UNIVERSIDAD NACIONAL DEL ALTIPLANO

FACULTAD DE INGENIERÍA ESTADÍSTICA E INFORMÁTICA

ESCUELA PROFESIONAL DE INGENIERÍA ESTADÍSTICA E INFORMÁTICA

SEGMENTACIÓN DE HOGARES CON INDICADORES SOCIOECONÓMICOS DEL DISTRITO DE MACUSANI - 2020

TESIS

PRESENTADA POR:

Bach. JUAN MANUEL CONDORI PERALTA

PARA OPTAR EL TÍTULO PROFESIONAL DE:

INGENIERO ESTADÍSTICO E INFORMÁTICO

PUNO – PERÚ

2024

NOMBRE DEL TRABAJO

SEGMENTACIÓN DE HOGARES CON INDICADORES SOCIOECONÓMICOS DEL DISTRITO DE MACUSANI - 2020

AUTOR

JUAN MANUEL CONDORI PERALTA

RECUENTO DE PALABRAS

17444 Words

RECUENTO DE CARACTERES

101377 Characters

RECUENTO DE PÁGINAS

99 Pages

TAMAÑO DEL ARCHIVO

1.9MB

FECHA DE ENTREGA

Jan 10, 2024 10:07 AM GMT-5

FECHA DEL INFORME

Jan 10, 2024 10:09 AM GMT-5

13% de similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para cada base

- 11% Base de datos de Internet
- 4% Base de datos de publicaciones
- Base de datos de Crossref
- Base de datos de contenido publicado de Cross
- 9% Base de datos de trabajos entregados

Excluir del Reporte de Similitud

- Material bibliográfico
- Material citado
- Material citado
- Coincidencia baja (menos de 10 palabras)

José P. Tito Lipa
Ing. Estadístico e Informático D.Sc.
CIP. 159645

M.Sc. Elqui Yeye Pari C.
CIP. N° 116626

DEDICATORIA

A mis queridos padres, Juan Isidro y Rosa Elsa, cuya amor, sacrificio y guía han sido la luz en mi camino. A ellos les debo no solo mi existencia, sino también la fortaleza y los valores que me han llevado a alcanzar este logro.

A mis hermanos, compañeros eternos de mi vida, quienes, con su apoyo incondicional y sus consejos sinceros, han sido pilares fundamentales en mi viaje. Su confianza y creencia en mis capacidades han sido una fuente constante de motivación.

Y a mis sobrinos, que, con su alegría y cariño inagotable, me han recordado siempre la importancia de perseguir mis sueños y la belleza de la vida más allá del ámbito académico.

Este trabajo es tanto mío como suyo, un testimonio de amor, unidad y perseverancia familiar.

Con todo mi cariño y gratitud.

Juan Manuel Condori Peralta

AGRADECIMIENTOS

En un momento importante de mi carrera académica, quiero agradecer a quienes han apoyado mi investigación. Mi gratitud a la Universidad Nacional del Altiplano de Puno, especialmente a la Escuela de Ingeniería Estadística e Informática, por una educación excepcional y un entorno propicio para mi desarrollo. Agradezco también a la Oficina de SISFO de la Municipalidad de Macusani por el acceso a datos clave para mi estudio. Un reconocimiento especial a mis docentes y al MSc. Elqui Yeye P.C., mi asesor, cuya orientación ha sido fundamental en todas las fases de mi investigación. Por último, agradezco a los miembros del jurado por sus valiosas revisiones y contribuciones, que han mejorado significativamente la calidad y rigor de mi trabajo.

Juan Manuel Condori Peralta

ÍNDICE GENERAL

Pág.

DEDICATORIA

AGRADECIMIENTOS

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE ANEXOS

ACRÓNIMOS

RESUMEN 14

ABSTRACT 15

CAPÍTULO I

INTRODUCCIÓN

1.1. PLANTEAMIENTO DEL PROBLEMA 17

1.2. FORMULACION DEL PROBLEMA 18

1.2.1. Problema general 18

1.2.2. Problemas específicos 18

1.3. JUSTIFICACIÓN DE LA INVESTIGACIÓN 18

1.4. OBJETIVOS DE LA INVESTIGACIÓN 19

1.4.1. Objetivo general 19

1.4.2. Objetivos específicos 19

1.5. HIPÓTESIS DE LA INVESTIGACIÓN 20

1.5.1. Hipótesis General. 20

CAPÍTULO II

REVISIÓN DE LITERATURA

2.1. ANTECEDENTES 21

2.1.1. Internacionales 21

2.1.2. Nacionales 23

2.1.3. Locales y regionales 26

2.2. MARCO TEÓRICO 28

2.2.1. Supervisado y No Supervisado 28

2.2.2. Supervisado 29

2.2.3. No supervisado 30

2.2.4. Reducción de Dimensionalidad. 30

2.2.5. Agrupamiento o clustering. 32

2.2.5.1. K-means 32

2.2.5.2. Agrupamiento Jerárquico 33

2.2.6. Clustering K- means 33

2.2.7. clustering K-medoids (PAM) 35

2.2.8. Clustering Jerárquico 36

2.2.8.1. clustering Jerárquico Aglomerativo 37

2.2.8.2. Clustering jerárquico divisivo 37

2.2.9. Métodos de enlace 38

2.2.10. Métodos para medir distancias 40

2.2.10.1. Distancia Euclidiana 41

2.2.10.2. Distancia Manhattan 42

2.2.10.3. Distancia Gower 43

2.2.11. Validación de Clústeres 44

2.2.11.1. Evaluación de la Tendencia del conjunto de datos. 44

2.2.11.2. Determinación del Número Óptimo de Clústeres 45

2.2.11.2.1. Método del Codo 46

2.2.11.2.2. Método del promedio de siluetas 47

2.2.11.3. Validación interna 49

2.2.11.3.1. Validación interna índice Davies-Bouldin: 49

2.2.11.3.2. Validación interna Indice Dunn: 50

2.2.11.3.3. Validación interna Indice conectividad: 51

2.2.11.3.4. Validación interna Indice Anchura de Silueta 52

2.2.12. Pre procesamiento de datos 53

2.2.12.1. Limpieza de Datos 53

2.2.12.2. Transformación de Datos 54

2.2.13. Análisis de correlación 54

2.3. MARCO CONCEPTUAL 56

2.3.1. Hogar 56

2.3.1.1. Composición 56

2.3.1.2. Funcionamiento Económico y Doméstico Conjunto 56

2.3.2. SISFOH 56

2.3.3. Vivienda 57

2.3.4. Tipo de seguro 57

2.3.4.1. Seguro Integral de Salud (SIS) 57

2.3.4.2. EsSalud 58

2.3.4.3. Seguros Privados 58

2.3.5. Nivel educativo 58

2.3.6. Tipo de combustible de cocina 58

2.3.7. Indicadores socioeconómicos 58

CAPÍTULO III

MATERIALES Y MÉTODOS

3.1. MATERIALES 60

3.2. METODOLOGÍA 60

3.2.1. Tipo de investigación 60

3.2.2. Diseño de investigación 60

3.2.3. Población y muestra 61

3.2.3.1. Población 61

3.2.3.2. Muestra 61

3.2.4. Lugar de investigación 61

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

4.1. RESULTADOS 63

4.1.1. Pre procesamiento de datos 63

4.1.2. Limpieza de Datos: 63

4.1.3. Transformación de Datos: 67

4.1.4. Evaluación de la Tendencia del conjunto de datos 69

4.1.5. Análisis de correlación 70

4.1.6. Primer objetivo especifico 71

4.1.7. Segundo objetivo especifico 73

4.1.7.1. Clustering K-means 73

4.1.7.2. Clustering K-medoids (PAM): 76

4.1.7.3. Clustering jerárquico 78

4.1.8. Encontrar del algoritmo más efectivo 80

4.1.9. Tercer objetivo especifico 81

4.2. DISCUSIÓN 82

V.CONCLUSIONES 84

VI. RECOMENDACIONES 85

VII. REFERENCIAS BIBLIOGRÁFICAS 86

ANEXOS 89

AREA: Estadística e informática

TEMA: Análisis multivariado, Big data, Minería de datos e investigación de mercados

FECHA DE SUSTENTACIÓN: 12 de enero del 2024

ÍNDICE DE TABLAS

Pág.

Tabla 1 Distribución del algoritmo K-means con K=2 73

Tabla 2 Distribución del algoritmo K-medoids (PAM) con K=2 76

Tabla 3 Distribución del algoritmo jerárquico con K=2 78

Tabla 4 Puntaje de índices de evaluación 80

ÍNDICE DE FIGURAS

Pág.

Figura 1 Clustering Jerárquico Aglomerativo 37

Figura 2 Clustering jerárquico divisivo 38

Figura 3 Gráfico del Método del Codo 47

Figura 4 Gráfico del método de la silueta promedio 48

Figura 5 Mapa de la provincia de Carabaya – Macusani 62

Figura 6 Distribución del distrito de Macusani (Urbano y Rural) 63

Figura 7 Porcentaje de datos faltantes por variables 64

Figura 8 Porcentaje de datos faltantes por variable 66

Figura 9 Estructura de los conjuntos de datos 67

Figura 10 Estructura de los conjuntos de datos 68

Figura 11 Análisis de correlación 70

Figura 12 Determinar número de clúster mediante visualización de dendograma. 71

Figura 13 Determinar del número óptimo de clúster con K-means con método del codo 72

Figura 14 Determinar del número óptimo de clúster con K-means con método de la silueta 73

Figura 15 Visualización de clústeres con algoritmo K-means con k=2 75

Figura 16 Visualización de Clústeres con algoritmo K-medoids(PAM) con k=2 77

Figura 17 Visualización de Clústeres con algoritmo jerárquico con k=2 79

ÍNDICE DE ANEXOS

Pág.

ANEXO 1 Matriz de consistencia 89

ANEXO 2 Caracterización de los clústeres 91

ANEXO 3 Hogares por clúster 93

ANEXO 4 Código en R- estudio 95

ANEXO 5 Declaración jurada de autenticidad de tesis 98

ANEXO 6 Autorización para el depósito de tesis en el Repositorio Institucional 99

ACRÓNIMOS

SISFOH:	Sistema de Focalización de Hogares.
IA:	Inteligencia Artificial.
PAM:	Partición Alrededor de Medoides.
SIS:	Seguro Integral de Salud.
INEI:	Instituto Nacional de Estadística e Informática.

RESUMEN

Esta investigación abordó la determinación del método para segmentar hogares utilizando indicadores socioeconómicos en el distrito de Macusani durante el año 2020. El estudio, basado en un conjunto de datos proporcionado por la oficina de SISFOH de la Municipalidad de Macusani, incluyó la totalidad de los 4,329 hogares del distrito Macusani donde incluye rural y urbano, centrándose específicamente en una muestra detallada de 344 hogares en rurales. Con un diseño de investigación descriptivo y no experimental, el proyecto comenzó con la preparación de los datos. La evaluación inicial del conjunto de datos, realizada mediante el test de Hopkins, arrojó un valor de 0.7246674, lo que confirmó de que el conjunto de datos muestra una tendencia de agrupamiento. la adecuación de los datos para análisis de clustering. Se procedió a determinar el número óptimo de clústeres, empleando métodos como dendogramas, el método del codo y el de la silueta, que colectivamente sugerían que la división en dos clústeres (k=2) era la más apropiada. Los análisis de clustering se realizaron utilizando técnicas variadas, incluyendo los algoritmos k-means, PAM y jerárquico. Se eligió el algoritmo jerárquico, Tiene los mejores puntajes en Connectivity, Dunn y Silhouette, lo que indica que crea clústeres bien definidos, compactos y separados. Donde el clúster 1 conformado 137 hogares y el clúster 2 con 207 hogares, donde el Clúster 1 se caracterizó por tener una mayor proporción de hogares con acceso al seguro de salud SIS, niveles de educación que alcanzan principalmente hasta la primaria, y una tendencia a la empleabilidad en sectores independientes. En contraste, el Clúster 2 se caracterizó por un mayor porcentaje de hogares con acceso al SIS, pero con una notable participación en el sector informal o doméstico, así como un predominio de viviendas construidas con materiales de piedra o madera.

Palabras clave: Análisis de clúster, jerárquico, k-means, Macusani, Segmentación.

ABSTRACT

This research addressed the determination of the method for segmenting households using socioeconomic indicators in the Macusani district during the year 2020. The study, based on a dataset provided by the SISFOH office of the Municipality of Macusani, included the entirety of the 4,329 households in the Macusani district, encompassing both rural and urban areas, with a specific focus on a detailed sample of 344 rural households. With a descriptive and non-experimental research design, the project began with data preparation. The initial assessment of the dataset, conducted using the Hopkins test, yielded a value of 0.7246674, confirming that the dataset exhibits a tendency for clustering. This confirmed the suitability of the data for cluster analysis. The optimal number of clusters was determined using methods such as dendrograms, the elbow method, and the silhouette method, which collectively suggested that dividing into two clusters (k=2) was most appropriate. Cluster analysis was conducted using various techniques, including k-means, PAM, and hierarchical algorithms. The hierarchical algorithm was chosen, as it has the best scores in Connectivity, Dunn, and Silhouette, indicating that it creates well-defined, compact, and separate clusters. Cluster 1 comprised 137 households and Cluster 2 consisted of 207 households, where Cluster 1 was characterized by a higher proportion of households with access to the SIS health insurance, education levels mainly up to primary school, and a tendency towards employment in independent sectors. In contrast, Cluster 2 was characterized by a higher percentage of households with access to SIS, but with a notable participation in the informal or domestic sector, and a predominance of houses built with stone or wood materials.

Keywords: Cluster Analysis, Hierarchical, k-means, Macusani, Segmentation

CAPÍTULO I

INTRODUCCIÓN

La diversidad y heterogeneidad de las realidades socioeconómicas en los departamentos han incrementado la complejidad de realizar análisis que sean universalmente válidos. Dada la falta de un criterio de clasificación basado en la teoría del desarrollo ampliamente aceptado, proponer una clasificación se torna imprescindible para impulsar el desarrollo territorial. Estas clasificaciones son cruciales para orientar las políticas públicas en áreas fundamentales como educación, salud, producción, seguridad, vivienda y transportes. Además, considerando que la clasificación es un proceso esencial en la ciencia, ya que permite ordenar los fenómenos para su mejor comprensión, este estudio asume una relevancia particular.

Por lo tanto, la investigación emprendida se centró en el análisis de los hogares del distrito de Macusani, con el objetivo de establecer una segmentación de estos utilizando indicadores socioeconómicos del periodo 2020. La información necesaria fue obtenida de la oficina de SISFOH de la municipalidad distrital de Macusani. El propósito fundamental fue conocer mejor a los hogares para optimizar la asignación de ayudas y la implementación de programas socioeconómicos por parte de la Municipalidad del distrito de Macusani, especialmente en el contexto de la pandemia de Covid-19, que reveló deficiencias en la distribución de beneficios a los hogares vulnerables.

La investigación se estructuró de la siguiente manera:

Capítulo I se centró en la definición del problema, estableciendo tanto los problemas generales como las específicas, y justificando la relevancia del estudio. Además, se establecieron los objetivos y las hipótesis de la investigación.

Capítulo II: Abordó los antecedentes del tema, considerando estudios a nivel internacional, nacional y local. Además, se revisaron las bases teóricas pertinentes al área de estudio.

Capítulo III: Presentó los materiales y métodos, describiendo el lugar de estudio, la población, muestreo y la metodología de investigación adoptada.

Capítulo IV: Expuso los resultados obtenidos y la discusión de los mismos.

Finalmente, se presentaron las conclusiones, recomendaciones y anexos pertinentes al estudio. Mediante este enfoque, la investigación buscó aportar conocimientos oportunos y relevantes para el desarrollo de estrategias eficaces en la distribución de recursos y ayudas en el distrito de Macusani.

1.1. PLANTEAMIENTO DEL PROBLEMA

En el distrito de Macusani, en la provincia de Carabaya, departamento de Puno, enfrentó un desafío significativo durante la crisis de COVID-19 en mayo de 2020, particularmente en la implementación de iniciativas de apoyo social. La municipalidad del distrito de Macusani tomó medidas proactivas para asistir a los hogares vulnerables, proporcionando una serie de ayudas como entrega de chips para estudiantes, bonos, alimentos básicos, entre otros. Sin embargo, se encontraron con un obstáculo considerable: la identificación precisa de los hogares en situación de vulnerabilidad. Durante este período crítico, la oficina de SISFOH (Sistema de Focalización de Hogares), encargada de la clasificación socioeconómica de los hogares, tenía su sistema en mantenimiento, lo que exacerbó las dificultades existentes. Aunque se recopiló información a través de encuestas por parte de la oficina de SISFOH, no se contaba con un mecanismo eficaz y preciso para determinar cuáles eran los hogares más necesitados. Esta situación puso de manifiesto la necesidad urgente de una segmentación efectiva de

hogares utilizando indicadores socioeconómicos específicos para el distrito de Macusani. La necesidad de una segmentación eficaz de hogares usando indicadores socioeconómicos específicos para Macusani era evidente, pero el distrito carecía de un método establecido para llevar a cabo dicha segmentación. Además, no se había determinado en cuántos grupos debían clasificarse estos hogares, ni se sabía si la segmentación propuesta sería la más efectiva. Estas incertidumbres planteaban interrogantes significativos sobre cómo se debería caracterizar a cada grupo dentro de la segmentación, para asegurar una distribución de recursos justa y eficiente.

1.2. FORMULACIÓN DEL PROBLEMA

1.2.1. Problema general

- ¿Cuál es el método para la segmentación de hogares con indicadores socioeconómico en el distrito de Macusani del 2020?

1.2.2. Problemas específicos

- ¿Cuál es el número óptimo de clústeres para la segmentación de hogares con indicadores socioeconómico en el distrito de Macusani del 2020?
- ¿Cuál es el algoritmo de análisis de clúster más efectivo para la segmentación de hogares con indicadores socioeconómico en el distrito de Macusani del 2020?
- ¿Cómo se caracteriza la segmentación de hogares con indicadores socioeconómico en el distrito de Macusani del 2020?

1.3. JUSTIFICACIÓN DE LA INVESTIGACIÓN

Las realidades socioeconómicas en los hogares del distrito de Macusani son cada vez más variadas y distintas. Esta diversidad se hizo más evidente con la llegada de la

pandemia del Covid-19, lo que resaltó la necesidad de segmentar los hogares del distrito. El propósito de esta segmentación es entender mejor las diferentes situaciones de los hogares para mejorar la forma en que se entregan las ayudas y se implementan los programas socioeconómicos de la Municipalidad Distrital de Macusani. Así surge la necesidad de esta investigación: para identificar los hogares más necesitados del distrito y utilizar los resultados como una herramienta para mejorar la asignación de beneficios a los hogares vulnerables, enfocándonos en saber a qué hogares dar el apoyo necesario para su mejora socioeconómica.

El conocimiento generado por este estudio ayudará a mantener informados a los encargados de la oficina SISFOH de la Municipalidad Provincial de Carabaya – Macusani y a la población del distrito de Macusani, contribuyendo a la toma de decisiones más efectivas. Finalmente, con los resultados obtenidos, se busca supervisar mejor a los hogares del distrito de Macusani y asegurar una distribución adecuada de los beneficios de hogares en situación de vulnerabilidad.

1.4. OBJETIVOS DE LA INVESTIGACIÓN

1.4.1. Objetivo general

- Determinar el método para la segmentación de hogares con indicadores socioeconómico en el distrito de Macusani del 2020.

1.4.2. Objetivos específicos

- Establecer el número óptimo de clústeres para la segmentación de hogares con indicadores socioeconómico en el distrito de Macusani del 2020.

- Encontrar el algoritmo de análisis de clúster más efectivo para la segmentación de hogares con indicadores socioeconómico en el distrito de Macusani del 2020.
- Caracterizar la segmentación de hogares con indicadores socioeconómico en el distrito de Macusani del 2020.

1.5. HIPÓTESIS DE LA INVESTIGACIÓN

1.5.1. Hipótesis General.

- El análisis de clúster es el método para la segmentación de hogares con indicadores socioeconómico en el distrito de Macusani del 2020.

CAPÍTULO II

REVISIÓN DE LITERATURA

2.1. ANTECEDENTES

2.1.1. Internacionales

Alvarado (2021) El proyecto de investigación se enfocó en establecer conglomerados en la provincia de Chimborazo en 2018, utilizando análisis de conglomerados jerárquicos para examinar 10 cantones con base en indicadores demográficos, económicos y sociales. Se empleó el software SPSS para identificar disparidades socioeconómicas entre los cantones. Como resultado, se formaron cuatro conglomerados, destacando a Riobamba por su heterogeneidad y otros tres conglomerados tras una depuración de variables. El estudio concluyó que las diferencias socioeconómicas en Chimborazo se deben a factores como la extensión territorial, la capacidad de generar valor agregado y el aprovechamiento de recursos locales.

Jackson et al. (2021) El estudio se enfocó en las escuelas oficiales de Panamá, con el objetivo de identificar aquellas que, a pesar de estar en contextos socioeconómicos desfavorables, lograron superar el promedio nacional en pruebas de lectura, matemáticas y ciencias para tercero y sexto grado. Utilizando un enfoque cuantitativo y el método de conglomerados de K medias, se analizaron 137 escuelas oficiales agrupadas según el nivel socioeconómico y cultural. Los resultados mostraron diferencias significativas en el rendimiento escolar asociadas a la condición socioeconómica, proporcionando bases para formular políticas educativas nacionales en Panamá. Las conclusiones resaltan la influencia

del entorno socioeconómico en el rendimiento académico, subrayando la necesidad de enfocar esfuerzos en las comunidades más desfavorecidas para mejorar la calidad educativa.

Parra (2020) La investigación se centró en el uso de técnicas de análisis multivariable para investigar las tasas de mortalidad en los departamentos de Colombia durante el año 2018, siguiendo la lista 6/67 de la Organización Panamericana de la Salud. El propósito principal fue descubrir similitudes y diferencias en los índices de mortalidad entre los departamentos y su conexión con aspectos socioeconómicos, aplicando métodos de análisis tanto jerárquicos como no jerárquicos, además de un análisis de correspondencias. Los resultados indicaron que las enfermedades del sistema circulatorio y las neoplasias eran las causas más comunes de muerte. Se observó que los departamentos con índices más elevados de estas enfermedades tenían generalmente mejores condiciones y calidad de vida, mientras que los departamentos con menores recursos evidenciaban tasas más bajas de estas enfermedades principales, pero altas en enfermedades infecciosas. El estudio utilizó un enfoque mixto, descriptivo y exploratorio, de corte transversal, analizando datos de mortalidad de los 33 departamentos de Colombia obtenidos del Sistema Integrado de Información de la Protección Social (SISPRO).

Finalmente, Chaparro et al.(2016) La investigación se enfocó en la clasificación de perfiles de alumnos de secundaria en Baja California, México, considerando su desempeño académico, situación socioeconómica, capital cultural y estructura familiar. Participaron 21,724 estudiantes y se empleó el análisis de clústeres K-medias para dividirlos en dos grupos diferenciados. Los hallazgos revelaron que el grupo con estudiantes de alto rendimiento se

caracterizaba por tener niveles socioeconómicos más elevados, un mayor capital cultural y una mayor participación familiar, en contraste con el grupo de bajo rendimiento, que mostraba tendencias opuestas en estas áreas. El estudio resaltó la influencia significativa de los factores familiares en la formación de los perfiles de los estudiantes y su conexión con el rendimiento escolar, enfatizando la necesidad de tomar en cuenta estos elementos para optimizar las políticas y estrategias en el ámbito educativo.

2.1.2. Nacionales

Piedra (2022) El estudio tuvo como finalidad mejorar las estrategias comerciales mediante la segmentación detallada de su mercado objetivo. Los propósitos del estudio incluyeron la creación de perfiles específicos para diferentes grupos de clientes, la determinación de la cantidad ideal de segmentos de compradores potenciales, y la evaluación comparativa de las metodologías de K-means y Vecinos más cercanos para la segmentación efectiva. Desde un enfoque metodológico, este análisis fue de carácter explicativo y cuantitativo, llevado a cabo de forma no experimental y transversal entre la población de Lima metropolitana y el Callao a principios de 2019. Se centró en individuos de 18 a 70 años pertenecientes a los niveles socioeconómicos A, B y C, con ingresos familiares desde S/. 3000, interesados en proyectos inmobiliarios. La muestra, extraída aleatoriamente de una base de datos de 6350 casos, consistió en 1100 encuestas telefónicas que incluyeron preguntas de naturaleza cuantitativa y cualitativa. El análisis se realizó mediante el software R, donde se decidió emplear el algoritmo K-means sobre Vecinos más cercanos. El resultado fue la identificación de cuatro categorías distintas de clientes, lo que permitió una

comunicación y ofertas más personalizadas y alineadas con las necesidades específicas de los clientes.

Calisaya (2021) El estudio se centró en clasificar y caracterizar los Centros de Educación Técnico-Productiva (CETPRO) públicos utilizando clúster bietápico, con el objetivo de identificar grupos basados en indicadores de calidad. Los objetivos abarcaron la determinación de atributos clave de los CETPRO en áreas como gestión, infraestructura, docencia y finanzas. Adoptando un enfoque cuantitativo no experimental, se recopiló información de todos los CETPRO públicos nacionales, sumando 704 instituciones, mediante cuestionarios enfocados en aspectos clave de calidad. El análisis incluyó una evaluación descriptiva y la limpieza de datos para asegurar la consistencia. Finalmente, el estudio identificó dos grupos principales de CETPRO, con 324 y 360 instituciones cada uno, diferenciados y descritos según criterios de calidad fundamentales.

Deza (2021)El estudio se centró en segmentar a los adolescentes infractores empleando el algoritmo PAM (Partición Alrededor de Medoides). Sus objetivos específicos incluyeron analizar exploratoriamente las características de estos adolescentes, determinar un número adecuado de agrupaciones y perfilar a los infractores. Adoptando un enfoque descriptivo, no experimental y transversal, el estudio se centró en los adolescentes infractores del Centro Juvenil, registrados en un censo a principios de 2016. La hipótesis principal sugería que estos adolescentes podrían ser segmentados en grupos basados en el tipo de delito cometido. Para realizar el estudio, se conformó una base de datos integrando registros de infractores y varias variables relacionadas con factores de riesgo y cuestiones del censo, utilizando 11 variables claves. El análisis descriptivo se llevó a cabo con el software SPSS, mientras que para el análisis cluster se usó el

software R, aplicando las técnicas de distancias de Gower y silueta para definir el número óptimo de clusters y la técnica PAM. El resultado final del estudio fue la identificación de dos segmentos principales de adolescentes infractores, diferenciados por sus perfiles demográficos y de riesgo. El primer segmento agrupaba a aquellos sin una familia nuclear y con antecedentes de consumo de drogas/alcohol, y el segundo a aquellos de familias mono parentales o bajo cuidado de terceros, viviendo en zonas con presencia de pandillas y bandas delictivas.

Chavez (2020). La investigación se orientó a definir el perfil de los nuevos estudiantes de una universidad pública mediante la aplicación de los algoritmos de segmentación K-prototypes y K-medoids. Los objetivos particulares fueron seleccionar el algoritmo de segmentación más adecuado mediante indicadores de validación interna de clusters, determinar el número ideal de grupos para este estudio y destacar las variables clave para caracterizar a los nuevos alumnos. Se analizaron datos de 690 aspirantes a la Universidad Nacional Agraria La Molina durante los ciclos académicos 2015-I y 2015-II, incluyendo variables tanto cuantitativas como cualitativas. Tras un preprocesamiento de datos, se emplearon técnicas de clustering para hallar el número óptimo de grupos y el algoritmo más apropiado. La validación de los clusters se llevó a cabo usando análisis univariados (ANOVA y prueba Chi cuadrado) y multivariados (algoritmo Boruta y árbol C5.0). Como resultado, se identificaron tres categorías distintas de alumnos, cada una con sus propias características: el Ingresante Previsto, el Ingresante en Proceso y el Ingresante en Inicio. Este hallazgo proporcionó una base sólida para el desarrollo de políticas educativas y estrategias de acompañamiento personalizado, que se alinean con el modelo educativo de la

universidad y buscan mejorar el rendimiento académico y la experiencia de los estudiantes.

Finalmente, Tang y Vargas (2016) El objetivo del estudio fue clasificar a los clientes de una tienda de electrodomésticos en Perú, enfocándose en varios puntos clave: identificar perfiles para los grupos hallados, crear estrategias de marketing personalizadas para cada perfil y sugerir nuevas variables para optimizar futuras segmentaciones. La metodología aplicada fue un análisis de conglomerados en dos etapas, adecuado para trabajar con datos tanto cuantitativos como cualitativos. Se partió de una base de datos inicial de 6284 clientes, que se redujo a 4980 tras eliminar datos atípicos y ausentes. El estudio identificó tres grupos principales: el primero con 1817 clientes (36.5%), el segundo con 1390 clientes (27.9%) y el tercero con 1773 clientes (35.6%). Basándose en estos grupos, se desarrollaron perfiles detallados y se diseñaron estrategias de marketing mix específicas para cada uno, con el fin de incrementar la lealtad de los clientes, así como las ventas y la presencia en el mercado de la tienda.

2.1.3. Locales y regionales

Flores (2020) El estudio se propuso agrupar las regiones peruanas en función de indicadores socioeconómicos. Sus objetivos específicos incluyeron la selección y evaluación de variables para definir grupos homogéneos, así como establecer criterios de agrupación basados en la proximidad de casos. La metodología aplicada fue descriptiva correlacional y no experimental, enfocada en datos del año 2018 que abarcaban 99 variables de 24 regiones. Se empleó un análisis clúster jerárquico multidimensional, resultando en seis grupos distintos de regiones: C1 (La Libertad, Lambayeque, Ancash y Piura), C2 (Madre de Dios y

Tumbes), C3 (Arequipa, Ica, Tacna y Moquegua), C4 (Lima), C5 (Loreto, Ucayali, Amazonas, Pasco y San Martin) y C6 (Cusco, Junín, Apurímac, Ayacucho, Puno, Huancavelica, Huánuco y Cajamarca). Este análisis reveló que los retos en el desarrollo económico y social de Perú exceden las barreras geográficas, destacando la diversidad y desequilibrio socioeconómico entre las regiones. La clusterización ayuda a entender mejor estas diferencias y facilita el desarrollo de estrategias regionales más efectivas.

Azañero (2020) El estudio se exploró las variaciones socioeconómicas y de salud entre las regiones de Perú. Se halló que ciertos indicadores, como partos institucionales y tasa de fecundidad, eran consistentes a nivel nacional, mientras que otros, como analfabetismo y PBI per cápita, variaban ampliamente. Se establecieron tres clústeres: el Clúster 1, con regiones de bajo riesgo y alta competitividad como Lima y Moquegua; el Clúster 2, abarcando departamentos con economías diversificadas como Loreto y Piura; y el Clúster 3, incluyendo áreas más empobrecidas y con desafíos en salud y educación, como Huánuco y Cajamarca. Este análisis subraya la necesidad de políticas diferenciadas para abordar las disparidades regionales en Perú.

Berrios (2023) El estudio se la aplicó minería de datos para analizar la información de reportes relacionados con la violencia contra las mujeres e integrantes del grupo familiar. Utilizó el algoritmo "K means" para crear clústeres individuales y el Análisis de Componentes Principales para los clústeres por variables. Se desarrollaron dos clústeres principales: el primero con el 47.83% de los reportes, centrado en casos que no incluyen violencia económica, y el segundo, con el 52.17% de los reportes, enfocado en casos donde la violencia económica es significativa. El Análisis de Componentes Principales contribuyó a la agrupación

de las variables en estos dos clústeres. Este enfoque fue validado por el Test de esfericidad de Bartlett y el índice KMO de 0.837, confirmando la interrelación significativa entre los reportes analizados.

Pacha (2018) El estudio se enfocó en la segmentación de alpacas Huacaya basada en características físicas de la calidad de su fibra. Las variables clave incluyeron el diámetro de la fibra, coeficiente de variación, longitud de mecha, índice de curvatura, fibra gruesa y factor de conformidad. Se utilizó un enfoque de Análisis Clúster Jerárquico para elegir grupos mediante un dendograma, seguido de un Análisis Clúster No Jerárquico (método K-means) para segmentar a nivel individual las alpacas Huacaya con características similares. Se identificaron tres grupos principales de alpacas, cada uno con características de medida distintas, pero homogéneos internamente y heterogéneos entre sí. La segmentación de los 208 individuos se basó en el método de enlace de promedios y la medida de distancia Euclídea. La validación se realizó mediante análisis de comparación de medias, revelando diferencias significativas entre los grupos. El estudio concluye destacando la heterogeneidad y homogeneidad encontradas en los clústeres segmentados.

2.2. MARCO TEÓRICO

2.2.1. Supervisado y No Supervisado

Patel (2019) El aprendizaje automático se divide principalmente en dos ramas: el aprendizaje supervisado y el aprendizaje no supervisado, con varias sub-ramas que conectan ambas. En el aprendizaje supervisado, el agente de IA (Inteligencia Artificial) utiliza etiquetas para mejorar su rendimiento en tareas específicas. Por ejemplo, en un filtro de spam de correo electrónico, se utilizan

etiquetas para identificar qué correos son spam. Estas etiquetas son esenciales para ayudar a la IA (Inteligencia Artificial) a diferenciar entre correos no deseados y otros. Por otro lado, en el aprendizaje no supervisado, no hay etiquetas disponibles. Esto hace que la tarea del agente de IA (Inteligencia Artificial) sea menos definida y su rendimiento más difícil de medir. Sin etiquetas, la IA (Inteligencia Artificial) intenta comprender la estructura subyacente de los datos, agrupándolos en diferentes categorías basadas en similitudes internas. Aunque este enfoque es más desafiante, puede ser más poderoso. Por ejemplo, en el caso del filtro de spam, una IA (Inteligencia Artificial) de aprendizaje no supervisado podría no solo identificar correos spam, sino también categorizar otros correos en grupos útiles como "importante", "familiar", "profesional", etc. La capacidad del aprendizaje no supervisado para descubrir patrones y categorías nuevas en datos futuros lo hace más flexible y adaptable que el aprendizaje supervisado. Esta es la fortaleza del aprendizaje no supervisado: su capacidad para encontrar patrones y relaciones no explícitas en los datos.

2.2.2. Supervisado

Patel (2019)El aprendizaje supervisado se enfoca principalmente en dos tipos de problemas: clasificación y regresión. La clasificación implica asignar elementos a categorías, siendo binaria si hay dos clases y multiclase si hay tres o más. Estos problemas se caracterizan por predecir categorías discretas. En contraste, la regresión se centra en predecir valores continuos, tratándose de problemas cuantitativos. Los algoritmos de aprendizaje supervisado varían en complejidad y buscan minimizar una función de costo o error, siempre con el objetivo de que el modelo se generalice efectivamente a nuevos casos no vistos anteriormente. La adecuación del algoritmo es crucial para reducir el error de

generalización. El modelo seleccionado debe coincidir en complejidad con la función real subyacente a los datos, lo cual es un desafío, ya que la función real es desconocida. Un modelo demasiado simple puede suba justar los datos, mientras que uno excesivamente complejo puede sobre ajustarlos, afectando la capacidad de generalización. Por lo tanto, la elección del algoritmo no siempre debe inclinarse hacia la mayor complejidad; a veces, una solución más simple es más efectiva. Entender las fortalezas, debilidades y suposiciones de cada algoritmo es esencial para aplicar con éxito el aprendizaje automático. Describiremos algunos de los algoritmos supervisados más comunes:

- Métodos Lineales.
- Métodos Basados en Vecindad.
- Estrategias de Análisis Utilizando Modelos de Árboles
- Métodos de Soporte Vectorial.
- Redes Neuronales.

2.2.3. No supervisado

Los algoritmos de aprendizaje no supervisado buscarán identificar la estructura intrínseca en los datos, tales como:

- Reducción de Dimensionalidad.
- Agrupamiento o clustering.

2.2.4. Reducción de Dimensionalidad.

Patel (2019) Los algoritmos de reducción de dimensionalidad, una categoría específica dentro del aprendizaje automático, se centran en transformar datos de alta dimensión en formatos de menor dimensión. Este proceso implica

filtrar características menos importantes mientras se retienen las más relevantes. Este enfoque facilita a las IA (Inteligencia Artificial) de aprendizaje no supervisado detectar patrones más claramente y manejar problemas complejos y de gran escala, especialmente aquellos que involucran imágenes, videos, audio y texto, de manera más eficiente. Algunos enfoques y métodos clave incluyen:

- Análisis de Componentes Principales (PCA) Este método busca una representación de baja dimensión de los datos, manteniendo la mayor variación posible. Reduce la dimensionalidad, pero preserva la estructura esencial, facilitando tareas como el agrupamiento.
- Descomposición en Valores Singulares (SVD): SVD reduce la dimensión de los datos transformando la matriz original de características en una de rango menor, capturando los elementos más importantes del espacio de características original.
- Proyección Aleatoria: Este algoritmo reduce la dimensionalidad proyectando los puntos de un espacio de alta dimensión a uno de menor dimensión, preservando la escala de distancias entre puntos.
- Aprendizaje de Manifold (Reducción de Dimensionalidad No Lineal): A diferencia de la proyección lineal, estos métodos realizan transformaciones no lineales. Ejemplos incluyen Isomap, que estima la distancia geodésica entre puntos, y t-SNE, que incrusta datos de alta dimensión en un espacio de dos o tres dimensiones para visualización.
- Aprendizaje de Diccionario: Este enfoque aprende una representación dispersa de los datos, donde cada instancia se puede reconstruir como una suma ponderada de elementos representativos binarios.

- Análisis de Componentes Independientes (ICA): ICA se utiliza para separar señales independientes mezcladas en los datos, lo cual es común en tareas de procesamiento de señales.
- Asignación de Dirichlet Latente (LDA): Utilizado principalmente en el análisis de texto, LDA descubre temas latentes en un documento, revelando la estructura oculta en un corpus de texto no estructurado.

Patel (2019) La reducción de dimensionalidad consiste en simplificar el conjunto original de características a uno más reducido, conservando únicamente las más relevantes. Este conjunto simplificado puede ser utilizado para aplicar otros algoritmos de aprendizaje no supervisado, facilitando la detección de patrones interesantes en los datos. Además, en casos donde se disponga de etiquetas, este proceso puede agilizar el entrenamiento de algoritmos de aprendizaje supervisado, ya que se utiliza esta matriz de características más compacta en lugar de la original, más compleja.

2.2.5. Agrupamiento o clustering.

Patel (2019)Esto se conoce como agrupamiento y se puede lograr con una variedad de algoritmos de aprendizaje no supervisado como:

2.2.5.1. K-means

Este algoritmo implica especificar el número de grupos (k) y asignar cada instancia a uno de estos grupos, optimizando la agrupación para minimizar la variación dentro de cada grupo. El proceso se realiza reasignando observaciones para reducir la distancia euclidiana entre cada observación y el centro de su grupo. Dado que k-means comienza con

asignaciones aleatorias, diferentes ejecuciones pueden producir resultados ligeramente distintos.

2.2.5.2. Agrupamiento Jerárquico

A diferencia de k-means, el agrupamiento jerárquico no requiere predefinir un número de grupos. Utilizando un enfoque basado en árboles, como el agrupamiento aglomerativo, se construye un dendrograma que representa gráficamente la similitud entre instancias. Las instancias o grupos más similares se unen primero, formando un árbol donde cortar el árbol en diferentes alturas determina el número de grupos. Cuanto más bajo se corta, más grupos se obtienen.

2.2.6. Clustering K- means

Según Hastie et al. (2016) fue desarrollado por MacQueen en 1967, se basa en representar cada grupo a través del punto central o la media de los datos que lo constituyen. Este enfoque, aunque efectivo, muestra sensibilidad frente a datos que son inusuales o desviados de la norma, conocidos como anómalos o atípicos. El principio central del método de K-medias es establecer conjuntos de manera que se reduzca al mínimo la variabilidad interna de cada uno, también conocida como variación intra-cluster. Hay diversas versiones de algoritmos de K-medias, y una de las más destacadas es el algoritmo de Hartigan-Wong de 1979. Este enfoque especifica la variabilidad interna de cada grupo como la suma de las distancias cuadradas euclidianas entre los puntos y el centroide de su respectivo grupo.

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

- C_k : Representa un clúster específico dentro del conjunto de datos.
- x_i : Son los puntos de datos que pertenecen a este clúster C_k.
- μ_k : Es el centroide del clúster C_k, el centroide es el punto promedio de todos los puntos de datos en C_k, calculado como la media de cada dimensión de los puntos de datos en el clúster.

Cada observación (x_i) se asigna a un grupo específico de manera que la suma de los cuadrados (SS) de la distancia de la observación a los centros de sus grupos asignados, μ_k, sea mínima. Definimos la variación total dentro de cada grupo de la siguiente manera:

$$Tvariacion = \sum_{k=1}^{k} W(C_k)$$

- $W(C_k)$: Es la suma de las diferencias al cuadrado entre los puntos de datos y el centroide del clúster C_k, Es una medida de la dispersión o variación de los puntos de datos dentro de un clúster específico.
- $\sum_{k=1}^{k}$: Esta es una sumatoria que se extiende sobre todos los clústeres en el conjunto de datos. Si hay k clústeres, entonces esta sumatoria se realiza desde $k = 1$ hasta k sumando la variación de cada clúster individual.
- $Tvariacion$ es una medida global de la dispersión o variación dentro de todos los clústeres del conjunto de datos. En el contexto del clustering, especialmente en métodos como K-means, el objetivo es minimizar $Tvariacion$, lo que implicaría que cada clúster es internamente coherente y los puntos dentro de cada clúster son similares entre sí.

Cuanto menor sea la variación total dentro de los clústeres, más precisos o bien definidos serán estos clústeres.

El algoritmo K-medias se puede describir:

- Paso 1: El analista determina la cantidad de clústeres (K) que se van a crear.
- Paso 2: Se seleccionan aleatoriamente k objetos del conjunto de datos como los centros iniciales de los grupos.
- Paso 3: Cada observación se agrupa con el centroide más próximo, utilizando la distancia euclidiana como criterio para medir la proximidad entre el objeto y el centroide.
- Paso 4: En cada uno de los k grupos, se actualiza el centroide tomando el promedio de todos los datos pertenecientes a ese grupo. El centroide de un grupo K se define como un vector de p dimensiones, donde cada dimensión representa la media de las observaciones de ese grupo para una de las p variables existentes, siendo p el número total de variables.
- Paso 5: La suma total de cuadrados dentro de cada grupo se reduce de manera iterativa. Esto implica la repetición de los pasos de asignación y actualización de centroides hasta que las asignaciones de grupo permanezcan estables o se llegue al límite establecido de iteraciones.

2.2.7. Clustering K-medoids (PAM)

Según Hastie et al. (2016)El algoritmo k-medoides es una técnica de agrupamiento que divide un conjunto de datos en k grupos, representando cada grupo por uno de sus puntos de datos, llamado medoide. Este método se distingue por su robustez, siendo menos afectado por ruidos y valores atípicos en

comparación con el clustering k-means. Requiere que el usuario especifique el número de grupos a formar y se beneficia del uso del método de la silueta para determinar este número óptimo. El algoritmo PAM (Partitioning Around Medoids) es la variante más común de k-medoides.

El algoritmo K-medoids (PAM) se puede describir:

- Paso 1: Selecciona k objetos para que actúen como medoides. Si ya se han proporcionado estos objetos, úsalos como medoides.
- Paso 2: Calcula la matriz de disimilitud si no se ha proporcionado previamente.
- Paso 3: Asigna cada objeto al medoide más cercano.
- Paso 4: Para cada agrupamiento, se examina si alguno de los objetos dentro del mismo disminuye la medida promedio de disimilitud. En caso afirmativo, se selecciona el objeto que logra la mayor reducción de dicho coeficiente para convertirse en el nuevo medoide del clúster.
- Paso 5: Si al menos un medoide ha cambiado, vuelve al paso 3; de lo contrario, finaliza el algoritmo.

2.2.8. Clustering Jerárquico

Según Los métodos de agrupamiento jerárquico, ya sean aglomerativos o divisivos, ofrecen una forma de organizar datos en clústeres de forma secuencial sin requerir especificaciones previas sobre el número de clústeres. Generan dendrogramas que representan gráficamente la estructura de clúster, donde cada nivel refleja un agrupamiento de datos. Estos dendrogramas son herramientas valiosas para identificar agrupamientos "naturales", aunque su interpretación

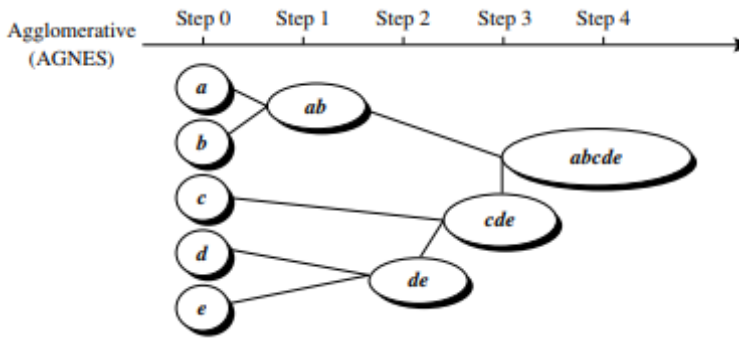
requiere cautela debido a la variabilidad en función de la metodología y los datos utilizados.

2.2.8.1. Clustering Jerárquico Aglomerativo

Según Han et al.(2011) Un método de clustering jerárquico aglomerativo emplea un enfoque ascendente. Inicia asignando a cada objeto su propio grupo y progresivamente va combinando estos grupos en otros más amplios. Este proceso continúa hasta que todos los objetos se agrupan en un único conjunto o se alcanzan determinadas condiciones de finalización, haciendo de este conjunto el núcleo de la jerarquía. Durante este proceso, identifica y une los dos grupos más similares en cada paso. Como en cada iteración se unen dos grupos, y cada uno contiene por lo menos un objeto, este método necesita un máximo de n iteraciones.

Figura 1

Clustering Jerárquico Aglomerativo



Nota: Han et al.(2011)

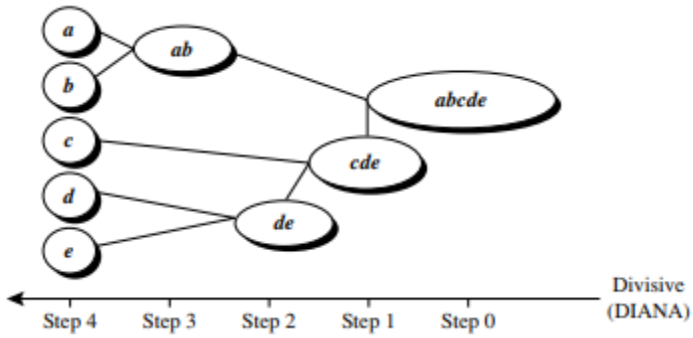
2.2.8.2. Clustering jerárquico divisivo

Según Han et al.(2011)Un enfoque de clustering jerárquico divisivo opera desde lo general a lo particular. Inicia con todos los objetos

en un solo conjunto, que es el punto de inicio de la jerarquía. Este conjunto inicial se subdivide en grupos más pequeños, y este proceso de subdivisión se repite sucesivamente. Este método sigue dividiendo los grupos hasta que cada uno en el nivel más bajo alcanza una coherencia adecuada, que puede ser tener un único objeto o que los objetos dentro de un grupo sean lo suficientemente similares. Tanto en los métodos jerárquicos aglomerativos como en los divisivos, es posible definir el número de grupos deseados como un límite para finalizar el proceso.

Figura 2

Clustering jerárquico divisivo



*Nota:*Han et al.(2011)

2.2.9. Métodos de enlace

Charrad et al. (2012) Para realizar el agrupamiento según el algoritmo de clustering jerárquico, es crucial definir cómo se mide la similitud entre dos grupos. Esto implica extender la noción de distancia entre pares de observaciones a pares de grupos, un proceso conocido como linkage. Se detallan cinco tipos comunes de linkage:

- Complete o Máximo: La distancia D_{ij} entre dos clústeres C_i y C_j es la máxima distancia entre dos puntos x e y, donde x pertenece al clúster C_i y y al clúster C_j.

$$D_{ij} = max_{x \in C_i, y \in C_j} d(x, y)$$

- Single o Mínimo: La distancia D_{ij} entre dos clústeres C_i y C_j es la distancia mínima entre dos puntos x e y, donde x pertenece al clúster C_iy y al clúster C_j.

$$D_{ij} = min_{x \in C_i, y \in C_j} d(x, y)$$

- Average: La distancia D_{ij} entre dos clústeres C_i y C_j es el promedio de las distancias entre el par de puntos x e y, donde x pertenece al clúster C_i y y al clúster C_j.

$$D_{ij} = sum_{x \in C_i, y \in C_j} \frac{d(x, y)}{n_i(n_j)}$$

donde n_i y n_j son respectivamente el número de elementos en los clústeres C_i y C_j . Este método tiene la tendencia a formar clústeres con la misma varianza y, en particular, con una varianza pequeña.

- Centroid: La distancia D_{ij} entre dos clústeres C_i y C_j es la distancia euclidiana al cuadrado entre los centros de gravedad de los dos clústeres, es decir, entre los vectores promedio de los dos clústeres, $\bar{x}_i$ y $\bar{x}_j$ respectivamente.

$$D_{ij} = \left\| \bar{x}_i - \bar{x}_j \right\|^2$$

Este método es más robusto que otros en términos de puntos aislados.

- Ward: El método de Ward se centra en reducir la varianza total dentro de los grupos. En cada etapa, fusiona aquellos pares de grupos que tengan la

menor distancia entre sí. Para aplicar este método, se identifica en cada paso el par de grupos cuya combinación conllevaría el incremento mínimo en la varianza total del grupo. Este incremento se mide como una distancia cuadrada ponderada entre los centroides de los grupos. En el método de Ward de mínima varianza, las distancias iniciales entre los grupos se calculan como la distancia euclidiana al cuadrado entre los puntos.

$$D_{ij} = \|x_i - y_i\|^2$$

Los métodos de linkage complete, average y Ward's minimum variance son comúnmente preferidos por su capacidad para generar dendrogramas equilibrados. Sin embargo, no hay un método superior; la elección depende del contexto específico del estudio. Por ejemplo, en genómica, el método de centroides es frecuentemente usado. Es importante siempre especificar la medida de distancia y el tipo de linkage utilizados en el clustering jerárquico, ya que estos pueden influir significativamente en los resultados.

2.2.10. Métodos para medir distancias

Según Kassambara (2017) La selección de métricas de distancia es un aspecto fundamental en el proceso de agrupamiento. Esta elección determina la manera en que se mide la similitud entre dos elementos (p, q) y afectará la configuración de los grupos. Los métodos convencionales para calcular estas medidas de distancia son las distancias Euclidiana y Manhattan, las cuales se describen de la siguiente manera:

2.2.10.1. Distancia Euclidiana

Según Kassambara (2017) La Distancia Euclidiana es la longitud del segmento de línea más corto entre dos puntos en un espacio euclidiano (espacio geométrico ordinario). Es la forma más común y directa de medir la distancia "en línea recta" entre dos puntos.

La fórmula para calcular la Distancia Euclidiana entre dos puntos p y q en un espacio n-dimensional es:

$$d_{eu}(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

Donde $p = (p_1, p_2 \dots, p_n)$ y $q = (q_1, q_2 \dots, q_n)$ son puntos en el espacio n-dimensional.

- n: Representa la dimensión del espacio en el que se encuentran los puntos p y q.
- p_i, q_i: Coordenadas de los puntos p y q en la i-ésima dimensión.
- La suma de los cuadrados de las diferencias de las coordenadas se toma para cada dimensión, y luego se saca la raíz cuadrada del total.

Esta fórmula proviene del Teorema de Pitágoras en geometría, que relaciona los lados de un triángulo rectángulo. En un espacio de dos dimensiones, la fórmula se reduce a la versión clásica del Teorema de Pitágoras.

Se utiliza en una variedad de contextos, especialmente en algoritmos de clustering como K-means, en análisis de componentes principales (PCA), y en general en cualquier situación donde se requiera medir distancias en un espacio continuo.

2.2.10.2. Distancia Manhattan

Según Kassambara (2017) La Distancia de Manhattan, también conocida como distancia de taxista o distancia de ciudad, mide la distancia entre dos puntos en una cuadrícula basada en una suma de sus diferencias absolutas a lo largo de los ejes.

La fórmula para calcular la Distancia de Manhattan entre dos puntos p y q en un espacio n-dimensional es:

$$d_{man}(p,q) = \sum_{i=1}^{n} |(p_i - q_i)|$$

Donde $p = (p_1, p_2 \dots, p_n)$ y $q = (q_1, q_2 \dots, q_n)$.

- n: Representa la dimensión del espacio en el que se encuentran los puntos p y q.
- p_i, q_i: Coordenadas de los puntos p y q en la i-ésima dimensión.
- La suma de las diferencias absolutas de las coordenadas se calcula a lo largo de cada dimensión.

Esta medida proviene de la geometría de la cuadrícula, donde solo se pueden realizar movimientos horizontales o verticales, como caminar por las calles de una ciudad con un diseño de cuadrícula.

Esta distancia es útil en aplicaciones donde los movimientos se restringen a una cuadrícula, como en planificación urbana, logística y optimización de rutas. También es relevante en ciertos tipos de análisis de datos donde las diferencias en cada dimensión son tratadas independientemente.

2.2.10.3. Distancia Gower

Según Gower (1971) La Distancia de Gower es una medida de similitud o disimilitud utilizada para conjuntos de datos mixtos, es decir, conjuntos de datos que incluyen variables de diferentes tipos, como nominales, ordinales, continuas y binarias. Fue desarrollada por J.C. Gower en 1971 y es particularmente útil en análisis de clustering y otras formas de análisis multivariante cuando se trabaja con tipos de datos heterogéneos.

La Distancia de Gower se calcula como un promedio ponderado de las diferencias individuales en cada variable. La fórmula general es:

$$d(i,j) = \frac{\sum_{k=1}^{n} w_{ijk} . d_{ijk}}{\sum_{k=1}^{n} w_{ijk}}$$

Donde:

- n: Es el número total de variables.
- d_{ijk}: Es la disimilitud en la variable k entre los objetos i y j.
- w_{ijk}: Es el peso asignado a la disimilitud en la variable k entre los objetos i y j, que puede ser 0 o 1 dependiendo de si esa variable es relevante para comparar esos dos objetos.

La Distancia de Gower es particularmente útil cuando los datos incluyen una mezcla de tipos de variables (cualitativos y cuantitativos), ya que puede manejar simultáneamente variables categóricas y numéricas. Es capaz de manejar datos faltantes ajustando los pesos para las comparaciones entre pares de objetos.Se utiliza en técnicas de clustering (como K-medoids) y otros métodos de análisis multivariante cuando los datos incluyen diferentes tipos de variables.

2.2.11. Validación de Clústeres

Según Kassambara (2017) La validación de clústeres consiste en medir la eficacia de los resultados obtenidos mediante técnicas de agrupamiento. Antes de emplear un algoritmo de clustering en un conjunto de datos, es crucial primero determinar si el conjunto de datos es adecuado para el agrupamiento y estimar el número potencial de clústeres. Posteriormente, se pueden aplicar métodos de agrupamiento jerárquico o de partición (definiendo previamente la cantidad de clusters). Para concluir, se utilizan distintas métricas, que se explican en esta parte, para evaluar la efectividad de los resultados del agrupamiento.

2.2.11.1.Evaluación de la Tendencia del conjunto de datos.

Lawson y Jurs (1990) La Estadística de Hopkins fue propuesta en su forma básica por el botánico Hopkins en 1954. Se concibió como una medida simple e intuitivamente atractiva para evaluar la tendencia al agrupamiento en conjuntos de datos. La estadística de Hopkins mide la tendencia de un conjunto de datos a agruparse, basada en la diferencia entre la distancia desde un punto real a su vecino más cercano (U) y la

distancia desde un punto artificial aleatorio hasta el punto de datos real más cercano (W).

Desde puntos artificiales uniformemente distribuidos en el espacio de datos hasta el punto de datos real más cercano (U_i). La estadística de Hopkins (H) se calcula como $H = \sum U_i / \sum U_i + \sum W_i$), donde un valor cercano a 1 indica una fuerte tendencia al agrupamiento, mientras que un valor cercano a 0.5 sugiere una distribución aleatoria y valores cercanos a 0 indican una dispersión regular.

- $H_0 =$ El conjunto de datos no muestra una tendencia al agrupamiento.
- $H_1 =$El conjunto de datos muestra una tendencia al agrupamiento.

La aceptación o rechazo de la hipótesis nula se basa en el valor calculado de la estadística de Hopkins. Un valor significativamente mayor que 0.5 tiende a rechazar la hipótesis nula, apoyando la hipótesis alternativa.

2.2.11.2. Determinación del Número Óptimo de Clústeres

Según Kassambara (2017) Establecer la cantidad ideal de grupos en un conjunto de datos es un aspecto crucial en el agrupamiento por partición, como en el caso del método k-means, donde el analista debe definir previamente el número de clústeres k a crear. No existe una respuesta concreta y definitiva para determinar el número óptimo de clústeres. Esta decisión es en gran parte subjetiva y varía según el método seleccionado para calcular las similitudes y los criterios usados en el

proceso de división. Un método común y sencillo es examinar el dendrograma creado a través del clustering jerárquico, lo que puede indicar un número apropiado de clusters. Sin embargo, esta técnica también se basa en la subjetividad. Los enfoques para determinar el número óptimo de clústeres se dividen en métodos directos. Estos métodos se enfocan en optimizar un criterio específico. Por ejemplo, la suma de cuadrados dentro de los clústeres o el valor medio de la silueta, conocidos como los métodos de codo y de silueta, respectivamente.

2.2.11.2.1. Método del Codo

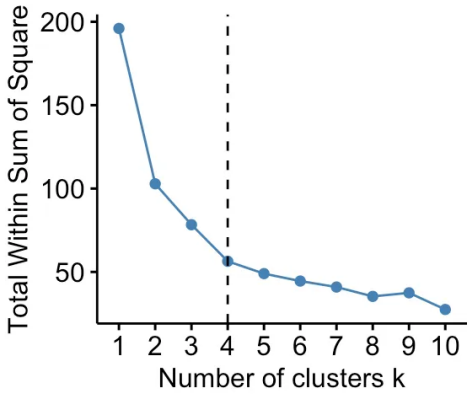
Según Kassambara (2017) La esencia de los métodos de partición, como en el agrupamiento k-means, reside en organizar los clústeres de tal manera que la variación total dentro de cada clúster, o la suma total de cuadrados dentro del clúster (WSS), sea lo más reducida posible. Esta WSS total es un indicador de la compacidad del agrupamiento, y el objetivo es minimizarla. El método del Codo considera la WSS total en relación con el número de clusters. La idea es seleccionar una cantidad de clusters donde la inclusión de un cluster adicional no resulte en una mejora significativa de la WSS total. La cantidad ideal de clústeres se puede determinar de la siguiente forma:

- Ejecutar el algoritmo de clustering (como el k-means) para una variedad de valores de k, por ejemplo, incrementando k desde 1 hasta 10 clústeres.
- Calcular para cada valor de k la suma total de cuadrados dentro de los clusters (wss).

- Trazar una gráfica de la (wss) en relación con el número de clusters k. Un punto donde se observe un cambio notable (como una "rodilla") en la gráfica se interpreta comúnmente como una señal del número óptimo de clústeres.

Figura 3

Gráfico del Método del Codo



Nota: Kassambara (2017)

En la figura 3 se muestra el grafico del método del codo para determinar el número óptimo de clústeres en un análisis de clustering. El eje Y representa la suma total de cuadrados dentro de los clústeres (Total Within sum of Square), y el eje X muestra el número de clústeres k, Según la figura 3, k=4.

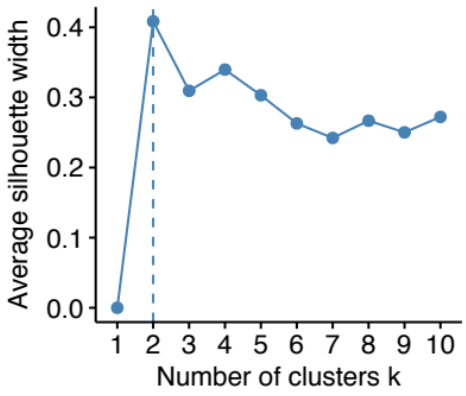
2.2.11.2.2. Método del promedio de siluetas

Este método calcula la silueta promedio de las observaciones para distintos valores de k. El número ideal de clusters k es aquel que resulta en el valor más alto de la silueta promedio dentro de un conjunto de posibles valores para k. Este algoritmo, al igual que el método del codo, puede implementarse de la siguiente forma:

- Aplicar un algoritmo de agrupamiento (como el k-means) probando una serie de valores de k, por ejemplo, desde 1 hasta 10 grupos.
- Calcular para cada k la silueta promedio de las observaciones.
- Dibujar la gráfica de la silueta promedio respecto al número de grupos k.
- El punto donde la silueta promedio alcanza su valor máximo indica el número óptimo de grupos.

Figura 4

Gráfico del método de la silueta promedio



Nota: Kassambara (2017)

La figura 4, el gráfico del ancho medio de silueta para determinar el número óptimo de clústeres en un análisis de clustering. El eje vertical Y representa el ancho medio de silueta (Average Silhouette Width), mientras que el eje horizontal X indica el número de clústeres k, de acuerdo con la Figura 4, el número óptimo de clústeres es k=2. Esta técnica fue introducida por Peter J. Rousseeuw en 1987.

2.2.11.3. Validación interna

2.2.11.3.1. Validación interna índice Davies-Bouldin:

Según Kassambara (2017) La idea detrás del índice Davies-Bouldin es evaluar qué tan compactos son los clusters internamente y qué tan separados están entre sí. Un valor menor del IDB indica una mejor calidad de agrupamiento donde con un valor de 0 siendo el ideal, lo que significaría que los clusters están perfectamente compactos y separados. La fórmula fue propuesta por David L. Davies y Donald W. Bouldin en 1979 y desde entonces se ha utilizado ampliamente en el análisis de clustering para evaluar la calidad de diferentes métodos de agrupamiento. La simplicidad y la efectividad del índice Davies-Bouldin lo han hecho popular en diversas aplicaciones de análisis de datos.

El algoritmo más eficiente es aquel que resulta en el valor más bajo, y se determina mediante el cálculo de:

$$IDB = \frac{1}{n}\sum_{i=1}^{n} max_{i\neq i}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_{j)}}\right)$$

Donde:

- n: representa el número de clústeres.
- σ_i: es la distancia promedio entre todos los elementos del clúster i y su centroide c_i.
- $d(c_i, c_j)$: es la distancia entre los centroides de los clústeres i y j.

En resumen, evalúa la separación entre clústeres. Un valor más bajo es mejor, indicando clústeres que están más separados y mejor definidos.

2.2.11.3.2. Validación interna Indice Dunn:

Según Kassambara (2017) El índice de Dunn es un método de validación interna utilizado para evaluar la calidad de un agrupamiento (o "clustering") en análisis de datos. Esta medida fue propuesta por J.C. Dunn en 1974. El índice de Dunn se enfoca en identificar conjuntos de datos que están compactos y bien separados. La teoría detrás del índice de Dunn se basa en dos principios fundamentales del agrupamiento eficaz:

- Compactación Interna: Los puntos dentro de un clúster deben estar lo más cerca posible unos de otros, lo que indica una alta densidad interna y, por tanto, una buena compactación. Esto se refleja en una baja dispersión o varianza dentro de cada clúster.
- Separación entre Clústeres: Los clústeres individuales deben estar lo más alejados posible entre sí. Una buena separación entre los clústeres sugiere que cada clúster es distintivo y bien definido, sin superposiciones significativas entre los clústeres.

El índice de Dunn se calcula como el cociente entre la distancia mínima entre los clústeres (que refleja la separación) y la máxima distancia intraclúster (que refleja la compactación). Matemáticamente, se formula como:

$$ID = \frac{Min_{1 \le i < j \le n} d(i,j)}{Max_{1 \le k \le n} d'(k)}$$

Donde:

- $d(i,j)$: representa la distancia entre los clústeres i y j (inter-clusters), la cual se mide como la distancia entre sus centroides.
- d'(k): es la distancia interna dentro del clúster k, que puede definirse como la distancia máxima entre pares de elementos dentro de ese clúster.

En decir se evalúa la compactación y separación de los clústeres. Un valor más alto es mejor, indicando clústeres bien definidos y separados.

2.2.11.3.3. Validación interna Indice conectividad:

Según Kassambara (2017)La conectividad en el contexto de análisis de clúster es una medida que evalúa qué tan bien se agrupan las observaciones en un conjunto de datos. Esta medida es particularmente relevante en el análisis de clúster porque ayuda a entender cómo están distribuidas las observaciones y si hay una tendencia natural a formar grupos o clústeres distintos.

$$CONN(C) = \sum_{i=1}^{N} \sum_{j=1}^{L} x_{i,nn_{i(j)}}$$

Donde:

- $CONN(C)$: Es la conectividad del conjunto de datos con respecto a una partición de clúster C.
- N: Es el número total de observaciones en el conjunto de datos.
- L: es el número de vecinos más cercanos considerados para cada observación.

- $x_{i,nn_{i(j)}}$: E s un valor que refleja la relación entre una observación i y su j-ésimo vecino más cercano. Este valor es 0 si i y su j-ésimo vecino más cercano están en el mismo clúster, y $1/j$ si están en clústeres diferentes.

La conectividad tiene un valor entre cero e infinito y debe ser minimizada donde Mide cuán cercanos están los elementos dentro de un clúster. Un valor más bajo es mejor. Indica que los elementos de un clúster están más cercanos entre sí.

2.2.11.3.4. Validación interna Indice Anchura de Silueta

Kassambara (2017) El Ancho de Silueta representa el promedio del coeficiente de Silueta para cada punto de datos. Este coeficiente evalúa qué tan adecuadamente se ha asignado una observación específica a su clúster, donde las observaciones correctamente agrupadas tienden a tener valores próximos a 1 y aquellas inadecuadamente agrupadas se acercan a -1. Para la observación i, este coeficiente se define de la siguiente manera:

$$S(i) = \frac{b_i - a_i}{max(b_i, a_i)}$$

Donde:

- a_i: Es la distancia promedio entre la observación i y todas las demás observaciones en el mismo clúster.
- b_i: Es la distancia promedio entre la observación i y las observaciones en el "clúster vecino más cercano".

Además, Mide cuán similar es un objeto a su propio clúster en comparación con otros clústeres. Los valores más altos indican una mejor asignación de clústeres

2.2.12. Pre procesamiento de datos

Consiste en una serie de pasos para preparar los datos para su análisis posterior. A continuación, se detallan las partes típicas de este proceso.

2.2.12.1. Limpieza de Datos

Tras identificar valores faltantes, errores y valores verdaderos (extremos o normales), los analistas deben decidir qué hacer con las observaciones problemáticas:

- **Mantenerlos Sin Cambios**: La opción más conservadora es aceptar los datos como una respuesta válida y no hacer cambios. En muestras grandes, una respuesta dudosa tiene menos efecto en el análisis; en muestras pequeñas, la decisión es más complicada.
- **Corregir los Datos:** Si se puede determinar la intención original del encuestado, corregir la respuesta (por ejemplo, tras consultar con el encuestador, queda claro que el encuestado se refería a la falta de ingresos en lugar de un exceso).
- **Eliminar los Datos:** Si los datos parecen ilógicos y el valor se desvía tanto de la norma que afectaría las estadísticas descriptivas o inferenciales. ¿Qué hacer? ¿Eliminar solo esa respuesta o todo el registro? Hay que recordar que, al eliminar datos, existe el riesgo de seleccionar datos de manera consciente o inconsciente para obtener resultados preferidos. Para entender el impacto de

eliminar un dato, se puede crear una variable binaria (1=registro sospechoso, 0=no sospechoso) y utilizarla como filtro en tablas dinámicas o filtrado en la tabla para comprender el impacto de datos potencialmente erróneos en los resultados finales.

- **Volver a Medir:** Si el tiempo y los recursos lo permiten, volver a medir los valores sospechosos o erróneos.

2.2.12.2. Transformación de Datos

El "Label Encoding" es una técnica común en el preprocesamiento de datos para convertir categorías nominales en valores numéricos. Si bien es efectiva para transformar datos categóricos en un formato que los modelos de análisis pueden procesar.

- **Asignación de Números a Categorías:** Cada categoría única dentro de una variable categórica se asigna a un número entero único. Esta asignación es generalmente arbitraria y no indica ninguna jerarquía o peso inherente entre las categorías.
- **Mantener la Singularidad:** Es crucial que la asignación de números mantenga la singularidad de las categorías originales. Cada número representa exclusivamente una categoría.

2.2.13. Análisis de correlación

Newbold et al. (2008)El coeficiente de correlación, que es la covarianza entre dos variables aleatorias ajustada por el producto de sus desviaciones estándar, proporciona una medida escalada de su asociación, limitada entre -1 y 1. A continuación se detalla la interpretación de esta medida:

- Un coeficiente de correlación de 0 señala la ausencia de cualquier asociación lineal entre dos variables. La independencia de las variables implica una correlación de 0, aunque la inversa no siempre es cierta.
- Una correlación positiva implica que altos (o bajos) valores en una variable tienden a coincidir con altos (o bajos) valores en la otra, lo que indica una dependencia lineal positiva. Una correlación de 1 denota una relación lineal positiva perfecta, donde el valor preciso de una variable puede predecir perfectamente el valor de la otra.
- Por el contrario, una correlación negativa significa que un alto valor en una variable probablemente corresponda a un bajo valor en la otra, reflejando una dependencia lineal negativa. Una correlación de -1 implica una relación lineal negativa perfecta, donde una variable puede predecir el valor opuesto de la otra con exactitud.

El coeficiente de correlación es más revelador que la covarianza ya que proporciona una indicación clara de la fuerza y la dirección de una relación lineal. Valores de correlación cercanos a 1 o -1 indican una relación fuerte, mientras que los valores cercanos a 0 muestran una relación más débil. A pesar de que el término "correlación" se utiliza comúnmente para sugerir cualquier tipo de asociación, es crucial reconocer que las variables con relaciones no lineales no se verán reflejadas con coeficientes de correlación próximos a 1 o -1. Este conocimiento es crucial para evitar confusiones entre correlaciones lineales y asociaciones no lineales entre variables.

2.3. MARCO CONCEPTUAL

2.3.1. Hogar

Según el Instituto Nacional de Estadística e Informática (INEI) Se refiere al grupo de personas que viven juntas en un espacio compartido y generalmente gestionan su economía de manera conjunta. Las características de un hogar incluyen:

2.3.1.1. Composición

Según el Instituto Nacional de Estadística e Informática (INEI) Puede estar formado por una persona viviendo sola (hogar unipersonal) o por varias personas, que pueden ser una familia (padres e hijos, parejas, etc.) o un grupo de individuos no relacionados que eligen vivir juntos.

2.3.1.2. Funcionamiento Económico y Doméstico Conjunto

Según el Instituto Nacional de Estadística e Informática (INEI) Los miembros del hogar suelen compartir responsabilidades y gastos, como los costos de alimentos, servicios y otros gastos domésticos. Según el Instituto Nacional de Estadística e Informática (INEI) Los miembros del hogar toman decisiones colectivas sobre aspectos de la vida diaria, crianza de los hijos, administración de recursos, entre otros.

2.3.2. SISFOH

El SISFOH (Sistema de Focalización de Hogares) es un sistema utilizado en países como Perú para identificar hogares en situación de vulnerabilidad. Recopila datos socioeconómicos para clasificar a los hogares y dirigir eficazmente

los programas sociales. Este sistema asegura que la asistencia gubernamental llegue a quienes más la necesitan. Es esencial para mejorar la precisión y eficacia de las políticas de asistencia social.

2.3.3. Vivienda

Según el Sistema de Focalización de Hogares (SISFOH) en Perú, una vivienda se refiere al espacio o estructura física donde reside un hogar o grupo de personas. Generalmente, el concepto de vivienda incluye aspectos como la estructura física (casa, departamento, habitación, etc.), las condiciones de habitabilidad (acceso a servicios básicos como agua, electricidad, saneamiento), y la ubicación geográfica.

2.3.4. Tipo de seguro

Según el Sistema de Focalización de Hogares (SISFOH) o instituciones similares suele referirse a la clasificación de los seguros de salud a los que tienen acceso los individuos o familias, especialmente en el contexto de los programas sociales. En el caso de Perú, los tipos de seguros de salud se clasifican principalmente en:

2.3.4.1. Seguro Integral de Salud (SIS)

Es un seguro de salud público destinado a personas que no tienen acceso a otro tipo de seguro y que están en situación de pobreza o pobreza extrema, según la clasificación del SISFOH. Este seguro cubre una amplia gama de servicios, incluyendo atención primaria, emergencias, hospitalizaciones, cirugías y medicamentos.

2.3.4.2. EsSalud

Es un seguro de salud para trabajadores formales y sus familias, financiado a través de contribuciones laborales. Ofrece una cobertura más amplia que el SIS y se accede a él a través del empleo formal.

2.3.4.3. Seguros Privados

Son ofrecidos por compañías de seguros privadas y suelen estar asociados a empleados de empresas privadas como un beneficio laboral o pueden ser contratados de manera individual. Ofrecen diferentes niveles de cobertura y acceso a una red de clínicas y hospitales privados.

2.3.5. Nivel educativo

Según el Sistema de Focalización de Hogares (SISFOH) en Perú, el nivel educativo se categoriza de la siguiente manera: Educación primaria, educación secundaria, educación técnica y educación Superior.

2.3.6. Tipo de combustible de cocina

Según la Oficina de Focalización de Hogares (SISFOH) de Perú se refiere a los materiales o recursos utilizados para generar energía térmica necesaria en la preparación de alimentos. Estos combustibles pueden incluir gas, leña, carbón, petróleo, electricidad, entre otros.

2.3.7. Indicadores socioeconómicos

Los indicadores socioeconómicos son herramientas de análisis estadístico que miden diversas facetas de la estructura económica y social de una población. Estos indicadores son fundamentales en estudios de economía, sociología, política

pública y desarrollo, y proporcionan una comprensión cuantitativa de las condiciones y tendencias de vida de individuos, hogares y comunidades. Estas medidas permiten evaluar aspectos como el nivel de ingresos, la calidad y accesibilidad de la educación, la naturaleza y seguridad del empleo, las condiciones de las viviendas, el acceso a servicios de salud, y la calidad general de vida. También incluyen índices de pobreza, desigualdad y otros factores socioeconómicos que reflejan la distribución de recursos y oportunidades en una sociedad. Los indicadores socioeconómicos son vitales para la formulación y evaluación de políticas, la planificación de programas de desarrollo y la investigación académica, proporcionando datos esenciales para comprender y abordar cuestiones relacionadas con la desigualdad social y económica.

CAPÍTULO III

MATERIALES Y MÉTODOS

Esta investigación se desarrolló empleando los materiales y métodos descritos a continuación.

3.1. MATERIALES

- Se empleó una computadora personal con un procesador AMD Ryzen 5 5600G a 3.90 GHz y 16 GB de RAM.
- Se empleó una impresora multifuncional Epson L5290.
- Se utilizo internet.
- Se utilizó el software R-estudio.
- Se utilizó una base de datos proveniente de la oficina del SISFOH de la Municipalidad de Macusani,

3.2. METODOLOGÍA

3.2.1. Tipo de investigación

El tipo de investigación es carácter descriptivo, se estableció la segmentación de hogares del distrito de Macusani a través de la descripción de sus variables socioeconómicas.

3.2.2. Diseño de investigación

El diseño de la investigación es no experimental. Esta elección metodológica se fundamenta en la naturaleza de los datos disponibles y en los objetivos específicos del estudio. Gracias a la colaboración de la oficina de

SISFOH (Sistema de Focalización de Hogares) de la Municipalidad Provincial de Carabaya – Macusani, se obtuvo acceso a una valiosa base de datos de hogares del distrito.

El carácter no experimental del estudio indica que no se realizaron manipulaciones deliberadas de variables ni se establecieron grupos de control; en lugar de ello, se analizaron las condiciones existentes de los hogares tal como se presentaron en la base de datos. Este enfoque permite un análisis detallado de las variables socioeconómicas en un momento específico, sin intervenir en el entorno natural de los hogares.

3.2.3. Población y muestra

3.2.3.1. Población

La población del estudio, que abarca a todos los 4,329 hogares del distrito de Macusani en 2020, incluye tanto zonas urbanas como rurales.

3.2.3.2. Muestra

La muestra seleccionada para un análisis más detallado se enfoca en 344 hogares del sector rural, elegidos para entender mejor las características. Y al centrarse en estos hogares rurales, se accede a datos más auténticos y representativos.

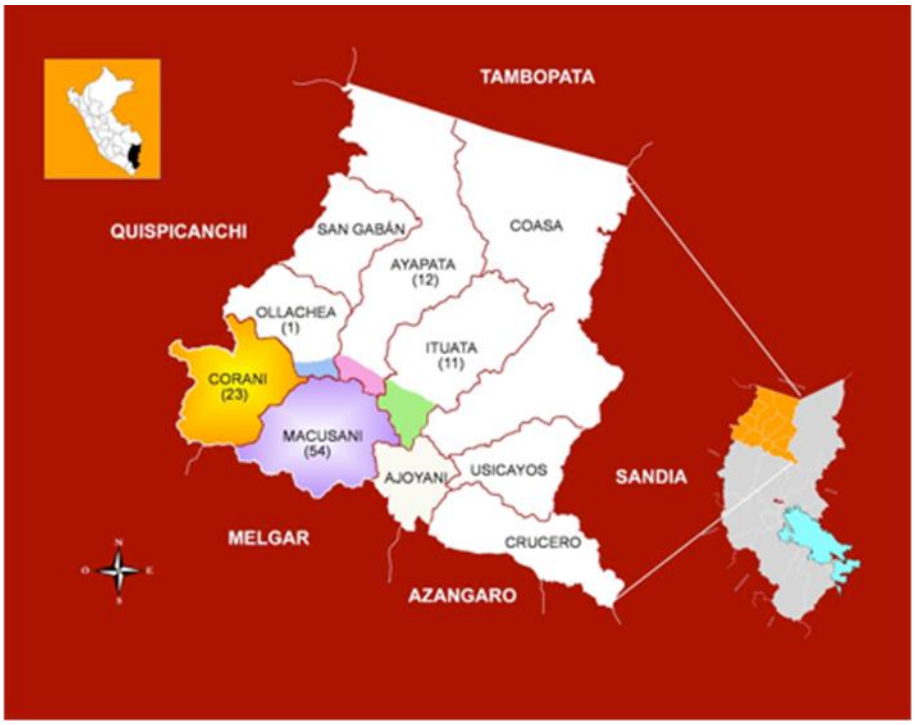
3.2.4. Lugar de investigación

La investigación se llevó a cabo utilizando la base de datos de la Oficina de SISFOH del distrito de Macusani durante el año 2020. Este distrito, ubicado en el sureste del Perú, en la provincia de Carabaya, sufrió un impacto

socioeconómico significativo como resultado de la pandemia de COVID-19. Sin embargo, la Municipalidad del distrito de Macusani no cuenta con una base de datos de hogares segmentada que facilite la entrega de ciertos beneficios a la población afectada. El distrito de Macusani es uno de los diez distritos que componen la provincia de Carabaya, ubicada en el departamento de Puno. Limita al norte con la provincia de Tambopata (Madre de Dios), al sur con las provincias de Melgar, Azángaro y San Antonio de Putina.

Figura 5

Mapa de la provincia de Carabaya – Macusani



Nota: obtenido del enlace: https://n9.cl/snqyl

CAPÍTULO IV

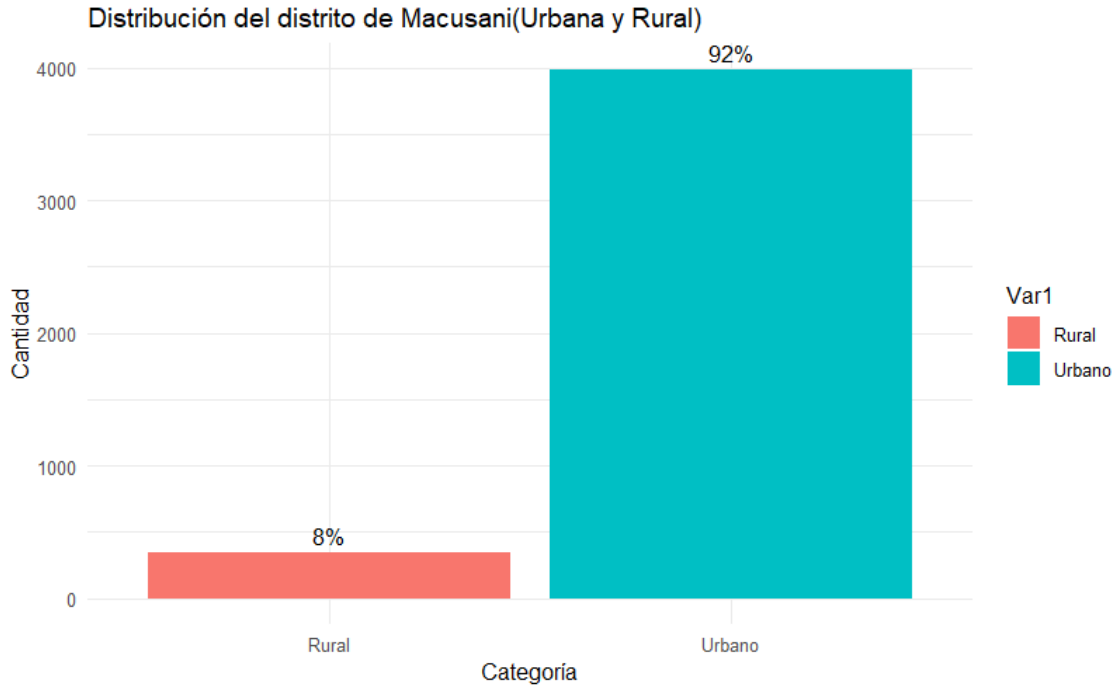
RESULTADOS Y DISCUSIÓN

4.1. RESULTADOS

4.1.1. Pre procesamiento de datos

Figura 6

Distribución del distrito de Macusani (Urbano y Rural)



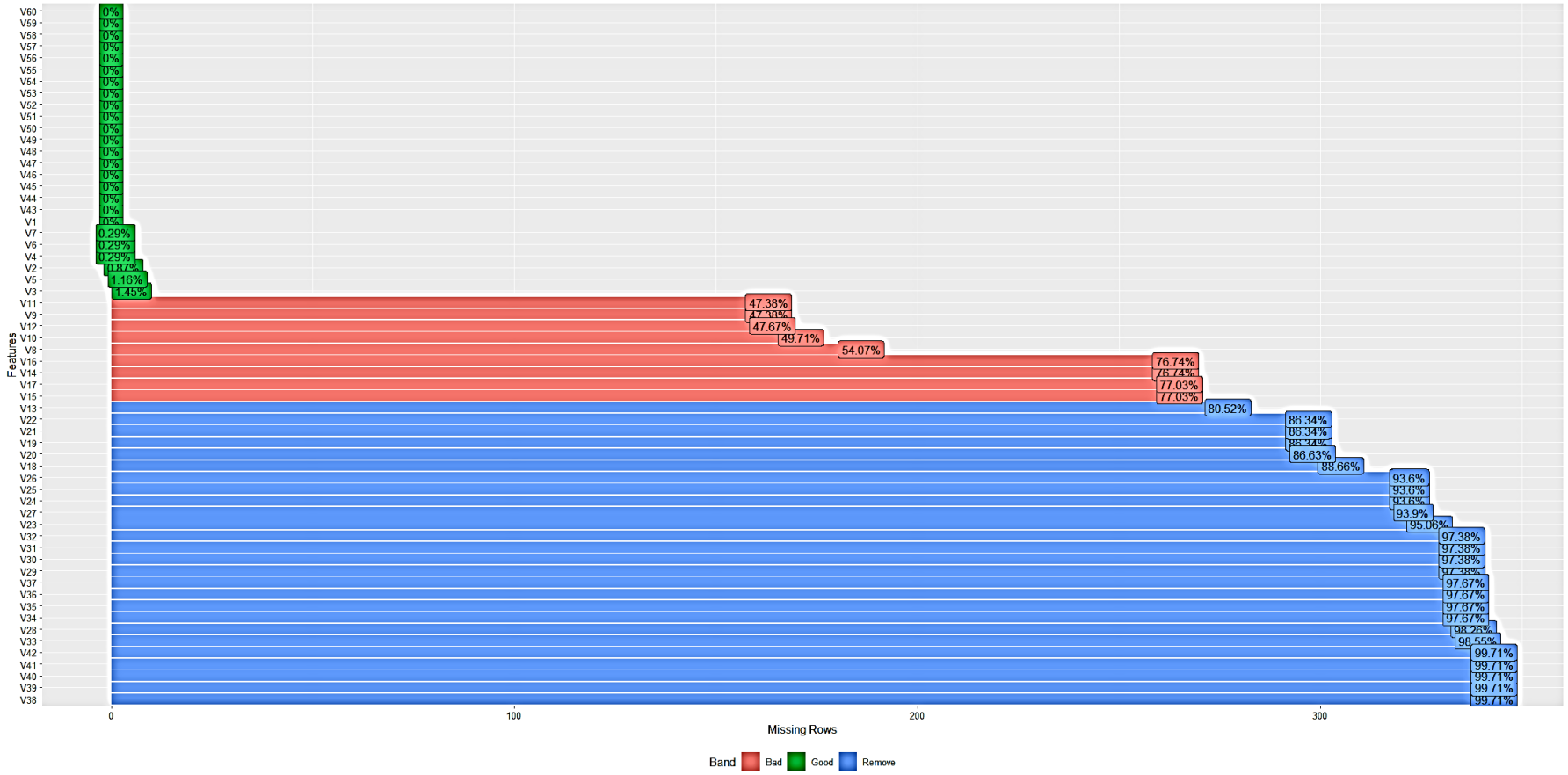
Nota: Resultado obtenido de la base de datos de SISFOH

En la Figura 6 presenta un diagrama de barras que muestra la distribución poblacional del distrito de Macusani, con un 92% de habitantes en la zona urbana y un 8% en la rural. La selección de 344 hogares rurales como muestra para el estudio facilita una comprensión más profunda y auténtica.

4.1.2. Limpieza de Datos:

Figura 7

Porcentaje de datos faltantes por variables



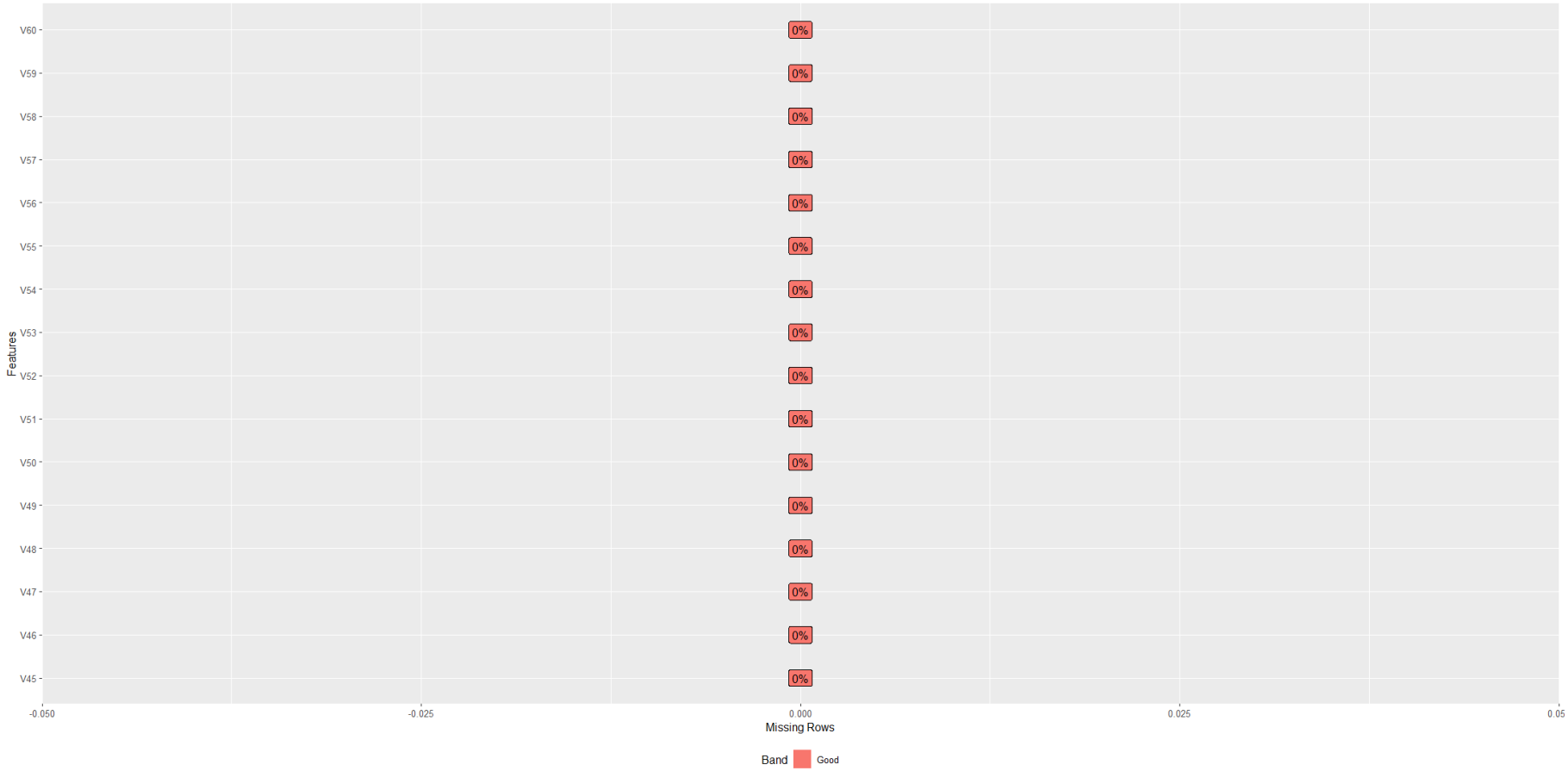
Nota: Resultado obtenido de la base de datos de SISFOH

La figura 7 se muestra que en el conjunto de datos hay diferentes cantidades de información faltante. En las variables del V8 a V17, los datos faltantes son más del 40%. Para las variables de V18 a V42, los datos que faltan son aún más, pasando del 80%, Pero en las variables de V1 a V7 y de V43 a V60 son menores al 2%.

Se decidió eliminar del conjunto de datos las variables con más porcentaje de datos faltantes. De esta forma, el análisis se centra únicamente en aquellas variables que presentan un 0% de ausencia de datos, asegurando así la integridad y la completitud de la información utilizada. También se han eliminado las variables V1, V43 y V44 del conjunto de datos porque contienen información no relevante para el análisis, como nombres o identificadores únicos, que no aportan al análisis. En la figura 8 se mostrará el conjunto de datos resultante, reflejando solo las variables con datos completos.

Figura 8

Porcentaje de datos faltantes por variable



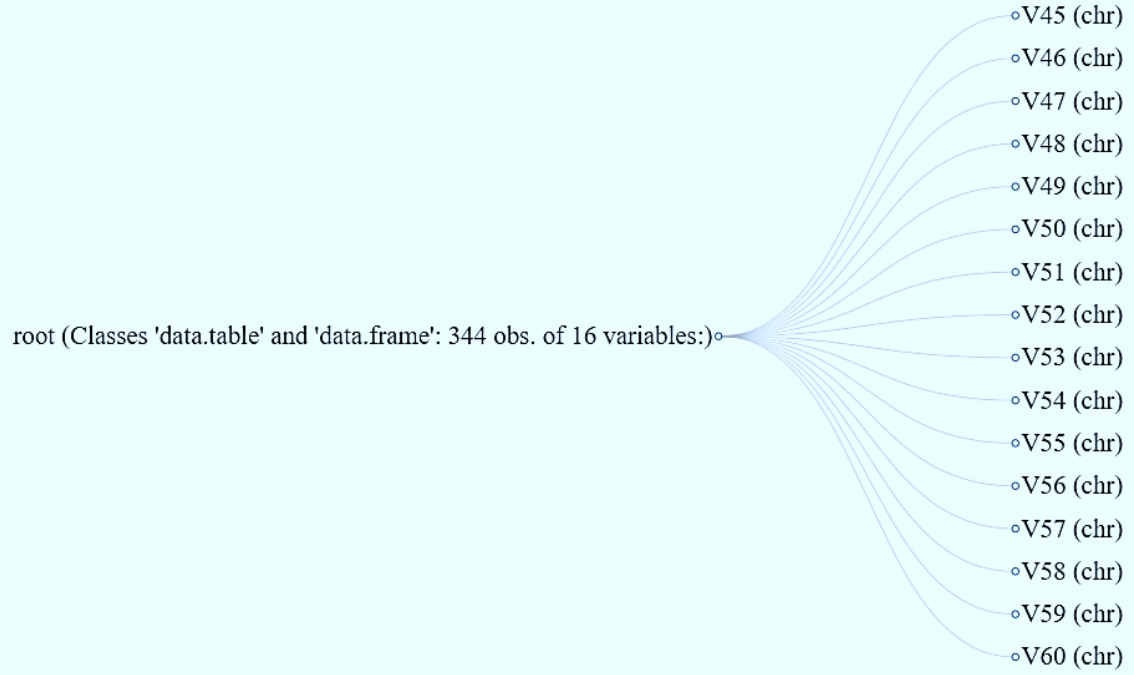
Nota: *Resultado obtenido de la base de datos de SISFOH*

La figura 8 muestra que todas las variables tienen 0%, que confirma que el conjunto de datos está completamente libre de datos faltantes.

4.1.3. Transformación de Datos:

Figura 9

Estructura de los conjuntos de datos

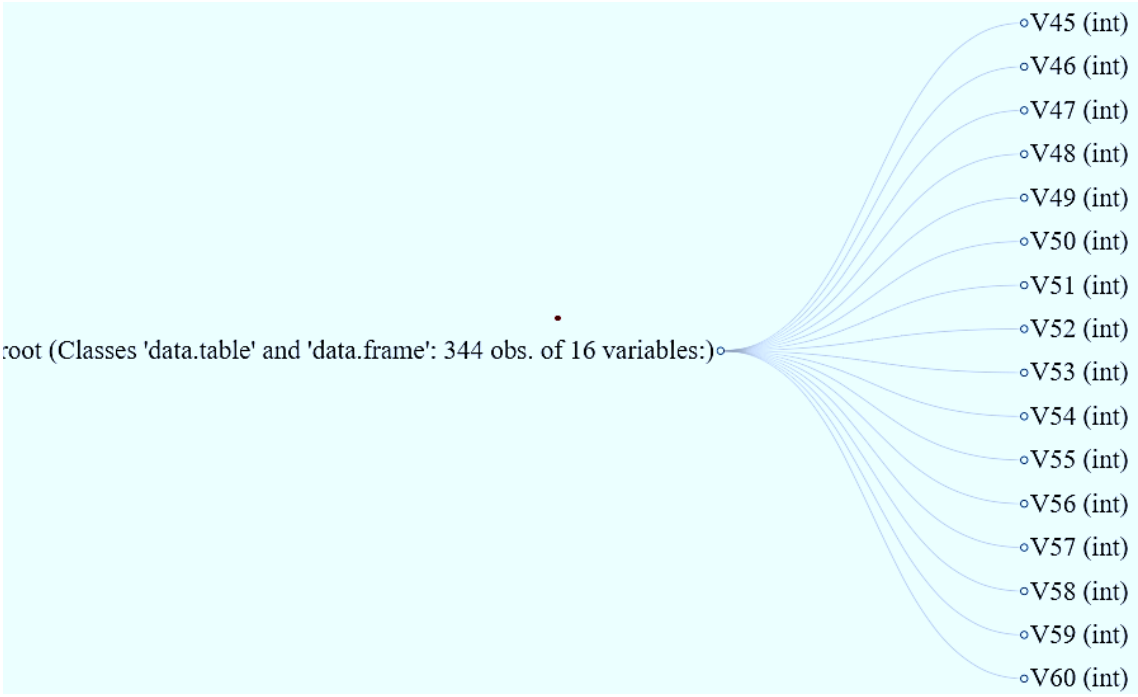


Nota: Resultado obtenido de la base de datos de SISFOH

La figura 9 muestra las dimensiones del conjunto de datos de 344 hogares y 16 variables y que todas las variables se presentan como datos cualitativos (chr). Para aplicar algoritmos de clúster como K-means, PAM, es necesario que estos datos sean cuantitativos(int), ya que tales métodos requieren variables numéricas para calcular distancias y realizar clústeres efectivos.

Figura 10

Estructura de los conjuntos de datos



Nota: *Resultado obtenido de la base de datos de SISFOH*

Después de utilizar el método Label Encoding la figura 10 muestra también las mismas dimensiones de 344 hogares y 16 variables del conjunto de datos con variables que se presentan como datos cuantitativos(int).

En el Anexo 2, se detalla la información referente a las variables y sus respectivas categorías, presentando como ejemplo la variable V45, que corresponde al "Tipo de seguro". Esta variable se clasifica en varias categorías, cada una representada por un número específico: el número 2 corresponde a "ESSALUD", el 4 a "NO TIENE", y el 3 a "SIS".

4.1.4. Evaluación de la Tendencia del conjunto de datos

Con un valor de 0.7246674, el resultado está más cerca de 1 que de 0.5 o de 0. Esto implica lo siguiente:

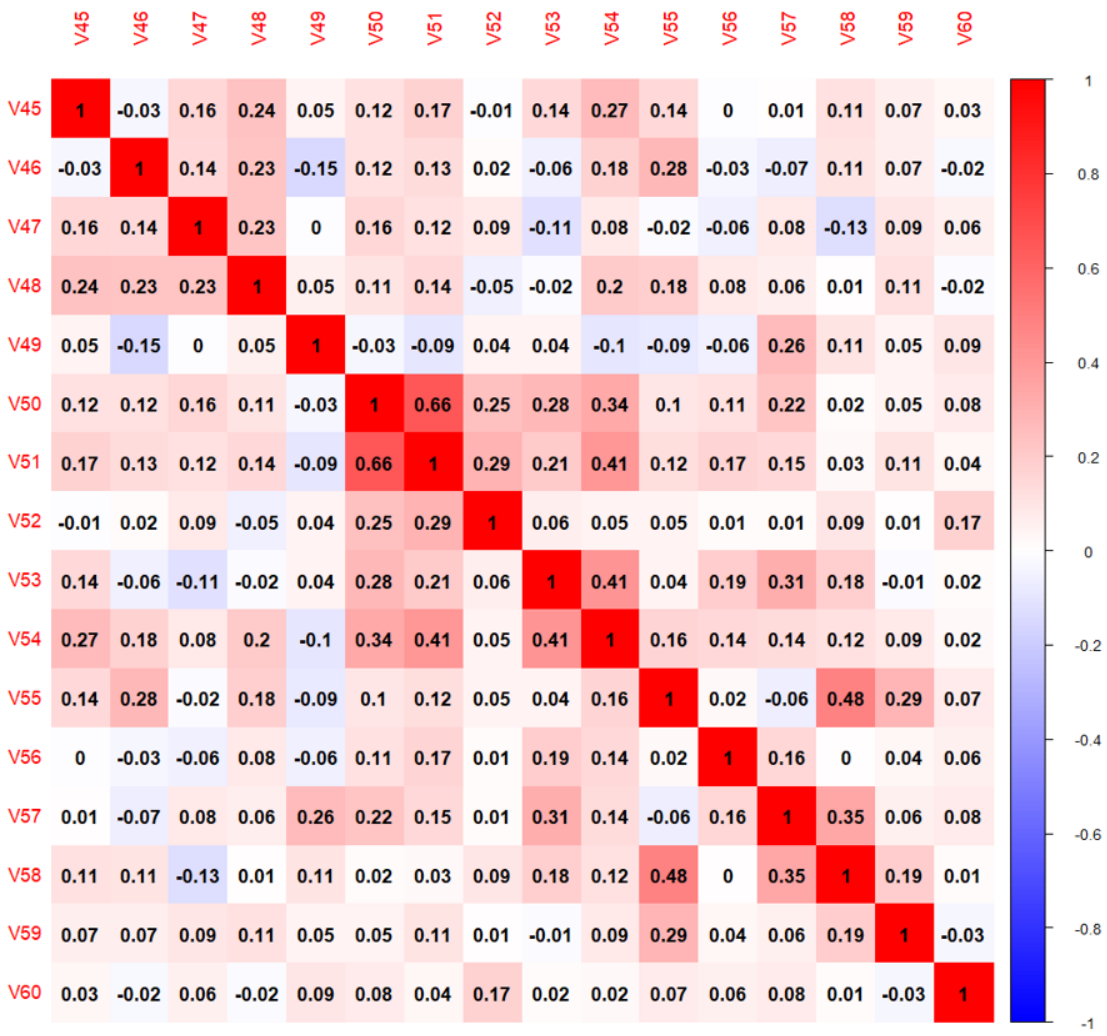
- $H_0 =$ El conjunto de datos no muestra una tendencia al agrupamiento.
- $H_1 =$El conjunto de datos muestra una tendencia al agrupamiento.

Según el resultado, es razonable rechazar la hipótesis nula H_0 que afirma que el conjunto de datos no muestra una tendencia al agrupamiento. el valor es de 0.7246674 es significativamente mayor que 0.5, lo que apoya la hipótesis alternativa H_1 de que el conjunto de datos muestra una tendencia al agrupamiento.

4.1.5. Análisis de correlación

Figura 11

Análisis de correlación



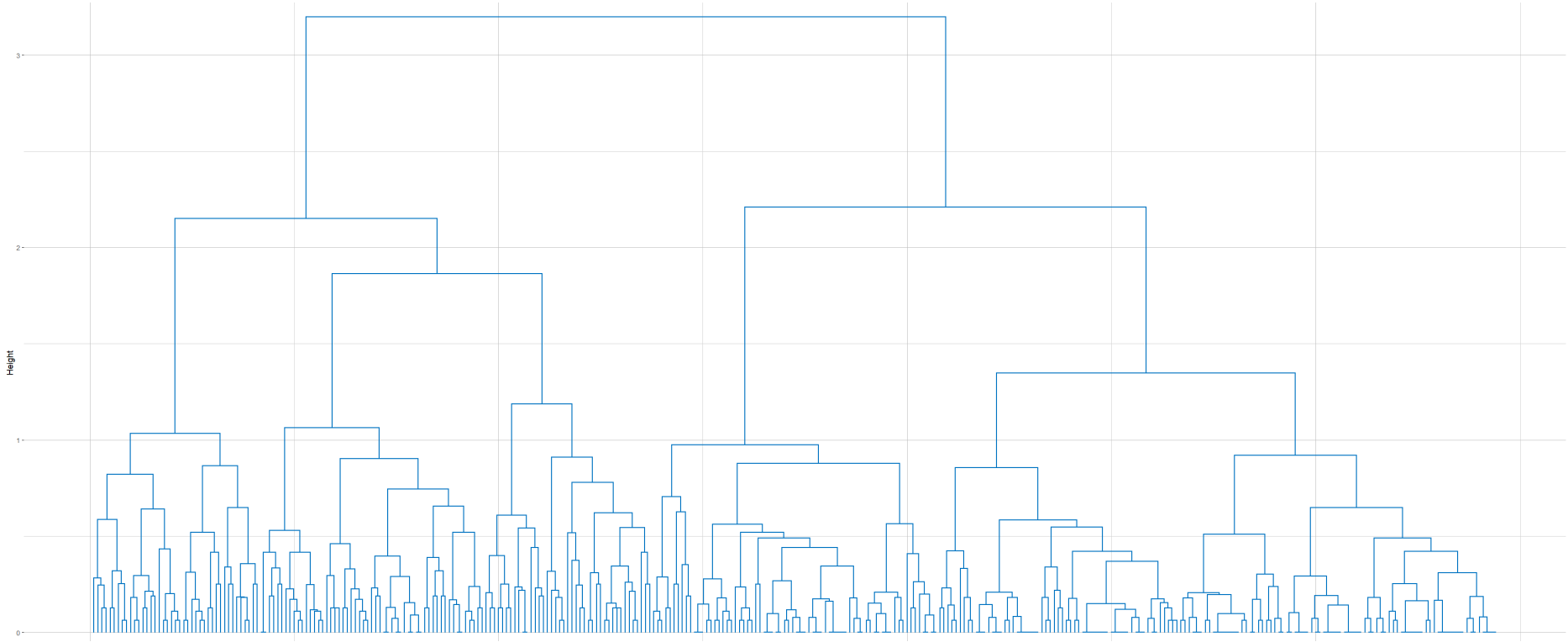
NOTA: Resultado obtenido de la base de datos de SISFOH

La figura 11 presenta una matriz de correlación para variables de "v45" a "v60", donde -1 y 1 representan correlaciones negativas y positivas perfectas. Las variables V50 y V51 destacan por su alta correlación positiva, pero serán retenidas para el análisis clúster.

4.1.6. Primer objetivo especifico

Figura 12

Determinar número de clúster mediante visualización de dendograma.

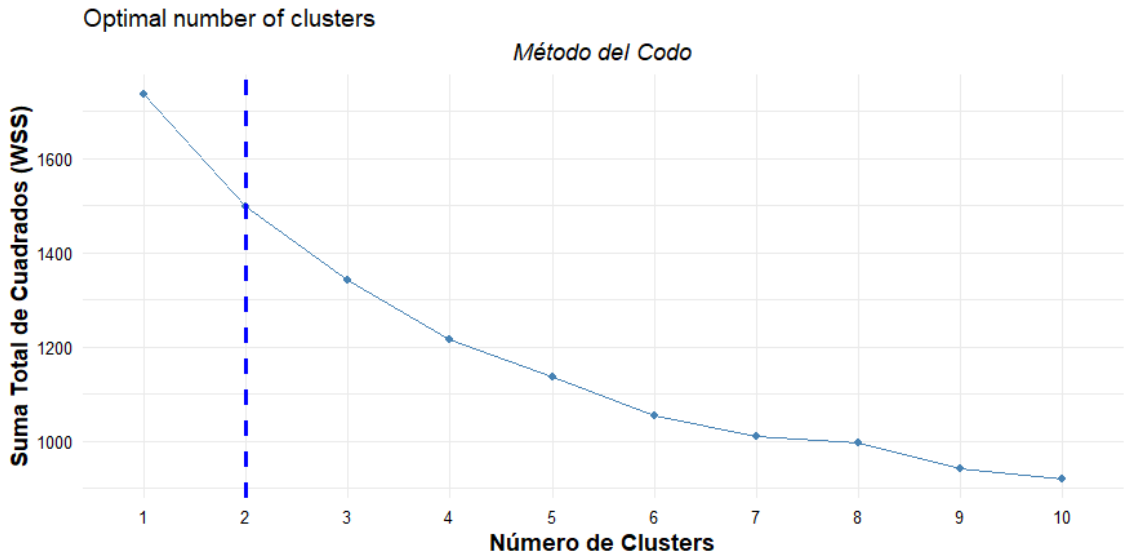


Nota: Resultado obtenido de la base de datos de SISFOH

En la figura 12, al observar las alturas de fusión en el dendrograma, se distinguen claramente dos clústeres principales, lo que indica dos grupos distintos en los datos. Estos clústeres se forman a una altura considerablemente mayor que las fusiones a nivel individual, sugiriendo una diferenciación significativa entre ellos. en el eje Y refleja la distancia o disimilitud entre los clústeres formados.

Figura 13

Determinar del número óptimo de clúster con K-means con método del codo

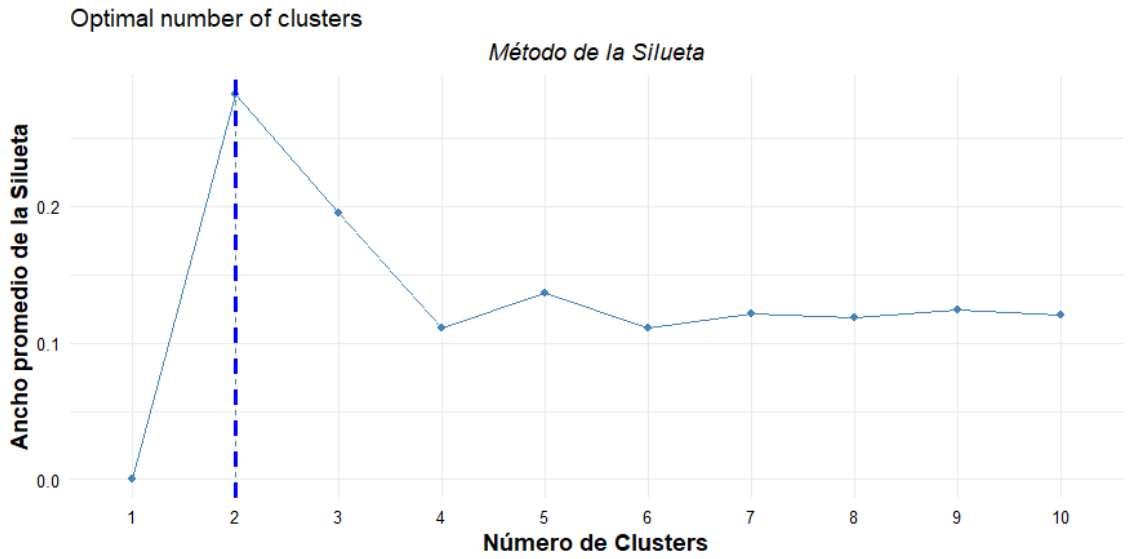


Nota: *Resultado obtenidos con métodos del codo*

En la figura 13, el Método del Codo muestra un punto de inflexión pronunciado en k=2, lo que indica que es el número óptimo de clústeres.

Figura 14

Determinar del número óptimo de clúster con K-means con método de la silueta.



Nota: Resultado obtenidos con métodos de la silueta

En la figura 14 El Método de la Silueta alcanza su valor más alto en k=2, corroborando que dos clústeres maximizan la cohesión interna y la separación entre clústeres. Estos métodos coinciden en que dos clústeres es la elección óptima para la agrupación de los datos presentados.

4.1.7. Segundo objetivo especifico

4.1.7.1. Clustering K-means

Tabla 1

Distribución del algoritmo K-means con K=2

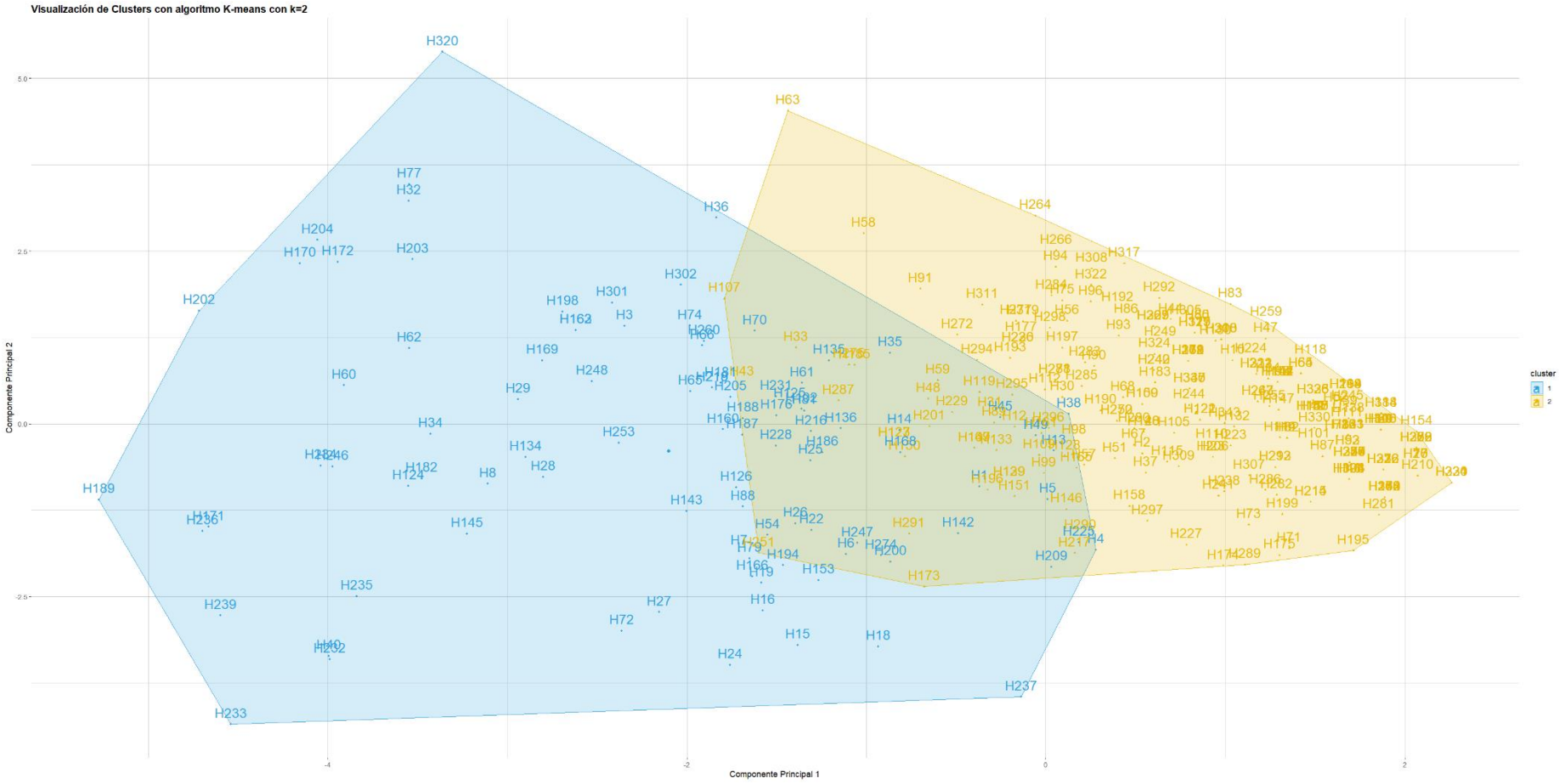
Clúster	Hogares	Porcentaje
1	246	72%
2	98	28%
Total	344	100%

Nota: Resultado del algoritmo K-means

La tabla 1 indica que, de un total de 344 hogares analizados, la mayoría 72% pertenece al clúster 1 con 246 hogares, mientras que el clúster 2 comprende 98 hogares, representando el 28% restante.

Figura 15

Visualización de clústeres con algoritmo K-means con k=2



Nota: *Resultado del algoritmo K-means*

En la figura 15, la visualización del algoritmo K-means con k=2 muestra dos clústeres. El clúster 1, con 246 hogares, está representado por colores azules, y el clúster 2, con 98 hogares, por color amarillos.

4.1.7.2. Clustering K-medoids (PAM):

Tabla 2

Distribución del algoritmo K-medoids (PAM) con K=2

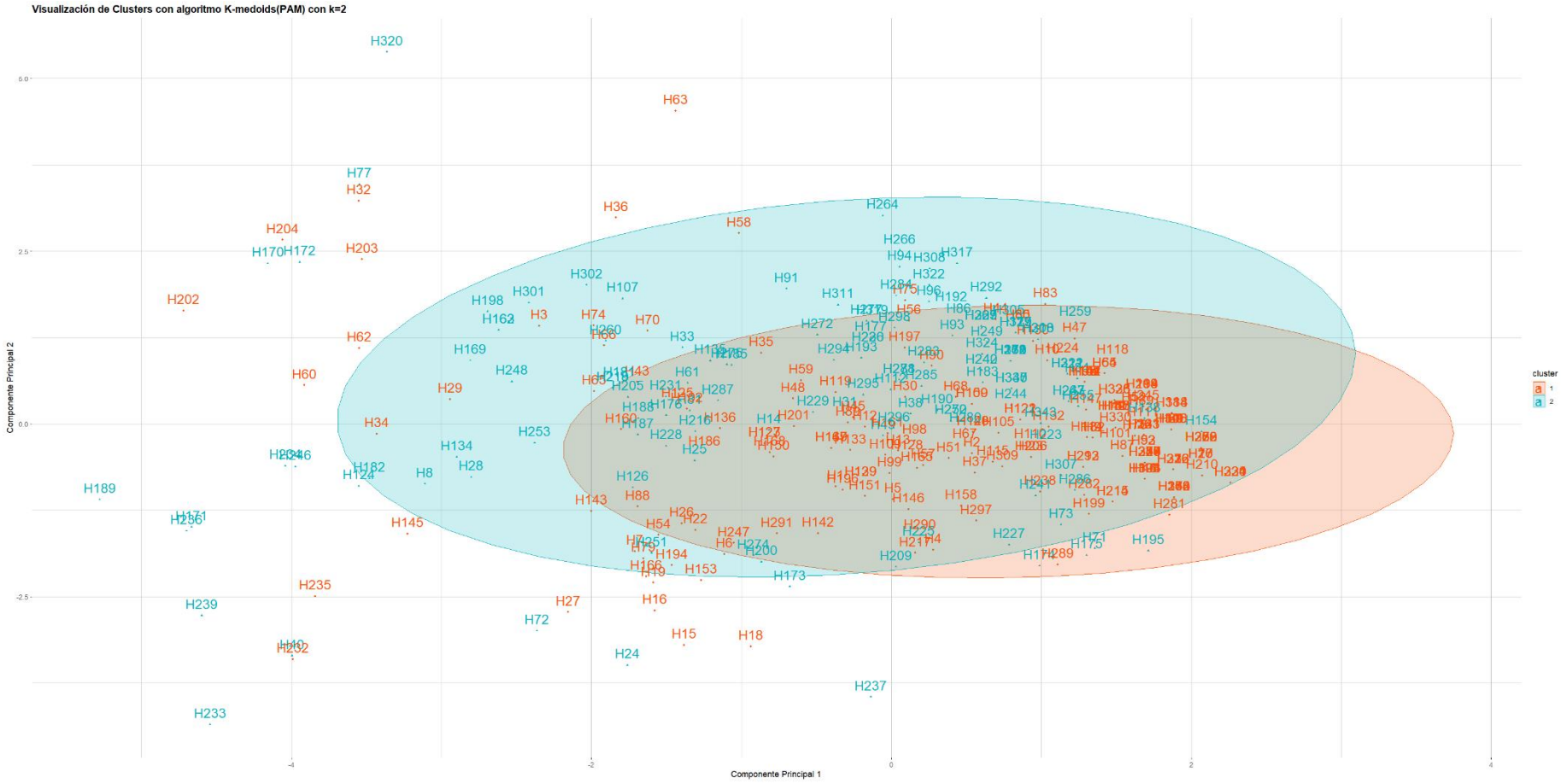
Clúster	Hogares	Porcentaje
1	205	60%
2	139	40%
Total	344	100%

Nota: Resultado del algoritmo K-medoids (PAM)

La tabla 2 indica que, de un total de 344 hogares analizados, la mayoría 60% pertenece al clúster 1 con 205 hogares, mientras que el clúster 2 comprende 139 hogares, representando el 40% restante.

Figura 16

Visualización de Clústeres con algoritmo K-medoids(PAM) con k=2



Nota: *Resultado del algoritmo K-medoids(PAM)*

En la figura 16, la visualización del algoritmo del K-medoids(PAM) con k=2 muestra dos clústeres. El clúster 1, con 205 hogares, está representado por colores anaranjados, y el clúster 2, con 139 hogares, por color verde.

4.1.7.3. Clustering jerárquico

Tabla 3

Distribución del algoritmo jerárquico con K=2

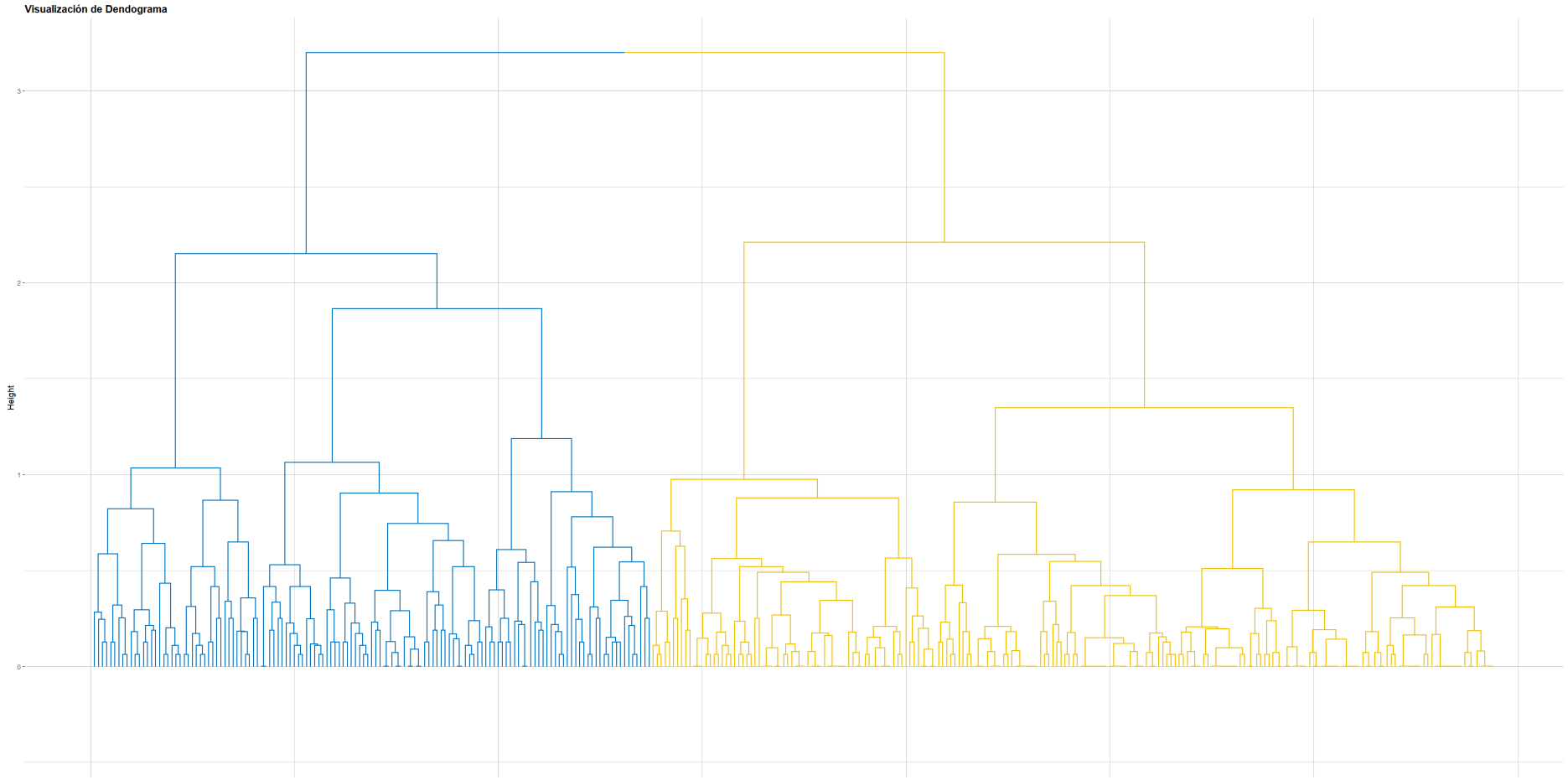
Clúster	Hogares	Porcentaje
1	137	40%
2	207	60%
Total	344	100%

Nota: Resultado del algoritmo jerárquico

La tabla 03 indica que, de un total de 344 hogares analizados, la mayoría con 60% pertenece al clúster 2 con 207 hogares, mientras que el clúster 1 comprende 137 hogares, representando el 40% restante.

Figura 17

Visualización de Clústeres con algoritmo jerárquico con k=2



Nota: *Resultado del algoritmo jerárquico*

En la figura 17, la visualización del dendograma con el algoritmo del jerarquico con k=2 muestra dos clústeres. El clúster 1, con 137 hogares, está representado por líneas azules, y el clúster 2, con 207 hogares, por líneas amarillas.

4.1.8. Encontrar del algoritmo más efectivo

Tabla 4

Puntaje de índices de evaluación

Método	Algoritmo	clúster	Puntaje
Connectivity	Jerárquico	2	3.8579
Connectivity	kmeans	2	49.5262
Connectivity	pam	2	154.302
Dunn	Jerárquico	2	0.4041
Dunn	kmeans	2	0.198
Dunn	pam	2	0.1443
Silhouette	Jerárquico	2	0.3727
Silhouette	kmeans	2	0.2731
Silhouette	pam	2	0.1112
DB	Jerárquico	2	2.781938
DB	kmeans	2	2.451226
DB	pam	2	3.4387481

Nota: Resultados de R estudio.

Basándonos en la tabla 4 con el método Connectivity el método Jerárquico tiene el valor más bajo (3.8579), lo que es bueno, Dunn Index el método Jerárquico también tiene el valor más alto (0.4041), lo que es bueno, Silhouette Score en aquí el método Jerárquico nuevamente supera a los demás con un valor más alto (0.3727) y Davies-Bouldin Index el método kmeans tiene el valor más bajo (2.451226048), lo cual es preferible. Considerando todas estas métricas, el método Jerárquico parece ser el mejor para tu caso. Tiene los mejores puntajes en Connectivity, Dunn y Silhouette, lo que indica que crea clústeres bien definidos, compactos y separados.

4.1.9. Tercer objetivo especifico

Según el anexo 2 el clúster 1 se caracteriza de la siguiente forma en cuanto al seguro de salud (V45), la mayoría de los hogares (56.93%) están inscritos en el SIS, seguido de un 26.28% que no tienen seguro y un 16.79% que cuentan con ESSALUD. Respecto al nivel educativo (V46), casi la mitad de los hogares (49.64%) tienen educación primaria como máximo nivel alcanzado, un 23.36% han completado la secundaria y un 11.68% tienen educación superior o técnica. Luego la mayoría de los hogares (65.69%) tienen un empleo contratado o son independientes (V47), mientras que un 23.36% se dedica al trabajo informal o doméstico. Solo un 5.84% está desempleado. En relación a los ingresos (V48), una abrumadora mayoría (88.32%) gana menos de 900 soles. La mayoría de los hogares (63.50%) no reciben ningún tipo de programa social (V49). En términos de materiales de construcción (V50), la mayoría de las viviendas (87.59%) están hechas de adobe o quincha. Para los techos (V51), la mayoría usa calamina o fibra (67.15%), y un 32.85% utiliza paja. El piso predominante es tierra (V52) en un 87.59% de los hogares. Sobre el acceso al agua (V53), un 66.42% obtiene agua de ríos o manantiales. La mayoría de los hogares (56.93%) utilizan río o acequia para el desagüe (V54). En cuanto a la posesión de teléfonos (V55), hay una distribución pareja entre aquellos que tienen celular sin internet y los que no tienen celular, ambos con un 48.91%. La mayoría de los hogares (84.67%) cuentan con solo una habitación (V56). La bosta es el principal combustible utilizado por un 63.50% (V57). Casi la mitad de los hogares (52.55%) tienen entre 1 y 5 electrodomesticos (V58). La mayoría no posee ningún tipo de vehículo (V59), con un 76.64% de los hogares. Finalmente, casi todos los hogares (98.54%) son de un piso o cabaña (V60).

Según el anexo 2 el clúster 2 se caracteriza de la siguiente forma para el seguro de salud (V45), un 58.45% están inscritos en el SIS, mientras que un 38.65% no tiene seguro y solo un 2.90% cuenta con ESSALUD. Respecto a la educación (V46), un 44.44% tiene educación primaria y un 27.54% secundaria. Solo un 5.80% tiene educación superior o técnica. En el ámbito laboral (V47), un 54.11% trabaja en el sector informal o doméstico, y un 39.61% son contratados o independientes. Un 98.07% de los hogares gana menos de 900 soles (V48). El 68.12% no recibe ningún programa social (V49). En cuanto a los materiales de construcción (V50), una gran mayoría (87.92%) de las viviendas están hechas de piedra o madera. Para los techos (V51), un 96.14% de los hogares usa paja. El 99.52% de los hogares tiene pisos de tierra (V52). En el acceso al agua (V53), un 87.92% se abastece de ríos o manantiales. Respecto al saneamiento (V54), un 92.27% utiliza río o acequia. La mayoría de los hogares (57.49%) no tiene celular (V55). En cuanto al número de habitaciones (V56), un 95.17% tiene solo una. El combustible más utilizado es la bosta (V57) por un 98.55% de los hogares. La mayoría de los hogares (65.22%) no tiene electrodomésticos (V58). Un 83.57% no tiene ningún tipo de vehículo (V59). y Todos los hogares son de un piso o cabaña (V60).

4.2. DISCUSIÓN

En nuestro estudio, al igual que en la investigación de Deza (2021), se empleó el método de la silueta para determinar el número óptimo de clústeres, concluyendo en ambos casos que dos es el número ideal. Esta coincidencia no solo valida nuestra elección metodológica, sino que también refuerza la confiabilidad de nuestros resultados en la segmentación socioeconómica. La consistencia entre ambos estudios subraya la solidez

de usar el método de la silueta en contextos de análisis de datos complejos y heterogéneos, como es el caso del distrito de Macusani.

Uno de los aspectos más destacados de nuestro estudio es la consistencia de nuestros hallazgos con investigaciones previas, particularmente con el trabajo realizado por Chavez (2020) En su estudio, Chavez también empleó el índice Davies-Bouldin para validar la segmentación de clústeres, encontrando resultados que respaldan la eficacia de este índice como una herramienta de evaluación confiable en el análisis de agrupamientos.

En nuestro estudio, a diferencia de la investigación de Pacha (2018)donde se utilizaron algoritmos de k-means y jerárquico para identificar tres clústeres, nosotros concluimos que dos clústeres son óptimos. Esta variación subraya cómo las diferencias en los contextos de datos o los criterios de segmentación pueden influir en la determinación del número ideal de clústeres, destacando la importancia de una selección metodológica adaptada a las características específicas de cada conjunto de datos.

V.CONCLUSIONES

- Se concluye que el método de análisis de clúster demostró ser una herramienta eficaz para segmentar los hogares en el distrito de Macusani según sus indicadores socioeconómicos en el año 2020.
- Se concluye que nuestro análisis indica que dos es el número óptimo de grupos para esta segmentación de hogares con indicadores socioeconómicos en el distrito de Macusani del 2020.
- Se concluye que algoritmo de clúster jerárquico ha demostrado ser el más efectivo para la segmentación de hogares con indicadores socioeconómicos en el distrito de Macusani del 2020.
- Con respecto a la caracterización de los clústeres según el anexo 2, el Clúster 1 se identificó por tener una mayor proporción de hogares con acceso al seguro de salud SIS, niveles de educación que alcanzan principalmente hasta la primaria, y una tendencia a la empleabilidad en sectores independientes. En contraste, el Clúster 2 se caracterizó por un mayor porcentaje de hogares con acceso al SIS, pero con una notable participación en el sector informal o doméstico, así como un predominio de viviendas construidas con materiales de piedra o madera.

VI. RECOMENDACIONES

- Sugerir estudios futuros en otros distritos o regiones para comparar y contrastar los patrones socioeconómicos. También sería valioso realizar estudios longitudinales para observar cómo cambian estos patrones con el tiempo.
- También recomendar análisis más detallados de variables particulares que mostraron variabilidad significativa entre los clústeres, como ingresos, acceso a servicios básicos o niveles educativos.
- También sugerir cómo los resultados pueden ser utilizados por los responsables de la formulación de políticas para diseñar e implementar programas más efectivos y dirigidos, especialmente para los grupos más vulnerables identificados en el estudio.
- Finalmente recomendar la aplicación de técnicas analíticas más avanzadas, como el aprendizaje automático y la inteligencia artificial, para mejorar la precisión y la eficiencia de la segmentación socioeconómica.

VII. REFERENCIAS BIBLIOGRÁFICAS

Aguirre, A. de, & Orfelinda, E. (2020). Diferenciación entre las regiones del país de la mortalidad materna según el contexto socioeconómico y de elementos asociados a la salud—2018. *Universidad Nacional del Altiplano*. https://repositorio.unap.edu.pe/handle/20.500.14082/14324

Alvarado, Y. (2021). *Análisis de conglomerados en la provincia de Chimborazo, periodo 2018*. Universidad Nacional de Chimborazo.

Berrios Mamani, L. Y. (2023). Minería de datos para explorar información que nos permita encontrar qué relación tienen los reportes atendidos por el Programa Nacional Contra la Violencia Familiar del año 2020. *Universidad Nacional del Altiplano*. https://repositorio.unap.edu.pe/handle/20.500.14082/19411

Chaparro Caso López, A. A., González Barbera, C., & Caso Niebla, J. (2016). Familia y rendimiento académico: Configuración de perfiles estudiantiles en secundaria. *Revista electrónica de investigación educativa*, *18*(1), 53-68.

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2012). *NbClust Package. An examination of indices for determining the number of clusters*. https://hal.science/hal-01126138

Chavez Valderrama, L. A. W. (2020). *Caracterización del perfil del ingresante de una Universidad Pública aplicando algoritmos clustering K-Prototypes y K- Medoids*. http://repositorio.lamolina.edu.pe/handle/20.500.12996/4633

Che Piu Deza, G. (2021). *Clasificación de los adolescentes infractores del centro juvenil de diagnóstico y rehabilitación de lima utilizando partición alrededor de medoides (PAM)*. http://repositorio.lamolina.edu.pe/handle/20.500.12996/4841

Flores Bermejo, G. B. (2020). Clusterización de las regiones del Perú, un análisis de interdependencia según indicadores socioeconómicos. *Universidad Nacional del Altiplano*. https://repositorio.unap.edu.pe/handle/20.500.14082/14427

Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, *27*(4), 857-871. https://doi.org/10.2307/2528823

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*.

Hastie, T., Tibshirani, R., & Friedman, J. (2016). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*.

Jackson, D., Arrocha, R., & Estrella Engelmann, J. (2021). El TERCE en poblaciones vulnerables de Panamá. *Investigación y Pensamiento Crítico*, *9*, 51-67. https://doi.org/10.37387/ipc.v9i3.264

Kassambara, M. A. (2017). *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*.

Lawson, R. G., & Jurs, P. C. (1990). New index for clustering tendency and its application to chemical problems. *Journal of Chemical Information and Computer Sciences*, *30*(1), 36-41. https://doi.org/10.1021/ci00065a010

Newbold, P. (s. f.). *Estadistica para administracion y economia*.

Pacha Phocco, O. Y. (2018). Segmentación de alpacas Huacaya de la calidad de fibra en el distrito de Cuyocuyo—Sandia, Puno—2016. *Universidad Nacional del Altiplano*. https://repositorio.unap.edu.pe/handle/20.500.14082/18297

Parra Perea, A. F. (2020). *Análisis multivariado para los departamentos de Colombia en el año 2018 según los criterios de mortalidad de la lista 6/67 y su relación con determinantes económicos*. https://repository.libertadores.edu.co/handle/11371/3593

Patel, A. A. (2019). *Hands-On Unsupervised Learning Using Python: How to Build Applied Machine Learning Solutions from Unlabeled Data*. O'Reilly Media, Inc.

Piedra Paravicino, C. A. (2022). *Segmentación de clientes potenciales del sector inmobiliario en Lima Metropolitana*. http://repositorio.lamolina.edu.pe/handle/20.500.12996/5921

Tang Bedoya, F. A., & Vargas Cuyan, C. (2016). Segmentación de clientes de una tienda de electrodomésticos utilizando el análisis de conglomerados. *Universidad Nacional Agraria La Molina*. http://repositorio.lamolina.edu.pe/handle/20.500.12996/2214

Tonconi Calisaya, C. A. (2021). *Identificación de perfiles de los Centros de Educación Técnico—Productiva Públicos usando indicadores de condiciones básicas de calidad mediante clúster bietápico*. http://repositorio.lamolina.edu.pe/handle/20.500.12996/4946

ANEXOS

ANEXO 1: Matriz de consistencia

Problema	**Objetivos**	**Hipótesis**	**Variables**	**Metodología**
General	**General**	**General**		
¿Cuál es el método para la segmentación de hogares con indicadores socioeconómico en el distrito de Macusani del 2020?	Determinar el método para la segmentación de hogares con indicadores socioeconómico en el distrito de Macusani del 2020.	El análisis de clúster es el método para la segmentación de hogares con indicadores socioeconómico en el distrito de Macusani del 2020	Variables Socioeconómicas	Tipo: Descriptivo
Específicos	**Específicos**			Diseño: No experimental transversal
¿Cuál es el número óptimo de clústeres para la segmentación de hogares con indicadores socioeconómico en el distrito de Macusani del 2020?	Establecer el número óptimo de clústeres para la segmentación de hogares con indicadores socioeconómico en el distrito de Macusani del 2020.			

¿Cuál es el algoritmo de análisis de clúster más efectivo para la segmentación de hogares con indicadores socioeconómico en el distrito de Macusani del 2020?	Encontrar el algoritmo de análisis de clúster más efectivo para la segmentación de hogares con indicadores socioeconómico en el distrito de Macusani del 2020.			
¿Cómo se caracteriza la segmentación de hogares con indicadores socioeconómico en el distrito de Macusani del 2020?	Caracterizar la segmentación de hogares con indicadores socioeconómico en el distrito de Macusani del 2020.			

ANEXO 2: Caracterización de los clústeres

Variables		Categorías	Clúster	
			1	2
V45: Tipo de seguro	2	ESSALUD	23 (16.79%)	6 (2.90%)
	4	NO TIENE	36 (26.28%)	80 (38.65%)
	3	SIS	78 (56.93%)	121 (58.45%)
V46: Educación	4	NO TIENE	21 (15.33%)	46 (22.22%)
	3	PRIMARIA	68 (49.64%)	92 (44.44%)
	2	SECUNDARIA	32 (23.36%)	57 (27.54%)
	1	SUPERIOR/TÉCNICO	16 (11.68%)	12 (5.80%)
V47: Estado Laboral	2	CONTRATADO/INDEPENDIENTE	90 (65.69%)	82 (39.61%)
	4	DESEMPLEADO	8 (5.84%)	9 (4.35%)
	3	INFORMAL/DOMESTICO	32 (23.36%)	112 (54.11%)
	1	NOMBRADO	7 (5.11%)	4 (1.93%)
V48: Ingreso Mensual	3	1500 a 900 SOLES	15 (10.95%)	2 (0.97%)
	2	2500 a 1500 SOLES	1 (0.73%)	2 (0.97%)
	4	MENOS DE 900 SOLES	121(88.32%)	203 (98.07%)
V49: Número Programas Beneficio	3	1 PROGRAMA/2 PROGRAMAS	48 (35.04%)	63 (30.43%)
	2	MAS DE 4 PROGRAMAS	0 (0.00%)	3 (1.45%)
	1	NO NECESITA	2 (1.46%)	0 (0.00%)
	4	NO TIENE	87 (63.50%)	141 (68.12%)
V50: Tipo de material de la pared	3	ADOBE/QUINCHA	120(87.59%)	25 (12.08%)
	1	LADRILLO Y CEMENTO	3 (2.19%)	0 (0.00%)
	4	PIEDRA/MADERA	14 (10.22%)	182 (87.92%)
V51: Tipo de material del techo	3	CALAMINA/FIBRA	92 (67.15%)	8 (3.86%)
	4	PAJA	45 (32.85%)	199 (96.14%)
V52: Tipo de material del suelo	3	CEMENTO	8 (5.84%)	0 (0.00%)
	2	MADERA	8 (5.84%)	0 (0.00%)
	1	PARQUET/CERÁMICA	1 (0.73%)	1 (0.48%)
	4	TIERRA	120(87.59%)	206 (99.52%)
V53: Fuente de agua	2	PILÓN USO COMÚN	5 (3.65%)	0 (0.00%)
	3	POZO	24 (17.52%)	25 (12.08%)
	1	RED PÚBLICA	17 (12.41%)	0 (0.00%)
	4	RÍO/MANANTIAL	91 (66.42%)	182 (87.92%)
V54: Tipo saneamiento	3	POZO CIEGO/LETRINA	42 (30.66%)	14 (6.76%)
	2	POZO SÉPTICO	3 (2.19%)	2 (0.97%)
	1	RED PÚBLICA	14 (10.22%)	0 (0.00%)
	4	RÍO/ACEQUIA	78 (56.93%)	191 (92.27%)
V55: Servicio de celular	2	CELULAR CON INTERNET	3 (2.19%)	6 (2.90%)
	3	CELULAR SIN INTERNET	67 (48.91%)	82 (39.61%)
	4	NO TIENE	67 (48.91%)	119 (57.49%)

V56: Cantidad de habitaciones	3	DOS	14 (10.22%)	3 (1.45%)
	1	MAS DE CUATRO	1 (0.73%)	1 (0.48%)
	2	TRES	6 (4.38%)	6 (2.90%)
	1	UNO	116(84.67%)	197 (95.17%)
V57: Tipo de combustible de cocina	4	BOSTA	87 (63.50%)	204 (98.55%)
	3	CARBÓN/LEÑA	3 (2.19%)	1 (0.48%)
	2	GAS	47 (34.31%)	2 (0.97%)
V58: Cantidad de electrodomésticos	3	5 A 1	72 (52.55%)	72 (34.78%)
	4	NO TIENE	65 (47.45%)	135 (65.22%)
V59: Tipo de vehículo	1	DE USO PARTICULAR	1 (0.73%)	1 (0.48%)
	3	MOTO LINEAL/BICICLETA	29 (21.17%)	32 (15.46%)
	4	NO TIENE	105(76.64%)	173 (83.57%)
	2	USO EMPRESARIAL/CARGA/SERVICIO	2 (1.46%)	1 (0.48%)
V60: Número de pisos	3	DOS PISOS	2 (1.46%)	0 (0.00%)
	4	UN PISO/CABAÑA	135(98.54%)	207(100.00%)

ANEXO 3: Hogares por clúster

Grupos	Total	Hogares
1	137	H1, H3, H4, H5, H6, H7, H8, H12, H13, H14, H15, H16, H18, H19, H22, H24, H25, H26, H27, H28, H29, H31, H32, H33, H34, H35, H36, H38, H40, H43, H45, H48, H49, H54, H57, H60, H61, H62, H65, H66, H67, H70, H71, H72, H74, H77, H79, H81, H88, H89, H90, H98, H100, H102, H107, H119, H120, H123, H124, H125, H126, H127, H128, H129, H133, H134, H135, H136, H139, H142, H143, H145, H148, H149, H150, H151, H153, H159, H160, H161, H162, H163, H165, H166, H167, H168, H169, H170, H171, H172, H174, H182, H185, H186, H187, H188, H189, H194, H196, H198, H200, H201, H202, H203, H204, H205, H209, H216, H218, H219, H227, H228, H229, H231, H232, H233, H234, H235, H236, H237, H239, H241, H246, H247, H248, H249, H251, H253, H260, H274, H287, H290, H291, H296, H301, H302, H309.
2	207	H2, H9, H10, H11, H17, H20, H21, H23, H30, H37, H39, H41, H42, H44, H46, H47, H50, H51, H52, H53, H55, H56, H58, H59, H63, H64, H68, H69, H73, H75, H76, H78, H80, H82, H83, H84, H85, H86, H87, H91, H92, H93, H94, H95, H96, H97, H99, H101, H103, H104, H105, H106, H108, H109, H110, H111, H112, H113, H114, H115, H116, H117, H118, H121, H122, H130, H131, H132, H137, H138, H140, H141, H144, H146, H147, H152, H154, H155, H156, H157, H158, H164, H173, H175, H176, H177, H178, H179, H180, H181, H183, H184, H190, H191, H192, H193, H195, H197, H199, H206, H207, H208, H210, H211, H212, H213, H214, H215, H217, H220, H221, H222, H223, H224, H225, H226, H230, H238, H240, H242, H243, H244, H245,

		H250, H252, H254, H255, H256, H257, H258, H259, H261, H262, H263, H264, H265, H266, H267, H268, H269, H270, H271, H272, H273, H275, H276, H277, H278, H279, H280, H281, H282, H283, H284, H285, H286, H288, H289, H292, H293, H294, H295, H297, H298, H299, H300, H303, H304, H305, H306, H307, H308, H310, H311, H312, H313, H314, H315, H316, H317, H318, H319, H320, H321, H322, H323, H324, H325, H326, H327, H328, H329, H330, H331, H332, H333, H334, H335, H336, H337, H338, H339, H340, H341, H342, H343, H344.
Total	344	

ANEXO 4: Código en R- estudio

```
#####################################################################
##############CARGAR LIBRERIA#########################################
#####################################################################
library(readxl)
library(ggplot2)
library(dplyr)
library(tidyr)
library(DataExplorer)
library(factoextra)
library(cluster)
library(factoextra)
library(clValid)
library(gridExtra)
library(ggcorrplot)
library(corrplot)
library(clustertend)
library(NbClust)
library(kableExtra)
library(compareGroups)
library(fpc)
library(RColorBrewer)
library(klaR)
#####################################################################
##############CARGAR DATOS DEL DISTRITO DE MACUSANI - RURAL#
#####################################################################
DATOS <- read_excel("C:/Users/jmcp1/OneDrive/Escritorio/DATOSDEMACUSANI6.xlsx")
DATOS<-data.frame(DATOS)
rownames(DATOS) <- DATOS[, 1]
DATOS <- DATOS[, -1]
datos <- data.frame(categoria = c(rep("Rural", 344), rep("Urbano", 3985)))
# Calcular los porcentajes
datos_agrupados <- as.data.frame(table(datos$categoria))
datos_agrupados$porcentaje <- round(datos_agrupados$Freq / sum(datos_agrupados$Freq) * 100, 0)
# Crear el gráfico de barras
ggplot(datos_agrupados, aes(x = Var1, y = Freq, fill = Var1)) +
geom_bar(stat = "identity") +
geom_text(aes(label = paste0(porcentaje, "%")), vjust = -0.5) +
labs(title = "Distribución del distrito de Macusani(Urbana y Rural)", x = "Categoría", y = "Cantidad") +
theme_minimal()
#####################################################################
##############PRE PROCESAMIENTO DE DATOS#############################
#####################################################################
##Limpieza de Datos
# Función para verificar si una columna tiene algún dato faltante
tiene_datos_faltantes <- function(columna) {
  any(is.na(columna))
}
DATOS <- Filter(function(columna) !tiene_datos_faltantes(columna), DATOS)
attach(DATOS)
#sacamos datos unicos
DATOS<- subset(DATOS, select = -c(V1,V43,V44))
DATOS2<-DATOS
plot_missing(DATOS)
DATOS2<-data.frame(DATOS2)
DATOS2 <- lapply(DATOS2, as.factor)
DATOS2<-data.frame(DATOS2)
create_report(DATOS)
DATOS$V45 <- match(DATOS$V45,  c("CLÍNICA PRIVADA", "ESSALUD", "SIS", "NO TIENE"))
DATOS$V46 <- match(DATOS$V46,  c("SUPERIOR/TÉCNICO", "SECUNDARIA", "PRIMARIA", "NO TIENE"))
DATOS$V47 <- match(DATOS$V47, c("NOMBRADO", "CONTRATADO/INDEPENDIENTE", "INFORMAL/DOMESTICO",
"DESEMPLEADO"))
DATOS$V48 <- match(DATOS$V48,  c("MAS DE 2500 SOLES", "2500 A 1500 SOLES", "1500 A 900 SOLES", "MENOS DE
900 SOLES"))
DATOS$V49 <- match(DATOS$V49,  c("NO NECESITA", "MAS DE 4 PROGRAMAS", "1 PROGRAMA/2 PROGRAMAS",
"NO TIENE"))
DATOS$V50 <- match(DATOS$V50, c("LADRILLO Y CEMENTO", "BLOQUE/SILLAR", "ADOBE/QUINCHA",
"PIEDRA/MADERA"))
DATOS$V51 <- match(DATOS$V51,  c("CONCRETO ARMADO", "MADERA/TEJA", "CALAMINA/FIBRA", "PAJA"))
DATOS$V52 <- match(DATOS$V52,  c("PARQUET/CERÁMICA", "MADERA", "CEMENTO", "TIERRA"))
DATOS$V53 <- match(DATOS$V53,  c("RED PÚBLICA", "PILÓN USO COMÚN", "POZO", "RÍO/MANANTIAL"))
DATOS$V54 <- match(DATOS$V54,  c("RED PÚBLICA", "POZO SÉPTICO", "POZO CIEGO/LETRINA", "RÍO/ACEQUIA"))
```

```
DATOS$V55 <- match(DATOS$V55,  c("TELÉFONO/MAS DE DOS CELULARES", "CELULAR CON INTERNET",
"CELULAR SIN INTERNET", "NO TIENE"))
DATOS$V56 <- match(DATOS$V56,  c("MAS DE 4", "TRES", "DOS", "UNO"))
DATOS$V57 <- match(DATOS$V57, c("ELECTRICIDAD/GAS", "GAS", "CARBÓN/LEÑA", "BOSTA"))
DATOS$V58 <- match(DATOS$V58,  c("MAS DE 12", " 6 A 11", "5 A 1", "NO TIENE"))
DATOS$V59 <- match(DATOS$V59, c("DE USO PARTICULAR", "USO EMPRESARIAL/CARGA/SERVICIO", "MOTO
LINEAL/BICICLETA", "NO TIENE"))
DATOS$V60 <- match(DATOS$V60,  c("MAS DE CUATRO PISOS", "TRES PISOS", "DOS PISOS", "UN PISO/CABAÑA"))
matriz_correlacion <- cor(DATOS, use = "complete.obs")
matriz_correlacion <- cor(DATOS, use = "complete.obs")
corrplot(matriz_correlacion, method = "color", col = colores, cl.lim = c(-1, 1), addCoef.col = "black")
#####################################################################
##############EVALUACION DE TENDENCIA DEL CONJUNTO DE DATOS#########
#####################################################################
get_clust_tendency(DATOS,n=40, gradient = list(low = "steelblue", high = "white"))
#####################################################################
##############DETERMINAR EL NUMERO OPRIMO DE CLUSTERES##############
#####################################################################
#Dendograma
gower_dist <- daisy(DATOS2, metric = "gower")
hc.res <- hclust(gower_dist, method = "ward.D2")
# Crear el dendrograma con etiquetas
fviz_dend(hc.res, k = 1, show_labels =FALSE,
          palette = "jco", as.ggplot = TRUE) +
  labs(title = "Visualización de Dendograma") +
  theme(text = element_text(size = 12),
        plot.title = element_text(size = 15, face = "bold"),
        panel.grid.major = element_line(size = 0.5, linetype = 'solid', color = "grey"),
        panel.grid.minor = element_line(size = 0.25, linetype = 'solid', color = "lightgrey"))
#metodo del codo
set.seed(123)
fviz_nbclust(DATOS, kmeans, method = "wss") +
  geom_vline(xintercept = 2, linetype = 2, color = "blue", size = 1) +
  labs(subtitle = "Método del Codo",
       x = "Número de Clusters",
       y = "Suma Total de Cuadrados (WSS)") +
  theme_minimal() +
  theme(plot.subtitle = element_text(hjust = 0.5, face = "italic", size = 12),
        axis.title.x = element_text(size = 12, face = "bold"),
        axis.title.y = element_text(size = 12, face = "bold"),
        axis.text.x = element_text(color = "black"),
        axis.text.y = element_text(color = "black"),
        legend.position = "none")  # Esconde la leyenda si es necesario
# Silhouette method
set.seed(123)
fviz_nbclust(DATOS, kmeans, method = "silhouette") +
  geom_vline(xintercept = 2, linetype = 2, color = "blue", size = 1) +
  labs(subtitle = "Método de la Silueta",
       x = "Número de Clusters",
       y = "Ancho promedio de la Silueta") +
  theme_minimal() +
  theme(plot.subtitle = element_text(hjust = 0.5, face = "italic", size = 12),
        axis.title.x = element_text(size = 12, face = "bold"),
        axis.title.y = element_text(size = 12, face = "bold"),
        axis.text.x = element_text(color = "black"),
        axis.text.y = element_text(color = "black"),
        legend.position = "none")
#####################################################################
##############CLASTERING DE PARTICION################################
#####################################################################
#KMEANS
set.seed(123)
km.res <- kmeans(DATOS, 2, nstart = 100)
fviz_cluster(km.res, data = DATOS, palette = c("#2E9FDF","#E7B800"),ellipse.type = "convex",
             ggtheme = theme_minimal(),
             #repel = TRUE,
             geom = c("point", "text"), pointsize = 1,labelsize = 20) +
  labs(title = "Visualización de Clusters con algoritmo K-means con k=2",
       x = "Componente Principal 1",
       y = "Componente Principal 2") +
  theme(text = element_text(size = 12),
        plot.title = element_text(size = 15, face = "bold"),
        panel.grid.major = element_line(size = 0.25, linetype = 'solid', color = "grey"),
        panel.grid.minor = element_line(size = 0.25, linetype = 'solid', color = "lightgrey"))

table(km.res$cluster)
```

```
#PAM
set.seed(123)
pam.res<- pam(DATOS,2)
fviz_cluster(pam.res, data = DATOS, palette = c("#FC4E07","#00AFBB"),ellipse.type = "t",
             ggtheme = theme_minimal(),
             geom =  c("point", "text"), pointsize = 1,labelsize = 20) +
  labs(title = "Visualización de Clusters con algoritmo K-medoids(PAM) con k=2",
       x = "Componente Principal 1",
       y = "Componente Principal 2") +
  theme(text = element_text(size = 12),
        plot.title = element_text(size = 15, face = "bold"),
        panel.grid.major = element_line(size = 0.5, linetype = 'solid', color = "grey"),
        panel.grid.minor = element_line(size = 0.25, linetype = 'solid', color = "lightgrey"))
table(pam.res$cluster)
#####################################################################
#############CLUSTERIN JERARQUICO#################################
#####################################################################
grp <- cutree(hc.res, k = 2)
# Crear el dendrograma con etiquetas
fviz_dend(hc.res, k = 2, show_labels =FALSE,
          palette = "jco", as.ggplot = TRUE) +
  labs(title = "Visualización de Dendograma") +
  theme(text = element_text(size = 12),
        plot.title = element_text(size = 15, face = "bold"),
        panel.grid.major = element_line(size = 0.5, linetype = 'solid', color = "grey"),
        panel.grid.minor = element_line(size = 0.25, linetype = 'solid', color = "lightgrey"))
table(grp)
#####################################################################
#############ESCOGER EL ALGORITMO#################################
#####################################################################
clmethods <- c("hierarchical","kmeans","pam")
intern <- clValid(DATOS, nClust = 2:10,
                  clMethods = clmethods, validation = "internal")
summary(intern)
library(clusterSim)
db.km<-index.DB(DATOS, km.res$cluster, centrotypes = "centroids")$DB
db.pam<-index.DB(DATOS, pam.res$clustering, centrotypes = "centroids")$DB
db.grp<-index.DB(DATOS, grp, centrotypes = "centroids")$DB
table(db.km,db.pam,db.grp)
#####################################################################
#############Caracterizar la segmentación de hogares ###############################
#####################################################################
DATOS_Hh<-cbind(DATOS,grp)
DATOS_H<-cbind(DATOS2,grp)
COMPARARH<-compareGroups(grp ~.,data = DATOS_H)
TABLE_H<-createTable(COMPARARH,digits = 2,
                     show.p.overall = FALSE)
TABLE_H
DATOS_H[] <- lapply(DATOS_H, as.factor)
colnames(DATOS_H)[colnames(DATOS_H) == "grp"] <- "Cluster"
clustergrp_1 <- which(grp == 1)
clustergrp_1
clustergrp_2 <- which(grp == 2)
clustergrp_2
create_report(DATOS_Hh
```

ANEXO 5: Declaración jurada de autenticidad de tesis.

Universidad Nacional del Altiplano Puno | Vicerrectorado de Investigación | Repositorio Institucional

DECLARACIÓN JURADA DE AUTENTICIDAD DE TESIS

Por el presente documento, Yo JUAN MANUEL CONDORI PERALTA,
identificado con DNI 70296281 en mi condición de egresado de:

☒ **Escuela Profesional,** ☐ **Programa de Segunda Especialidad,** ☐ **Programa de Maestría o Doctorado**
INGENIERIA ESTADISTICA E INFORMATICA,
informo que he elaborado el/la ☒ **Tesis** o ☐ **Trabajo de Investigación** denominada:
" SEGMENTACION DE HOGARES CON INDICADORES SOCIOECONOMICOS DEL DISTRITO DE MACUSANI - 2020 "

Es un tema original.

Declaro que el presente trabajo de tesis es elaborado por mi persona y **no existe plagio/copia** de ninguna naturaleza, en especial de otro documento de investigación (tesis, revista, texto, congreso, o similar) presentado por persona natural o jurídica alguna ante instituciones académicas, profesionales, de investigación o similares, en el país o en el extranjero.

Dejo constancia que las citas de otros autores han sido debidamente identificadas en el trabajo de investigación, por lo que no asumiré como suyas las opiniones vertidas por terceros, ya sea de fuentes encontradas en medios escritos, digitales o Internet.

Asimismo, ratifico que soy plenamente consciente de todo el contenido de la tesis y asumo la responsabilidad de cualquier error u omisión en el documento, así como de las connotaciones éticas y legales involucradas.

En caso de incumplimiento de esta declaración, me someto a las disposiciones legales vigentes y a las sanciones correspondientes de igual forma me someto a las sanciones establecidas en las Directivas y otras normas internas, así como las que me alcancen del Código Civil y Normas Legales conexas por el incumplimiento del presente compromiso

Puno 5 de ENERO del 2024

FIRMA (obligatoria)

Huella

ANEXO 6: Autorización para el depósito de tesis en el Repositorio Institucional.

Universidad Nacional del Altiplano Puno | Vicerrectorado de Investigación | Repositorio Institucional

AUTORIZACIÓN PARA EL DEPÓSITO DE TESIS O TRABAJO DE INVESTIGACIÓN EN EL REPOSITORIO INSTITUCIONAL

Por el presente documento, Yo JUAN MANUEL CONDORI PERALTA,
identificado con DNI 70296281 en mi condición de egresado de:

☒ **Escuela Profesional,** ☐ **Programa de Segunda Especialidad,** ☐ **Programa de Maestría o Doctorado**

INGENIERIA ESTADISTICA E INFORMATICA,

informo que he elaborado el/la ☒ **Tesis** o ☐ **Trabajo de Investigación** denominada:

"SEGMENTACIÓN DE HOGARES CON INDICADORES SOCIOECÓNOMICOS DEL DISTRITO DE MACUSANI - 2020"

para la obtención de ☐**Grado,** ☒ **Título Profesional** o ☐ **Segunda Especialidad.**

Por medio del presente documento, afirmo y garantizo ser el legítimo, único y exclusivo titular de todos los derechos de propiedad intelectual sobre los documentos arriba mencionados, las obras, los contenidos, los productos y/o las creaciones en general (en adelante, los "Contenidos") que serán incluidos en el repositorio institucional de la Universidad Nacional del Altiplano de Puno.

También, doy seguridad de que los contenidos entregados se encuentran libres de toda contraseña, restricción o medida tecnológica de protección, con la finalidad de permitir que se puedan leer, descargar, reproducir, distribuir, imprimir, buscar y enlazar los textos completos, sin limitación alguna.

Autorizo a la Universidad Nacional del Altiplano de Puno a publicar los Contenidos en el Repositorio Institucional y, en consecuencia, en el Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto, sobre la base de lo establecido en la Ley N° 30035, sus normas reglamentarias, modificatorias, sustitutorias y conexas, y de acuerdo con las políticas de acceso abierto que la Universidad aplique en relación con sus Repositorios Institucionales. Autorizo expresamente toda consulta y uso de los Contenidos, por parte de cualquier persona, por el tiempo de duración de los derechos patrimoniales de autor y derechos conexos, a título gratuito y a nivel mundial.

En consecuencia, la Universidad tendrá la posibilidad de divulgar y difundir los Contenidos, de manera total o parcial, sin limitación alguna y sin derecho a pago de contraprestación, remuneración ni regalía alguna a favor mío; en los medios, canales y plataformas que la Universidad y/o el Estado de la República del Perú determinen, a nivel mundial, sin restricción geográfica alguna y de manera indefinida, pudiendo crear y/o extraer los metadatos sobre los Contenidos, e incluir los Contenidos en los índices y buscadores que estimen necesarios para promover su difusión.

Autorizo que los Contenidos sean puestos a disposición del público a través de la siguiente licencia:

Creative Commons Reconocimiento-NoComercial-CompartirIgual 4.0 Internacional. Para ver una copia de esta licencia, visita: https://creativecommons.org/licenses/by-nc-sa/4.0/

En señal de conformidad, suscribo el presente documento.

Puno 5 de ENERO del 2024

FIRMA (obligatoria) Huella