UNIVERSIDAD NACIONAL DE SAN ANTONIO ABAD DEL CUSCO

ESCUELA DE POSGRADO

MAESTRÍA EN ESTADÍSTICA

TESIS

APLICACIONES DE LOS MÉTODOS DE ANÁLISIS DE CLÚSTER Y CORRESPONDENCIA EN EL ESTUDIO DE RESULTADOS DE EXAMEN DE ADMISIÓN DE LA UNSAAC, 2022

PARA OPTAR AL GRADO ACADÉMICO DE MAESTRO EN ESTADÍSTICA

AUTOR

Br. LUZ MARINA CATUNTA GUILLEN

ASESOR

Mtro. ARTURO ZUÑIGA BLANCO

CODIGO ORCID

0000-0002-8576-3415

CUSCO – PERÚ

2024

INFORME DE ORIGINALIDAD

(Aprobado por Resolución Nro.CU-303-2020-UNSAAC)

El que suscribe, **Asesor** del trabajo de investigación/tesis titulada: "Aplicaciones de los métodos de análisis de clúster y correspondencia en el estudio de resultados de examen de admisión de la UNSAAC, 2022"

presentado por: Luz Marina Catunta Guillen con DNI Nro.: 40591687 presentado por: ..... con DNI Nro.: ..... para optar el título profesional/grado académico de Maestro en estadística

Informo que el trabajo de investigación ha sido sometido a revisión por 2 veces, mediante el Software Antiplagio, conforme al Art. 6° del ***Reglamento para Uso de Sistema Antiplagio de la UNSAAC*** y de la evaluación de originalidad se tiene un porcentaje de 10 %.

Evaluación y acciones del reporte de coincidencia para trabajos de investigación conducentes a grado académico o título profesional, tesis

Porcentaje	Evaluación y Acciones	Marque con una (X)
Del 1 al 10%	No se considera plagio.	X
Del 11 al 30 %	Devolver al usuario para las correcciones.	
Mayor a 31%	El responsable de la revisión del documento emite un informe al inmediato jerárquico, quien a su vez eleva el informe a la autoridad académica para que tome las acciones correspondientes. Sin perjuicio de las sanciones administrativas que correspondan de acuerdo a Ley.	

Por tanto, en mi condición de asesor, firmo el presente informe en señal de conformidad y **adjunto** la primera página del reporte del Sistema Antiplagio.

Cusco, 07 de marzo de 2024

Firma

Post firma Arturo Zuniga Blanco

Nro. de DNI 46452024

ORCID del Asesor 0000-0002-8576-3415

Se adjunta:

1. Reporte generado por el Sistema Antiplagio.
2. Enlace del Reporte Generado por el Sistema Antiplagio: **oid:** 27259:338007099

NOMBRE DEL TRABAJO

Luz Marina Catunta-APLICACIONES DE LOS MÉTODOS DE ANÁLISIS DE CLÚSTER Y CORRESPONDENCIA.docx

AUTOR

LUZ MARINA CATUNTA GUILLEN

RECUENTO DE PALABRAS

18694 Words

RECUENTO DE CARACTERES

99545 Characters

RECUENTO DE PÁGINAS

96 Pages

TAMAÑO DEL ARCHIVO

925.8KB

FECHA DE ENTREGA

Mar 7, 2024 6:18 AM GMT-5

FECHA DEL INFORME

Mar 7, 2024 6:20 AM GMT-5

10% de similitud general

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

- 10% Base de datos de Internet
- 0% Base de datos de publicaciones
- Base de datos de Crossref
- Base de datos de contenido publicado de Crossref
- 4% Base de datos de trabajos entregados

Excluir del Reporte de Similitud

- Material bibliográfico
- Material citado
- Material citado
- Coincidencia baja (menos de 15 palabras)
- Bloques de texto excluidos manualmente

DEDICATORIA

A Dios, fuente inagotable de sabiduría y guía, agradezco por iluminar mi camino durante este arduo viaje académico. A mis queridos padres, cuyo amor incondicional y apoyo constante han sido mi mayor fortaleza, les dedico este logro a su sacrificio y dedicación que han sido la inspiración que me impulsa a alcanzar mis metas; A Santiago por su apoyo incondicional, a mis hermanos Edgar y Graciela, quienes me impulsaron a continuar con mis estudios y a mi amado hijo Samín por ser mi motor y motivo de seguir creciendo más académicamente y profesionalmente.

Luz Marina Catunta Guillen

AGRADECIMIENTO

Quiero expresar mi sincero agradecimiento al Mtro. Arturo Zuñiga Blanco, mi asesor, cuya orientación y apoyo incondicional han sido esenciales para el éxito de esta tesis. Su dedicación y conocimiento han sido una inspiración constante en mi camino académico.

Asimismo, deseo agradecer a la Escuela de Posgrado de la Universidad San Antonio Abad del Cusco por brindarme la oportunidad de realizar este estudio y por ofrecer un entorno propicio para el crecimiento académico. La calidad de la educación y los recursos proporcionados han sido clave en mi formación.

Luz Marina Catunta Guillen

RESUMEN

El objetivo de la presente investigación fue la de analizar las APLICACIONES DE LOS MÉTODOS DE ANÁLISIS DE CLÚSTER Y CORRESPONDENCIA EN EL ESTUDIO DE RESULTADOS DE EXAMEN DE ADMISION DE LA UNSAAC, 2022, se aplicaron técnicas estadísticas del análisis de conglomerados como el algoritmo bietapico, pam (partition around medoids) y clara (Clustering Large Applications) para poder cumplir con el objetivo trazado, la investigación tuvo un enfoque cuantitativo, de alcance descriptivo, mientras que el diseño de investigación es no experimental, de tipo transversal; así mismo la población de estudio son los estudiantes que ya egresaron de la educación básica regular quienes postularon en los diferentes semestres de la universidad San Antonio Abad del Cusco.

Se han utilizado 3 algoritmos para poder conglomerar a los individuos, como el algoritmos de PAM (Partitioning Around Medoids) y CLARA (Clustering Large Applications) y el algoritmo bietápico, se calculó una matriz de distancias con la metodología de distancias mixtas de gower, en tanto se decidió utilizar el algoritmo bietápico por tener mejor medida de silueta de cohesión y separación; se determinó que 4 clústeres eran adecuados para describir los perfiles de los postulantes donde se observó que el clúster 1, tiene prevalencia de alumnos que no ingresaron cuando postularon, su edad promedio es 18.49 años y su nota promedio fue de 7.44, su procedencia en su mayoría es del Cusco, de sexo femenino, procedencia de colegio nacional y mayormente postulan al grupo D. En el clúster 2, tiene prevalencia de alumnos que si lograron una vacante en la UNSAAC, con edad promedio de 20.7 años y su nota promedio fue de 12.58, su procedencia mayormente es del Cusco, de sexo masculino, procedencia de colegio nacional; en el clúster 3, tiene prevalencia de alumnos que no ingresaron cuando postularon, su edad promedio es 18.49 años y su nota promedio fue de 7.67, su procedencia en su mayoría es del Cusco, de sexo masculino, procedencia de colegio particular y mayormente postulan al grupo A; en el clúster 4, tiene prevalencia de alumnos que no ingresaron cuando postularon, su edad promedio es 18.50 años y su nota promedio fue de 7.75, su procedencia en su mayoría es del Cusco, de sexo femenino, procedencia de colegio particular y mayormente postulan al grupo A.

Palabras clave: Perfil, Postulantes, Algoritmo bietápico, PAM, CLARA.

ABSTRACT

The objective of the present research was to analyze the APPLICATIONS OF THE CLUSTER AND CORRESPONDENCE ANALYSIS METHODS IN THE STUDY OF UNSAAC ADMISSION EXAM RESULTS, 2022, statistical techniques of cluster analysis such as the bietapic algorithm were applied, pam (partition around medoids) and clara (Clustering Large Applications) were applied in order to meet the objective set, the research had a quantitative approach, descriptive in scope, while the research design is non-experimental, cross-sectional; Likewise, the study population are the students who have already graduated from regular basic education who applied to the different semesters of the San Antonio Abad del Cusco University.

Three algorithms have been used to cluster the individuals, such as the PAM (Partitioning Around Medoids) and CLARA (Clustering Large Applications) algorithms and the bietapic algorithm, a distance matrix was calculated with the methodology of mixed distances of Gower, while it was decided to use the bietapic algorithm for having a better measure of cohesion and separation silhouette; It was determined that 4 clusters were adequate to describe the profiles of the applicants, where it was observed that Cluster 1, has a prevalence of students who did not enter when they applied, their average age is 18. 49 years and their average grade was 7.44, they are mostly from Cusco, are female, come from a national school, and mostly apply to group D. Cluster 2, has a prevalence of students who entered when they applied, their average age is 20.7 years old and their average grade was 12.58, they are mostly from Cusco, male, from a national school; Cluster 3, has a prevalence of students who did not enter when they applied, their average age is 18.49 years old and their average grade was 7.67, they are mostly from Cusco, are male, come from a private school and mostly apply to group A; in Cluster 4, there is a prevalence of students who did not enter when they applied, their average age is 18.50 years and their average grade was 7.75, they are mostly from Cusco, are female, come from a private school and mostly apply to group A.

Key words: profile, applicants, bietapic algorithm, PAM, CLARA.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

CAPÍTULO I

PLANTEAMIENTO DEL PROBLEMA

1.1. SITUACIÓN PROBLEMÁTICA

En el contexto educativo peruano, la selección de estudiantes para ingresar a universidades nacionales es un proceso crucial que influye en la calidad y diversidad de la población estudiantil (Cuenca, 2015). Sin embargo, existe una falta de comprensión detallada sobre los perfiles de los postulantes y los factores que impactan en la toma de decisiones durante el proceso de selección (Rojas, 2020).

A pesar de los esfuerzos por garantizar la equidad y la diversidad en el acceso a la educación superior, persisten desafíos en la identificación y evaluación de los perfiles de los postulantes a una universidad nacional en el Perú (Gairín & Palmeros, 2018). Este problema se manifiesta en la falta de información detallada sobre aspectos clave, como antecedentes académicos, habilidades extracurriculares, experiencias de vida y desafíos enfrentados por los postulantes (Clavijo & Bautista-Cerro, 2020).

La evaluación sumativa tiene como función determinar el grado de logro o aprovechamiento que un alumno ha obtenido en relación con los objetivos fijados para una etapa (Pérez Morales, 2007). Se realiza habitualmente, por tanto, al final de un proceso de enseñanza-aprendizaje por ejemplo quinto de secundaria, y se vincula a las decisiones de ingreso a las universidades o culminación de un curso. Está totalmente ligada a la denominada evaluación del aprovechamiento (Paredes, 2017).

La Universidad Nacional de San Antonio Abad del Cusco en su reglamento de admisión estipula ingresos por concurso: de admisión, primera oportunidad, por

centro preuniversitario y la modalidad especial, las pruebas de admisión se centran en las evaluaciones de conocimiento de alternativa múltiple (UNSAAC, 2018).

Para toda institución académica es de vital importancia analizar el perfil de los estudiantes postulantes, donde algunos de ellos serán admitidos y otros no, por lo tanto, se debe aplicar técnicas estadísticas para este fin, una de ellas es el análisis de clúster y correspondencia (Chavez L. , 2020).

La metodología de análisis de clústeres se fundamenta en la ingeniosa noción de reunir un conjunto de observaciones en una cantidad predeterminada de clústeres o grupos. Este enfoque estratégico busca dotar de coherencia y orden a la información, facilitando la comprensión y extracción de patrones significativos en el conjunto de datos. Al agrupar de manera inteligente las observaciones, se abre la puerta a la revelación de conexiones y tendencias que podrían pasar desapercibidas en un análisis individual. En esencia, el análisis de clústeres se erige como un medio eficaz para explorar la estructura inherente en conjuntos de datos complejos, ofreciendo una perspectiva que va más allá de la mera observación individual. Este agrupamiento se basa en la idea de distancia o similitud entre las observaciones; en cambio el análisis de correspondencia permite describir perfiles según los grupos constituidos por el clúster (Areválo & Pérez Gonzales, 2018).

1.2. FORMULACION DEL PROBLEMA

1.2.1. PROBLEMA GENERAL

¿Cuál es el perfil de los conglomerados de estudiantes postulantes a la UNSAAC obtenidos de las aplicaciones de los métodos de análisis de clúster y correspondencia?

1.2.2. PROBLEMA ESPECÍFICOS

a) ¿Qué método de clúster es el más óptimo para determinar el perfil de los estudiantes postulantes a la UNSAAC?
b) ¿Qué variables presentan mayor repercusión en la elaboración de conglomerados de estudiantes postulantes a la UNSAAC?
c) ¿Qué perfil presentan los estudiantes con mejores puntajes en el examen de admisión de la UNSAAC, obtenidos del análisis de correspondencia?

1.3. JUSTIFICACION DE LA INVESTIGACIÓN

La realización de este trabajo académico es importante por varias razones que abordan las necesidades actuales y futuras del sistema educativo. A continuación, se presenta la justificación del estudio:

La equidad en la educación es un principio fundamental para garantizar que todos los individuos, independientemente de su origen socioeconómico, tengan igualdad de oportunidades en el acceso a la educación superior.

La comprensión detallada de los perfiles de los postulantes permitirá a las universidades peruanas optimizar sus procesos de admisión. Al identificar los factores más influyentes, las instituciones podrán ajustar sus criterios de evaluación para asegurar la selección de estudiantes con habilidades y aptitudes diversas, enriqueciendo así el ambiente académico.

Los resultados de esta investigación proporcionarán una base sólida para el desarrollo de políticas educativas informadas. Las autoridades educativas y las instituciones podrán utilizar estos hallazgos para diseñar estrategias que fomenten la inclusión y la diversidad, abordando de manera efectiva los desafíos identificados en el proceso de selección.

Este estudio aportará conocimientos significativos al campo de la educación superior en el Perú. La falta de investigaciones detalladas sobre los perfiles de los postulantes en contextos específicos puede limitar el diseño de intervenciones efectivas. Este trabajo académico busca llenar ese vacío y servir como referencia para futuras investigaciones y mejoras en el sistema educativo.

El estudio permitirá formar conglomerados de estudiantes de acuerdo con los resultados en el examen de admisión y también conglomerados con malos resultados; en base a estas se elabora los perfiles que servirá de base para la toma de decisiones.

1.4. OBJETIVOS DE LA INVESTIGACIÓN

1.4.1. OBJETIVO GENERAL

Analizar el perfil de los conglomerados de estudiantes postulantes a la UNSAAC obtenidos de las aplicaciones de los métodos de análisis de clúster y correspondencia.

1.4.2. OBJETIVOS ESPECÍFICOS

a) Comparar los métodos de Conglomerados en base a los índices de validación para determinar el perfil de los estudiantes postulantes A LA UNSAAC

b) Determinar las variables presentan mayor repercusión en la elaboración de conglomerados de estudiantes postulantes a la UNSAAAC

c) Describir el perfil presentan los estudiantes con mejores puntajes en el examen de admisión de la UNSAAC, obtenidos del análisis de correspondencia

CAPÍTULO II

MARCO TEÓRICO CONCEPTUAL

2.1. ANÁLISIS DE AGRUPAMIENTO O CLUSTERING

El Análisis de agrupamiento, conocido como Análisis de conglomerados, es una técnica estadística multivariada cuyo propósito es agrupar un conjunto de objetos, tratando de lograr la máxima homogeneidad en cada grupo y la mayor diferencia entre los grupos (Elguera, 2018).

El término "Análisis de Clúster" abarca una diversidad de métodos muy útiles que se emplean para generar clasificaciones significativas. Este enfoque polifacético ofrece una gama amplia de herramientas que pueden ser aprovechadas con flexibilidad, permitiendo la creación de agrupaciones discernibles. Al utilizar el Análisis de Clúster, se abre un abanico de posibilidades para estructurar y organizar datos de manera efectiva, facilitando la identificación de patrones y relaciones relevantes. Este recurso multifacético se erige como un aliado valioso en la tarea de dar sentido y orden a conjuntos de información diversos. Más concretamente, un método clúster es un procedimiento estadístico multivariante que comienza con un conjunto de datos conteniendo información sobre una muestra de entidades e intenta reorganizarlas en grupos relativamente homogéneos a los que llamaremos clústeres (Chavez L. , 2020).

En el fascinante mundo del Análisis de Clúster o conglomerado, se destaca la singularidad de contar con escasa o nula información previa sobre la estructura de las categorías. Este aspecto lo diferencia notablemente de los métodos multivariantes de asignación y discriminación, donde la comprensión previa es esencial. Aquí, lo único disponible es una colección de observaciones, siendo el desafío operacional descubrir

la intrincada estructura de las categorías a la que se ajustan estas observaciones. En esencia, el Análisis de Clúster se convierte en una herramienta intrigante para explorar y revelar la complejidad inherente en conjuntos de datos, sin tener prejuicios previos sobre la disposición de las categorías. Más concretamente, el objetivo es ordenar las observaciones en grupos tales que el grado de asociación natural es alto entre los miembros del mismo grupo y bajo entre miembros de grupos diferentes. (Cuadras, 2019)

Aunque la estructura de las categorías suele ser un misterio inicialmente, es común contar con algunas ideas sobre las características deseables e indeseables al definir un esquema de clasificación específico. A pesar de la incertidumbre inicial sobre la disposición de las categorías, a menudo se poseen ciertas percepciones acerca de los atributos que se consideran ideales o inaceptables al establecer un determinado sistema de clasificación. Este enfoque flexible permite adaptarse a la complejidad del análisis, permitiendo la incorporación de criterios claros y relevantes en el proceso de clasificación. En términos operacionales, el investigador es informado suficientemente sobre el problema, de tal forma que puede distinguir entre buenas y malas estructuras de categorías cuando se encuentra con ellas (Everitt & Torsten Hothorn, 2011).

Entonces, ¿por qué no enumerar todas las posibilidades y elegir la más atrayente?

El número de formas en las que se pueden clasificar m observaciones en k grupos es un número de Stirling de segunda especie (Everitt B. , 2011)

$$\mathbb{S}_m^{(k)} = \frac{1}{k!}\sum_{i=0}^{k}(-1)^{k-1}\binom{k}{i}i^m$$

El problema se complica aún más por el hecho de que usualmente el número de grupos es desconocido, por lo que el número de posibilidades es suma de números

de Stirling; así, por ejemplo, en el caso de m observaciones tendríamos que el número total de posibles clasificaciones sería (Everitt B. , 2011)

$$\sum_{j=1}^{m} \mathbb{S}_m^{(j)}$$

que es un número excesivamente grande, por lo que el número de posibles clasificaciones puede ser enorme.

2.2. CLASIFICACIÓN DE LAS TÉCNICAS DE CONGLOMERADOS.

La clasificación que se presentará se refiere a algunas de las diversas técnicas de agrupamiento disponibles. Como se evidenciará, esta clasificación es bastante extensa, dada la diversidad de métodos existentes. Es importante señalar que no todos los procedimientos mencionados serán abordados en detalle; nos concentraremos únicamente en los más comúnmente utilizados en aplicaciones prácticas. Estos son los métodos para los cuales se cuenta con un mayor nivel de experiencia y que suelen ser implementados en los paquetes estadísticos disponibles. Cabe destacar que el uso efectivo de cualquier técnica de agrupamiento requiere un hardware robusto y un software especializado, ya que el desarrollo práctico de dichas técnicas no es viable sin un ordenador potente y un programa informático adecuado. (Peña, 2002).

En líneas generales, podemos identificar dos amplias clasificaciones de métodos de clústeres: los enfoques jerárquicos y los enfoques no jerárquicos.

2.2.1 Los Métodos de agrupamiento

Dentro de la literatura, encontramos una diversidad de métodos y algoritmos diseñados para la agrupación de objetos. La elección de un algoritmo específico de agrupamiento suele depender de factores como la cantidad y naturaleza de los datos

disponibles, así como del propósito para el cual se va a aplicar. A continuación, presentamos una categorización de los algoritmos de agrupamiento disponibles.:

1. ***Métodos jerárquicos:***

   Se apoyan en un procedimiento secuencial para la creación de grupos o clústeres, estableciendo jerarquías entre los objetos o individuos. Estos métodos pueden dividirse en enfoques aglomerativos o divisivos. En el método aglomerativo, se comienza con un número de clústeres igual a la totalidad de los objetos, y estos se van uniendo de manera progresiva basándose en una métrica de distancias. Al concluir este proceso, se obtiene un único clúster que abarca todos los objetos. Las técnicas para la formación de los clústeres son: Enlace simple (vecino más cercano), Enlace completo (vecino más alejado), Enlace promedio, Enlace centroide, Enlace Ward. (Everitt B. , 2011)

2. ***Método de divisiones:***

   Su método secuencial para la creación de clústeres difiere del enfoque aglomerativo. Comienza con un único grupo o clúster que abarca todos los objetos en su totalidad. Posteriormente, se procede a dividir (en un sentido descendente) estos grupos en subgrupos, tomando en cuenta aquellos que están más distantes entre sí, hasta alcanzar "n" grupos, cada uno compuesto por un solo objeto (Areválo & Pérez Gonzales, 2018). Uno de los algoritmos más empleados en este contexto es el Algoritmo de Howard-Harris (Cuadras, 2019).

3. ***Métodos basados en particiones:***

   Este método de clúster, no jerárquico, opera con un número predeterminado de clústeres, denotado como "k", un valor conocido y parametrizado. Los "k" grupos se generan mediante un proceso iterativo que implica la formación de "k"

particiones. La metodología se centra en distribuir (mover) los objetos entre los "k" grupos, minimizando las distancias entre los objetos dentro de cada grupo en relación con su centroide (que puede ser la media, mediana, moda, medoides, entre otros). El proceso se inicia seleccionando de manera aleatoria "k" objetos que actuarán como centroides iniciales para cada clúster. En cada vuelta del proceso, se asigna el objeto al clúster más afín (el de menor distancia en relación con el centroide), recalculando así el nuevo centroide de cada clúster. El proceso culmina cuando todos los objetos han sido asignados, siguiendo un criterio de detención basado en la convergencia (Cuadras, 2019), agrupa los métodos no jerárquicos en cuatro familias:

a. ***Métodos de reasignación***

   Posibilitan que un objeto asignado a un grupo en una etapa específica del proceso pueda ser reasignada a otro grupo en una fase posterior, siempre y cuando esto optimice el criterio de selección. La conclusión del proceso ocurre cuando ya no hay objetos cuya reasignación pueda mejorar el resultado obtenido (Ng & Han, 2002). Estos métodos engloban:

   - El método K-Medias.
   - El Quick-Cluster análisis.
   - El método de Forgy.

b. ***Métodos de búsqueda de la densidad***

   Entre estas metodologías se incluyen aquellas que ofrecen una aproximación tipológica y una aproximación probabilística. En el primer enfoque, los grupos se crean identificando las áreas con una mayor concentración de individuos. Dentro de este contexto, sobresalen:

   - El análisis modal de Wishart.

- El método Taxmap.
- El método de Fortin.

En la segunda categoría, se parte de la premisa de que las variables siguen una distribución de probabilidad en la cual los parámetros experimentan variaciones entre distintos grupos. El objetivo es identificar individuos que pertenezcan a la misma distribución. Dentro de los métodos de este enfoque, se destaca la metodología de las combinaciones de Wolf (Ng & Han, 2002).

c. Métodos directos.

Facilitan la clasificación simultánea de tanto individuos como variables. Uno de los algoritmos más reconocidos en este conjunto es el Block-Clustering (Chavez L. , 2020).

d. Métodos de reducción de dimensiones.

Estas metodologías se centran en la exploración de factores en el espacio de los individuos, asignando a cada factor la representación de un grupo. Son reconocidos bajo la denominación de Análisis Factorial tipo Q (Chavez L. , 2020).

2.2.2. Etapas en análisis de clúster

2.2.2.1. Elección de las variables

La selección inicial del conjunto específico de características utilizadas para describir a cada individuo establece un marco de referencia crucial para la formación de agrupaciones o clústeres. Esta elección posiblemente refleje la perspectiva del investigador sobre el propósito de la clasificación. Por lo

tanto, la primera interrogante respecto a la elección de variables radica en su relevancia para el tipo de clasificación buscado. Es esencial tener en cuenta que la elección inicial de variables constituye, por sí misma, una categorización de los datos, para la cual solo existen directrices matemáticas y estadísticas limitadas.

Otro aspecto crucial a tener en cuenta es la determinación del número de variables a utilizar. En diversas situaciones, es probable que el investigador cometa el error de emplear un exceso de medidas, lo que podría generar diversos inconvenientes, ya sea a nivel computacional o porque estas variables adicionales puedan opacar la estructura de los grupos.

En numerosas aplicaciones, las variables que describen los objetos a clasificar pueden no medirse en las mismas unidades. De hecho, es posible que existan variables de tipos completamente diversos, algunas de naturaleza categórica, otras ordinales e incluso algunas que presenten una escala de tipo intervalo (Zuniga-Jara, et.al, 2022).

Es evidente que sería incorrecto tratar como equivalentes, en algún aspecto, medidas como el peso expresado en kilos, la altura en milímetros y la evaluación de la ansiedad en una escala de cuatro puntos. En el caso de variables cuantitativas, la solución común implica estandarizar las variables antes del análisis, calculando las desviaciones estandar a partir de todos los individuos. Sin embargo, algunos autores argumentan que esta técnica podría tener desventajas significativas al diluir las diferencias entre grupos en las variables más discriminatorias; como alternativa, sugieren

utilizar la desviación estándar entre grupos para la estandarización (Manrique, 2016).

Cuando las variables presentan tipos diversos, es común transformar todas las variables en binarias antes de calcular las similitudes. Esta estrategia brinda claridad, pero a costa de sacrificar información. Una opción más atractiva consiste en emplear un coeficiente de similitud que pueda incorporar de manera sensible la información de variables heterogéneas, como el propuesto por Gower en 1971, al cual nos referiremos más adelante (Elguera, 2018). Para variables mixtas, también existe la posibilidad de realizar análisis por separado e intentar sintetizar los resultados a partir de los diversos estudios (Elguera, 2018).

2.2.2.2. Elección de la medida de asociación.

La mayoría de las técnicas de agrupamiento demandan la definición de una medida de asociación que permita evaluar la cercanía entre los objetos en estudio. En el contexto de un Análisis de Clúster de individuos, dicha proximidad suele expresarse en términos de distancias, mientras que el Análisis de Clúster por variables generalmente implica el uso de medidas como coeficientes de correlación. Algunas de estas medidas poseen interpretaciones claras, mientras que otras resultan difíciles de describir debido a su naturaleza subjetiva (Everitt B. , 2011).

Subrayamos la clasificación de las medidas para variables e individuos, siendo algunas de ellas de uso general. Esta categorización se ha delineado, principalmente, con la consideración de llevar a cabo prácticas informáticas mediante el paquete estadístico BMDP, el cual cuenta con dos

capítulos específicos: uno destinado al Análisis de Clúster por variables y otro enfocado en individuos, es esencial tener presente la relevancia de los tipos de datos utilizados, ya sean categóricos o no (Tonconi, 2021).

2.2.2.3. Elección de la técnica clúster a emplear en el estudio.

En los últimos años, se ha observado un abanico considerable y diverso de métodos de agrupamiento, con concepciones que varían significativamente. En una primera instancia, se clasifican en jerárquicos y no jerárquicos. La distinción fundamental radica en que, en los métodos jerárquicos, las asignaciones de individuos a los clústeres creados permanecen inalterables durante todo el proceso, sin permitir reasignaciones posteriores a clústeres distintos. En contraste, los métodos no jerárquicos posibilitan estas reasignaciones. Además, en los métodos jerárquicos, el investigador debe extraer sus propias conclusiones, mientras que en los métodos no jerárquicos, el número final de clústeres suele estar predefinido, aunque dentro de este enfoque se han desarrollado técnicas que ofrecen cierta flexibilidad en la determinación final del número de clústeres, con el objetivo de evitar posibles alteraciones en los resultados definitivos (Areválo & Pérez Gonzales, 2018).

Entonces, en ciertos problemas prácticos, la elección del método a utilizar puede resultar bastante intuitiva, dependiendo principalmente de la naturaleza de los datos y de los objetivos finales. No obstante, en otros casos, la elección puede no ser tan evidente. Lo que siempre resulta recomendable en aplicaciones prácticas es no limitarse a un único procedimiento, sino explorar un amplio espectro de posibilidades y

comparar los resultados obtenidos con cada uno de ellos. De esta manera, si los resultados finales son coherentes entre sí, podemos extraer conclusiones más robustas sobre la estructura inherente de los datos. En caso contrario, la falta de consistencia en los resultados podría indicar que los datos con los que estamos trabajando no siguen una estructura claramente definida.

2.2.2.3. Validación de los resultados e interpretación de estos.

Esta parte marca el cierre de la secuencia lógica que caracteriza la ejecución de una investigación mediante un método de agrupamiento. Sin duda, es la fase más crucial, ya que es aquí donde se extraerán las conclusiones del estudio (Arbin, Suhailayani, Zafirah, & Othman, 2015).

Existen diversos métodos propuestos para validar un procedimiento de agrupamiento. Por ejemplo, cuando se trabaja con métodos jerárquicos, surgen dos interrogantes:

a) ¿Hasta qué punto la estructura final refleja las similitudes o diferencias entre los objetos en estudio?

b) ¿Cuál es la cantidad óptima de clústeres que mejor representa la estructura inherente de los datos?

La respuesta predominante ante la primera interrogante se basa en la utilización del coeficiente de correlación cofenético, propuesto por Sokal y Rohlf en 1962. Este coeficiente evalúa la correlación entre las distancias iniciales, derivadas de los datos originales, y las distancias finales que han surgido al unirse los individuos durante el desarrollo del método. Valores

elevados de este coeficiente indicarán que el proceso no ha provocado una alteración significativa en la estructura original de los datos.

En cuanto a los métodos no jerárquicos, las preguntas anteriores tienden a perder relevancia, centrándose en cambio en validar los resultados mediante el examen de la homogeneidad de los grupos formados durante el desarrollo del método. Algunos autores han propuesto la aplicación de técnicas multivariantes como el análisis multivariante de la varianza (Elguera, 2018). También se sugiere realizar múltiples análisis de la varianza (ANOVA) sobre cada variable en cada clúster, una opción incluida en BMDP.

Estos procedimientos plantean desafíos significativos y no deben considerarse como definitivos. Una técnica comúnmente empleada, conocida como remuestreo, implica tomar múltiples submuestras de la muestra original y realizar el análisis en cada una de ellas. Si al repetir el análisis se obtienen soluciones aproximadamente idénticas y similares a la obtenida con la muestra principal, se podría inferir que la solución encontrada puede ser válida, aunque este no sería un argumento suficiente para tomar una decisión definitiva (Areválo & Pérez Gonzales, 2018). Sin embargo, este método resulta más útil cuando se emplea de manera inversa; es decir, si las soluciones obtenidas en las diversas submuestras no muestran cierta similitud, entonces es evidente que se debe cuestionar la estructura obtenida con la totalidad de la muestra.

2.3. DISTANCIAS Y SIMILARIDADES

2.3.1. Distancias

Definición: Considerando un conjunto de elementos, U, finito o infinito. Una función $d: U\text{x}U \rightarrow \mathbb{R}$ se denomina una distancia métrica si, para cada par de elementos x e y pertenecientes a U, se cumple:

1. $d(x,y) \geq 0$
2. $d(x,y) = 0 \iff x = y$
3. $d(x,y) = d(y,x)$
4. $d(x,z) \leq d(x,y) + d(y,z)\ , \forall z \in U$

Ampliando la noción tradicional de distancia expuesta previamente, ciertos autores proponen definiciones de distancias métricas que admiten valores negativos. En consecuencia, una función de distancia métrica se configura como una función $d: U\text{x}U \rightarrow \mathbb{R}$ que satisface los siguientes axiomas.

1. $d(x,y) \geq d_0$
2. $d(x,y) = d_0 \iff x = y$
3. $d(x,y) = d(y,x)$
4. $d(x,z) \leq d(x,y) + d(y,z)\ , \forall z \in U$

donde d_0 puede ser inferior a cero. Esta definición se sustenta en la idea de que, al contar con una función distancia métrica d, es posible derivar otra, d', a partir de ella mediante la expresión $d'\,(x,y) = d(x,y) - d_0$. Se demuestra de manera sencilla que d' constituye una distancia métrica, conforme a la definición previamente expuesta.

2.3.2. Similaridades.

De manera análoga a las distancias, presentamos la siguiente definición de similitud:

Definición: Considerando un conjunto de elementos, U, ya sea finito o infinito, Una función $s : U \times U \rightarrow \mathbb{R}$ se llama similaridad si cumple las siguientes propiedades: $\forall x, y \in U$

1. $s(x, y) \leq s_0$

2. $s(x, x) = s_0$

3. $s(x, y) = s(y, x)$

donde s_0 es un número real finito arbitrario.

Definición: Una función s, que cumple con las condiciones previamente establecidas, recibe la denominación de similitud métrica si, adicionalmente, satisface:

1. $s(x, y) \leq s_0 \Longrightarrow x = y$

2. $|s(x, y) + s(y, z)|\, s(x, z) \geq s(x, y)s(y, z),\ \forall z \in U$

Observamos que la segunda parte de la definición previa se refiere al hecho de que solo dos elementos idénticos poseen la máxima similitud. En los siguientes segmentos, presentaremos algunas de las distancias y similitudes más comunes en la práctica. Consideraremos, en términos generales, m individuos sobre los cuales se han medido n variables $X_1, \ldots, X_n$. Esto nos proporciona $m\, x\, n$ datos que organizaremos en una matriz de dimensiones $m\, x\, n$.

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{pmatrix}$$

La fila i-ésima de la matriz X alberga los valores de cada variable correspondientes al i-ésimo individuo, mientras que la columna j-ésima exhibe los valores relacionados con la j-ésima variable a lo largo de todos los individuos de la muestra. Aunque distinguiremos entre medidas de asociación para individuos y variables, desde el punto de vista técnico, estas son aplicables tanto a individuos como a variables (simplemente considerando estas medidas en un espacio n-dimensional o m-dimensional, es decir, transponiendo la matriz) (Elguera, 2018).

2.4. MEDIDAS DE ASOCIACIÓN ENTRE VARIABLES.

Para combinar variables, es esencial contar con medidas numéricas que describan las relaciones entre ellas. La premisa fundamental de todas las técnicas de agrupamiento radica en que las medidas numéricas de asociación sean comparables; en otras palabras, si la medida de asociación entre un par de variables es 0,72 y la de otro par es 0,59, entonces el primer par presenta una asociación más sólida que el segundo. Naturalmente, cada medida refleja una asociación en un contexto específico, por lo que es crucial seleccionar una medida adecuada para el problema particular en cuestión (Elguera, 2018).

1. Coseno del ángulo de vectores

Se tiene dos variables X_i y X_j, obtenidas de m individuos, y sean x_i y x_j los conjuntos de vectores en los que las k-ésimas componentes señalan el valor de la variable correspondiente en el k-ésimo individuo:

$$x_i = (x_{1i}, \dots, x_{mi})' \quad ; \quad x_j = \left(x_{1j}, \dots, x_{mj}\right)'$$

Como es conocido, el producto escalar de dos vectores es:

$$x_i' x_j = \sum_{l=1}^{m} x_{li} x_{lj}$$

Lo que en el ámbito estadístico se denomina como la suma de productos cruzados entre x_i y x_j, mientras que el producto escalar de un vector consigo mismo, la norma al cuadrado del vector, se identifica como la suma de cuadrados. De esta manera, se expresa:

$$x_i' x_j = \|x_i\| \left\|x_j\right\| \cos(\beta)$$

donde β representa al ángulo entre los vectores x_i y x_j.

Al analizar la figura, la distancia desde el origen (O) hasta B equivale a $\|x_i\|$ cos(β), siendo este valor la proyección ortogonal de x_i sobre x_j. De esta

manera, el producto escalar puede entenderse como el resultado del producto entre la longitud del vector x_j y la longitud de la proyección de x_i sobre x_j.

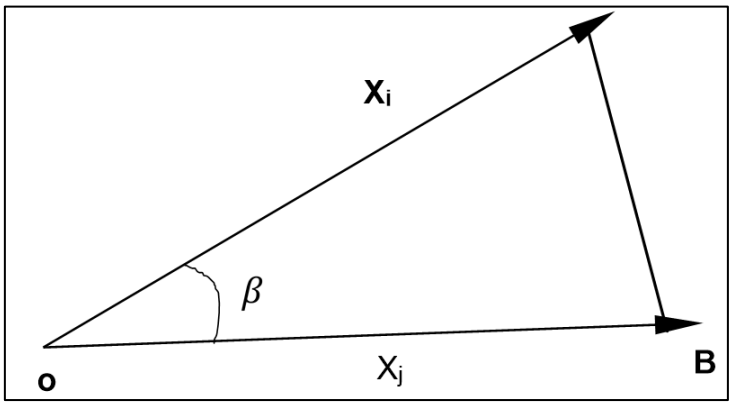


Figura 1. Ángulo entre vectores

A partir de la ecuación anterior se tiene

$$\cos(\beta) = \frac{x_i' x_j}{\|x_i\|\|x_j\|} = \frac{\sum_{l=1}^{m} x_{li}\, x_{lj}}{\left(\sum_{l=1}^{m} x_{li}^2 \sum_{l=1}^{m} x_{lj}^2\right)^{\frac{1}{2}}}$$

El coseno del ángulo se presenta como una métrica de similitud entre x_i y x_j, adquiriendo valores en el rango de -1 a 1, de acuerdo con la desigualdad de Schwarz. No solo eso, sino que también es la medida más idónea para evaluar si son paralelos dos vectores, ya que dos vectores se consideran paralelos cuando el coseno del ángulo entre ellos es uno en términos absolutos. Esta medida resulta independiente, salvo el signo, de la longitud de los vectores en consideración. Desde una perspectiva algebraica, consideremos dos escalares, b y c, se define:

$$\hat{x}_i = bx_i \;;\; \hat{x}_j = cx_j \;;\; b, c \neq 0$$

entonces:

$$\cos(\hat{x}_i\,, \hat{x}_j) = \frac{\hat{x}_i' \hat{x}_j}{\|\hat{x}_i\|\|\hat{x}_j\|} = \frac{\sum_{l=1}^{m} bx_{li}\, cx_{lj}}{\left(\sum_{l=1}^{m} b^2 x_{li}^2 \;\sum_{l=1}^{m} c^2 x_{lj}^2\right)^{\frac{1}{2}}}$$

$$= \frac{bc \sum_{l=1}^{m} x_{li}\, x_{lj}}{|bc| \left(\sum_{l=1}^{m} x_{li}^2 \ \sum_{l=1}^{m} x_{lj}^2\right)^{\frac{1}{2}}} Sgn(bc) \cos\left(x_i, x_j\right)$$

Por tanto, el coseno entre x_i y x_j es invariante ante homotecias, excepto un eventual cambio de signo.

2.5. MEDIDAS DE ASOCIACIÓN ENTRE INDIVIDUOS

Las métricas empleadas para evaluar la similitud son reconocidas como distancias y se utilizan de manera intercambiable con el término métrica. Una distancia más elevada señala que los objetos están más distantes entre sí. Usualmente, antes de aplicar estas medidas de distancia, se lleva a cabo un proceso de estandarización o tipificación de los datos con antelación, con el propósito de eliminar la influencia de las unidades de medida de las variables en el análisis (Luy, Salvatierra, Rengifo, & Rivera, 2021). Sean $x_i = \left(x_{i1}, x_{i2}, \dots, x_{ip}\right)$ y $x_j = \left(x_{j1}, x_{j2}, \dots, x_{jp}\right)$

dos vectores de observaciones de p-dimensional de variables, entonces algunas medidas de similitud son las siguientes:

Para variables cuantitativas:

1. Distancia euclidiana: Es la más conocida y utilizada. Mide la distancia geométrica entre los vectores. Se define como:

$$d\left(x_i, x_j\right) = \sqrt{\sum_{l=1}^{p} \left(x_{il} - x_{jl}\right)^2}$$

2. Distancia Manhattan: Es la sumatoria en valor absoluto de las distancias entre las componentes de los vectores. Se define como:

$$d(x_i, x_j) = \sum |x_{ik} - x_{jk}|$$

3. Distancia de Minkowski: Se generaliza las dos anteriores, las cuales se obtienen, respectivamente, haciendo $m = 2$ y $m = 1$. Se puede utilizar para todo valor real $m \geq 1$. Se define como:

$$d(x_i, x_j) = \left[\sum_{k=1}^{p} (x_{ik} - x_{jk})^m\right]^{\frac{1}{m}}$$

4. Distancia Chevychev: Es la sumatoria en valor absoluto de las distancias entre las componentes de los vectores. Se define como:

$$d(x_i, x_j) = \lim_{k \to \infty} \left(\sum |x_{ik} - x_{jk}|^k\right)^{1/k}$$

5. Distancia de Mahalanobis: Es la sumatoria en valor absoluto de las distancias entre las componentes de los vectores ponderando por la matriz de covariancias. Se define como:

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{p} (\bar{x}_i - \bar{x}_j)' S^{-1} (\bar{x}_i - \bar{x}_j)}$$

2.6. MÉTODOS JERÁRQUICOS DE ANÁLISIS CLÚSTER

Los métodos conocidos como jerárquicos tienen la finalidad de agrupar clusters para crear uno nuevo o separar alguno ya existente para generar dos nuevos, de manera que, al llevar a cabo repetidamente este proceso de aglomeración o división, se minimice alguna distancia o se maximice alguna medida de similitud. Los métodos jerárquicos se dividen en dos categorías: aglomerativos

y disociativos. Cada una de estas categorías engloba una amplia variedad de variantes.

1. Los métodos aglomerativos, también conocidos como ascendentes, comienzan el análisis con tantos grupos como individuos haya. A partir de estas unidades iniciales se van formando grupos, de forma ascendente, hasta que al final del proceso todos los casos tratados están englobados en un mismo conglomerado.

2. Los métodos disociativos, también llamados descendentes, constituyen el proceso inverso al anterior. Comienzan con un conglomerado que engloba a todos los casos tratados y, a partir de este grupo inicial, a través de sucesivas divisiones, se van formando grupos cada vez más pequeños. Al final del proceso se tienen tantas agrupaciones como casos han sido tratados.

Para aclarar conceptos, detengámonos un momento en los métodos aglomerativos. Supongamos que tenemos el conjunto de individuos de la muestra, dando lugar al nivel $K = 0$, con n grupos. En el siguiente nivel, se agruparán los dos individuos con mayor similitud (o menor distancia), resultando así en n - 1 grupos. Luego, siguiendo la misma estrategia, en el nivel subsiguiente se agruparán los dos individuos (o clusters ya formados) con la menor distancia o mayor similitud, y así sucesivamente. De esta manera, en el nivel L tendremos $n - L$ grupos formados. Si continuamos con este proceso, llegaremos al nivel $L = n - 1$, donde solo habrá un grupo compuesto por todos los individuos de la muestra (Areválo & Pérez Gonzales, 2018).

La particularidad de este enfoque para la creación de nuevos grupos radica en que, una vez que dos clústeres se agrupan en un nivel determinado, permanecen jerárquicamente unidos para los niveles subsiguientes. Los métodos jerárquicos

posibilitan la construcción de un árbol de clasificación conocido como dendrograma, que visualiza de manera gráfica el proceso de unión, indicando qué grupos se fusionan, en qué nivel específico ocurre y el valor de la medida de asociación entre los grupos cuando se agrupan (denominado nivel de fusión). En resumen, la operativa general de estos métodos es bastante sencilla. En los métodos aglomerativos, se comienza con tantos grupos como individuos existan. Luego, se elige una medida de similitud y se agrupan los dos grupos o clústeres con mayor similitud. Así se continúa hasta que:

1. Se forma un solo grupo.
2. Se alcanza el número de grupos prefijado
3. Mediante una contrastación de significancia, se evidencia que existen fundamentos estadísticos para no proseguir con la agrupación de clústeres, ya que aquellos más similares no muestran la homogeneidad suficiente para establecer una misma categorización.

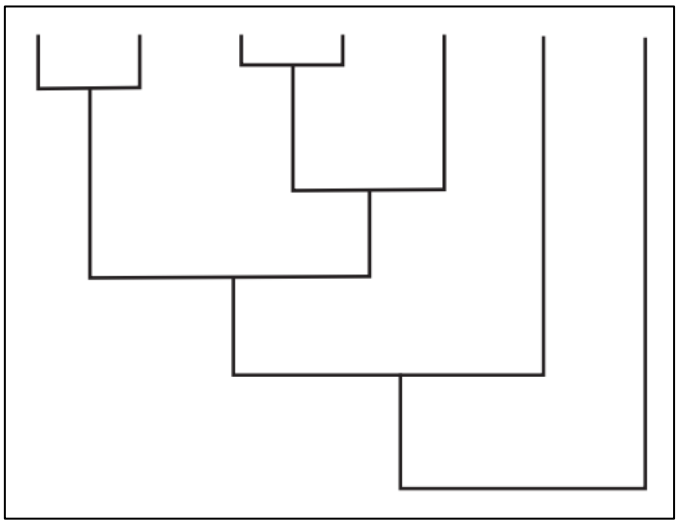


Figura 2. Dendrograma

2.7. MÉTODOS JERÁRQUICOS AGLOMERATIVOS

A continuación, vamos a exponer diversas estrategias que pueden emplearse al unir clústeres en las diferentes etapas o niveles de un procedimiento jerárquico. Ninguno de estos métodos garantiza una solución óptima para todos los posibles problemas, ya que los resultados pueden variar según la elección del método. El buen juicio del investigador, la comprensión del problema en cuestión y la experiencia orientarán hacia la elección del método más apropiado. En cualquier caso, es recomendable utilizar varios procedimientos con el objetivo de contrastar los resultados y extraer conclusiones, ya sea que los resultados coincidan entre métodos diferentes o no.

2.7.1. Estrategia de la distancia mínima o similitud máxima.

Esta técnica se conoce en la literatura anglosajona como "amalgamación simple" (single linkage). En este enfoque, la distancia o similitud entre dos clústeres se determina por la distancia mínima (o similitud máxima) entre sus elementos constituyentes. Así, si tras efectuar la etapa K-ésima, tenemos ya formados $n - K$ clusters, la distancia entre los clusters C_i (con n_i elementos) y C_j (con n_j elementos) seria:

$$d(C_i, C_j) = \underset{\substack{x_l \in C_i \\ x_m \in C_j}}{\text{Min}} \{d(x_l, x_m)\} \qquad l = 1, \dots, n_i \; ; m = 1, \dots, n_j$$

mientras que la similitud, si estuviéramos empleando una medida de tal tipo, entre los dos clusters sería:

$$d(C_i, C_j) = \underset{\substack{x_l \in C_i \\ x_m \in C_j}}{\text{Min}} \{s(x_l, x_m)\} \qquad l = 1, \dots, n_i \; ; m = 1, \dots, n_j$$

Con ello, la estrategia seguida en el nivel $K + 1$ será:

1. En el caso de emplear distancias, se unirán los clústeres C_i y C_j si

$$d(C_i, C_j) = \min_{\substack{i_1, j_i = 1, \dots, n-k \\ i_1 \neq j_i}} \{d(C_{i_1}, C_{j_1})\} =$$

$$= \min_{\substack{i_1, j_i = 1, \dots, n-k \\ i_1 \neq j_i}} \left\{ \min_{\substack{x_l \in C_{i_1} \\ x_m \in C_{j_1}}} \{d(x_l, x_m)\} \right\} \quad l = 1, \dots, n_{i_1};\ \ m = 1, \dots, n_{j_1}$$

En el caso de emplear similitudes, se unirán los clústeres C_i y C_j si

$$s(C_i, C_j) = \max_{\substack{i_1, j_i = 1, \dots, n-k \\ i_1 \neq j_i}} \{s(C_{i_1}, C_{j_1})\} =$$

$$= \max_{\substack{i_1, j_i = 1, \dots, n-k \\ i_1 \neq j_i}} \left\{ \max_{\substack{x_l \in C_{i_1} \\ x_m \in C_{j_1}}} \{s(x_l, x_m)\} \right\} \quad l = 1, \dots, n_{i_1};\ \ m = 1, \dots, n_{j_1}$$

donde, como es natural, se ha seguido la norma general de maximizar las similitudes o bien minimizar las distancias.

2.7.2. Estrategia de la distancia máxima o similitud mínima.

En este enfoque, también denominado el método de amalgamación completa (complete linkage), se evalúa la distancia o similitud entre dos clústeres teniendo en cuenta sus elementos más diferentes. En otras palabras, la distancia o similitud entre clústeres se determina por la distancia máxima (o similitud mínima) entre sus componentes. Así pues, al igual que en la estrategia anterior, si estamos ya en la etapa K-ésima, y por lo tanto hay ya formados n − K clusters, la distancia y similitud entre los clusters C_i y C_j (con n_i y n_j elementos respectivamente), serán:

$$d(C_i, C_j) = \max_{\substack{x_l \in C_{i_1} \\ x_m \in C_{j_1}}} \{d(x_l, x_m)\} \qquad l = 1, \dots, n_i\ ;\ \ m = 1, \dots, n_j$$

$$s(C_i, C_j) = \min_{\substack{x_l \in C_{i_1} \\ x_m \in C_{j_1}}} \{s(x_l, x_m)\} \qquad l = 1, \dots, n_i \; ; \; m = 1, \dots, n_j$$

Y con ello, la estrategia seguida en el siguiente nivel, $k + 1$, será:

1. En el caso de emplear distancias, se unirán los clusters C_i y C_j si

$$d(C_i, C_j) = \min_{\substack{i_1, j_i = 1, \dots, n-k \\ i_1 \neq j_i}} \{d(C_{i_1}, C_{j_1})\} =$$

$$= \min_{\substack{i_1, j_i = 1, \dots, n-k \\ i_1 \neq j_i}} \left\{ \max_{\substack{x_l \in C_{i_1} \\ x_m \in C_{j_1}}} \{d(x_l, x_m)\} \right\} \quad l = 1, \dots, n_{i_1}; \; m = 1, \dots, n_{j_1}$$

2. En el caso de emplear similitudes, se unirán los clusters C_i y C_j si

$$s(C_i, C_j) = \max_{\substack{i_1, j_i = 1, \dots, n-k \\ i_1 \neq j_i}} \{s(C_{i_1}, C_{j_1})\} =$$

$$= \max_{\substack{i_1, j_i = 1, \dots, n-k \\ i_1 \neq j_i}} \left\{ \min_{\substack{x_l \in C_{i_1} \\ x_m \in C_{j_1}}} \{s(x_l, x_m)\} \right\} \quad l = 1, \dots, n_{i_1}; \; m = 1, \dots, n_{j_1}$$

2.8. ALGORITMO DE PARTICIÓN ALREDEDOR DE MEDOIDES (PAM)

Para implementar el método de k-medoids, se emplea comúnmente el algoritmo conocido como Partición Alrededor de Medoides (PAM). Este algoritmo utiliza una estrategia de búsqueda codiciosa que, aunque no garantiza la solución óptima, es más rápida que una búsqueda exhaustiva. Para identificar k agrupamientos, el modelo PAM asigna un objeto representativo, llamado medoide, a cada clúster (Arora, Deepali, & Varshney, 2016). El medoide es el objeto que se encuentra más centrado dentro del clúster. Una vez seleccionados los medoides, cada objeto no elegido se agrupa

con el medoide al que es más similar. La búsqueda de los k medoides inicia con una selección arbitraria de k objetos (Elguera, 2018). En cada iteración, se realiza un intercambio entre un objeto seleccionado Oi y un objeto no seleccionado Oj solo si dicho intercambio mejora la calidad del agrupamiento (Arbin, etal, 2015). El algoritmo puede ser descrito:

1. Inicialización. Seleccionar k objetos de los n puntos como el medoide inicial.

2. Sea O_j un objeto no seleccionado y O_i es un medoide (objeto seleccionado). Se calcula la medida de disimilaridad o distancia $d_{ij} = d(O_j, O_i)$

3. Se indica que O_j pertenece al clúster representado por O_i, si $d_{ij} = d(O_j, O_i) = Min\_Medoids(O_j, O_i)$ es valor el mínimo sobre todos los medoids Oe

4. Se calcula la ganancia total obtenida seleccionando el objeto j: $GT_j = \sum_i d_{ji}$

5. Si el costo total de la configuración aumentó, deshacer el intercambio e ir al paso 2). De lo contrario determinar los nuevos medoides, ir al paso 2).

6. Se termina cuando no hay más objetos que agrupar.

Ventajas y desventajas del algoritmo PAM

Las principales ventajas que se puede indicar con el método PAM son las siguientes:

- El algoritmo PAM, es una alternativa robusta frente a los k-medias para dividir un conjunto de datos en grupos. (Kassambara, 2017)

- Los clústeres formados por el algoritmo PAM tienen por medoide a una observación que tiene la menor distancia con respecto a las demás observaciones de su clúster y la mayor distancia posible con otros medoides, asegurando grupos más homogéneos.

Las principales desventajas con el método PAM son las siguientes:

6. PAM exige que el usuario esté familiarizado con los datos y determine el número adecuado de clústeres a generar.
7. En situaciones donde se manejan conjuntos de datos extensos, el algoritmo PAM podría requerir un uso excesivo de memoria o un tiempo de cálculo prolongado en la computadora. En tales casos, la función CLARA se presenta como una alternativa más eficiente.

2.9. ALGORITMO DE CLUSTERING LARGE APPLICATIONS (CLARA)

Es una técnica de segmentación que se explora en los métodos de partición en el presente trabajo de investigación, este se centra en la identificación de medoides para el conjunto completo de datos. CLARA opera considerando una muestra reducida de datos con un tamaño predefinido (tamaño de muestra) y aplica el algoritmo K-medoides para generar un conjunto óptimo de medoides para esa muestra. La evaluación de la calidad de los medoides obtenidos se realiza mediante la disimilitud promedio entre cada objeto en el conjunto de datos general y el medoide de su respectivo grupo, definido como la función de costo. La agrupación final se determina por el conjunto de medoides que minimiza el costo (Everitt & Torsten Hothorn, 2011).

Este método fusiona la noción de K-medoides con el remuestreo para adaptarse a conjuntos de datos extensos. En lugar de intentar identificar los medoides utilizando todos los datos de manera simultánea, CLARA opta por

seleccionar una muestra aleatoria de un tamaño específico y aplicar el algoritmo K-medoides para determinar las agrupaciones óptimas según esa muestra. Al emplear estos medoides, se procede a agrupar las observaciones de todo el conjunto de datos. La eficacia de los medoides obtenidos se mide mediante la suma total de las distancias entre cada observación del conjunto de datos y su respectivo medoide (suma total de distancias intra-cluster) (Ng & Han, 2002).

CLARA realiza iteraciones de este procedimiento un número predefinido de veces para disminuir cualquier sesgo del muestreo. Los clústeres finales se eligen entre aquellos generados por los medoides que han logrado la menor suma total de distancias. A continuación, se detallan los pasos del algoritmo CLARA (Elguera, 2018).

Algoritmo

1. Se realiza una partición aleatoria del conjunto de datos en n partes de tamaño uniforme, siendo n una elección del analista.
2. Para cada una de las n partes:

   a) Emplear el algoritmo K-medoides y determinar quiénes son los K-medoides.

   b) Utilizar los medoides obtenidos en el paso anterior para agrupar todas las observaciones del conjunto de datos.

   c) Calcular la suma global de distancias entre cada observación del conjunto de datos y su respectivo medoide (suma total de distancias intra-clusters).

3. Seleccionar como segmentación final aquel que ha conseguido la menor suma total de distancias intra-clusters en el paso 2.c.

Usos y beneficios

- Es adaptable a extensas cantidades de datos.
- Aplica procedimientos de muestreo y agrupamiento un número definido de veces para reducir al mínimo el sesgo de muestreo.

Restricciones

- Si el centroide de la muestra no está entre los mejores k-centroides, el algoritmo no logrará identificar la mejor agrupación, lo que resulta en una pérdida de eficiencia.
- Cuando el tamaño de la muestra es insuficiente, la efectividad del algoritmo disminuye; sin embargo, si el tamaño de la muestra es excesivo, el rendimiento del algoritmo se ve afectado negativamente.

Mejoras

Ng y Han (2002) presentaron sus hallazgos acerca del algoritmo CLARANS (Clustering Large Applications based on Randomized Search), que fusiona técnicas de muestreo con K-medoides. El proceso de agrupamiento se visualiza como una exploración en un gráfico, donde cada nodo representa una solución potencial, es decir, un conjunto de K-medoides. Al reemplazar un medoide, se obtiene un nuevo agrupamiento llamado vecino del agrupamiento actual. CLARANS elige un nodo y lo compara con un número predeterminado de vecinos, buscando un mínimo local. Si se identifica un vecino mejor (es decir, con un error cuadrático inferior), CLARANS se traslada al nodo del vecino y reinicia el proceso; de lo contrario, el agrupamiento actual se considera un óptimo local. Si se alcanza un óptimo local, CLARANS comienza con un nuevo nodo seleccionado aleatoriamente para buscar un nuevo óptimo local (Pastrán & Gongora, 2021).

2.10. CLÚSTER BIETAPICO

La esencia fundamental de esta técnica radica en lograr un agrupamiento eficiente en conjuntos que no se conocen de antemano, pero que son sugeridos por la propia naturaleza de los datos. En contraste con otros métodos de agrupamiento, el cluster bietápico se distingue por su método automático de selección del número óptimo de conjuntos, la capacidad de crear modelos de conjuntos que incluyen variables tanto continuas como categóricas, y su capacidad para analizar conjuntos de datos extensos (Elguera, 2018).

La génesis del método "TwoStep" se encuentra en el algoritmo BIRCH (Zhang et al., 1996), diseñado para realizar análisis de clústeres. Este método se compone de dos fases: en primer lugar, se lleva a cabo un proceso de preclusterización en el conjunto completo de registros, agrupándolos en numerosos subclústeres más pequeños; a continuación, estos subclústeres se agrupan mediante un algoritmo de clústering jerárquico hasta alcanzar el número deseado de clústeres. Siguiendo esta metodología, dado que el número de elementos a procesar es considerablemente menor que el número total original de registros y requiere un análisis para todos ellos, este algoritmo se destaca por su eficiencia operativa en términos de costos. El método Two-Step proporciona dos tipos de medidas según el tipo de variables presentes en la matriz, las cuales son:

- Distancia Euclídea
- Distancia Máxima Verosimilitud

Además, maneja dos criterios de agrupamiento, el AKAIKE y el SCHWARTZ:

AKAIKE: $AIC = -2 * \ln L(\theta) + 2k$

SCHWARTZ: $BIC = -2 * \ln L(\theta) + (\ln(n) * k)$

El modelo que presenta el valor más bajo de BIC (criterio de información bayesiano) se considera el más óptimo para explicar los datos del análisis, logrando esta explicación con el menor número posible de parámetros.

2.10. ANTECEDENTES DE LA INVESTIGACIÓN

2.10.1. ANTECEDENTES INTERNACIONALES

Los autores Huapaya, Lizarralde, & Arona, (2011) en una conferencia expusieron la siguiente investigación "Propuesta para construir perfiles cognitivos en la evaluación del estudiante", esta línea de investigación estudia métodos para analizar el estado del conocimiento de estudiantes universitarios y desarrolla sistemas computacionales para implementar y probar los resultados alcanzados. Ante la incertidumbre inherente en la evaluación del estado del conocimiento, empleamos la lógica difusa para proponer modelos de perfiles cognitivos. Se examina el nivel de conocimiento a través de tres perfiles: individual, colectivo y colaborativo (Huapaya etal, 2011). El perfil individual se desarrollará utilizando lógica difusa; el perfil colectivo mediante agrupamiento difuso, y para el perfil colaborativo se aplicará el cuestionario de Cantwell y Andrews (Actitudes hacia el Trabajo en Grupo). Los tres perfiles conformarán el nivel del conocimiento medido cuantitativa y cualitativamente de cada estudiante en un momento dado.

(Zuniga-Jara, Zuniga-Soria, & Soria-Barreto, 2022) publicaron su investigación que intitula "Taxonomía de las carreras de medicina en Chile" en este estudio se realiza una taxonomía de las carreras de medicina en Chile, usando un enfoque basado en dendrogramas y componentes principales para visualizar

las observaciones en un espacio de dos dimensiones, con el objeto de identificar agrupamientos uniformes o atípicos. Se examinaron 11 variables, incluyendo 5 relacionadas con los estudiantes y 6 asociadas a la institución o universidad. Como resultado, se identificaron dos dimensiones de clasificación: 1) un indicador de la calidad de las instituciones y de sus estudiantes, y 2) un indicador del costo anual y del perfil socioeconómico de los estudiantes. En un nivel inicial de disimilitud, se observan dos grandes grupos de carreras de medicina: 1) con un perfil tradicional y regional, con mejores indicadores promedio de calidad institucional, y 2) asociado a instituciones privadas más recientes, con sede central en Santiago de Chile y con estudiantes de un nivel socioeconómico más elevado. Se concluye que es posible caracterizar las carreras de medicina en Chile mediante solo dos dimensiones.

En (Sankar, 2011), se aplica un algoritmo de agrupación a la información demográfica para identificar la agrupación de clientes. En la fase inicial, se procede a depurar y analizar los datos del cliente, identificando patrones mediante diversos parámetros. Posteriormente, en la segunda fase, se lleva a cabo un perfilado de los datos, utilizando técnicas de agrupamiento para distinguir a los clientes con riesgo bajo y alto. Los resultados experimentales evidencian que el enfoque propuesto genera patrones más útiles a partir de conjuntos de datos extensos. Se logra identificar tres segmentos de clientes: aquellos de alto beneficio, alto valor y bajo riesgo, mediante una de las técnicas de agrupamiento, IBM I-Miner. Se destaca el clúster de bajo riesgo y alto valor, que representa el 10-20 por ciento de los clientes y genera el 80% de los ingresos.

En (Arora, Deepali, & Varshney, 2016), se aplica los dos algoritmos de agrupamiento basados en particiones que son los más populares K-Means y K-Medoids se evalúan en el conjunto de datos transaccional. Los resultados de la comparación revelan que K-Medoids supera a K-Means en el tiempo dedicado a la selección de valores iniciales y en la complejidad espacial de la superposición de clústeres. Además, K-Medoids destaca en cuanto al tiempo de ejecución, mostrando una menor sensibilidad a los valores atípicos y una capacidad para reducir el ruido en comparación con K-Means, gracias a su enfoque de minimizar la suma de las diferencias entre los objetos de datos.

En (Arbin, Suhailayani, Zafirah, & Othman, 2015) se realizó el análisis K-Means y KMedoids Con varios conjuntos de datos. Se llevaron a cabo análisis con diversos parámetros y atributos de los datos, comparando exhaustivamente ambos algoritmos para identificar sus respectivas fortalezas y debilidades. Se realizaron estudios detallados para evaluar la correlación de los datos con los algoritmos y determinar la relación entre ellos. Ambos enfoques implementados arrojaron resultados satisfactorios, con un error cuadrático medio inferior al 3%. No obstante, en la mayoría de los conjuntos de datos, se observó que K-Medoids destacó como la opción más eficaz para la agrupación de datos.

2.10.2. ANTECEDENTES NACIONALES

(Chavez L. , 2020) realizo un trabajo que intitula “Caracterización del perfil del ingresante de una universidad pública aplicando algoritmos clustering k-prototypes y k-medoids”, en el presente trabajo de investigación se realizó un estudio comparativo de algoritmos no supervisados para la caracterización del perfil del ingresante de una universidad pública respecto a sus variables

sociodemográficas, económicas y de rendimiento académico utilizando algoritmos de segmentación K-prototypes y K-medoids, con el fin de generar conocimientos valiosos y útiles para lograr una mejor comprensión de la diversidad de universitarios que ingresan y con ello conocer el tipo de estudiante que la institución forma. Se llevó a cabo el procesamiento previo de los datos y la implementación de algoritmos de agrupamiento, considerando tanto variables cuantitativas como cualitativas. Se determinó el número óptimo de conglomerados y el algoritmo más apropiado mediante índices de validación interna. La validación de los clústeres obtenidos se realizó de manera univariada (mediante análisis de varianza o ANOVA y prueba Chi cuadrado) y multivariada (con algoritmo Boruta y árbol C5.0). Por último, se identificaron las variables más relevantes para caracterizar el perfil de los ingresantes. La investigación reveló la existencia de tres tipos de alumnos: Ingresante previsto, Ingresante en proceso y el Ingresante en inicio. Cada categoría presenta peculiaridades que proporcionarán a los responsables de las políticas educativas y, especialmente, a los profesores consejeros, información valiosa sobre el tipo de alumno desde su ingreso a la universidad. Este conocimiento facilitará la implementación de políticas educativas, como el fomento del acompañamiento especializado, sistemático e integral, alineándose con el paradigma de aprendizaje propuesto en el Modelo Educativo de la universidad.

(Tonconi, 2021) realizo una investigación que intitula “Identificación de perfiles de los centros de educación técnico-productiva públicos usando indicadores de condiciones básicas de calidad mediante clúster bietápico” donde el obsjetivo del estudio fue de: Se buscó identificar grupos entre los Centros de Educación Técnico-Productiva (CETPRO) públicos utilizando la técnica de clúster

bietápico, con el objetivo de segmentar y perfilizar estos centros según sus indicadores de condiciones básicas de calidad. La investigación se basó en datos recopilados en 2019 de todos los CETPRO públicos a nivel nacional, inicialmente contando con información de 704 instituciones educativas. Tras realizar un análisis exploratorio y limpieza de datos, que incluyó la eliminación de valores atípicos y datos faltantes, se trabajó con un conjunto de datos de 684 instituciones. Se empleó el análisis de clúster bietápico, una técnica que permite trabajar con variables tanto cuantitativas como categóricas. Los resultados de la aplicación de esta técnica revelaron la existencia de 2 conglomerados: el primero comprendía 324 CETPRO (47.4%), mientras que el segundo conglomerado estaba conformado por 360 (52.6%). Posteriormente, se procedió a describir los perfiles de cada conglomerado, identificando las características clave en relación con las variables asociadas a las 5 condiciones básicas de calidad establecidas.

(Luy, Salvatierra, Rengifo, & Rivera, 2021) realizaron un artículo de investigación que titula “El clúster jerárquico en la segmentación de los logros del aprendizaje de matemática y comunicación” La presente investigación se fundamenta en el análisis y el comportamiento del componente: Productos e Impactos de la educación peruana a partir de los datos alojados en el organismo educativo, la detección de los conglomerados fue gracias a la teoría de Clúster jerárquico quien tiene por finalidad identificar al porcentaje de los estudiantes de los departamentos que conforman los clústers. Dadas las características y el análisis realizado, la investigación se sitúa en el contexto del enfoque cuantitativo, adoptando un enfoque descriptivo exploratorio. La técnica de recopilación de datos se llevó a cabo mediante el análisis

documental, utilizando los procesos de construcción de árboles de clasificación a través del dendrograma generado por el software estadístico SPSS. Este proceso permitió la identificación de grupos conformados por departamentos, considerando el porcentaje de estudiantes que alcanzan competencias en áreas específicas como comunicación y matemáticas (Luy, Salvatierra, Rengifo, & Rivera, 2021). De acuerdo con los resultados del clúster, los porcentajes de los estudiantes que presentan dificultad de lograr los objetivos de aprendizaje esperados se ubican los departamentos de Huancavelica, Huánuco, Madre de Dios, Apurímac y San Martín, mientras que los estudiantes con mayor porcentaje de logros.

(Manrique, 2016) realizo el trabajo de investigación “Segmentación de clientes de Corporación Lindley de la región Lima mediante el análisis Cluster Bietápico en octubre de 2016”, el objetivo de este trabajo de investigación es describir el método del Análisis Clúster Bietápico y su aplicación para segmentar la cartera de clientes de Corporación Lindley de la Región Lima en octubre de 2016. La técnica de Análisis Clúster Bietápico se presenta como una herramienta de exploración diseñada para revelar las agrupaciones naturales, o conglomerados, presentes en un conjunto de datos. Este método combina enfoques jerárquicos y no jerárquicos (de partición), permitiendo el análisis simultáneo de variables de distintos tipos, ya sean categóricas o continuas. Además, su versatilidad lo hace apto para abordar conjuntos de datos extensos. El análisis de conglomerados Bietápico encuentra aplicaciones en diversas áreas, destacando su relevancia en la Inteligencia Comercial, donde la segmentación de clientes ha sido motivo de discusión y estudio (Manrique, 2016).

CAPÍTULO III

HIPÓTESIS Y VARIABLES

3.1. HIPOTESIS

3.1.1. HIPÓTESIS GENERAL.

Existen cuatro perfiles o conglomerados de estudiantes postulantes a la UNSAAC obtenidos de las aplicaciones de los métodos de análisis de clúster y correspondencia.

3.1.2. HIPÓTESIS ESPECÍFICAS

a) El método partition around medoids (PAM) presenta mejores índices de validación para determinar el perfil de los estudiantes postulantes A LA UNSAAC
b) Las variables procedencia y condiciones económicas presentan mayor repercusión en la elaboración de conglomerados de estudiantes postulantes a la UNSAAAC
c) Los estudiantes con mejores puntajes en el examen de admisión de la UNSAAC presentan edades inferiores al promedio, condiciones económicas bajas y proceden de colegios estatales.

3.2. IDENTIFICACIÓN DE VARIABLES E INDICADORES

Variable: Perfiles de los estudiantes postulantes a la universidad

3.3. OPERACIONALIZACION DE VARIABLES

Tabla 1. Operacionalización de variable

Variable	**Definición conceptual**	**Definición Operacional**	**Indicadores**
Perfiles de los estudiantes postulantes a la universidad	En la educación superior su sentido es un poco más amplio y se le entiende como el conjunto de características referidas a conocimientos, habilidades, valores y actitudes que se demandan de un estudiante para acceder a una carrera determinada	La elaboración de un perfil sólido ayuda a los comités de admisión a entender quién es el estudiante más allá de sus calificaciones, permitiendo tomar decisiones más informadas sobre la aceptación en la universidad	Nota en el examen de admisión
			Edad
			Grupo al que postula
			Sexo
			Tipo de colegio
			Condición de ingreso o no
			Procedencia

CAPÍTULO IV

METODOLOGÍA

4.1. AMBITO DE ESTUDIO: LOCALIZACION POLITICA Y GEOGRAFICA

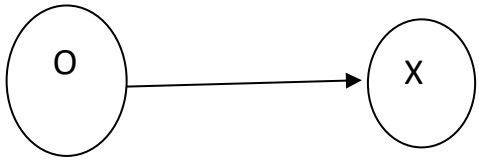
El presente estudio se realizará en la Universidad Nacional de San Antonio Abad del Cusco, ubicada en el distrito, Provincia y Departamento del Cusco.

4.2. TIPO Y NIVEL DE INVESTIGACION.

La investigación es de tipo aplicada, puesto que su objetivo es profundizar el conocimiento respecto al resultado del examen de admisión.

El nivel de investigación es descriptivo, debido a que se describirá las características de los estudiantes con buenos y malos resultados del examen de admisión.

El diseño de investigación es No experimental(observacional), transversal (estudio en un solo momento), descriptiva.



Dónde:

O= Estudiantes

X= Resultado del examen de admisión

4.3. UNIDAD DE ANÁLISIS

Postulantes a la Universidad Nacional de San Antonio Abad del Cusco en los diferentes exámenes de admisión ordinaria desde el 2021-I hasta el 2023-1

4.4. POBLACIÓN DE ESTUDIO

En base a la información proporcionada por la oficina de admisión de la Universidad Nacional de San Antonio Abad del Cusco, se estima que por semestre postulan alrededor de 4000 estudiantes.

4.5. TAMAÑO DE MUESTRA

La muestra considerada para poder evaluar en la investigación estuvo conformada por 24810 postulantes a la casa de estudios de la universidad Nacional de San Antonio Abad del Cusco la cual tiene la siguiente distribución de datos, según el cuadro siguiente por periodo de postulación.

Tabla 2. Postulantes a la UNSAAC por Semestre

	Frecuencia	Porcentaje
2021-1	3974	16,0
2021-2	4215	17,0
2022-1	6316	25,5
2022-2	5895	23,8
2023-1	4410	17,8
Total	**24810**	**100,0**

4.6. TÉCNICAS DE SELECCIÓN DE MUESTRA

La muestra utilizada es censal, la cual es no probabilística y no aleatorio, no hubo específicamente un procedimiento de selección debido a que se trabajó con toda la información proporcionada por la oficina de admisión de la Universidad.

4.7. TÉCNICAS DE RECOLECCIÓN DE INFORMACIÓN

En el presente trabajo de investigación se utilizó la técnica documental, para recopilar información del perfil y resultado del examen de admisión de los estudiantes.

El instrumento utilizado para el recojo de datos es la ficha de registro de datos documental.

4.8. TÉCNICAS DE ANÁLISIS E INTERPRETACIOÓN DE LA INFORMACIÓN.

La información se organizará en una base de datos en una hoja Excel, este se procesó mediante el lenguaje de programación libre R y en el SPSS en una versión de prueba de 30 días.

4.9. TÉCNICAS PARA DEMOSTRAR LA VERDAD O FALSEDAD DE LA HIPÓTESIS PLANTEADA

Tabla 3. Técnicas para demostrar la hipótesis

Hipótesis de investigación	**Hipótesis estadística**	**Nivel de significancia**	**Prueba estadística**
Existen cuatro perfiles o conglomerados de estudiantes postulantes a la UNSAAC obtenidos de las aplicaciones de los métodos de análisis de clúster y correspondencia.	**Ho: No existe** cuatro perfiles o conglomerados de estudiantes postulantes a la UNSAAC obtenidos de las aplicaciones de los métodos de análisis de clúster y correspondencia **Ha: E**xisten cuatro perfiles o conglomerados de estudiantes postulantes a la UNSAAC obtenidos de las aplicaciones de los métodos de análisis de clúster y correspondencia	**5%**	Análisis de clúster o conglomerados

CAPÍTULO V

RESULTADOS

5.1. COMPARACIÓN DE ALGORITMOS DE CLÚSTER CON EL ÍNDICE DE SEPARACIÓN Y COHESIÓN

Tabla 4. Comparación de Algoritmos

Algoritmo	Índice	Numero óptimo de clúster
Bietápico	0.374	4
Pam	0.2169	4
Clara	0.3516	8

Para poder generar los índice se han utilizado 3 algoritmos para poder conglomerar a los individuos, es así que el primero se probó en el SPSS 25 versión trial por tener un algoritmo (bietápico) muy bien implementado para trabajar tanto con variables cuantitativas y cualitativas el cual dio un índice de silueta y cohesión en 0.4 donde el numero ideal de clusters fue 4; así mismo como la base de datos tenía tanto datos cualitativos como cuantitativos para poder usar los algoritmos de PAM (*Partitioning Around Medoids*) y CLARA (*Clustering Large Applications*) en el programa R, se calculó una matriz de distancias con la metodología de distancias mixtas de gower y con ello poder usar los otros algoritmos obteniéndose índices de 0.2169 con 4 clústeres y en el algoritmo clara se obtuvo 0.3516 con 8 clústeres, obviamente son los más altos dentro de cada algoritmo, como se puede evidenciar en las siguientes líneas.

Salida SPSS

Tabla 5. Medida de silueta de cohesión - Bietápico

Category	Medida de silueta de cohesión y separación	V3
1	0,374	0,4

Código y salida del programa R

a) Indice de separación y cohesión con el método PAM

```
> library(cluster)
> gower_dist <- daisy(data_nueva,metric = "gower")
>
> # pam
> library(fpc)
> a=pamk(gower_dist,criterion="asw")
> a$crit
 [1] 0.0000000 0.1731278 0.1715935 0.2169045 0.1992636 0.2015927 0.1
883585
 [8] 0.1804238 0.1973837 0.1919144
> a$nc
[1] 4
```

b) Índice de separación y cohesión con el método CLARA

```
> # clara
> b=pamk(gower_dist,criterion="asw",usepam=FALSE)
> b$crit
 [1] 0.0000000 0.2993381 0.2444897 0.3370344 0.3422099 0.3442859 0.3
038037
 [8] 0.3516095 0.3186699 0.2580091
> b$nc
[1] 8
```

5.2. RESUMEN DEL ALGORITMO BIETAPICO

Tabla 6. Distribución de clúster

		N	% de combinado	% del total
Clúster	1	9446	38,1%	38,1%
	2	3024	12,2%	12,2%
	3	6165	24,8%	24,8%
	4	6175	24,9%	24,9%
Total		24810		100,0%

Se observa, que el Clúster 1 tiene 9,446 observaciones, lo que representa el 38.1% del total de 24,810 observaciones, mientras que en el Clúster 3 y 4 tienen respectivamente 24.8% y 24.9%.

Tabla 7. Centroides

		Nota		Edad	
		Media	Desviación	Media	Desviación
Clúster	1	7,8538	3,23974	18,49	1,705
	2	12,5808	2,57538	20,77	2,166
	3	7,6751	3,30879	18,49	1,712
	4	7,7523	3,31439	18,50	1,715
	Combinado	8,3603	3,56869	18,77	1,921

Se observa que en las variables cuantitativas nota y edad en los diferentes clústeres una variación en el centroide; es así como podemos resaltar el centroide del clúster 2 $(nota\ promedio;\ edad\ promedio) = (\bar{x}_{21}, \bar{x}_{22}) = (12.58; 20.77)$ se evidencia que la nota promedio que alcanza este grupo es de 12.58 puntos, y la edad del postulante es de 20.77 años, este clúster parece conformar a los alumnos ingresantes.

Tabla 8. Distribución Clúster y Grupo de postulación

		Grupo A		Grupo B		Grupo C		Grupo D	
		fi	%	fi	%	fi	%	fi	%
Clúster	1	2601	36,4%	2267	37,1%	1820	39,4%	2758	39,8%
	2	1144	16,0%	487	8,0%	533	11,5%	860	12,4%
	3	1042	14,6%	1987	32,5%	1185	25,6%	1951	28,1%
	4	2366	33,1%	1364	22,3%	1082	23,4%	1363	19,7%
	Combinado	7153	100,0%	6105	100,0%	4620	100,0%	6932	100,0%

fi: frecuencia absoluta o conteo; %: porcentaje

En el Clúster 1, hay 2,601 postulaciones al grupo A, lo que representa el 36.4%
En el Clúster 2 existe un 16.0%, en el Clúster 3 existe un 14.6%, en el último Clúster 4, existe el 33.1%.
De la misma manera interpretaremos para los demás grupos de postulación, por ejemplo, en el Clúster 1, hay 2,267 postulaciones al grupo B, lo que representa el 37.1%, en el Clúster 2 existe un 8.0%, en el Clúster 3 existe un 32.5%, en el último Clúster 4, existe el 22.3%.

Tabla 9. Distribución Clúster y sexo del postulante

		Femenino		Masculino	
		Frecuencia	Porcentaje	Frecuencia	Porcentaje
Clúster	1	4732	38,6%	4714	37,6%
	3	1360	11,1%	1664	13,3%
	3	6165	50,3%	0	0,0%
	4	0	0,0%	6175	49,2%
	Combinado	12257	100,0%	12553	100,0%

Del total de postulantes del sexo femenino la mitad se encuentra distribuida en el clúster 3, seguido del clúster 1 con un 38.6%, mientras que del total de postulantes de sexo masculino la mitad se centra en el clúster 4, seguido de un 37.6% en el clúster 1.

Tabla 10. Distribución Clúster y tipo Colegio

		Particular		Nacional	
		Frecuencia	Porcentaje	Frecuencia	Porcentaje
Clúster	1	0	0,0%	9446	81,2%
	2	878	6,7%	2146	18,4%
	3	6165	46,8%	0	0,0%
	4	6133	46,5%	42	0,4%
	Combinado	13176	100,0%	11634	100,0%

Del total de postulantes de colegio particular el 46.8% se encuentra distribuida en el clúster 3, seguido del clúster 4 con un 46.5%, mientras que del total de postulantes de colegios nacionales el 81.2% se centra en el clúster 1, seguido de un 18.4% en el clúster 2.

Tabla 11. Distribución Clúster y Condición

		Ingreso		No ingreso	
		Frecuencia	Porcentaje	Frecuencia	Porcentaje
Clúster	1	0	0,0%	9446	43,4%
	2	3024	99,8%	0	0,0%
	3	7	0,2%	6158	28,3%
	4	0	0,0%	6175	28,4%
	Combinado	3031	100,0%	21779	100,0%

Del total de alumnos postulantes que ingresaron el 99.8% se ubica en el clúster 2; del mismo modo se observa que del total de alumnos que no ingresaron a la universidad el 43.4% está ubicado en el clúster 1, hay casi un empate con 28.3 y 28.4% para el clúster 3 y 4.

5.3. PERFILES DE LOS ALUMNOS POSTULANTES POR CADA CLUSTER O CONGLOMERADO

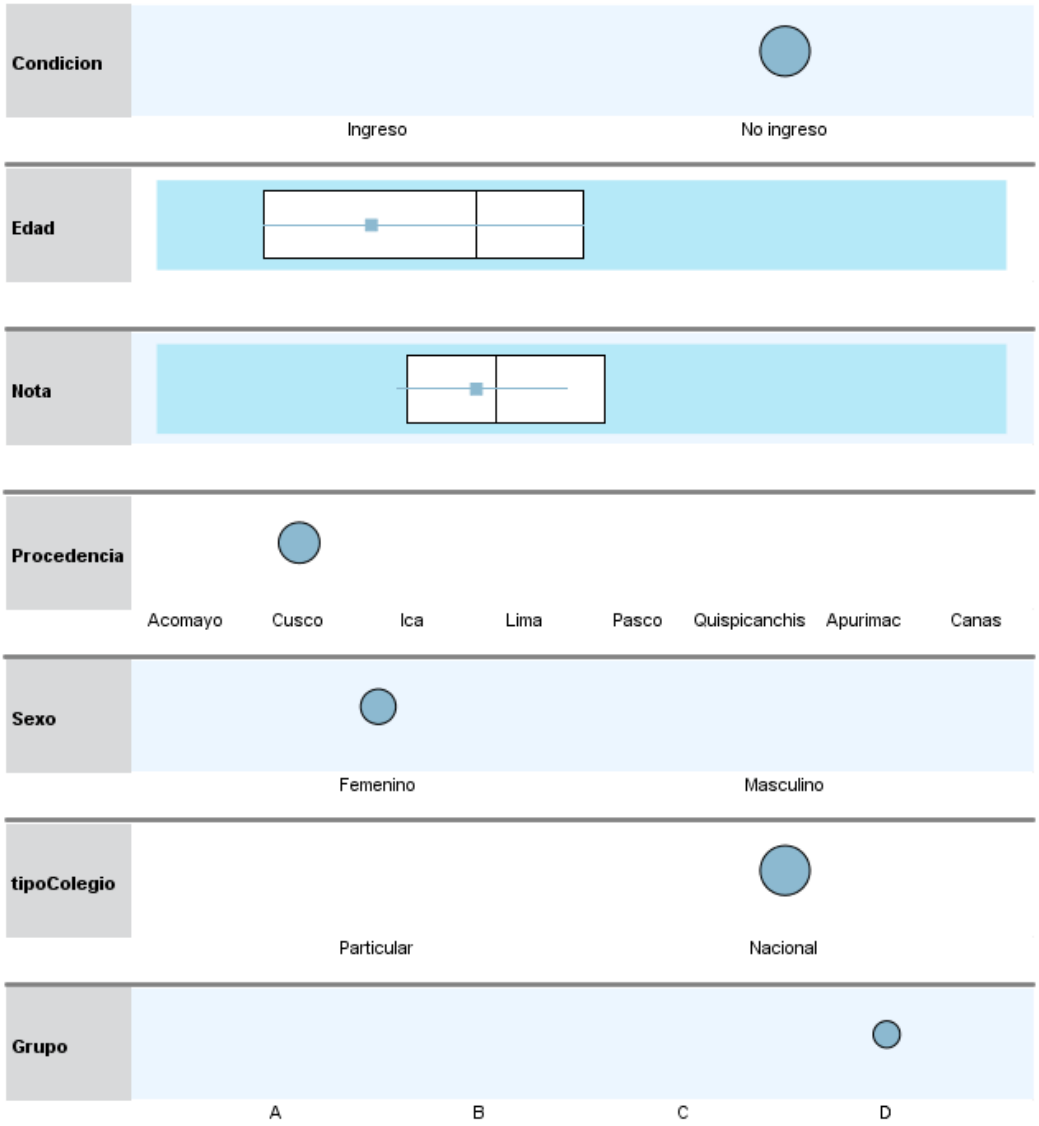


Figura 3. Conglomerado 1

Se observa que en este conglomerado o clúster 1, tiene prevalencia de alumnos que no ingresaron cuando postularon, su edad promedio es 18.49 años y su nota promedio fue de 7.44, su procedencia en su mayoría es del Cusco, de sexo femenino, procedencia de colegio nacional y mayormente postulan al grupo D.

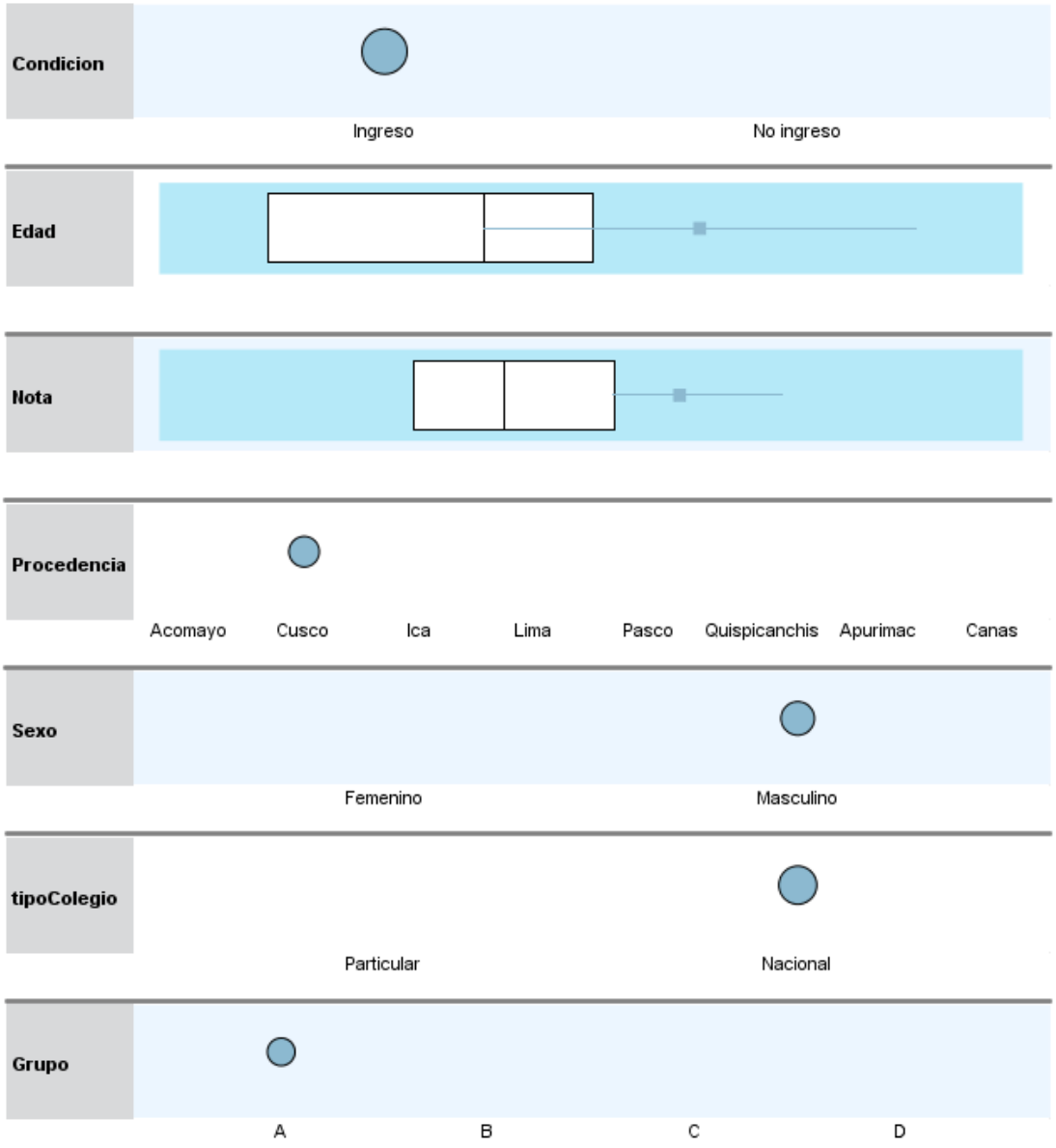


Figura 4. Conglomerado 2

Se observa que en este conglomerado o clúster 2, tiene prevalencia de alumnos que ingresaron cuando postularon, su edad promedio es 20.7 años y su nota promedio fue de 12.58, su procedencia en su mayoría es del Cusco, de sexo masculino, procedencia de colegio nacional y mayormente postulan al grupo A, obviamente que hay ingresantes a todos los grupos, pero hay más ingresantes al grupo A, donde se sitúa la mayor cantidad de escuelas profesionales en la universidad.

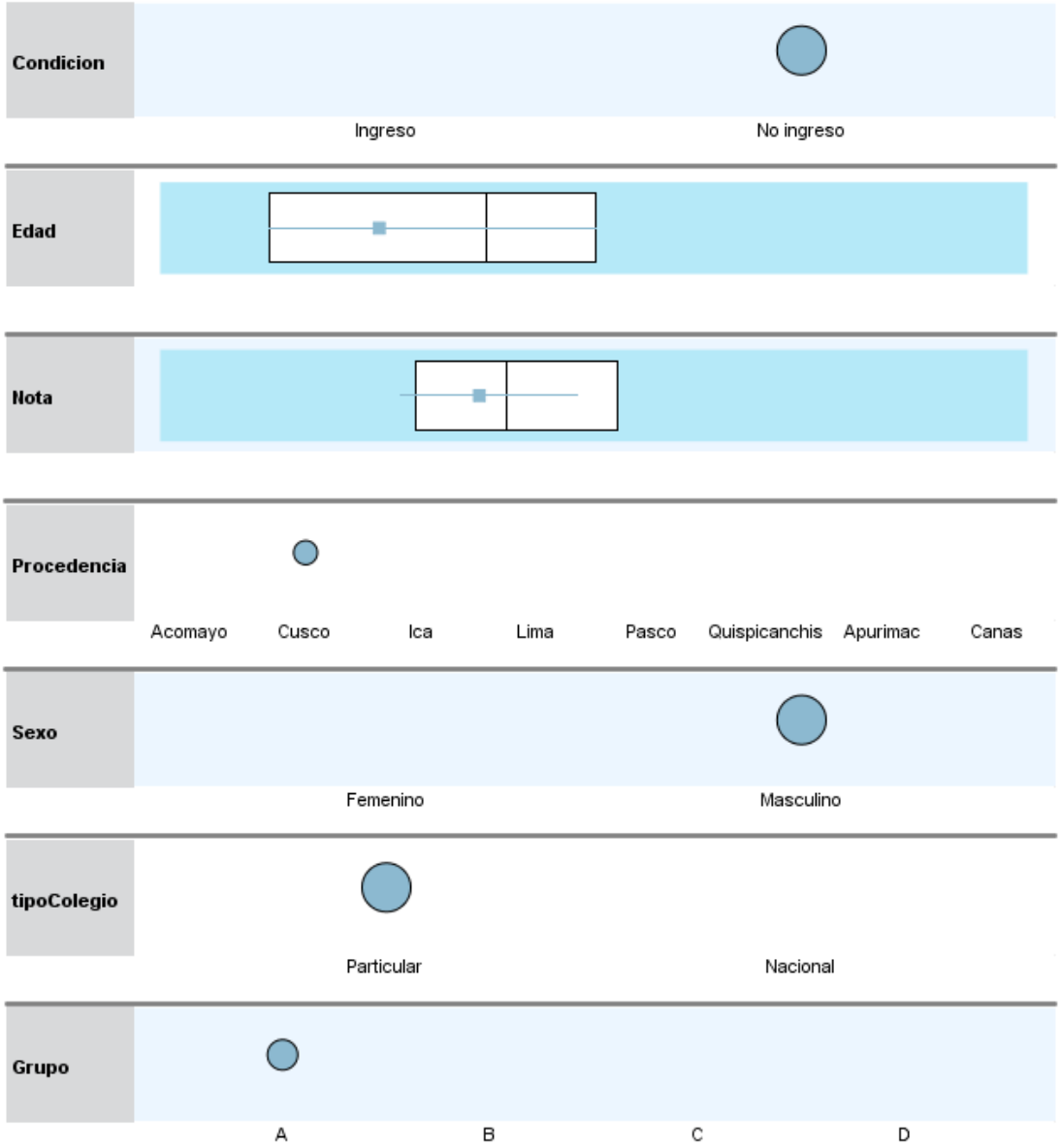


Figura 5. Conglomerado 3

Se observa que en este conglomerado o clúster 3, tiene prevalencia de alumnos que no ingresaron cuando postularon, su edad promedio es 18.49 años y su nota promedio fue de 7.67, su procedencia en su mayoría es del Cusco, de sexo masculino, procedencia de colegio particular y mayormente postulan al grupo A, donde se sitúa la mayor cantidad de escuelas profesionales en la universidad.

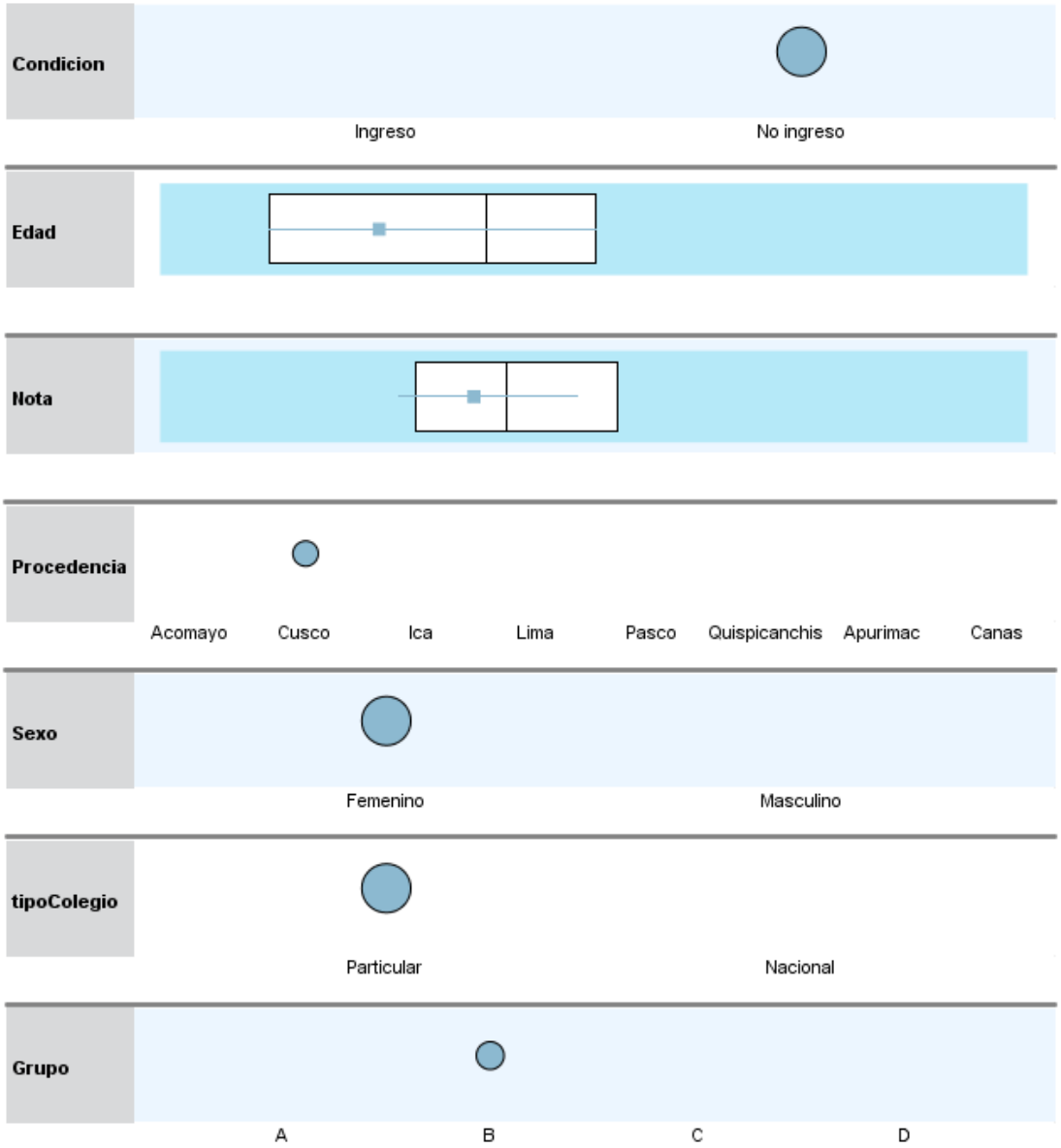


Figura 6. Conglomerado 4

Se observa que en este conglomerado o clúster 4, tiene prevalencia de alumnos que no ingresaron cuando postularon, su edad promedio era de 18.50 años y su nota promedio fue de 7.75, su procedencia en su mayoría es del Cusco, de sexo femenino, procedencia de colegio particular y mayormente postulan al grupo B, donde se sitúa la mayor cantidad de escuelas profesionales en la universidad.

5.4. ANÁLISIS DE CORRESPONDENCIA ENTRE EL GRUPO DE POSTULACIÓN Y CLÚSTER DE PERTENENCIA

Tabla 12.Tabla de correspondencias

	clúster				
Grupo	Clúster 1	Clúster 2	Clúster 3	Clúster 4	Margen activo
A	2601	1144	1042	2366	7153
B	2267	487	1987	1364	6105
C	1820	533	1185	1082	4620
D	2758	860	1951	1363	6932
Margen activo	9446	3024	6165	6175	24810

Se observa de la tabla la distribución de los postulantes a la Universidad Nacional de San Antonio Abad del Cusco según el clúster en el que fue agrupado y el grupo al que postulo, se observa que más de 7 mil estudiantes postulan al grupo A, seguido de estudiantes que postulan al grupo D, en tercer lugar se encuentra el grupo B y con menor cantidad de postulantes el grupo C; así mismo en el clúster 1 hay mayor concentración de estudiantes, seguido del clúster 4 y 3, y por último el clúster 2 con menor cantidad de estudiantes agrupados.

Tabla 13. Perfiles de fila

	clúster				
Grupo	Clúster 1	Clúster 2	Clúster 3	Clúster 4	Margen activo
A	,364	,160	,146	,331	1,000
B	,371	,080	,325	,223	1,000
C	,394	,115	,256	,234	1,000
D	,398	,124	,281	,197	1,000
Masa	,381	,122	,248	,249	

Del total de alumnos postulantes al grupo A, el 36.4% se encuentra en el clúster 1, seguido de un 33.1% en el clúster 2, así mismo del total de postulantes al grupo B, el 37.1% se encuentra en el clúster 1 y el 32.5% en el clúster 3, de la misma manera se evidencio que la mayoría de los postulantes están el primer conglomerado o clúster.

Tabla 14. Perfiles de columna

	clúster				
Grupo	Clúster 1	Clúster 2	Clúster 3	Clúster 4	Masa
A	,275	,378	,169	,383	,288
B	,240	,161	,322	,221	,246
C	,193	,176	,192	,175	,186
D	,292	,284	,316	,221	,279
Margen activo	1,000	1,000	1,000	1,000	

En el clúster 1, se encontró que el 29.2% está en el grupo D, seguido del 27.5% que se encuentra en el grupo A, se puede observar que en el clúster 2, el 37.8% postuló al grupo A y 28.4% al grupo D.

El clúster 3, esta mayormente representado por postulantes al grupo B y grupo D; mientras que en el clúster 4, se encuentra en mayores proporciones postulantes al grupo A, B y D.

Tabla 15. Resumen de correspondencia

					Proporción de inercia	
Dimensión	Valor singular	Inercia	Chi cuadrado	Sig.	Contabilizado	Acumulado
1	,189	,036			,922	,922
2	,054	,003			,075	,997
3	,010	,000			,003	1,000
Total		,039	960,119	,000[a]	1,000	1,000

En principio de la prueba chi cuadrado se observa que el sig=0.000<0.05, lo cual indica existe una asociación entre el grupo al que postula y el clúster de pertenencia, además la proporción de la inercia para la primera dimensión es de 92.2% lo cual indicaría que con una dimensión sería suficiente para explicar el comportamiento de las variables grupo de postulación y clúster de agrupación.

Tabla 16. Puntos de fila generales

		Puntuación en dimensión			Contribución				
					Del punto en la inercia de dimensión		De la dimensión en la inercia del punto		
Grupo	Masa	1	2	Inercia	1	2	1	2	Total
A	,288	-,658	,065	,024	,661	,023	,997	,003	1,000
B	,246	,435	,319	,010	,246	,464	,866	,133	1,000
C	,186	,085	-,055	,000	,007	,011	,699	,086	,785
D	,279	,240	-,311	,005	,085	,502	,673	,324	,997
Total activo	1,000			,039	1,000	1,000			

En la tabla anterior se puede evidenciar que para la dimensión 1 el grupo A es el que más influye, mientras que para la segunda dimensión el grupo D y B son los que más influyen, así mismo se tienen las coordenadas en el biplot como por ejemplo para el grupo A su ubicación en el biplot es (-0,658; 0,065) para el grupo B (0,435; 0,319) y así sucesivamente para los demás puntos.

Tabla 17. Puntos de columna generales

		Puntuación en dimensión			Contribución				
					Del punto en la inercia de dimensión		De la dimensión en la inercia del punto		
clúster	Masa	1	2	Inercia	1	2	1	2	Total
Clúster 1	,381	,050	-,131	,001	,005	,121	,311	,607	,917
Clúster 2	,122	-,507	-,413	,007	,166	,385	,837	,158	,996
Clúster 3	,248	,641	,086	,019	,540	,034	,994	,005	,999
Clúster 4	,249	-,468	,316	,012	,288	,460	,885	,115	1,000
Total activo	1,000			,039	1,000	1,000			

En la tabla anterior se puede evidenciar que para la dimensión 1 el clúster 3, es el que más influye, mientras que para la segunda dimensión el clúster 4 y 2 son los que más influyen, así mismo se tienen las coordenadas en el biplot como por ejemplo para

el clúster 1, su ubicación en el biplot es (0,050; -,0.131) para el clúster B (-0,507; -0,413) y así sucesivamente para los demás puntos.

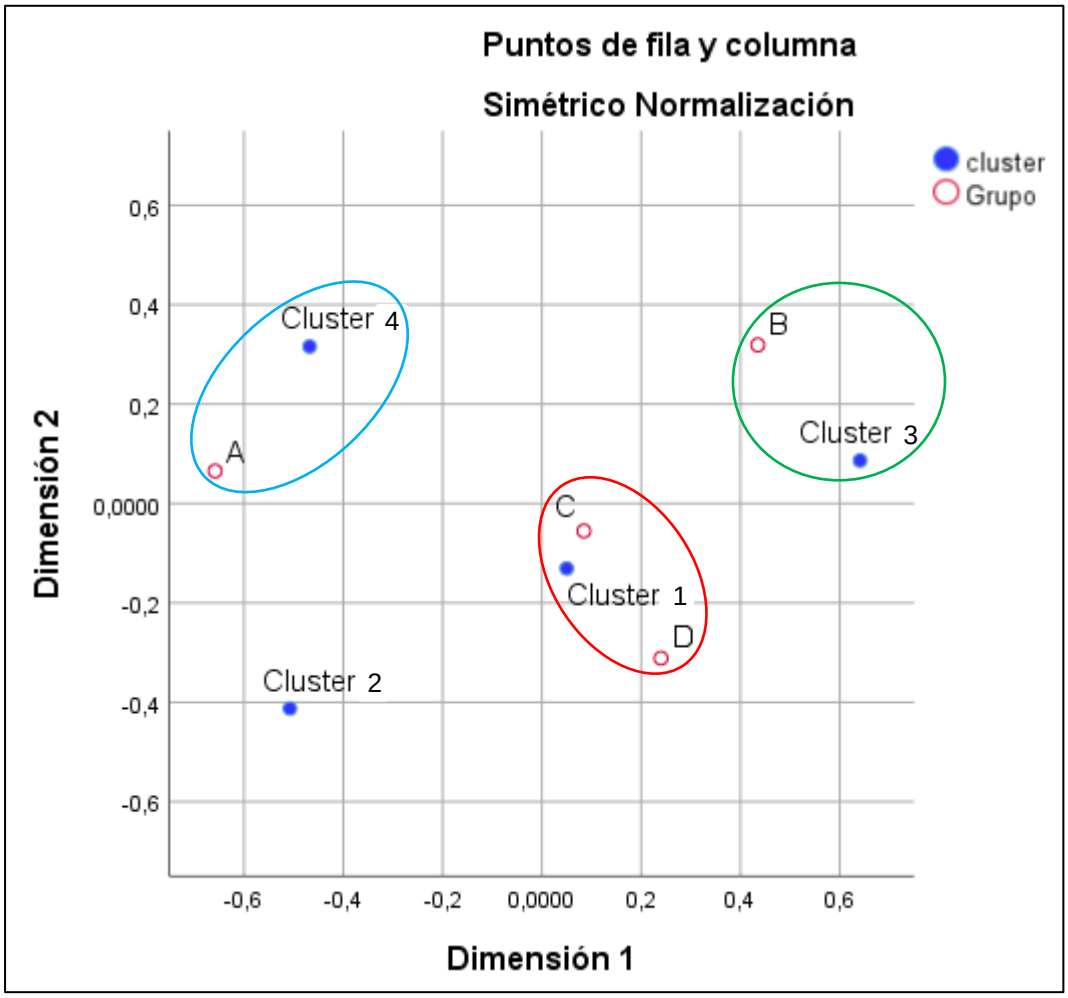


Figura 7. Biplot grupo y clúster

En la figura se observa que al superponer los puntos del clúster y grupo al que postulan los estudiantes se observa coincidencia del clúster 3, más cercano al grupo B, así mismo el clúster 1, conglomera mucho mas alumnos que postulan al grupo C y D, por otro lado se observa el clúster 4, quien conglomera a los postulantes del grupo A; sin embargo a los postulantes declarados en el clúster 2, no se le identifica con algún grupo en particular.

5.5. ANÁLISIS DE CORRESPONDENCIA ENTRE PROCEDENCIA Y CLÚSTER DE PERTENENCIA

Tabla 18. Procedencia y clúster

	Clúster				Margen activo
	Clúster 1	Clúster 2	Clúster 3	Clúster 4	
Acomayo	0	90	225	230	545
Ancash	1	1	2	7	11
Anta	722	137	70	88	1017
Apurímac	735	170	549	612	2066
Arequipa	0	15	93	98	206
Ayacucho	2	6	21	20	49
Cajamarca	0	1	6	2	9
Calca	201	110	352	338	1001
Canas	3	105	270	318	696
Canchis	771	399	643	668	2481
Chachapoyas	0	0	2	1	3
Chiclayo	2	2	1	5	10
Chumbivilcas	133	45	169	180	527
Cusco	6417	1382	1652	1468	10919
Espinar	151	32	76	81	340
Extranjero	0	0	0	5	5
Huancavelica	0	1	1	3	5
Huancayo	0	2	0	3	5
Ica	1	0	10	9	20
Junin	0	2	5	0	7
La convencion	59	71	330	330	790
Lambayeque	0	0	0	1	1
Lima	5	34	163	178	380
Loreto	0	0	2	1	3
Madre d Dios	11	17	143	123	294
Moquegua	0	0	9	2	11
Paruro	71	33	96	76	276
Pasco	0	0	0	3	3
Paucartambo	62	71	172	198	503
Piura	0	1	2	1	4
Puno	9	38	334	277	658
Quispicanchis	19	128	410	438	995
San Martin	1	0	0	0	1
Tacna	0	2	16	4	22

Ucayali	0	1	1	2	4
Urubamba	70	128	340	405	943
Margen activo	9446	3024	6165	6175	24810

Se observa de la tabla la distribución de los postulantes a la Universidad Nacional de San Antonio Abad del Cusco según el clúster en el que fue agrupado y la procedencia, se observa que más de 10 mil postulantes son de la ciudad del Cusco, seguido de los postulantes que provienen de Canchis, Apurímac, Anta y Calca.

Tabla 19. Perfiles de fila

	Clúster				
Procedencia	Clúster 1	Clúster 2	Clúster 3	Clúster 4	Margen activo
Acomayo	,000	,165	,413	,422	1,000
Ancash	,091	,091	,182	,636	1,000
Anta	,710	,135	,069	,087	1,000
Apurímac	,356	,082	,266	,296	1,000
Arequipa	,000	,073	,451	,476	1,000
Ayacucho	,041	,122	,429	,408	1,000
Cajamarca	,000	,111	,667	,222	1,000
Calca	,201	,110	,352	,338	1,000
Canas	,004	,151	,388	,457	1,000
Canchis	,311	,161	,259	,269	1,000
Chachapoyas	,000	,000	,667	,333	1,000
Chiclayo	,200	,200	,100	,500	1,000
Chumbivilcas	,252	,085	,321	,342	1,000
Cusco	,588	,127	,151	,134	1,000
Espinar	,444	,094	,224	,238	1,000
Extranjero	,000	,000	,000	1,000	1,000
Huancavelica	,000	,200	,200	,600	1,000
Huancayo	,000	,400	,000	,600	1,000
Ica	,050	,000	,500	,450	1,000
Junín	,000	,286	,714	,000	1,000
Masa	,459	,125	,208	,208	

Del total de alumnos postulantes de la ciudad del Cusco, el 58.8% se encuentra en el clúster 1, asimismo en el caso de los postulantes de Canchis representando el 31.1% que se encuentra en el clúster 1, seguido del clúster 4; por otro lado, de los

postulantes de Apurímac el 35.6% se encuentra ubicado en el clúster 1 y el 29,6% en el clúster 4.

Del total de postulantes de Anta, el 71% pertenece al clúster 1, mientras que los postulantes de Calca, el 35.2% está en el clúster 3, y el 33.8% está ubicado en el clúster 4.

Tabla 20. Perfiles de columna

	Clúster o Conglomerado				
Procedencia	**Clúster 1**	**Clúster 2**	**Clúster 3**	**Clúster 4**	**Masa**
Acomayo	,000	,036	,054	,056	,027
Ancash	,000	,000	,000	,002	,001
Anta	,079	,055	,017	,021	,051
Apurímac	,080	,068	,132	,148	,104
Arequipa	,000	,006	,022	,024	,010
Ayacucho	,000	,002	,005	,005	,002
Cajamarca	,000	,000	,001	,000	,000
Calca	,022	,044	,085	,082	,050
Canas	,000	,042	,065	,077	,035
Canchis	,084	,160	,155	,162	,125
Chachapoyas	,000	,000	,000	,000	,000
Chiclayo	,000	,001	,000	,001	,001
Chumbivilcas	,015	,018	,041	,044	,026
Cusco	,702	,553	,398	,355	,548
Espinar	,017	,013	,018	,020	,017
Extranjero	,000	,000	,000	,001	,000
Huancavelica	,000	,000	,000	,001	,000
Huancayo	,000	,001	,000	,001	,000
Ica	,000	,000	,002	,002	,001
Junín	,000	,001	,001	,000	,000
Margen activo	1,000	1,000	1,000	1,000	

Del total de postulantes que se ubican en el clúster 1, el 70.2% es de la ciudad del Cusco, seguido de Canchis 8.4%; en el clúster 2, el 55.3% son postulantes de la ciudad del Cusco, seguido de Canchis con un 16%, en el resto de clúster es similar la agrupación.

Tabla 21. Resumen de correspondencia

Dimensión	Valor singular	Inercia	Chi cuadrado	Sig.	Proporción de inercia		Valor singular de confianza	
					Contabilizado	Acumulado	Desviación estándar	Correlación 2
1	,411	,169			,955	,955	,006	-,077
2	,072	,005			,030	,985	,007	
3	,052	,003			,015	1,000		
Total		,177	3529,733	,000[a]	1,000	1,000		

En principio de la prueba chi cuadrado se observa que el sig=0.000<0.05, lo cual indica existe una asociación entre la procedencia y el clúster de pertenencia, además la proporción de la inercia para la primera dimensión es de 95.5% lo cual indicaría que con una dimensión sería suficiente para explicar el comportamiento de las variables procedencia y clúster de agrupación.

Tabla 22.Puntos de fila generales

Procedencia	Masa	Puntuación en dimensión		Inercia	Contribución				
					Del punto en la inercia de dimensión		De la dimensión en la inercia del punto		
		1	2		1	2	1	2	Total
Acomayo	,027	-1,460	,469	,024	,142	,083	,981	,018	,999
Ancash	,001	-1,351	-,590	,001	,002	,003	,629	,021	,651
Anta	,051	,851	,076	,015	,090	,004	,986	,001	,987
Apurímac	,104	-,424	-,489	,010	,045	,342	,806	,189	,995
Arequipa	,010	-1,625	-,566	,011	,066	,046	,978	,021	,999
Ayacucho	,002	-1,395	,006	,002	,012	,000	,992	,000	,992
Cajamarca	,000	-1,492	,098	,001	,002	,000	,622	,000	,623
Calca	,050	-,885	-,148	,016	,096	,015	,985	,005	,990
Canas	,035	-1,480	,279	,032	,186	,038	,990	,006	,996
Canchis	,125	-,432	,398	,011	,057	,273	,869	,130	,999
Chachapoyas	,000	-1,704	-1,195	,000	,001	,003	,715	,062	,777
Chiclayo	,001	-,787	,645	,000	,001	,003	,387	,046	,433
Chumbivilcas	,026	-,762	-,443	,007	,037	,071	,944	,056	1,000
Cusco	,548	,434	,011	,043	,251	,001	,999	,000	,999
Espinar	,017	-,106	-,356	,000	,000	,030	,333	,658	,991
Extranjero	,000	-1,892	-1,871	,001	,002	,012	,386	,067	,452
Huancavelica	,000	-1,454	,659	,000	,001	,002	,698	,025	,724

Huancayo	,000	-1,129	2,784	,001	,001	,027	,261	,280	,540
Ica	,001	-1,577	-1,343	,001	,006	,025	,866	,111	,976
Junín	,000	-1,145	2,178	,001	,001	,023	,257	,164	,420
Total activo	1,000			,177	1,000	1,000			

En la tabla anterior se puede evidenciar que para la dimensión 1 Cusco, canas Acomayo son las provincias que más influyen en la dimensión, mientras que para la segunda dimensión influyen más Apurímac y Canchis, así mismo se tienen las coordenadas en el biplot como por ejemplo para Cusco su ubicación en el biplot es 0,548; 0,434) para Canchis (0,125; -0,432) y así sucesivamente para los demás puntos.

Tabla 23. Puntos de columna generales

		Puntuación en dimensión			Contribución				
					Del punto en la inercia de dimensión		De la dimensión en la inercia del punto		
Clúster	Masa	1	2	Inercia	1	2	1	2	Total
Clúster 1	,459	,651	-,104	,080	,472	,069	,996	,004	1,000
Clúster 2	,125	,007	,708	,005	,000	,868	,000	,996	,996
Clúster 3	,208	-,662	-,062	,039	,222	,011	,960	,001	,962
Clúster 4	,208	-,778	-,136	,053	,306	,053	,973	,005	,978
Total activo	1,000			,177	1,000	1,000			

En la tabla anterior se puede evidenciar que para la dimensión 1 el clúster 1, es el que más influye, mientras que para la segunda dimensión el clúster 2, es el que más influye, así mismo se tienen las coordenadas en el biplot como por ejemplo para el clúster 1, su ubicación en el biplot es (0,459; 0,651) para el clúster 2, su ubicación en el biplot es (0,125; 0,007) y así sucesivamente para los demás puntos.

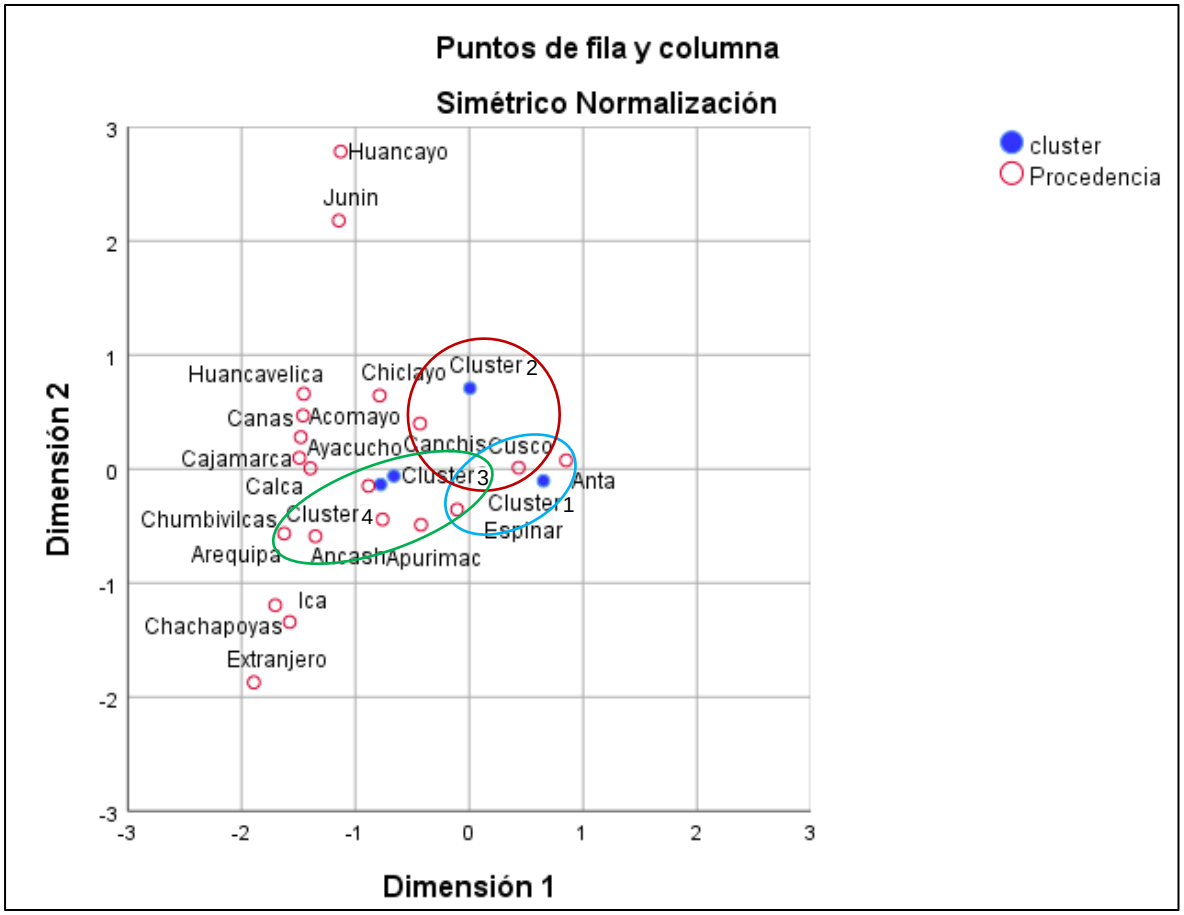


Figura 8. Biplot procedencia y Clúster

En la figura se observa que al superponer los puntos del clúster y la procedencia de los estudiantes se observa coincidencia del clúster 1, más cercano a las provincias de Cusco, Anta y Espinar, así mismo el clúster 2, esta más próximo a Canchis y Cusco, así mismo el clúster 3 y 4 se encuentra más próximo a Cusco, Ancash, Apurímac, Espinar, Calca y Chumbivilcas.

5.6. VALIDACIÓN DE LOS CLÚSTERES

5.6.1. Pruebas de independencia chi cuadrado de Pearson.

Tabla 24. Sexo vs Clúster

			Clúster				
			Clúster 1	Clúster 2	Clúster 3	Clúster 4	Total
Sexo	Femenino	Recuento	4732	1360	6165	0	12257
		%	38,6%	11,1%	50,3%	0,0%	100,0%
	Masculino	Recuento	4714	1664	0	6175	12553
		%	37,6%	13,3%	0,0%	49,2%	100,0%
Total		Recuento	9446	3024	6165	6175	24810
		%	38,1%	12,2%	24,8%	24,9%	100,0%
Chi-cuadrado de Pearson =12368,824					pvalor=0.000		

Ho: No existe asociación entre el sexo y el clúster de pertenencia

H1: Existe asociación entre el sexo y el clúster de pertenencia

Debido a que el p.valor = 0.000<0.05, se rechaza la Ho, en tanto se puede confirmar que existe evidencia para afirmar que Existe asociación entre el sexo y el clúster de pertenencia, se puede observar que las postulantes de sexo femenino se encuentran en mayor proporción en el clúster 3 y 1; por otro lado los de sexo masculino en el clúster 4 y 1

Tabla 25. Tipo de colegio vs Clúster

			Clúster				
			Clúster 1	Clúster 2	Clúster 3	Clúster 4	Total
Tipo Colegio	Particular	Recuento	0	878	6165	6133	13176
		%	0,0%	6,7%	46,8%	46,5%	100,0%
	Nacional	Recuento	9446	2146	0	42	11634
		%	81,2%	18,4%	0,0%	0,4%	100,0%
Total		Recuento	9446	3024	6165	6175	24810
		%	38,1%	12,2%	24,8%	24,9%	100,0%
Chi-cuadrado de Pearson =22140,519					pvalor=0.000		

Ho: No existe asociación entre el tipo de colegio y el clúster de pertenencia

H1: Existe asociación entre el tipo de colegio y el clúster de pertenencia

Debido a que el p.valor = 0.000<0.05, se rechaza la Ho, en tanto se puede confirmar que existe evidencia para afirmar que Existe asociación entre el tipo de colegio y el clúster de pertenencia, se puede observar que aquellos que son de colegios particulares se encuentran en mayores proporciones en el clúster 3 y 4; mientras que los que se encuentran en el clúster 1 y 2 son mayormente de colegios nacionales.

Tabla 26. Condición vs Clúster

			Clúster				
			Clúster 1	Clúster 2	Clúster 3	Clúster 4	Total
Condición	Ingreso	Recuento	0	3024	7	0	3031
		%	0,0%	99,8%	0,2%	0,0%	100,0%
	No ingreso	Recuento	9446	0	6158	6175	21779
		%	43,4%	0,0%	28,3%	28,4%	100,0%
Total		Recuento	9446	3024	6165	6175	24810
		%	38,1%	12,2%	24,8%	24,9%	100,0%
Chi-cuadrado de Pearson =24744.802					pvalor=0.000		

Ho: No existe asociación entre la condición y el clúster de pertenencia

H1: Existe asociación entre la condición y el clúster de pertenencia

Debido a que el p.valor = 0.000<0.05, se rechaza la Ho, en tanto se puede confirmar que existe evidencia para afirmar que Existe asociación entre la condición y el clúster de pertenencia, se puede observar que aquellos que ingresaron casi la totalidad fue ubicado en el clúster 2, y el resto que no ingreso está ubicado en los diferentes clústeres.

5.6.2. Comparación de notas según clúster (kruskall-wallis)

Tabla 27. Resumen de prueba de hipótesis Kruskall Wallis

	Hipótesis nula	Prueba	Sig.	Decisión
1	Las medianas de Nota son las mismas entre las categorías de clúster.	Prueba de la mediana para muestras independientes	,000	Rechazar la hipótesis nula.

Se muestran significaciones asintóticas. El nivel de significación es de ,05.

N total	24.810
Mediana	8,000
Estadístico de contraste	3.796,830
Grados de libertad	3
Sig. asintótica (prueba bilateral)	,000

De la prueba estadistica de Kruskall-Wallis(sig o pvalor = 0.00<0.05) se evidencio que si existen diferencias de las notas en los diferentes clusteres, es asi que observando el grafico de cajas se evidencia que el clúster 2, en cuanto a nota se refiere presente mayor nota promedio en comparacion al resto, esto puede deberse por que por informacion anterior se vio que este clúster aglomeradaba mayormente a los ingresantes a la universidad.

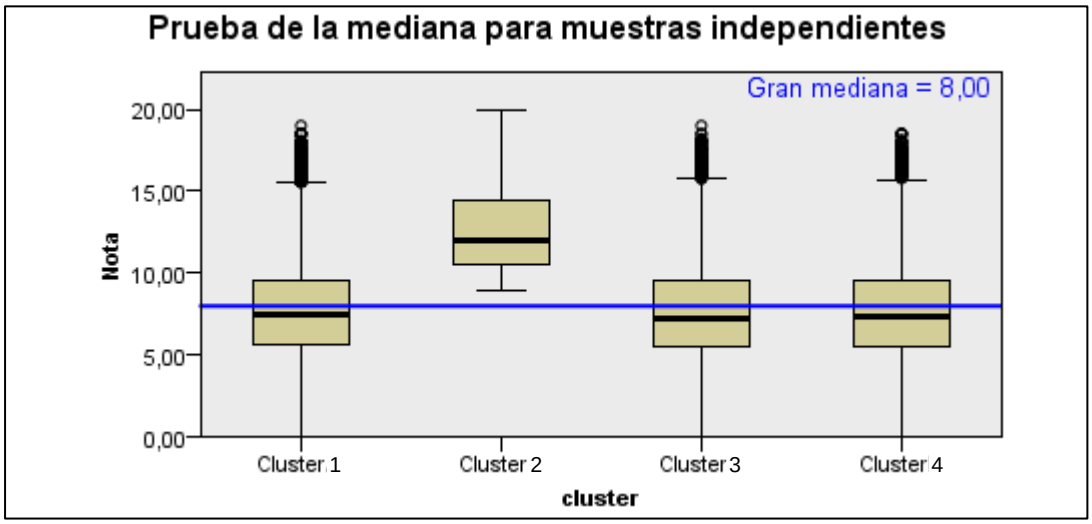
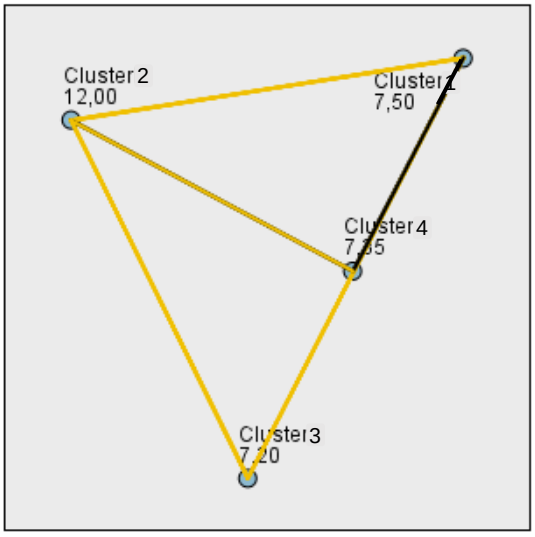


Figura 9. Gráfico de cajas de la nota por Clúster

Comparaciones entre parejas de cluster



Cada nodo muestra la mediana de la muestra de cluster.

Muestra 1-Muestra 2	Estadístico de contraste	Sig.	Sig. ajust.
Cluster 3-Cluster 4	1,054	,305	1,000
Cluster 3-Cluster 1	11,955	,001	,003
Cluster 3-Cluster 2	3.629,044	,000	,000
Cluster 4-Cluster 1	4,964	,026	,155
Cluster 4-Cluster 2	3.655,471	,000	,000
Cluster 1-Cluster 2	4.067,059	,000	,000

Figura 10. Comparación por pares

5.6.3. Comparacion de la edad según cluster de pertenencia

Tabla 28. Resumen de prueba de hipótesis- kruskall Wallis

	Hipótesis nula	Prueba	Sig.	Decisión
1	Las medianas de Edad son las mismas entre las categorías de clúster.	Prueba de la mediana para muestras independientes	,000	Rechazar la hipótesis nula.

Se muestran significaciones asintóticas. El nivel de significación es de ,05.

N total	24.810
Mediana	19,000
Estadístico de contraste	1.626,785
Grados de libertad	3
Sig. asintótica (prueba bilateral)	,000

De la prueba estadistica de Kruskall-Wallis(sig o pvalor = 0.00<0.05) se evidencio que si existen diferencias de la edad en los diferentes clusteres, es asi que observando el grafico de cajas se evidencia que el clúster 2, en cuanto a edad se refiere presenta mayor nota promedio en comparacion al resto, esto puede deberse por que por informacion anterior se vio que este clúster aglomeradaba mayormente a los ingresantes a la universidad, lo explicado se puede observar en los graficos posteriores y pruebas de comparaciones por pares.

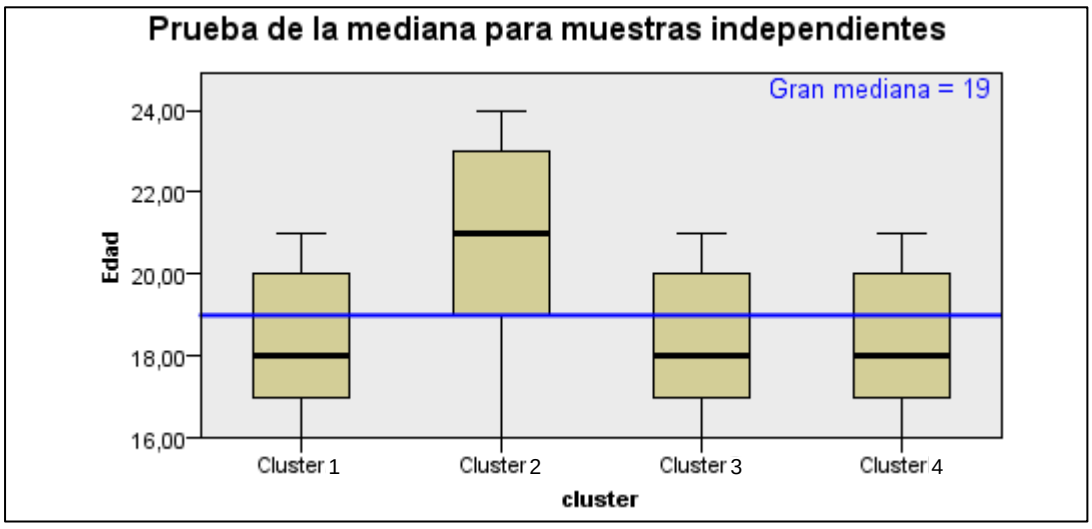
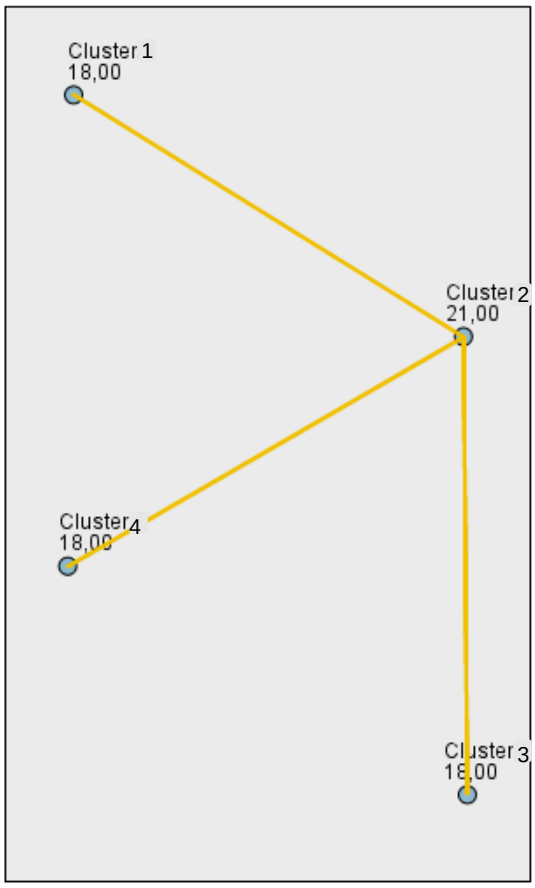


Figura 11. Gráfico de cajas de la edad por Clúster

Comparaciones entre parejas de cluster



Cada nodo muestra la mediana de la muestra de cluster.

Muestra 1-Muestra 2	Estadístico de contraste	Sig.	Sig. ajust.
Cluster 1-Cluster 3	,090	,764	1,000
Cluster 1-Cluster 4	,155	,694	1,000
Cluster 1-Cluster 2	1.358,052	,000	,000
Cluster 3-Cluster 4	,398	,528	1,000
Cluster 3-Cluster 2	1.199,094	,000	,000
Cluster 4-Cluster 2	1.148,803	,000	,000

Figura 12. Comparación por pares

5.7. CLÚSTER PARA ALUMNOS INGRESANTES A LA UNIVERSIDAD

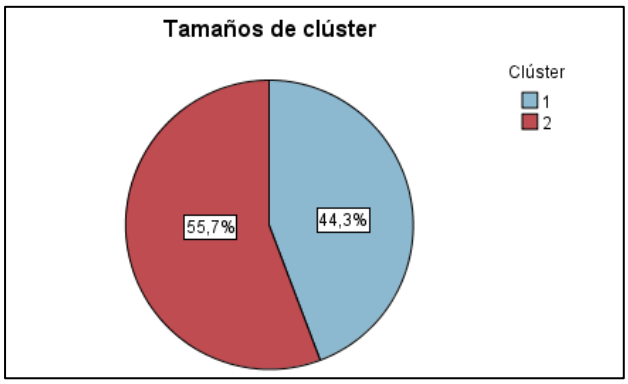


Figura 13. Tamaños de clúster

Se observa en la figura 13, que el algoritmo del clúster Bietápico divide en dos perfiles a los ingresantes a la universidad, considerando el 44.3% de sus ingresantes en un primer clúster y el 55.7% en el clúster 2, y a continuación empezaremos a detallar las características de cada uno de estos grupos.

Tabla 29. Descriptivos de la nota por clúster

Nota		Clúster 1	Clúster 2
Media		12.4123	12.4808
95% de intervalo de confianza para la media	Límite inferior	12.2485	12.3316
	Límite superior	12.5762	12.6300
Mediana		12.0000	12.0750
Varianza		5.364	5.597
Desviación		2.31601	2.36570
Mínimo		9.00	9.00
Máximo		19.60	19.50
Rango		10.60	10.50
Rango intercuartil		3.40	3.60
Asimetría		0.758	0.506
Curtosis		0.100	-0.619

De la tabla 29, se observa que respecto a la nota de ingreso, en ambos clúster el promedio de nota es similar , con un 95% de confianza se puede indicar que la media de la nota del clúster va desde 12.248 hasta 12.576 para los que han sido agrupados en el clúster 1, mientras que en el otro clúster la nota varia desde 12.33 hasta 12.63 datos que son similares, así mismo comparando la mediana, desviación el mínimo y máximo se llega a la conclusión de que la nota no es un decisor fuerte para poder clasificar a los ingresantes en los 2 perfiles, los datos se pueden visualizar en la figura 14.

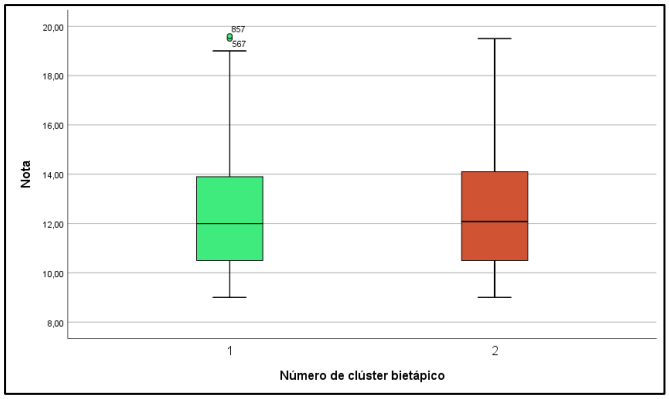


Figura 14. Gráfico de cajas de las notas por clúster

Tabla 30. Descriptivos de la edad por clúster

Edad		Clúster 1	Clúster 2
Media		20.75	20.73
95% de intervalo de confianza	Límite inferior	20.60	20.60
para la media	Límite superior	20.90	20.86
Mediana		21.00	21.00
Varianza		4.574	4.211
Desviación		2.139	2.052
Mínimo		16	16
Máximo		24	24
Rango		8	8
Rango intercuartil		3	3
Asimetría		-0.328	-0.309
Curtosis		-0.466	-0.421

De la tabla 30, se observa que respecto a la edad, en ambos clúster el promedio de edad es similar , con un 95% de confianza se puede indicar que la media de la edad del clúster va desde 20.6 hasta 20.90 para los que han sido agrupados en el clúster 1, mientras que en el otro clúster la nota varia desde 20.6 hasta 20.86 datos que son similares, así mismo comparando la mediana, desviación el mínimo y máximo se llega a la conclusión de que la edad no es un decisor fuerte para poder clasificar a los ingresantes en los 2 perfiles, los datos se pueden visualizar en la figura 15.

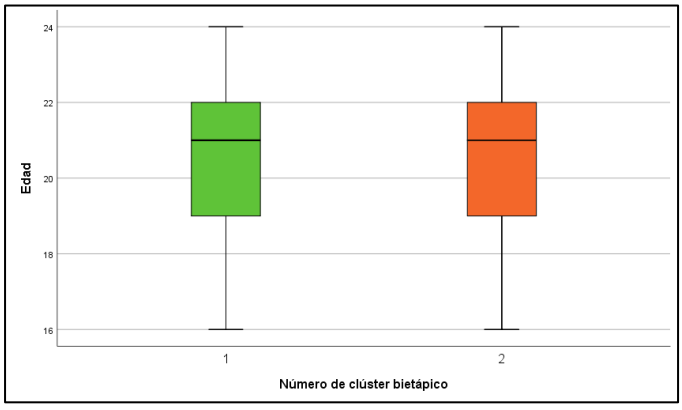


Figura 15. Gráfico de cajas de la edad por clúster

Tabla 31. Grupo y Número de clúster bietápico

			Número de clúster Bietápico		
			1	2	Total
Grupo	A	fi	202	552	754
		%	26,2%	57,0%	43,4%
	B	fi	131	53	184
		%	17,0%	5,5%	10,6%
	C	fi	130	140	270
		%	16,9%	14,5%	15,5%
	D	fi	307	223	530
		%	39,9%	23,0%	30,5%
Total		fi	770	968	1738
		%	100,0%	100,0%	100,0%

Se observa que en el clúster 1, esta mas representado por los ingresantes al grupo D, seguido de un 26.2% que ingresaron al grupo A; por otro lado, en el clúster 2 esta mas representado por los postulantes al grupo A, seguido ya también por el grupo D, se podría indicar que el cluste1 asociado a grupo D y clúster 2 al grupo A.

Tabla 32. Sexo y Número de clúster Bietápico

			Número de clúster bietápico		
			1	2	Total
Sexo	Femenino	fi	770	0	770
		%	100,0%	0,0%	44,3%
	Masculino	fi	0	968	968
		%	0,0%	100,0%	55,7%
Total		fi	770	968	1738
		%	100,0%	100,0%	100,0%

En el clúster 1 se observa que el 100% son del sexo femenino, mientras que en el clúster 2 se observa la totalidad de varones ingresantes a la universidad, esta variable si es una variable decisora por que nos permite diferenciar entre los diferentes perfiles.

Tabla 33. Tipo Colegio y Número de clúster Bietápico

			Número de clúster bietápico		Total
			1	2	
Tipo Colegio	Particular	fi	231	238	469
		%	30,0%	24,6%	27,0%
	Nacional	fi	539	730	1269
		%	70,0%	75,4%	73,0%
Total		fi	770	968	1738
		%	100,0%	100,0%	100,0%

El primer y segundo clúster están más representado por ingresantes que provienen de colegios nacionales que en los colegios particulares, esto conlleva a pensar que la universidad Nacional de San Antonio Abad del Cusco alberga más a estudiantes de colegios nacionales

Tabla 34. Procedencia y clúster bietápico

			Número de clúster bietápico		Total
			1	2	
Procedencia	Acomayo	fi	7	4	11
		%	0,9%	0,4%	0,6%
	Anta	fi	11	24	35
		%	1,4%	2,5%	2,0%
	Apurímac	fi	7	17	24
		%	0,9%	1,8%	1,4%
	Arequipa	fi	3	0	3
		%	0,4%	0,0%	0,2%
	Ayacucho	fi	1	1	2
		%	0,1%	0,1%	0,1%
	Calca	fi	8	8	16
		%	1,0%	0,8%	0,9%
	Canas	fi	2	7	9
		%	0,3%	0,7%	0,5%
	Canchis	fi	21	180	201
		%	2,7%	18,6%	11,6%
	Chiclayo	fi	0	2	2
		%	0,0%	0,2%	0,1%
	Chumbivilcas	fi	5	0	5
		%	0,6%	0,0%	0,3%
	Cusco	fi	656	666	1322
		%	85,2%	68,8%	76,1%
	Espinar	fi	1	7	8
		%	0,1%	0,7%	0,5%
	La convención	fi	4	5	9
		%	0,5%	0,5%	0,5%

	Lima	fi	1	4	5
		%	0,1%	0,4%	0,3%
	Madre d Dios	fi	1	2	3
		%	0,1%	0,2%	0,2%
	Paruro	fi	7	4	11
		%	0,9%	0,4%	0,6%
	Paucartambo	fi	5	10	15
		%	0,6%	1,0%	0,9%
	Puno	fi	6	3	9
		%	0,8%	0,3%	0,5%
	Quispicanchis	fi	13	10	23
		%	1,7%	1,0%	1,3%
	Urubamba	fi	11	14	25
		%	1,4%	1,4%	1,4%
Total		fi	770	968	1738
		%	100,0%	100,0%	100,0%

Tanto en el clúster 1 como en el clúster 2 se encuentran los estudiantes ingresantes que provienen en su mayoría de la ciudad del Cusco, es de notar que en el clúster 1 donde están las ingresantes de sexo femenino casi la totalidad son de Cusco, sin embargo, en el clúster 2 que es mayormente de sexo masculino hay presencia de ingresantes de la ciudad del Cusco, seguido de ingresantes de Canchis, Anta, Apurímac.

Figura 16. Clúster 1 para ingresantes

En resumen, en el clúster 1, se evidencia es un grupo donde están en su mayoría los ingresantes de sexo femenino, el 39.9% ingresantes al grupo D, el 85.2% son provenientes de la ciudad del Cusco, el 75.4% de colegios nacionales, en cuanto a nota promedio es de 12.45 y la edad promedio es de 20.73 años.

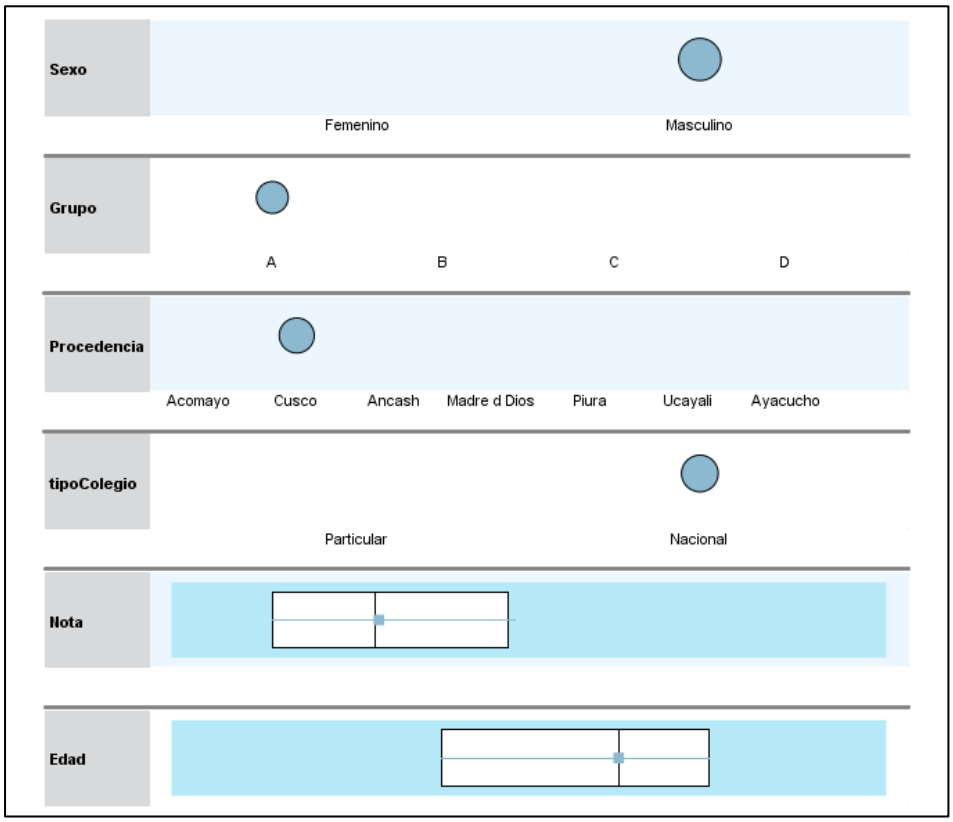


Figura 17. Clúster 2 para ingresantes

En resumen, en el clúster 2, se evidencia es un grupo donde están en su mayoría los ingresantes de sexo masculino, el 57% ingresantes al grupo A, el 68.8% son provenientes de la ciudad del Cusco, con porcentajes menores provienen de Canchis, Anta, Apurímac y demás provincias, el 70% de colegios nacionales, en cuanto a nota promedio es de 12.48 y la edad promedio es de 20.75 años.

DISCUSIÓN DE RESULTADOS

(Huapaya, Lizarralde, & Arona, 2011) indica que al analizar el nivel del conocimiento de los estudiantes encontró tres perfiles: individual, colectivo y colaborativo, en nuestro caso la investigación nos brindó 4 perfiles de alumnos postulantes los cuales tienen diferencias debido a sus características.

(Zuniga-Jara, Zuniga-Soria, & Soria-Barreto, 2022) conglomero las carreras profesionales de medicina de distintas universidades usando un enfoque basado en dendrogramas identifico agrupamientos. Después de analizar 11 variables, 5 asociadas a los estudiantes y otras 6 asociadas a la institución o universidad. Como resultado, se obtuvieron dos dimensiones de clasificación: 1) un proxy de la calidad de las instituciones y de sus estudiantes, y 2) un proxy del costo anual y del perfil socioeconómico de los estudiantes. A un primer nivel de disimilitud aparecen dos grandes grupos de carreras de medicina: 1) perfil tradicional, regional, con mejores indicadores de calidad institucional en promedio, y 2) perfil de instituciones privadas jóvenes, con casa central en Santiago de Chile y con estudiantes de mayor nivel socioeconómico. Se concluye que es posible caracterizar las carreras de medicina; de la misma manera que nosotros utilizamos una técnica de clustering (bietápico) para poder desarrollar nuestro propósito, si bien es cierto nosotros no usamos los dendogramas por la cantidad de sujetos a evaluar, pero el criterio de agrupar es el mismo y los resultados difieren puesto que a nosotros nos resulto 4 grupos de perfiles de estudiantes.

En (Arora, Deepali, & Varshney, 2016) se aplicaron las técnicas más populares K-Means y K-Medoids sus resultados de la comparación muestran que el tiempo empleado en la selección de los valores iniciales y la complejidad espacial de la

superposición del clúster es mucho mejor en K-Medoids que en K-Means. Además, K-Medoids es mejor en términos de tiempo de ejecución, no sensible a valores atípicos y reduce el ruido en comparación con KMeans, ya que minimiza la suma de las diferencias de los objetos de datos, estos resultados se asemejan a los de (Arbin, Suhailayani, Zafirah, & Othman, 2015); Nosotros no compartimos la misma conclusión debido que al comparar los algoritmos bietapico, PAM y CLARA logramos un mejor índice de cohesión y separación con el algoritmo mietapico esto debido a la gran cantidad de datos que nosotros utilizamos y el hecho de tener tanto variables cualitativas como cuantitativas hizo que este algoritmo tenga un mejor performance

(Chavez L. , 2020) cuando caracterizo el perfil del ingresante de una universidad pública aplicando algoritmos clustering k-prototypes y k-medoides", lograron identificar 3 tipos de alumnos: Ingresante previsto, Ingresante en proceso y el Ingresante en inicio; cada uno con características peculiares, las cuales permitirán a los responsables de las políticas educativas y en especial a los profesores consejeros saber el tipo de alumno que tienen a su cargo desde que ingresa a la universidad y empezar con ello políticas educativas como el emprendimiento del acompañamiento especializado, sistemático e integral; buscando la realización del paradigma del aprendizaje que la universidad se ha propuesto en su Modelo Educativo; la diferencia con nuestro estudio es que nosotros analizamos a todos los postulantes mientras que Chávez analizo solo a los ingresantes, nosotros obtuvimos 4 perfiles de postulantes y el método usado fue el bietápico y no el K-medoides.

CONCLUSIONES

1. Se han utilizado 3 algoritmos para poder conglomerar a los individuos, con el algoritmo (bietápico) dio un índice de silueta y cohesión en 0.37 donde el numero ideal de clusters fue 4; así mismo en los algoritmos de PAM (Partitioning Around Medoids) y CLARA (Clustering Large Applications) en el programa R, se calculó una matriz de distancias con la metodología de distancias mixtas de gower y con ello poder usar los otros algoritmos obteniéndose índices de 0.2169 con 4 clústeres en PAM y en el algoritmo CLARA se obtuvo 0.3516 con 8 clústeres, En tanto se decidió utilizar el algoritmo bietápico por tener mejor medida de silueta de cohesión y separación.

2. Se observa que en el conglomerado o Clúster 1, tiene prevalencia de alumnos que no ingresaron cuando postularon, su edad promedio es 18.49 años y su nota promedio fue de 7.44, su procedencia en su mayoría es del Cusco, de sexo femenino, procedencia de colegio nacional y mayormente postulan al grupo D. en el clúster 2, tiene prevalencia de alumnos que ingresaron cuando postularon, su edad promedio es 20.7 años y su nota promedio fue de 12.58, su procedencia en su mayoría es del Cusco, de sexo masculino, procedencia de colegio nacional y mayormente postulan al grupo A, obviamente que hay ingresantes a todos los grupos, pero hay más ingresantes al grupo A, donde se sitúa la mayor cantidad de escuelas profesionales en la universidad; en el clúster 3, tiene prevalencia de alumnos que no ingresaron cuando postularon, su edad promedio es 18.49 años y su nota promedio fue de 7.67, su procedencia en su mayoría es del Cusco, de sexo masculino, procedencia de

colegio particular y mayormente postulan al grupo A; en el conglomerado 4, tiene prevalencia de alumnos que no ingresaron cuando postularon, su edad promedio es 18.50 años y su nota promedio fue de 7.75, su procedencia en su mayoría es del Cusco, de sexo femenino, procedencia de colegio particular y mayormente postulan al grupo A, donde se sitúa la mayor cantidad de escuelas profesionales en la universidad.

3. Se pudo observar después de realizar el análisis de validación de los clústeres mediante pruebas de correspondencias, pruebas de independencia chi cuadrado de Pearson y pruebas de kruskall Wallis, que todas las variables consideradas en el estudio son importantes a la hora de construir los perfiles de los postulantes a una universidad como son, la nota del examen, la edad, el grupo al que postula, la procedencia, sexo, tipo de colegio y grupo al que postula.

4. Realizado el clúster solo para los ingresantes, se pudo observar que se agruparon en dos perfiles, en el clúster 1, se evidencio que es un grupo de sexo femenino, el 39.9% ingresantes al grupo D, el 85.2% son provenientes de la ciudad del Cusco, el 75.4% de colegios nacionales, en cuanto a nota promedio es de 12.45 y la edad promedio es de 20.73 años; mientras que en el clúster 2, se evidencia un grupo masculino, el 57% ingresantes al grupo A, el 68.8% son provenientes de la ciudad del Cusco, con porcentajes menores provienen de Canchis, Anta, Apurímac y demás provincias, el 70% provienen de colegios nacionales, en cuanto a nota promedio es de 12.48 y la edad promedio es de 20.75 años; se propone trabajar más en la promoción de la universidad a los estudiantes de provincias y dar a conocer que carreras no cubren vacantes y cuál es su campo de acción.

RECOMENDACIONES

Para posteriores investigaciones se recomienda:

1. Investigar la efectividad en la formación de estudiantes de educación básica regular que pretenden ingresar a una Universidad Nacional como lo es la Universidad Nacional San Antonio Abad del Cusco, debido a las bajas notas registradas.
2. En los colegios fortalecer el tema psicológico específicamente para abordar la ansiedad frente a las evaluaciones y la ansiedad social en materias como matemáticas y lenguaje, con el objetivo de mejorar su desempeño de los estudiantes a la hora de postular a la universidad.
3. A la universidad se recomienda tener en cuenta los perfiles encontrados en la presente investigación y así poder tener mayor amplitud a la hora de convocar a exámenes de admisión.
4. Implementar un programa de orientación específico para estudiantes de provincias que brinde información detallada sobre el proceso de postulación, becas disponibles y recursos de apoyo académico y emocional. Establecer una ruta para acceder a fondos de becas destinado exclusivamente a estudiantes de provincias, con criterios que tengan en cuenta las circunstancias económicas y las necesidades específicas de esta población; por otro lado, desarrollar campañas de concientización que destaquen los beneficios de estudiar en una universidad nacional, haciendo hincapié en las oportunidades académicas, la red de contactos y el prestigio asociado.

REFERENCIAS BIBLIOGRÁFICAS

1. Arbin, N., Suhailayani, N., Zafirah, N., & Othman, Z. (2015). Comparative Analysis between K-Means and K-Medoids for Statistical Clustering. *3rd International Conference on Artificial Intelligence, Modelling & Simulation (AIMS)*, (págs. 117-121). doi:10.1109/AIMS.2015.82.

2. Areválo, J., & Pérez Gonzales, S. (2018). El análisis de conglomerados como herramienta para evaluar el rendimiento académico: una experiencia en la universidad. *Revista Espacios*, 30-37.

3. Arora, P., Deepali, & Varshney, S. (2016). Analysis of K-Means and K-Medoids Algorithm for Big Data. *Physics Procedia, 78*, 507-512.

4. Cerda, G., Pérez, C., Aguilar, M., & Aragón, E. (2018). Algunos factores asociados al desempeño académico en matemáticas y sus proyecciones en la formación docente. *Educ Pesqui, São Paulo*, 1-19. doi:https://doi.org/10.1590/S1678-4634201706155233

5. Chavez, L. (2020). *Caracterización del perfil del ingresante de una Universidad Pública aplicando algoritmos clustering K-Prototypes y K- Medoids.* Lima, Perú: Universidad Nacional Agraria La Molina.

6. Chavez, L. (2020). *Caracterización del perfil del ingresante de una universidad pública aplicando algoritmos clustering K-prototypes y k-medoides.* Lima, Perú: Universidad Nacional Agraria la Molina.

7. Clavijo, R., & Bautista-Cerro, M. (2020). La educación inclusiva. Análisis y reflexiones en la educación superior ecuatoriana. *Alteridad, 15*(1), 113-124. doi:https://doi.org/10.17163/alt.v15n1.2020.09

8. Cuadras, C. (2019). *Nuevos métodos de análisis multivariante.* Barcelona, España: CMC Editions.

9. Cuenca, R. (2015). *La educación universitaria en el Perú: democracia, expansión y desigualdades.* Lima: IEP Instituto de Estudios Peruanos.

10. Elguera, R. (2018). *"Segmentación de clientes de un casino utilizando el algoritmo partición alrededor de medoides (PAM) con datos Mixtos.* Lima, Perú: Universidad Nacional Agraria La Molina.

11. Everitt, B. (2011). *Cluster analysis 5th Editions.* Londres: Wiley.

12. Everitt, B., & Torsten Hothorn. (2011). *An Introduction to Applied Multivariate Analysis with R.* Londres: Springer.

13. Gairín, J., & Palmeros, G. (2018). *Políticas y prácticas para la equidad en la educación superior.* Madrid: Wolters Kluwer España S.A.

14. Huapaya, C., Lizarralde, F., & Arona, G. (2011). Propuesta para Construir Perfiles Cognitivos en la Evaluación del Estudiante. *XIII Workshop de Investigadores en Ciencias de la Computación* (págs. 920-924). Argentina: Red de Universidades con Carreras en Informática (RedUNCI).

15. Ibarra, M., & Michalus, J. (2010). Análisis del rendimiento academico mediante un modelo logit. *Ingeniería Industrial*, 47-56.

16. Kassambara, A. (2017). *Practical Guide to cluster Analysis in R.* -: STHDA.

17. Luy, C., Salvatierra, A., Rengifo, R., & Rivera, J. (2021). El clúster jerárquico en la segmentación de los logros del aprendizaje de matemática y comunicación. *Laplage em Revista (International), 7*(3), 166-176. doi:https://doi.org/10.24115/S2446-6220202173C1513p.166-176

18. Manrique, C. (2016). *Segmentación de clientes de Corporación Lindley de la región Lima mediante el análisis Cluster Bietápico en octubre de 2016.* Lima, Perú: Universidad Nacional Mayor de San Marcos.

19. Moreira, T. (2009). Factores endógenos y exógenos asociados al rendimiento en matemática: un análisis multinivel. *Educación*, 61-80.

20. Ng, R., & Han, J. (2002). Clarans: A method for clustering objects for spatial data mining. *IEEE transactions on knowledge and data engineering, 14*(5), 1003-1016.

21. Paredes, E. (2017). De la evaluación diagnóstica a la evaluación sumativa: logros y fracasos en los aprendizajes. *Congreso internacional de educación evaluación 2016*, 1507-1518.

22. Pastrán, L., & Gongora, S. (2021). *Algoritmo de selección y validación del metodo de clusterizacion optimo para datos no supervisados.* Bogotá D.C.: Universidad Tecnológica de Pereira.

23. Peña, D. (2002). *Análisis de datos multivariantes.* Madrid, España: McGraw Hill Academic Press.

24. Pérez Morales, J. (2007). *La evaluación como instrumento de mejora de la calidad del aprendizaje. Propuesta de intervención psicopedagógica para el aprendizaje del idioma Ingles.* Cataluña, España: Universidad de Girona.

25. Rodríguez, C. (2011). *Componentes principales y regresión logística: Analizando el rendimiento académico de los estudiantes en matemática prebásica.* Mayaguez: Universidad de Puerto Rico.

26. Rojas, A. (2020). Factores que afectan el ingreso a la educación superior de los egresados de la I.E. Silvania. *Revista PACA 10*, 51-64.

27. Sankar, R. (2011). Customer Data Clustering Using Data Mining Technique. *International Journal of Database Management Systems, 3*(4), 1-11.

28. Tonconi, C. (2021). *"Identificación de perfiles de los Centros de educación técnico-productiva públicos usando indicadores de condiciones básicas de calidad mediante Clúster Bietápico.* Lima, Perú: Universidad Nacional Agraria La Molina.

29. UNSAAC. (2018). *Reglamento de admisión a la UNSAAC.* Cusco: Resolución CU-307-2018-UNSAAC.

30. Zamora-Aray, J. (2020). Las actitudes hacia la matemática, el desarrollo social, el nivel educativo de la madre y la autoeficacia como factores asociados al rendimiento académico en la matemática. *Uniciencia*, 74-87.

31. Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: An Efficient Data Clustering Method for Very Large. *ACM SIGMOD Record, 25*(2), 103-114. doi:https://doi.org/10.1145/235968.233324

32. Zuniga-Jara, S., Zuniga-Soria, J., & Soria-Barreto, K. (2022). Taxonomía de las carreras de medicina en Chile. *Formación universitaria, 15*(6), 59-70. doi:http://dx.doi.org/10.4067/S0718-50062022000600059

ANEXOS

ANEXO 1

MATRIZ DE CONSISTENCIA

APLICACIONES DE LOS MÉTODOS DE ANÁLISIS DE CLÚSTER Y CORRESPONDENCIA EN EL ESTUDIO DE RESULTADOS DE EXAMEN DE ADMISION DE LA UNSAAC, 2022

PROBLEMA GENERAL	**OBJETIVO GENERAL**	**HIPOTESIS GENERAL**	**VARIABLE DE ESTUDIO**
¿Cuál es el perfil de los conglomerados de estudiantes postulantes a la UNSAAC obtenidos de las aplicaciones de los métodos de análisis de clúster y correspondencia? **PROBLEMAS ESPECÍFICOS** a) ¿Qué método de clúster es el óptimo para determinar el perfil de los estudiantes postulantes a la UNSAAC? b) ¿Qué variables presentan mayor repercusión en la elaboración de conglomerados de estudiantes postulantes a la UNSAAC? c) ¿Qué perfil presentan los estudiantes con mejores puntajes en el examen de admisión de la UNSAAC, obtenidos del análisis de correspondencia?	Analizar el perfil de los conglomerados de estudiantes postulantes a la UNSAAC obtenidos de las aplicaciones de los métodos de análisis de clúster y correspondencia. **OBJETIVOS ESPECIFICOS** a) Comparar los métodos de Conglomerados en base a los índices de validación para determinar el perfil de los estudiantes postulantes A LA UNSAAC b) Determinar las variables presentan mayor repercusión en la elaboración de conglomerados de estudiantes postulantes a la UNSAAAC c) Describir el perfil presentan los estudiantes con mejores puntajes en el examen de admisión de la UNSAAC, obtenidos del análisis de correspondencia.	Existen cuatro perfiles o conglomerados de estudiantes postulantes a la UNSAAC obtenidos de las aplicaciones de los métodos de análisis de clúster y correspondencia. **HIPÓTESIS ESPECÍFICOS** a) El método partition around medoids (PAM) presenta mejores índices de validación para determinar el perfil de los estudiantes postulantes A LA UNSAAC b) Las variables procedencia y condiciones económicas presentan mayor repercusión en la elaboración de conglomerados de estudiantes postulantes a la UNSAAAC c) Los estudiantes con mejores puntajes en el examen de admisión de la UNSAAC presentan edades inferiores al promedio, condiciones económicas bajas y proceden de colegios estatales.	Variable: perfil de postulantes a la universidad Nacional de San Antonio Abad del Cusco **VARIABLE INTERVINIENTES** Edad Tipo de colegio Procedencia Nota Condición (ingreso) Sexo Grupo postula

ANEXO 2

CÓDIGO DE R

```
library(foreign)
datos=read.spss("data_final.sav",
                    use.value.labels=TRUE,          max.value.labels=Inf,
to.data.frame=TRUE)
attach(datos)

# Tabla de medias
med<-aggregate(x = datos[,2:3],by = list(datos$Cluster),FUN = mean)
med

# Describir variables
par(mfrow=c(1,2))
for (i in 2:3) {
  boxplot(datos[,i]~datos$Cluster, main=names(datos[i]), type="l")
}
par(mfrow=c(1,1))

library(ggpubr)
my_cols <- c("#0D0887FF", "#6A00A8FF", "#B12A90FF",
             "#E16462FF", "#FCA636FF", "#F0F921FF")
data= table(Grupo,)
data <- prop.table(data, margin = 2)
data= as.data.frame(data)
ggballoonplot(data )

ggballoonplot(data, x = "Grupo", y = "TSC_353",
              size = "Freq", fill = "Freq") +
  scale_fill_gradientn(colors = my_cols) +
  guides(size = FALSE)

# sexo vs cluster
data= table(Sexo,TSC_353)
data <- round(prop.table(data, margin = 2),3)*100
data= as.data.frame(data)
ggballoonplot(data )

ggballoonplot(data, x = "Sexo", y = "TSC_353",
              size = "Freq", fill = "Freq",show.label = TRUE) +
  scale_fill_gradientn(colors = my_cols) +
  guides(size = FALSE)
```

```
### Comparar entre distintos algoritmos
library(caret)
set.seed(3456)
trainIndex <- createDataPartition(datos$Condicion, p = 0.8,
                                  list = FALSE,
                                  times = 1)
head(trainIndex)
data_nueva <- datos[ trainIndex,]
data_nueva = data_nueva[,2:8]
set.seed(1680)
# calculo de las distancias por medio del metodo Gower de la funcion daisy

library(cluster)
gower_dist <- daisy(data_nueva,metric = "gower")

# PAM CLUSTER
library(fpc)
a=pamk(gower_dist,criterion="asw")
a$crit
a$nc

# CLARA CLUSTER
b=pamk(gower_dist,criterion="asw",usepam=FALSE)
b$crit
b$nc

# metodo jerarquico
d=res=agnes(gower_dist,method="ward")
d$ac
```