Major Project Mid-Term Report on

**UNDERSTANDING STRATIFICATION IN NEPALI SOCIETY USING CLUSTER ANALYSIS**

****

***Submitted to***

**Department of Computer and IT Engineering**

**Everest Engineering College**

**Sanepa-2,Lalitpur**

*In partial fulfillment for the award of the degree*

*Of*

**BACHELOR OF COMPUTER ENGINEERING**

*Submitted By*

**Bibek poudel 20070283**

**Hari Narayan Chaudhary 20070294**

**Prajwol Bista 20070299**

**Rikesh Lama Golay 20070306**

**10/01/2024**

**TABLE OF CONTENTS**

[LIST OF FIGURES iii](#_Toc169787030)

[Chapter 1: INTRODUCTION 1](#_Toc169787031)

[1.1 Background 1](#_Toc169787032)

[Chapter 2: LITERATURE REVIEW 2](#_Toc169787033)

[Chapter 3: METHODOLOGY 5](#_Toc169787034)

[3.1 Introduction 5](#_Toc169787035)

[3.2 Block Diagram 6](#_Toc169787036)

[3.3 Methodology 7](#_Toc169787037)

[3.3.1. Obtaining required data from US-AID (DATASET): 7](#_Toc169787038)

[3.3.2. Pre-processing (Cleaning): 7](#_Toc169787039)

[3.3.3 Redefining the values: 8](#_Toc169787040)

[3.3.4 Giving numeric values: 8](#_Toc169787041)

[3.3.5 Post Analysis (Inference): 10](#_Toc169787042)

[Chapter 4: RESULT AND ANALYSIS 11](#_Toc169787043)

[4.1 Histogram of Indices 11](#_Toc169787044)

[4.2 Elbow method 11](#_Toc169787045)

[4.3 Cluster Centroids 12](#_Toc169787046)

[REFERENCES 13](#_Toc169787047)

[APPENDIX 14](#_Toc169787048)

# LIST OF FIGURES

[Figure 3.2: Block Diagram](#_Toc136949780) 6.

Figure 4.1: Histogram according to index…………………………………………9.

Figure 4.2: Number of Clusters……………………………………………………9.

Figure 4.3: Centroids of Cluster………………………………………………….10.

# Chapter 1: INTRODUCTION

Social stratification in simple terms refers to the way society is organized into different layers or groups based on certain characteristics. These characteristics can include things like wealth, education, job status, or even where people live.

Understanding social stratification is crucial because it directly affects how fair and equal a society is. It helps governments make better policies to ensure everyone gets a fair shot at opportunities like education and jobs. This knowledge is key for creating a society where everyone, no matter their background, has an equal chance to succeed. It also helps in promoting fairness, breaking stereotypes, and resolving conflicts. Recognizing social stratification is essential for using resources wisely, fostering social mobility, and making a country more competitive globally. In essence, it's about treating everyone fairly and ensuring everyone has a fair chance in life.[1]

## Background

Nepal, a country rich in cultural and ethnic diversity, has a long history of social stratification primarily influenced by the caste system, which has been deeply

embedded in its social fabric for centuries. This system, rooted in Hinduism, traditionally categorized people into hierarchical groups based on their birth. Despite legal efforts to abolish caste-based discrimination, its impacts remain evident in contemporary Nepali society, influencing various aspects of life, including economic opportunities, educational attainment, and social mobility.

In recent decades, Nepal has undergone significant political and economic changes, transitioning from a monarchy to a federal democratic republic. These changes, along with increased urbanization, globalization, and economic development, have introduced new dimensions to social stratification. Economic disparities, access to education, and occupational diversity have started to shape societal divisions alongside traditional caste hierarchies.

The Nepal Demographic Health Survey 2022 is a comprehensive study conducted every five years to gather information about the demographic and health-related aspects of the population in Nepal.

The Nepal Demographic Health Survey 2022 is conducted by selecting different households from across the country to represent the whole population. Trained interviewers ask people in these households questions about various topics like their age, health, and family. [2]

# Chapter 2: LITERATURE REVIEW

The main objective of statistics is to make inferences about population based on information obtained from sample. Sample from a study is collect to obtain data that are relavant to the subject to be study. Therefore, process of collecting data is a step that needs to be take into accounts. Data collection using proper sampling methods will provide better result accuracy. One of the most popular probability sampling method is stratified random sampling. In stratified random sampling, data is group into similar characteristics within strata and heterogenous characteristics between strata. [1]

The 2022 Nepal Demographic and Health Survey (2022 NDHS) was implemented by New ERA under the aegis of the Ministry of Health and Population of Nepal. The sampling frame used for the 2022 NDHS is an updated version of the frame of the Nepal Population and Housing Census (NPHC) conducted in 2011, provided by the Central Bureau of Statistics. The smallest administrative unit in Nepal is the sub-ward. The census frame includes a complete list of Nepal’s 36,020 sub-wards. Each sub-ward has a residence type (urban or rural) and a measure of size is the number of households. Nepal is divided into seven provinces: Province 1, Madhesh Province, Bagmati Province, Gandaki Province, Lumbini Province, Karnali Province, and Sudurpashchim Province. Each province is divided into districts, districts into municipalities, municipalities into wards, and wards into sub-wards. Nepal has 77 districts, which include a total of 753 (local level) municipalities. Of the municipalities, 293 are urban and 460 are rural. [2]

Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. There are lots of libraries available, but the most popular and important Python libraries for working on data are Numpy, Matplotlib, and Pandas. Once you downloaded your data set and named it as a .csv file, you need to load it into a pandas DataFrame to explore it. manual review of the data set is important, to avoid mistakes in the data analysis and the modeling process. With a filtered data set explored, you need to create a matrix of dependent variables and a vector of independent variables. To clean the data set, you need to handle missing values and categorical features. In most cases, categorical values are discrete and can be encoded as dummy variables, assigning a number for each category. These are the very basic steps required to work through a large data set, cleaning and preparing the data for any Data Science project. There are other forms of data cleaning that you might find useful. [3]

For determining K(numbers of clusters) we use Elbow method. Elbow Method is a technique that we use to determine the number of centroids(k) to use in a k-means clustering algorithm. In this method to determine the k-value we continuously iterate for k=1 to k=n (Here n is the hyperparameter that we choose as per our requirement). For every value of k, we calculate the within-cluster sum of squares (WCSS) value. WCSS - It is defined as the sum of square distances between the centroids and each points. As we have got the number of clusters, so we can now train the model on the dataset.To train the model, we will use the same two lines of code as we have used in the above section, but here instead of using i, we will use 5, as we know there are 5 clusters that need to be formed. The code is given below:

#training the K-means model on a dataset

kmeans = KMeans(n\_clusters=5, init='k-means++', random\_state= 42)

y\_predict= kmeans.fit\_predict(x)  [4]

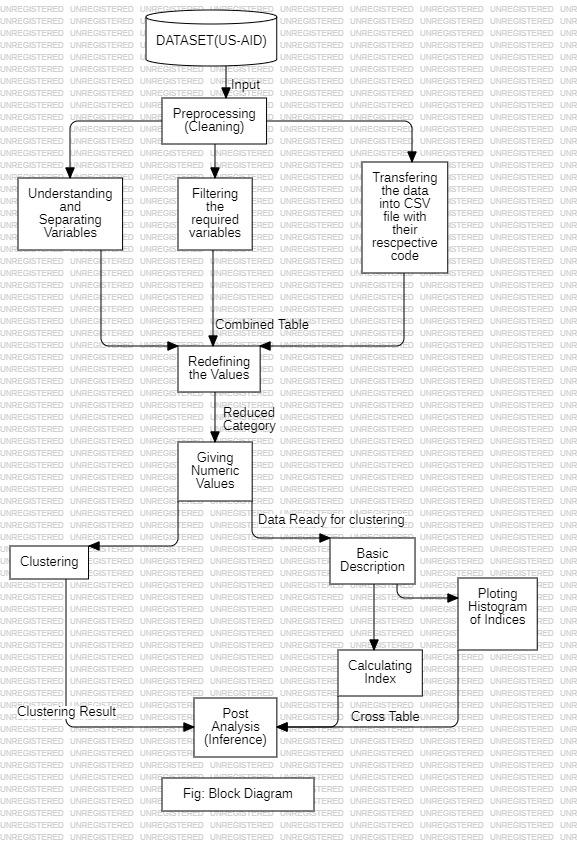
# 

# Chapter 3: METHODOLOGY

## 3.1 **Introduction**

This study employs cluster analysis to explore and understand the social stratification in Nepali society. Cluster analysis is a statistical method that groups individuals based on similarities across various characteristics, allowing for the identification of distinct social clusters. We will analyze data from national surveys and censuses, focusing on demographic, economic, and cultural variables. By leveraging software tools such as R and Python, we aim to reveal patterns and groupings that highlight the complex layers of social inequality. This methodological approach provides a nuanced perspective on how traditional and modern factors contribute to social stratification in Nepal.

## 3.2 Block Diagram



**Fig 3.2: Block Diagram**

## 3.3 Methodology

### 3.3.1. Obtaining required data from US-AID (DATASET):

We knew from our supervisor there is already a survey done by the US-AID so we just need that data for our project. We sent an email to the US-AID office and they responded with data files after a few weeks.

### 3.3.2. Pre-processing (Cleaning):

The study primarily relies on publicly available data sources, including the US-AID survey, The Nepal Demographic Health Survey 2022 is conducted by selecting different households from across the country to represent the whole population. Which were merged, processed, and normalized to conduct the analysis.

1. **Understanding and separating variables:**

Since there was data from 1987,1996,2001,…….. 2022. We selected the year 2022 and it consists of thousands of variables. Our challenge was to pick only those variables that affect stratification in our society. It was challenging because we had to pick it manually in MS Word. After analyzing all the data we came to know there exists a file name called “NPHR81SP” which only includes household records.

1. **Filtering the required variables:**

After separating variables in MS-WORD. We categorized those variables from MS-WORD into three categories i.e. Health, Wealth, and Education. Besides these, we also included a category named “others” where variables like a province, district, etc are included.

1. **Transfering the Data into CSV files with their respective code:**

Another challenge was to convert the SAV file to a CSV file. We tackle these challenges by writing code in Python. We used the “pyreadstat” library of Python to read the SAV file in VS Code and eventually convert it into a CSV file.

### 3.3.3 Redefining the values:

There are different data variables like type of toilet facilities, source of drinking water, types of cooking fuel, and type of cooking stove which contain categorical values that make k-means clustering difficult to apply so we updated those values into two categories. Eg. For types of toilets are traditional and modern, the same goes for cooking fuel and stoves, and sources of drinking water, which are safe and unsafe.

### 3.3.4 Giving numeric values:

Here, we replaced those updated values with numeric values like

|  |  |
| --- | --- |
| Safe = 1 | Unsafe = 0 |
| Modern = 1 | Traditional = 0 |

#### **3.3.4.1. Basic Description:**

This block contains calculating indices block and plotting a histogram of indices combined.

##### **a) Calculating indices:**

For calculating the Health index:

There is a value like 0 and 1 from the above-updated data, we calculate the index by adding all these values.

Eg.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Types of toilet facilities | Source of drinking water | Types of cooking fuel | Type of cooking stove | Health index  (=SUM(D1:D2;D3;D4)) |
| 1 | 1 | 0 | 0 | 2 |

For calculating the Education index:

The highest Education Level contains data values which are mentioned below. For Basic Education level index value is calculated as the number of years required to reach class 5 (for eg: Nursary=1, LKG=2, UKG=3, Class 1=4, Class 2=5, Class 3=6, Class 4=7, Class 5=8 years)

Eg.

|  |  |
| --- | --- |
| Highest Education Level | Education Index  (In years) |
| 0 (No Education) | 0 |
| 1 (Basic) | 8 (Up to Class 5) |
| 2 (Secondary) | 13 (Up to Class 10) |
| 3 (Higher) | 19 (Up to Bachelor) |

##### **b) Plotting histogram of Indices:**

In this step, our team tried to understand variables and their relationship. we plot a histogram of three indices (Health index, Wealth index, Education index). From the diagrams of histograms, we concluded some results.

#### **3.3.4.2. Clustering:**

Clustering is the task of dividing data sets into a certain number of clusters in such a manner that the data points belonging to a cluster have similar characteristics. There are several Clustering algorithms available such as K-Means Clustering, Hierarchical Clustering, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), etc. From those algorithms, we are doing clustering from the K-Means Clustering algorithm.

**a) K-Means Clustering:**

K-means clustering is an unsupervised learning algorithm that is used to solve clustering problems in machine learning or data science. In this topic, we will learn what is K-means clustering algorithm is, how the algorithm works, along with the Python implementation of k-means clustering.

**b) Elbow method:**

The Elbow Method is a technique used to determine the optimal number of clusters (K) in K-means clustering. We used this method to determine the number of clusters according to our data.

**c) Centroids:**

K-means clustering is a technique to group data into K clusters, where each cluster is represented by a centroid, which is the average of all points in the cluster. Centroids are essential in K-means clustering, helping to form and refine clusters.

**Finding Centroid:**

We used the in-built function in Python to find the centroids of each cluster:kmeans\_clusters\_centers

### 3.3.5 Post Analysis (Inference):

After clustering, we conclude that there are 5 clusters which include the Education, Health, and Wealth indexes where each cluster is represented by a centroid, which is the average of all points in the cluster concerning their indexes. It concludes that the higher the mean value have higher the level of status in Nepalese society.

# Chapter 4: RESULT AND ANALYSIS

## 4.1 Histogram of indices

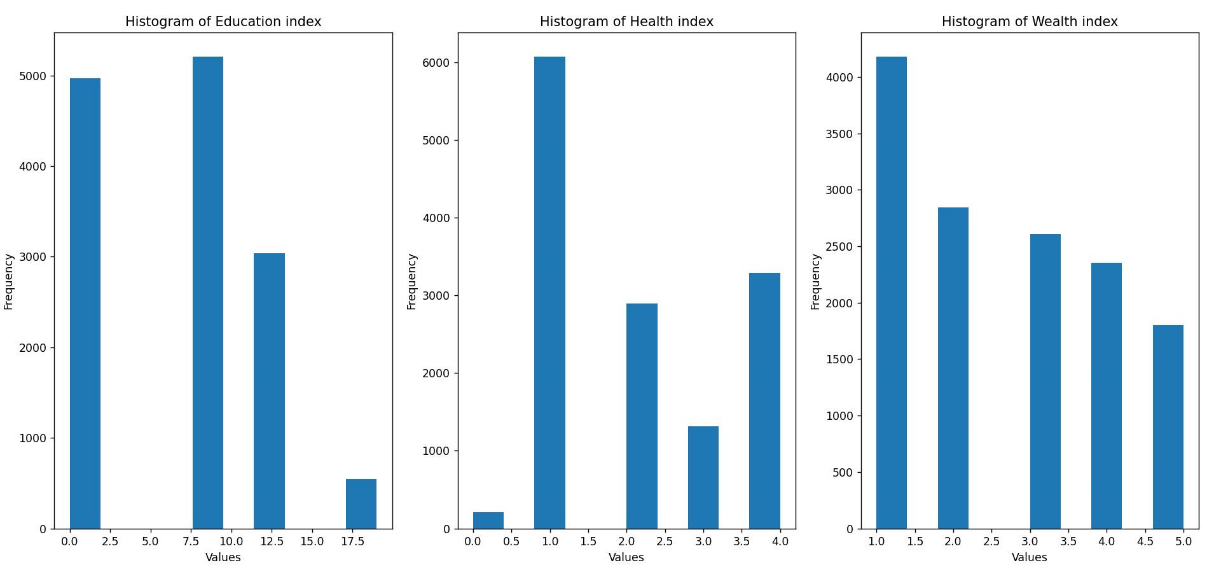


Fig 4.1: Histogram according to index

This is the histogram of the Education index, Health index, and Wealth index

## 4.2 Elbow method

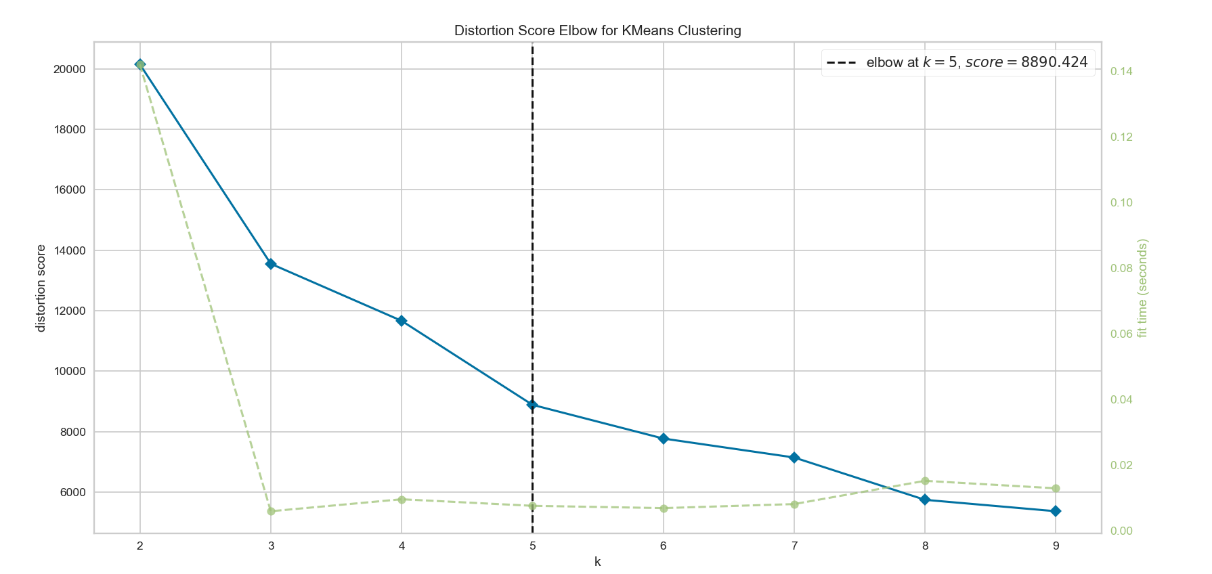


Fig 4.2: Number of Clusters

Using the elbow method we have calculated cluster number for applying the K-means algorithm. From the above diagram we can see there is 5 cluster number k=5 for this project.

## 4.3 Cluster Centroids

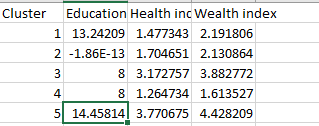


Fig 4.3: Centroids of Cluster

There are 5 clusters concluded for our projects according to our variables:

Cluster 1: consists of households that have higher literate people followed by wealth index and lower index of health so according to our research province having these characteristics is Karnali province source = Cental beuro of Statistics (CBS), Nepal Human Development Report.

Cluster 2: consists of a region that has an education index nearly equal to zero and a higher health and wealth index so, according to the research districts such as Humla, mugu, jumla and kalikot, etc are included.

Cluster 5:

# REFERENCES

1. Haron, N.H. (2022) *Stratified sampling using cluster analysis*, *The 5th Innovation and Analytics Conference & Exhibition (IACE 2021)*. Available at: https://www.academia.edu/86693570/Stratified\_sampling\_using\_cluster\_analysis (Accessed: 10 January 2024).
2. (No date) *Nepal demographic and Health Survey 2022: Key indicators report*. Available at: https://mohp.gov.np/uploads/Resources/Nepal%20Demographic%20and%20Health%20Survey%202022%20Key%20Indicators%20Report.pdf (Accessed: 10 January 2024).
3. Sciforce (2019) *Data Cleaning and preprocessing for Beginners*, *Medium*. Available at: https://medium.com/sciforce/data-cleaning-and-preprocessing-for-beginners-25748ee00743 (Accessed: 20 June 2024).
4. Tan, P.-N. (no date) *Introduction to data mining*, *Google Books*. Available at: https://books.google.com/books/about/Introduction\_to\_Data\_Mining.html?id=Wx4NPK4qHXsC (Accessed: 20 June 2024).
5. (No date a) *Dhsprogram*. Available at: https://www.dhsprogram.com/pubs/pdf/FR379/FR379.pdf (Accessed: 20 June 2024).
6. (No date a) *Technical notes - the united nations*. Available at: https://data.un.org/\_Docs/HDR\_2011\_EN\_TechNotes.pdf (Accessed: 20 June 2024).

# APPENDIX

Fig A:

Fig B:

Fig C: