences prior to thesis examination.

v. Publish one article from your research findings in a refereed journal prior to thesis examination.

For any further clarification please contact the undersigned.

Yours Sincerely,

DIRECTORATE OF POSTGRADUATE STUDIES
MURANG'A UNIVERSITY OF TECHNOLOGY
17 JUL 2023
P. O. Box 75 - 10200, MURANG'A
Tel: 0771463515, Email: info@mut.ac.ke

Prof. Geoffrey Muchiri, PhD
Director, Postgraduate Studies
GM/PIM

Copy to:

1. Registrar – Academics and Students Affairs
2. Dean, School of Pure, Applied and Health Sciences

Appendix 4: Research Permit

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Ref No: 450348

Date of Issue: 11/August/2023

RESEARCH LICENSE

This is to Certify that Ms.. Jane Wangui Runo of Murang'a University of Technology, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev.2014) in Muranga on the topic: MODELING SELF-HELP GROUPS' IMPACT ON MEMBERS' LIVELIHOODS IN MURANG'A EAST SUB-COUNTY USING MACHINE LEARNING for the period ending : 11/August/2024.

License No: NACOSTI/P/23/28418

450348

Applicant Identification Number

Director General
NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.

Appendix 5: R Codes

```
getwd()
data=read.csv("C:/Users/Jane/Desktop/SELFHELPGROUPSDATA.csv",header=TRUE)
data
#Find total observations in the data
nrow(data)
summary(data)
attach(data)
sum(is.na(data))
na.omit(data)
class(data)

#split into training and test set (80% for training)
#Trainset
#Setting the seed for reproducibility
library(caret)
library(caTools)
set.seed(970)
split=sample.split(data,SplitRatio=0.8)
Trainset=subset(data,split==TRUE);Trainset
Testset=subset(data,split==FALSE);Testset
write.csv(Trainset,"Trainset.csv")
#Testset
write.csv(Testset,"Testset.csv")
class(Trainset)
class(Testset)
attach(Trainset)
#Sums the missing values
sapply(data,function(x)sum(is.na(x)))

category=factor(c("A1","A2","A3","A4","A5"));category
Wealth_status=factor(c("H1","H2","H3","H4","H5"));Wealth_status
dependent=factor(c("C1","C2","C3","C4","C5","D1","D2","D3","D4","D5","E1","E2","E3","E4","E5","
length(Wealth_status)
length(dependent)

#Convert wealth_status dependent variable to binary(0 for wealth_status not improved,1 for weal
#Set the threashold
threashold=4;threashold
Trainset$H_binary=ifelse(Trainset$H1>=4,"improved","not improved");H_binary
# Remove the last value from H1
Trainset$H1 <- Trainset$H1[-length(Trainset$H1)];Trainset$H1
```

```
Trainset$H_binary=ifelse(Trainset$H1>=4,"improved","not improved");H_binary
# Remove the last value from H1
Trainset$H1 <- Trainset$H1[-length(Trainset$H1)];Trainset$H1
# Add a value to the end of the H column to match H_binary
sum(is.na(Trainset))
na.omit(H)
length(H)
length(H_binary)
Trainset=data.frame(H_binary=H_binary,H=H,H=H,H=H,H=H,H=H,H=H)
length(Trainset$H)
(nrow(Trainset))
sum(is.na(Trainset$H_binary))
Trainset=na.omit(Trainset)

#Recreate the data frame with consistent lengths and data types
Trainset=data.frame(H_binary=factor(H_binary),H1=as.numeric(H1),H2=as.numeric(H2),H3=as.numeric

#Fitting a binary logistic Regression Model]
length(C)
length(D)
length(H_binary)
length(H)
length(E)
length(F6)
length(G)
nrow(Trainset)
summary(Trainset)
attach(Trainset)
names(Trainset)
sum(is.na(Trainset))
na.omit(Trainset)
class(Trainset)
length(Trainset)
length(C1)
length(D1)
length(E1)
length(F1)
length(G1)
length(H_binary)
logitmodel=glm(H_binary~C+D+E+F6+G,data=Trainset,family=binomial(link="logit"));logitmodel
summary(logitmodel)
```

```
# Extract the coefficients
coefficients <- coef(logitmodelsig);coefficients
coefficients <- coef(logitmodel);coefficients

#Odds ratio for the coefficients
(exp(coef(logitmodel)))

#ODDS ratio can be retrieved in a beautiful tidy formatted table using the tidy( ) function of
#install.packages("broom")
# Load the broom package
install.packages("broom")
library(broom)
tidy(logitmodel, exponentiate = TRUE, conf.level = 0.95);tidy
logitmodel=glm(H_binary~C+D+E+F6+G,data=Trainset,family=binomial(link="logit"));logitmodel
summary(logitmodel)

#Making predictions using the entire data
predicted_probabilities <- predict(logitmodel, newdata = data, type = "response");predicted_pro

threshold <- 0.8
predicted_class <- ifelse(predicted_probabilities >= threshold, 1, 0);predicted_class
table(predicted_class)

#Making predictions using the testset
predicted_probabilitiesTest <- predict(logitmodel, newdata = Testset, type = "response");predic

threshold <- 0.8
predicted_classTest <- ifelse(predicted_probabilitiesTest >= threshold, 1, 0);predicted_class
table(predicted_classTest)

#Making predictions using the Trainset
predicted_probabilities <- predict(logitmodel, newdata = Trainset, type = "response");predicted

threshold <- 0.8
predicted_class <- ifelse(predicted_probabilities >= threshold, 1, 0);predicted_class
table(predicted_class)

# Extract the coefficients
coefficients <- coef(logitmodel);coefficients
```

```r
#Odds ratio for the coefficients
(exp(coef(logitmodel)))
#ODDS ratio can be retrieved in a beautiful tidy formatted table using the tidy( ) function of
#install.packages("broom")
# Load the broom package
install.packages("broom")
library(broom)
tidy(logitmodel, exponentiate = TRUE, conf.level = 0.95);tidy
logitmodel=glm(H_binary~C+D+E+F6+G,data=Trainset,family=binomial(link="logit"));logitmodel
summary(logitmodel)

# Calculate 95% confidence intervals for coefficients
confidence_intervals <- confint(logitmodel)
confidence_intervals

#Confusion matrix Trainset
confusion_matrix <- table(predicted_class,Trainset$H_binary);confusion_matrix
confusion_matrix

#Confusion matrix Testset
confusion_matrix <- table(predicted_classTest,Testset$H_binary)
confusion_matrix

Accuracy = sum(diag(confusion_matrix)) / sum(confusion_matrix);Accuracy
length(predicted_class)
```

```r
#Demographic Factors
Trainset=data.frame(H_binary=factor(H_binary),Age=factor(Age),Gender=factor(Gender),Marital_sta
train_levels=levels(Trainset$A1)
logitmodel=glm(H_binary~Age+Gender+Marital_status+Education_level+source_of_income,data=Trainse
summary(logitmodel)

# Create a data frame with the factors
colnames(Trainset)
names(Trainset)
factor_data=Trainset[,c("B1","B2","B3","B4","B5","B6","B7","B8","B9")]
# Perform PCA for factors affecting performance of SHGs
pca_result <- prcomp(factorsdata, scale = TRUE);pca_result
plot(pca_result)
# Biplot (plot of the first two principal components)
biplot(pca_result, scale = 0)

# Summary of PCA results
summary(pca_result)

# Variance explained by each principal component
pca_result$sdev^2 / sum(pca_result$sdev^2)
loadings <- pca_result$rotation[, 1];loadings
loadings_data <- data.frame(Variable = variable_names, Loading = loadings);loadings_data

data=read.csv("C:/Users/Jane/Desktop/SHGSDATA-DEMOGRAPHIC.csv",header=TRUE);data

#split into training and test set (80% for training)
#Trainset
#Setting the seed for reproducibility
library(caret)
library(caTools)
set.seed(970)
split=sample.split(data,SplitRatio=0.8)
Trainset=subset(data,split==TRUE);Trainset
Testset=subset(data,split==FALSE);Testset
names(Trainset)

#Fit a binary logistic Regression Model on the training set for Sociodemographic variables
logitmodeldemographic=glm(H_binary~Age+Gender+Marital_status+Education_level+source_of_income,d
summary(logitmodeldemographic)
```

```
#SVM MODEL
data=read.csv("C:/Users/Jane/Desktop/SHGSDATA-SVM.csv",header=TRUE);data
#split into training and test set (80% for training)
#Trainset
#Setting the seed for reproducibility
library(caret)
library(caTools)
set.seed(970)
split=sample.split(data,SplitRatio=0.8)
Trainset=subset(data,split==TRUE);Trainset
Testset=subset(data,split==FALSE);Testset
write.csv(Trainset,"Trainset.csv")
#Testset
write.csv(Testset,"Testset.csv")
class(Trainset)
class(Testset)
attach(Trainset)
#Fit the SVM Model
class(H_binary)
class(Trainset)
print(Trainset)
print(H_binary)
print(Trainset)
print(H_binary)
summary(Trainset)
str(Trainset)
table(Trainset$H_binary)

install.packages("e1071")
library(e1071)
# Train the SVM model
svm_model <- svm(H_binary ~ C+D+E+F6+G, data = Trainset, kernel = "radial", probability = TRUE,
summary(svm_model)

# Making predictions Trainset
predicted_probabilities <- predict(svm_model, data =Trainset,probability = TRUE);predicted_prob
print(predicted_probabilities)
table(predicted_probabilities)

# Making predictions Testset
predicted_probabilitiesTest <- predict(svm_model, newdata =Testset);predicted_probabilitiesTest
print(predicted_probabilitiesTest)
table(predicted_probabilitiesTest)
```

```
#Confusion matrix Trainset
length(predictions)
length(Trainset$H_binary)
confusion_matrix <- table(predictions, Trainset$H_binary)
confusion_matrix

#Confusion matrix Testset
length(predicted_probabilitiesTest)
length(Testset$H_binary)
confusion_matrixTest <- table( predicted_probabilitiesTest,Testset$H_binary)
confusion_matrixTest

# No. of support vectors
svM_indices <- svm_model$index;svM_indices
num_support_vectors <- length(svM_indices);num_support_vectors

# Access kernel parameters
kernel_params <- svm_model$cost;kernel_params
```

```r
#NAIVE BAYES MODEL
data=read.csv("C:/Users/Jane/Desktop/SHGSDATA-SVM.csv",header=TRUE);data
#split into training and test set (80% for training)
#Trainset
#Setting the seed for reproducibility
library(caret)
library(caTools)
set.seed(970)
split=sample.split(data,SplitRatio=0.8)
Trainset=subset(data,split==TRUE);Trainset
Testset=subset(data,split==FALSE);Testset
write.csv(Trainset,"Trainset.csv")
#Testset
write.csv(Testset,"Testset.csv")
class(Trainset)
class(Testset)
attach(Trainset)
# Install and load the required package
if (!requireNamespace("e1071", quietly = TRUE)) {
  install.packages("e1071")
}
library(e1071)

# Fit Naive Bayes Model
nb_model <- naiveBayes(H_binary ~ C + D + E + F6 + G, data = Trainset);nb_model

# Make Predictions Trainset
predictions <- predict(nb_model, Trainset);predictions
table(predictions)

# Make Predictions Testset
predictionsTest <- predict(nb_model, Testset);predictionsTest
table(predictionsTest)


# For example, using a confusion matrix
confusion_matrixnb <- table(predictionsTest, Testset$H_binary)
print(confusion_matrixnb)
```

```r
#HYPOTHESIS
library(psych)

# Calculate Cronbach's alpha
alpha_result <- alpha(data, check.keys = TRUE)

# Print the result
print(alpha_result)

# Create contingency tables for LR, NBA, and SVM
LR_confusion_matrix <- matrix(c(165, 13, 6, 20), nrow = 2, byrow = TRUE);LR_confusion_matrix
NBA_confusion_matrix <- matrix(c(190, 24, 4, 3), nrow = 2, byrow = TRUE);NBA_confusion_matrix
SVM_confusion_matrix <- matrix(c(186, 20, 8, 7), nrow = 2, byrow = TRUE);SVM_confusion_matrix

# Perform chi-square test for LR
LR_chi_square <- chisq.test(LR_confusion_matrix)
print(LR_chi_square)

# Perform chi-square test for NBA
NBA_chi_square <- chisq.test(NBA_confusion_matrix)
print(NBA_chi_square)

# Perform chi-square test for SVM
SVM_chi_square <- chisq.test(SVM_confusion_matrix)
print(SVM_chi_square)



## LR AND SVM
# Sample proportions
p1 <- 0.8804  # Proportion for method 1 (e.g., LR)
p2 <- 0.9234  # Proportion for method 2 (e.g., NBA)
p3 <- 0.8462  # Proportion for method 3 (e.g., SVM)

# Sample sizes
n1 <- 206  # Sample size for method 1
n2 <- 221  # Sample size for method 2

# Calculate pooled proportion
p_pooledRS <- (p1 * n1 + p2 * n2) / (n1 + n2)
```

```r
# Calculate standard error of the difference in proportions
SERS <- sqrt(p_pooledRS * (1 - p_pooledRS) * (1 / n1 + 1 / n2))

# Calculate the test statistic
zRS <- (p1 - p2) / SERS

# Set the significance level (alpha)
alpha <- 0.05

# Calculate the critical value (two-tailed test)
critical_value <- qnorm(1 - alpha)

# Print the test statistic and critical value
cat("Test Statistic (z):", zRS, "\n")
cat("Critical Value:", critical_value, "\n")

# Compare the test statistic to the critical value
if (abs(zRS) > critical_value) {
  cat("Reject the null hypothesis. There is a statistically significant difference in proportic
} else {
  cat("Fail to reject the null hypothesis. There is insufficient evidence to conclude a differe
}

# Calculate the p-value
p_valueRS <- 2 * pnorm(-abs(zRS))

# Print the p-value
cat("p-value:", p_valueRS, "\n")


## LR AND NBA
# Sample proportions
p1 <- 0.8804  # Proportion for method 1 (e.g., LR)
p2 <- 0.9234  # Proportion for method 2 (e.g., NBA)
p3 <- 0.8462  # Proportion for method 3 (e.g., Logistic Regression)

# Sample sizes
n1 <- 206  # Sample size for method 1
n2 <- 221  # Sample size for method 2
```

```r
# Sample sizes
n1 <- 206  # Sample size for method 1
n2 <- 221  # Sample size for method 2

# Calculate pooled proportion
p_pooledRL <- (p1 * n1 + p3 * n2) / (n1 + n2)

# Calculate standard error of the difference in proportions
SERL <- sqrt(p_pooledRL * (1 - p_pooledRL) * (1 / n1 + 1 / n2))

# Calculate the test statistic
zRL <- (p1 - p2) / SERL

# Set the significance level (alpha)
alpha <- 0.05

# Calculate the critical value (two-tailed test)
critical_value <- qnorm(1 - alpha)

# Print the test statistic and critical value
cat("Test Statistic (z):", zRL, "\n")
cat("Critical Value:", critical_value, "\n")

# Compare the test statistic to the critical value
if (abs(zRL) > critical_value) {
  cat("Reject the null hypothesis. There is a statistically significant difference in proportio
} else {
  cat("Fail to reject the null hypothesis. There is insufficient evidence to conclude a differe
}

# Calculate the p-value
p_valueRL <- 2 * pnorm(-abs(zRL))

# Print the p-value
cat("p-value:", p_valueRL, "\n")

## LR AND SVM
# Sample proportions
p1 <- 0.8804  # Proportion for method 1 (e.g., LR)
p2 <- 0.9234  # Proportion for method 2 (e.g., NBA)
p3 <- 0.8462  # Proportion for method 3 (e.g., Logistic Regression)
# Sample sizes
```

```r
# Sample sizes
n1 <- 221  # Sample size for method 1
n2 <- 221  # Sample size for method 2

# Calculate pooled proportion
p_pooledSL <- (p2 * n1 + p3 * n2) / (n1 + n2)

# Calculate standard error of the difference in proportions
SESL <- sqrt(p_pooledSL * (1 - p_pooledSL) * (1 / n1 + 1 / n2))

# Calculate the test statistic
zSL <- (p1 - p2) / SESL

# Set the significance level (alpha)
alpha <- 0.05

# Calculate the critical value (two-tailed test)
critical_value <- qnorm(1 - alpha)

# Print the test statistic and critical value
cat("Test Statistic (z):", zSL, "\n")
cat("Critical Value:", critical_value, "\n")

# Compare the test statistic to the critical value
if (abs(zSL) > critical_value) {
  cat("Reject the null hypothesis. There is a statistically significant difference in proportio
} else {
  cat("Fail to reject the null hypothesis. There is insufficient evidence to conclude a differe
}

# Calculate the p-value
p_valueSL <- 2 * pnorm(-abs(zSL))

# Print the p-value
cat("p-value:", p_valueSL, "\n")
```

APPENDIX 6: Publication

The paper listed below has been published from this Thesis.

Runo, J. W., Anapapa, A., & Nyarige, E. (2024). Modeling self-help groups' impact on livelihoods in Murang'a East Sub-County: A logistic regression approach. *Asian Journal of Probability and Statistics, 26*(3), 1-12.