UNIVERSIDAD NACIONAL DE SAN ANTONIO ABAD DEL CUSCO

ESCUELA DE POSGRADO

MAESTRÍA EN ESTADÍSTICA

TESIS

PERFILES DE RENDIMIENTO ACADÉMICO EN LA EVALUACIÓN CENSAL DE ESTUDIANTES EN EL DEPARTAMENTO DE CUSCO USANDO MINERÍA DE DATOS

PARA OPTAR AL GRADO ACADÉMICO DE MAESTRO EN ESTADÍSTICA

AUTOR
Br. MARCO ANTONIO TORRES VALVERDE

ASESOR
Mtro. DAVID RODRIGUEZ QUISPE
CÓDIGO ORCID: 0009-0002-9166-5579

CUSCO - PERÚ

2026

Universidad Nacional de San Antonio Abad del Cusco

INFORME DE SIMILITUD

(Aprobado por Resolución Nro.CU-321-2025-UNSAAC)

El que suscribe, **el Asesor** DAVID RODRIGUEZ QUISPE**,** quien aplica el software de detección de similitud al trabajo de investigación/tesis titulada: PERFILES DE RENDIMIENTO ACADÉMICO EN LA EVALUACIÓN CENSAL DE ESTUDIANTES EN EL DEPARTAMENTO DE CUSCO USANDO MINERÍA DE DATOS

presentado por: Br. MARCO ANTONIO TORRES VALVERDE con DNI Nro.: 46373563

para optar el Título Profesional/ Grado Académico de MAESTRO EN ESTADÍSTICA

Informo que el trabajo de investigación ha sido sometido a revisión por TURNITIN DOS veces, mediante el software similitud, conforme al Artículo 6° del Reglamento para Uso del Sistema de Detección de Similitud en la UNSAAC y de la evaluación de originalidad se tiene un porcentaje de 9% de similitud.

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 sobrepasa el porcentaje aceptado de similitud.	9%
Del 11 al 30%	Devolver al usuario para las subsanaciones.	
Mayores a 31%	El responsable de la revisión del documento emite un informe al inmediato jerárquico, conforme al reglamento, quien a su vez eleva el informe al Vicerrectorado de Investigación 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 de Detección de Similitud.

Cusco, 22 de agosto de 2026

Firma

Post firma: DAVID RODRIGUEZ QUISPE

Nro. de DNI: 42484199

ORCID del Asesor: 0009-0002-9166-5579

Se adjunta:

1. Reporte Generado por el Sistema de Detección de Similitud.
2. Enlace del Reporte Generado por el Sistema Detección de Similitud: https://unsaac.turnitin.com/viewer/submissions/oid:27259:616356146?locale=es-MX

MARCO ANTONIO TORRES VALVERDE

RENDIMIENTO ACADÉMICO EN LA EVALUACIÓN CENSAL DE ESTUDIANTES EN EL DEPARTAMENTO DE CUSCO USANDO MI...

Universidad Nacional San Antonio Abad del Cusco

Detalles del documento

Identificador de la entrega
trn:oid:::27259:616356146

Fecha de entrega
22 ago 2026, 7:49 p.m. GMT-5

Fecha de descarga
22 ago 2026, 7:58 p.m. GMT-5

Nombre del archivo
RENDIMIENTO ACADÉMICO EN LA EVALUACIÓN CENSAL DE ESTUDIANTES EN EL DEPARTAMENTO ....pdf

Tamaño del archivo
3.5 MB

135 páginas

25.141 palabras

143.408 caracteres

9% Similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para ca...

Filtrado desde el informe

- Bibliografía
- Texto citado
- Texto mencionado
- Coincidencias menores (menos de 12 palabras)

Fuentes principales

8% Fuentes de Internet
2% Publicaciones
6% Trabajos entregados (trabajos del estudiante)

Marcas de integridad

N.º de alerta de integridad para revisión

Texto oculto
5 caracteres sospechosos en N.º de páginas
El texto es alterado para mezclarse con el fondo blanco del documento.

Los algoritmos de nuestro sistema analizan un documento en profundidad para buscar inconsistencias que permitirían distinguirlo de una entrega normal. Si advertimos algo extraño, lo marcamos como una alerta para que pueda revisarlo.

Una marca de alerta no es necesariamente un indicador de problemas. Sin embargo, recomendamos que preste atención y la revise.

UNIVERSIDAD NACIONAL DE SAN ANTONIO ABAD DEL CUSCO
ESCUELA DE POSGRADO

INFORME DE LEVANTAMIENTO DE OBSERVACIONES A TESIS

Dr. LEONCIO SOLIS QUISPE, Director de la Escuela de Posgrado, nos dirigimos a usted en condición de integrantes del jurado evaluador de la tesis intitulada PERFILES DE RENDIMIENTO ACADÉMICO EN LA EVALUACIÓN CENSAL DE ESTUDIANTES EN EL DEPARTAMENTO DE CUSCO USANDO MINERÍA DE DATOS del / de la Br. MARCO ANTONIO TORRES VALVERDE. Hacemos de su conocimiento que el (la) sustentante ha cumplido con el levantamiento de las observaciones realizadas por el Jurado el día CINCO DE JUNIO DEL 2026.

Es todo cuanto informamos a usted fin de que se prosiga con los trámites para el otorgamiento del grado académico de MAESTRO EN ESTADÍSTICA.

Cusco, 07 AGOSTO DEL 2026

MTRO. YELI SANDRA SANCHEZ FARFAN
Primer Replicante

MTRO. CARMEN LAIME LLOCLLA
Segundo Replicante

DRA. NELLY MARIA SALAZAR PEÑA
Primer Dictaminante

MTRO. GILBERT ALBERTO MONZON DIAZ
Segundo Dictaminante

DEDICATORIA

Dedico este trabajo en primer lugar a Dios por darme la vida y salud, a mis padres Vicente y Encarnación. Por el amor inmenso que me dieron, por sus sacrificios incansables, lucha constante por la familia. En seguida a mi esposa Nilda e hijos Kritell Illary y Zedrich, por siempre estar ahí conmigo en cada caída, en cada fracaso y también en los momentos más felices de mi vida. También a mis hermanos, Rocio, Yhon, Nelson, Nohemi, Marcia y Fabricio. Por siempre estar ahí conmigo, y finalmente a toda mi familia Torres y Valverde. Por siempre estar juntos a mí y por las palabras de aliento y haberme acompañado en este camino, por la alegría que traen a mi vida, ustedes representan la familia que me sostiene y me impulsa a ser mejor cada día y ser una fuente de motivación, del mismo modo a las personas que creyeron en mí, para que este sueño se haga realidad, a los docentes de la escuela de posgrado de la UNSAAC, por guiarme con paciencia y por compartir su sabiduría, por todo lo que soy, por todo lo que me han dado, gracias de corazón.

AGRADECIMIENTO

Quisiera comenzar expresando mi más sincero agradecimiento a mí asesor de tesis, el Mtro. David Rodríguez Quispe, cuya experiencia, paciencia y apoyo constante fueron fundamentales para la realización de este trabajo. Su guía no solo me proporcionó claridad académica, sino también motivación en momentos de duda. Su confianza en mí me impulsó a seguir adelante y superar los desafíos.

A mi familia, especialmente a mis padres, les agradezco profundamente su amor incondicional y su apoyo constante. A mi esposa e Hijos, Su fe en mí ha sido el motor que me permitió completar este camino. A mis hermanos, por sus palabras de aliento, Sin ustedes, este logro no habría sido posible.

A la Universidad Nacional de San Antonio Abad del Cusco, gracias por brindarme la oportunidad de crecer académica y profesionalmente.

A mis amigos y compañeros, gracias por su compañía y apoyo en los momentos de estrés y alegría. Ustedes fueron mi red de contención y su amistad me ayudó a mantener el ánimo en los momentos más duros. Cada uno de ustedes contribuyó a que este proceso fuera más llevadero y significativo.

A todos, gracias por ser parte de este sueño tan anhelado para mí.

Índice general

Índice de tablas

Índice de figuras

Resumen

El rendimiento escolar es un indicador crítico de la calidad y equidad educativa en el Perú; sin embargo, los reportes promedio oficiales de la Evaluación Censal de Estudiantes (ECE) ocultan las brechas complejas y la profunda heterogeneidad socioeconómica, cultural y territorial del departamento de Cusco, dificultando el diseño de intervenciones focalizadas. Frente a esta problemática, el objetivo fue identificar los perfiles de rendimiento en lectura de los estudiantes de 2.º grado de educación secundaria en la ECE del departamento de Cusco, usando minería de datos. El estudio adoptó un enfoque cuantitativo, de nivel exploratorio-descriptivo y diseño no experimental-transversal, empleando información censal de la ECE 2019. El análisis combinó el Análisis de Correspondencias Múltiples (ACM) y el Clustering Jerárquico sobre Componentes Principales (HCPC) en el software R. Los resultados evidenciaron perfiles estudiantiles diferenciados. El perfil de bajo rendimiento, correspondiente a los niveles Previo al Inicio e Inicio, se concentró principalmente en estudiantes con nivel socioeconómico muy bajo, lengua materna quechua y asistencia a instituciones educativas de gestión estatal, tanto en ámbitos rurales como urbanos. En contraste, el perfil de rendimiento aceptable, integrado por los niveles En Proceso y Satisfactorio, se asoció con estudiantes de niveles socioeconómicos medios y altos. Se concluye que el rendimiento en lectura se vincula con la interacción conjunta del nivel socioeconómico, la lengua materna, el área geográfica y la gestión de la institución educativa, evidenciando que el área urbana no garantiza un rendimiento favorable cuando persisten condiciones de vulnerabilidad socioeconómica y barreras lingüísticas.

Palabras clave: Rendimiento, Nivel Socioeconómico, Lengua, Gestión.

Abstract

Academic performance is a critical indicator of educational quality and equity in Peru; however, the official average reports of the Census Evaluation of Students (ECE) conceal the complex gaps and the profound socioeconomic, cultural, and territorial heterogeneity of the department of Cusco, hindering the design of targeted interventions. In response to this issue, the objective was to identify reading performance profiles among second-grade secondary school students assessed through the ECE in the department of Cusco using data mining techniques. The study adopted a quantitative approach, with an exploratory-descriptive level and a non-experimental, cross-sectional design, using census data from the 2019 ECE. The analysis combined Multiple Correspondence Analysis (MCA) and Hierarchical Clustering on Principal Components (HCPC) using R software. The results revealed differentiated student profiles. The low-performance profile, corresponding to the levels Before Beginning and Beginning, was concentrated mainly among students with a very low socioeconomic status, Quechua as their mother tongue, and attendance at public educational institutions in both rural and urban areas. In contrast, the acceptable-performance profile, comprising the levels In Progress and Satisfactory, was associated with students from middle- and high-socioeconomic backgrounds. It is concluded that reading performance is linked to the combined interaction of socioeconomic status, mother tongue, geographical area, and school management, demonstrating that urban areas do not guarantee favorable performance when conditions of socioeconomic vulnerability and linguistic barriers persist.

Keywords: Reading, Socioeconomic Status, Longue, Management.

INTRODUCCIÓN

La educación constituye uno de los pilares fundamentales para el desarrollo de un país, ya que fortalece las capacidades de la población. El rendimiento escolar constituye un indicador fundamental para evaluar tanto la calidad del proceso educativo como la equidad en el acceso a las oportunidades de aprendizaje.

En el Perú, el Ministerio de Educación realiza un seguimiento periódico de estos logros mediante la Evaluación Censal de Estudiantes (ECE). Sin embargo, los informes oficiales suelen presentar promedios generales o estadísticas agregadas que, con frecuencia, ocultan las particularidades propias de cada contexto local. En un territorio con una marcada diversidad social, cultural y geográfica, como el departamento de Cusco, un análisis sustentado únicamente en valores promedio restringe la identificación de las brechas existentes entre estudiantes que pertenecen a realidades diferenciadas.

Ante esta situación, las herramientas estadísticas tradicionales, basadas en promedios o en análisis de carácter convencional, resultan insuficientes para representar la complejidad y heterogeneidad de los datos. En respuesta a esta limitación, la presente investigación propone la aplicación de técnicas de minería de datos y aprendizaje no supervisado mediante la integración secuencial del Análisis de Correspondencias Múltiples (ACM) y el Clustering Jerárquico sobre Componentes Principales (HCPC). Estas metodologías posibilitan reducir la dimensionalidad de las variables cualitativas de la ECE para favorecer su representación espacial y, posteriormente, identificar agrupamientos estables de estudiantes que comparten características y niveles de rendimiento similares.

En este marco, el estudio tuvo como propósito analizar las relaciones e interacciones entre el nivel socioeconómico, la lengua materna, el área

geográfica y la gestión del colegio con el rendimiento en lectura de los estudiantes de segundo grado de secundaria evaluados en el departamento de Cusco. A través de este análisis, se identificaron y caracterizaron tanto los perfiles de rendimiento aceptable como los perfiles de rendimiento bajo en lectura.

Los resultados de esta investigación ofrecen evidencia científica y metodológica útil para la gestión de las políticas públicas de la región. Al identificar con precisión en qué grupos y UGEL se concentran las mayores desventajas en lectura, la Dirección Regional de Educación (DRE) y las UGEL del Cusco cuentan con una herramienta clave para diseñar intervenciones educativas focalizadas, orientando los recursos materiales y los programas de educación intercultural bilingüe de manera urgente hacia los perfiles estudiantiles de mayor vulnerabilidad.

I. PLANTEAMIENTO DEL PROBLEMA

1.1. Situación problemática

El sistema educativo peruano enfrenta diversas limitaciones estructurales y persistentes desafíos en términos de calidad y equidad, los cuales se reflejan en los resultados de evaluaciones nacionales e internacionales del aprendizaje. Estas dificultades impactan directamente en el rendimiento académico de los estudiantes y evidencian la necesidad de fortalecer los procesos de enseñanza y aprendizaje en distintos niveles del sistema educativo.

En el ámbito internacional, los resultados del Programa para la Evaluación Internacional de Estudiantes (PISA) 2018, menos de la mitad de los estudiantes de 15 años alcanzo el nivel mínimo aceptable (nivel 2 o superior) en las competencias de lectura y ciencias; y solo el 40% lo hizo en matemáticas. Los estudiantes con desempeño excelente (niveles 5 y 6) fueron menos del 1% en todas las áreas.

Es importante precisar que las evaluaciones PISA y la Evaluación Censal de Estudiantes (ECE) responden a marcos conceptuales, poblaciones objetivo y propósitos distintos, por lo que sus resultados no son directamente comparables. Mientras que PISA evalúa a estudiantes de 15 años mediante una muestra representativa internacional, la ECE es una evaluación nacional y censal, alineada al currículo peruano y aplicada a grados específicos del sistema educativo. En este sentido, los resultados de PISA se emplean únicamente como un referente contextual del panorama educativo general del país, mientras que el análisis empírico del rendimiento académico

desarrollado en la presente investigación se fundamenta exclusivamente en la información proveniente de la ECE.

En el ámbito nacional, la Evaluación Censal de Estudiantes 2019, aplicada por el Ministerio de Educación, evidenció que, en el área de Matemática para el 2.º grado de educación secundaria, únicamente el 17.70% de los estudiantes alcanzó el nivel Satisfactorio, en tanto que alrededor del 17.30% se ubicó en el nivel En Proceso. Estos resultados muestran que una proporción importante de estudiantes no alcanza los aprendizajes esperados para su grado, lo que pone de manifiesto la necesidad de profundizar en el análisis de los factores asociados al rendimiento académico.

Bajo este panorama, resulta pertinente centrar el análisis en contextos regionales específicos, debido a que el rendimiento académico presenta un comportamiento heterogéneo a lo largo del territorio peruano. En este sentido, el departamento de Cusco constituye un escenario de especial interés por su amplia diversidad geográfica, social y cultural, así como por la coexistencia de ámbitos urbanos y rurales donde persisten brechas educativas. Estas particularidades hacen de Cusco un contexto propicio para el estudio de patrones diferenciados de rendimiento académico y de las desigualdades educativas.

De igual manera, el 2.º grado de educación secundaria representa una etapa clave dentro de la trayectoria escolar, puesto que corresponde al proceso de consolidación de competencias esenciales en áreas como Matemática y Comprensión Lectora, además de preceder niveles de mayor exigencia académica. Por ello, el análisis del rendimiento en este grado permite identificar de manera temprana perfiles académicos diferenciados y los

factores asociados que podrían incidir en el desempeño posterior de los estudiantes.

Tradicionalmente, la evaluación del rendimiento académico se ha sustentado en puntajes globales o valores promedio. No obstante, este enfoque presenta limitaciones al no considerar la heterogeneidad de la población estudiantil ni la influencia de los diversos factores sociales, económicos, familiares y escolares que condicionan el desempeño académico. Como consecuencia, se dificulta la identificación de grupos de estudiantes con características semejantes y se restringe el diseño de intervenciones educativas focalizadas y sustentadas en evidencia.

En este contexto, surge la necesidad de identificar patrones y perfiles de rendimiento académico en los estudiantes de 2.º grado de educación secundaria que participaron en la Evaluación Censal de Estudiantes del departamento de Cusco, mediante el análisis integrado de variables académicas y contextuales. En dicha evaluación, el área de Lectura registró que aproximadamente el 22.70% de los estudiantes se ubicó en el nivel Previo al Inicio, el 43% en el nivel En Inicio, el 22.60% en el nivel En Proceso y únicamente el 11.70% alcanzó el nivel Satisfactorio. Por su parte, en el área de Matemáticas, el 35.80% se situó en el nivel Previo al Inicio, el 31.90% en el nivel En Inicio, el 16.10% en el nivel En Proceso y el 16.10% en el nivel Satisfactorio. Estas distribuciones muestran una elevada variabilidad en los niveles de logro dentro de la región.

La identificación de perfiles diferenciados de estudiantes con desempeños exitosos y no exitosos permitirá obtener una visión más precisa y

comprensiva de la situación educativa regional, superando el análisis basado únicamente en promedios.

Para superar estas limitaciones, la presente investigación propone un enfoque metodológico integrado de Minería de Datos y aprendizaje no supervisado, utilizando de manera combinada el Análisis de Correspondencias Múltiples (ACM) y el Clustering Jerárquico sobre Componentes Principales (HCPC). Esta integración metodológica permitirá, en primer lugar, reducir la alta dimensionalidad de las variables categóricas de la ECE para representarlas espacialmente en dimensiones continuas (ACM), y, en segundo lugar, agrupar a los estudiantes de manera estable en perfiles homogéneos de rendimiento y contexto (HCPC), descubriendo así estructuras y patrones latentes de desigualdad que los promedios globales suelen ocultar.

Los resultados obtenidos podrán constituir un insumo técnico relevante para el diseño de estrategias de gestión académica y políticas educativas orientadas a la mejora del rendimiento académico en el departamento de Cusco.

1.2. Formulación del problema

Problema general

¿Cuáles son los perfiles de rendimiento académico en lectura en los estudiantes de 2.º grado de educación secundaria en la Evaluación Censal de Estudiantes en el departamento de Cusco, usando minería de datos?

Problemas específicos

- ¿Cómo se relacionan el nivel socioeconómico, la lengua materna, el área y la gestión educativa con el rendimiento en lectura de los estudiantes del departamento de Cusco?
- ¿Qué características de nivel socioeconómico lengua materna, área y gestión distinguen a los estudiantes con un perfil de rendimiento bajo en lectura en el departamento de Cusco?

1.3. Justificación de la investigación

La Evaluación Censal de Estudiantes (ECE), aplicada por el Ministerio de Educación, constituye una de las principales fuentes de información para el monitoreo del rendimiento académico de los estudiantes del sistema educativo peruano. En particular, la ECE 2019, aplicada de manera censal a los estudiantes de 2.º grado de educación secundaria, proporciona información sobre el desempeño en la competencia de lectura, así como un conjunto amplio de variables socioeconómicas, académicas e institucionales. Los informes oficiales derivados de esta evaluación presentan, en general, estadísticas descriptivas agregadas a nivel nacional, regional y por características generales de los estudiantes y de las instituciones educativas.

Si bien estos reportes ofrecen una visión general del estado de los aprendizajes, su carácter predominantemente agregado restringe la comprensión de la heterogeneidad del rendimiento académico y dificulta la identificación de patrones complejos derivados de la interacción entre múltiples factores contextuales. En regiones con una elevada diversidad social, cultural y territorial, como el departamento de Cusco, un análisis

sustentado exclusivamente en promedios o porcentajes globales resulta insuficiente para comprender las desigualdades educativas existentes y orientar el diseño de intervenciones educativas focalizadas.

Desde la perspectiva metodológica, la presente investigación se sustenta en la aplicación de técnicas de Minería de Datos y análisis multivariado exploratorio, las cuales permiten identificar estructuras latentes y perfiles diferenciados de estudiantes a partir de grandes volúmenes de información censal. A diferencia de los enfoques estadísticos convencionales, orientados principalmente a comparaciones bivariadas o a modelos explicativos basados en supuestos específicos, las técnicas de clustering y el Análisis de Correspondencias Múltiples posibilitan el análisis simultáneo de múltiples variables categóricas, favoreciendo la exploración de la estructura conjunta de las características socioeconómicas, académicas e institucionales en relación con el rendimiento académico en lectura. De este modo, la investigación aporta una contribución metodológica al evidenciar la utilidad de estas técnicas para el análisis del rendimiento educativo en el ámbito regional.

Desde una perspectiva teórica, la presente investigación se fundamenta en el enfoque que concibe el rendimiento académico como un fenómeno multidimensional, resultado de la interacción entre factores individuales, familiares, escolares y contextuales. Bajo este enfoque, la identificación de perfiles de estudiantes mediante técnicas de agrupamiento facilita la comprensión de la heterogeneidad del desempeño académico y permite examinar las desigualdades educativas desde una perspectiva integral,

superando los enfoques fragmentados que analizan cada factor de forma independiente.

Asimismo, la investigación posee una importante justificación social, ya que sus resultados permiten visibilizar patrones persistentes de bajo rendimiento académico vinculados con condiciones estructurales de vulnerabilidad social, económica y territorial. En este sentido, la identificación de perfiles de estudiantes con bajo desempeño en la competencia de lectura proporciona evidencia útil para orientar el diseño de políticas públicas y estrategias de intervención más focalizadas, contribuyendo potencialmente a disminuir las brechas educativas y fortalecer la equidad del sistema educativo regional.

La elección del departamento de Cusco como ámbito de estudio responde a sus particularidades territoriales y socioculturales, entre las que destacan su pronunciada heterogeneidad geográfica, la coexistencia de zonas urbanas y rurales, y la presencia significativa de estudiantes cuya lengua materna es el quechua. Estas características convierten a Cusco en un escenario de especial relevancia para el análisis de las desigualdades educativas y la identificación de perfiles diferenciados de rendimiento académico, permitiendo generar evidencia empírica contextualizada que contribuya a la toma de decisiones en los distintos niveles de gestión educativa.

Finalmente, desde el ámbito académico, esta investigación aporta al campo de la investigación educativa mediante la aplicación de técnicas de Minería de Datos sobre información censal, ampliando las alternativas metodológicas disponibles para el estudio del rendimiento académico y proporcionando un enfoque susceptible de ser replicado en otros contextos regionales del país.

1.4. Objetivos de la investigación

1.4.1. Objetivo general

Identificar los perfiles de rendimiento en lectura en los estudiantes de 2.º grado de educación secundaria en la Evaluación Censal de Estudiantes en el departamento de Cusco, usando minería de datos.

1.4.2. Objetivos específicos

- Analizar la relación entre el nivel socioeconómico, la lengua materna, el área y la gestión de la institución educativa con el rendimiento en lectura de los estudiantes en el departamento de Cusco.
- Caracterizar el nivel socioeconómico, la lengua materna, el área y la gestión de los estudiantes que presenta un perfil de rendimiento bajo en lectura en el departamento de Cusco.

II. MARCO TEÓRICO CONCEPTUAL

2.1. Bases teóricas

Rendimiento académico

El rendimiento académico es el resultado del proceso de aprendizaje, generado por la intervención pedagógica del docente y manifestado en el estudiante. No se trata de un producto derivado de una sola aptitud, sino más bien de un resultado que emerge de la combinación (nunca completamente identificada) de diversos elementos que influyen en, y provienen de, la persona que aprende, como factores institucionales, pedagógicos, psicosociales y sociodemográficos (Montero Rojas, Villalobos Palma, y Valverde Bermúdez, 2007).

Perfil de rendimiento académico

En el ámbito de la investigación educativa, el concepto de perfil de rendimiento académico se utiliza para describir configuraciones o patrones característicos de estudiantes que comparten atributos similares en términos de desempeño académico y condiciones contextuales asociadas. A diferencia de los enfoques tradicionales, fundamentados en promedios o distribuciones marginales, el análisis basado en perfiles reconoce la heterogeneidad inherente a la población estudiantil y posibilita la identificación de grupos diferenciados que comparten combinaciones específicas de características académicas, socioeconómicas e institucionales (Everitt, Landau, Leese, y Stahl, 2011).

Rendimiento académico como fenómeno multidimensional

El rendimiento académico constituye un constructo complejo y multidimensional que no puede atribuirse exclusivamente a las capacidades

cognitivas individuales del estudiante, sino que es el resultado de la interacción de diversos factores personales, familiares, escolares y contextuales. La evidencia científica ha demostrado que variables como el nivel socioeconómico, el contexto institucional, las características lingüísticas y el entorno social ejercen una influencia significativa sobre los resultados de aprendizaje, particularmente en competencias fundamentales como la lectura. En concordancia con esta perspectiva, las evaluaciones educativas a gran escala conciben el rendimiento académico como un fenómeno sistémico, en el que las oportunidades de aprendizaje a las que acceden los estudiantes condicionan los niveles de logro alcanzados (OECD, 2019).

Capital cultural y desigualdades educativas

Desde la perspectiva de la sociología de la educación, la teoría del capital cultural propuesta por Bourdieu ofrece un marco explicativo para comprender las desigualdades persistentes en el rendimiento académico entre los distintos grupos sociales. De acuerdo con este enfoque, los estudiantes no ingresan al sistema educativo en igualdad de condiciones, debido a que disponen de diferentes niveles de capital cultural adquiridos en su entorno familiar, los cuales se expresan en aspectos como el dominio del lenguaje, las prácticas culturales y la familiaridad con los códigos propios de la institución escolar. En consecuencia, estas diferencias inciden de manera directa en el desempeño académico y tienden a reproducirse a lo largo del sistema educativo, favoreciendo a los estudiantes provenientes de contextos socioeconómicos más favorecidos (Bourdieu, 1986).

Enfoque de capacidades y equidad educativa

El enfoque de capacidades, desarrollado por Amartya Sen, proporciona un marco teórico relevante para analizar el rendimiento académico desde una perspectiva de equidad y justicia social. Este enfoque plantea que el desarrollo humano debe evaluarse en función de las oportunidades reales que tienen las personas para desarrollar sus capacidades, más allá de los resultados observables. En el ámbito educativo, el bajo rendimiento académico no debe interpretarse únicamente como una limitación individual del estudiante, sino como la consecuencia de restricciones estructurales que limitan sus oportunidades de aprendizaje, tales como la pobreza, la exclusión lingüística o la baja calidad de los servicios educativos (Sen, 1999).

Minería de datos

“La minería de datos puede definirse inicialmente como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos” (Pérez lópez y Santín González, 2007, p.1).

Inteligencia artificial (IA)

“La rama de la ciencia de la computación que se ocupa de la automatización de la conducta inteligente” (Luger y Stubblefield, 1993).

“Un campo de estudio que se enfoca en la explicación y emulación de la conducta inteligente en función de procesos computacionales” (Schalkoff, 1990).

El campo de la inteligencia artificial (IA) abarca un amplio conjunto de áreas de investigación y aplicación, entre las cuales sobresalen: la búsqueda y optimización de soluciones, el desarrollo de sistemas expertos, el

procesamiento del lenguaje natural, la robótica, el aprendizaje automático, la lógica formal y el tratamiento de la incertidumbre mediante la lógica difusa.

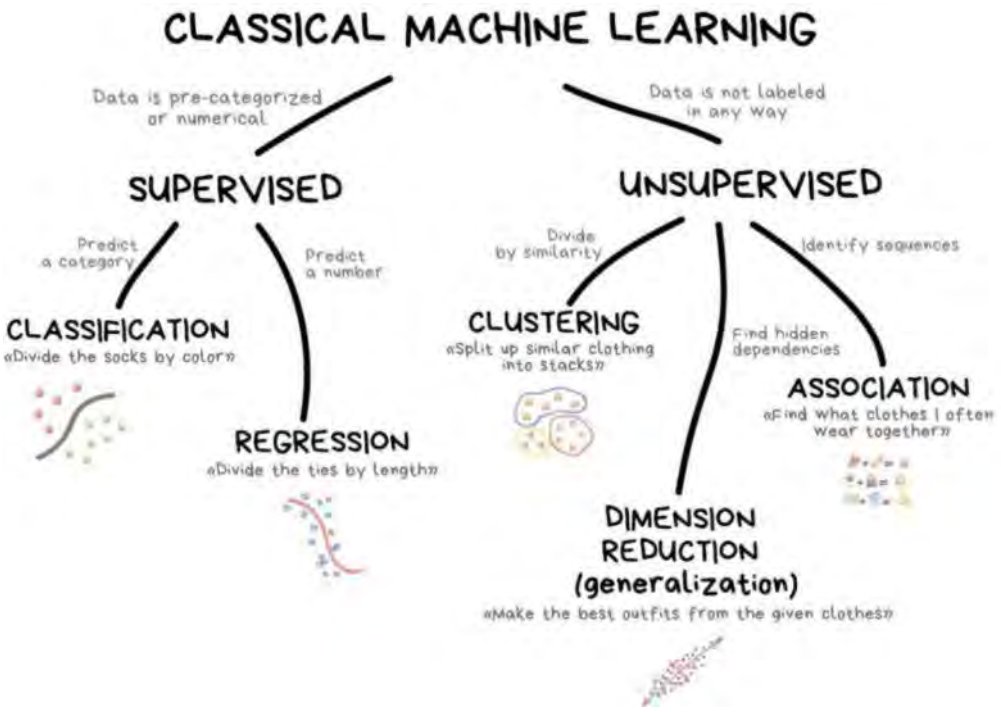
Aprendizaje de máquinas (Machine Learning)

El aprendizaje automático o aprendizaje automatizado o aprendizaje de máquinas (del inglés, Machine Learning). Lewis (2017) afirma que es una colección de algoritmos que generan información a partir de los datos. Esa información podría ser utilizada por humanos u otras máquinas para tomar una decisión.

El aprendizaje automático clásico a menudo se divide en dos categorías: aprendizaje supervisado y no supervisados.

Figura 1

Aprendizaje supervisado y no supervisado



Nota. De Machine Learning para todos, de forma simple e com exemplos!, por P. C. Tebaldi, 2019, DataGeeks (https://www.datageeks.com.br/machine-learning/).

Aprendizaje supervisado

En este enfoque, la máquina cuenta con un supervisor o maestro que le proporciona las respuestas correctas, indicándole, por ejemplo, si una imagen corresponde a un gato o a un perro. El maestro previamente ha clasificado (etiquetado) los datos en estas categorías, y la máquina utiliza dichos ejemplos para aprender. Este tipo de aprendizaje resulta más eficiente bajo la guía de un maestro, por lo que se aplica con frecuencia en contextos reales. Dentro de este enfoque se distinguen dos tipos principales de tareas: la clasificación, que consiste en predecir la categoría a la que pertenece un objeto, y la regresión, orientada a estimar un valor específico dentro de una escala numérica.

Clasificación: este tipo de aprendizaje requiere necesariamente la presencia de un maestro, ya que los datos deben estar etiquetados con características que permitan a la máquina asignar correctamente cada observación a una clase específica. Prácticamente cualquier tipo de información puede clasificarse: usuarios según sus intereses, textos según idioma o temática (relevante para los motores de búsqueda), piezas musicales por género (como en las listas de reproducción de Spotify) e incluso correos electrónicos. Entre los algoritmos más utilizados para este propósito se encuentran Naive Bayes**,** árboles de decisión**,** regresión logística**,** k-vecinos más cercanos y máquinas de vectores de soporte.

Regresión: se trata de un tipo de análisis similar a la clasificación, con la diferencia de que, en lugar de predecir una categoría, se estima un valor numérico. Algunos ejemplos incluyen la estimación del precio de un automóvil según su kilometraje, el nivel de tráfico en función de la hora del

día o el volumen de demanda en relación con el crecimiento de una empresa. Este enfoque resulta especialmente adecuado cuando la variable de interés depende del tiempo. Entre los algoritmos más utilizados se encuentran la regresión lineal y la regresión polinómica.

Aprendizaje no supervisado

Agrupación: consiste en un tipo de clasificación sin categorías predefinidas. En este enfoque, los algoritmos de agrupamiento buscan identificar objetos con características similares y agruparlos dentro de una misma clase. De esta manera, los elementos que comparten mayor similitud se integran en un mismo conjunto. En ciertos algoritmos, además, es posible determinar previamente el número exacto de clústeres o grupos que se desea obtener.

Reducción de dimensionalidad: estos métodos surgieron como una necesidad de los científicos de datos que debían identificar patrones significativos dentro de grandes volúmenes de información numérica. Cuando las representaciones gráficas tradicionales, como las ofrecidas por Excel, resultaron insuficientes, se recurrió a técnicas que permitieran a las máquinas descubrir relaciones subyacentes entre las variables. De este modo, se desarrollaron los métodos de reducción de dimensiones, orientados a simplificar los datos sin perder la información esencial.

Aprendizaje de reglas de asociación: comprende un conjunto de métodos utilizados para analizar comportamientos secuenciales o conjuntos de eventos, como los patrones de compra, la automatización de estrategias de marketing y otras aplicaciones similares. Cuando se dispone de una secuencia de datos y se busca identificar relaciones o patrones recurrentes,

algoritmos como Apriori**,** Eclat y FP-Growth resultan especialmente apropiados para su análisis.

Algoritmos de Machine Learning

K-means

El algoritmo K-means (MacQueen, 1967) clasifica las observaciones en un número predefinido de k clusters, con el objetivo de minimizar la suma de las varianzas internas de dichos clusters.

Consideremos $C_1, \dots, C_k$ como los conjuntos que agrupan los índices de las observaciones correspondientes a cada uno de los clústeres. Por ejemplo, el conjunto C_1 contiene los índices de las observaciones que pertenecen al clúster 1. La notación empleada para expresar que la observación i forma parte del clúster k es: $i \in C_k$. Todos estos conjuntos satisfacen dos propiedades fundamentales:

$C_1 \cup C_2 \cup \dots \cup C_k = \{1, \dots, n\}$. Significa que toda observación se asigna a uno de los k clusters.

$C_1 \cap C_{k'} = \emptyset$ para todo $K \neq K'$. Esto implica que los clústeres son mutuamente excluyentes, es decir, ninguna observación puede pertenecer simultáneamente a más de un clúster.

Existen dos medidas ampliamente utilizadas para cuantificar la varianza interna dentro de un clúster ($W(C_k)$), son:

- Corresponde a la suma de las distancias euclidianas al cuadrado entre cada observación x_i y el centroide μ de su respectivo clúster, lo que equivale a la suma de cuadrados dentro del clúster.

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

- Se calcula como la suma de las distancias euclidianas al cuadrado entre todos los pares de observaciones pertenecientes a un mismo clúster, dividida entre el número total de observaciones que conforman dicho clúster.

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

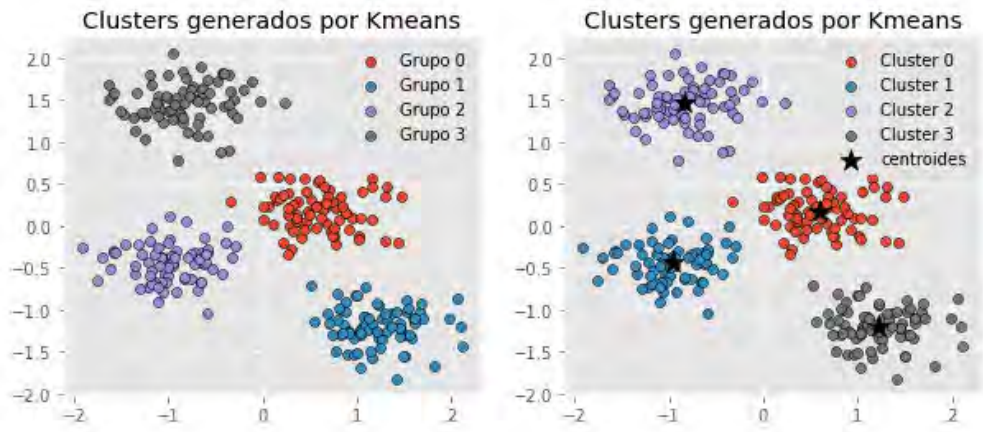
Minimizar de manera exacta la suma total de la varianza interna, representada por $\sum_{k=1}^{k} W(C_K)$, constituye un proceso altamente complejo, dado el gran número de combinaciones posibles en las que las n observaciones pueden distribuirse entre K grupos. Por esta razón, el algoritmo k-means no busca necesariamente el mínimo global, sino que obtiene una solución aproximada que corresponde a un óptimo local. El proceso del algoritmo es el siguiente:

1. Definir previamente el número k de clústeres que se desea obtener en el proceso de agrupamiento.
2. Elegir de manera aleatoria k observaciones del conjunto de datos para utilizarlas como centroides iniciales del algoritmo.
3. Asignar cada observación al centroide más próximo, de acuerdo con la distancia euclidiana u otra métrica de similitud seleccionada.
4. Recalcular el centroide de cada uno de los k clústeres formados en el paso anterior, promediando las posiciones de las observaciones que pertenecen a cada grupo.
5. Repetir los pasos 3 y 4 de forma iterativa hasta que las asignaciones de las observaciones a los clústeres permanezcan estables o se alcance el número máximo de iteraciones establecido.

Este algoritmo garantiza que, en cada iteración, la varianza interna total de los clústeres disminuya progresivamente, permitiendo así que la solución converja hacia un óptimo local. Dado que el algoritmo k-means no evalúa todas las posibles distribuciones de observaciones, sino solo una parte de ellas, los resultados dependen de la asignación aleatoria inicial (paso 2). Por ello, resulta esencial ejecutar el algoritmo en múltiples ocasiones (generalmente entre 25 y 50 veces), utilizando diferentes asignaciones iniciales aleatorias, y seleccionar aquella ejecución que presente la menor varianza total obtenida.

Figura 2

Datos simulados con cuatro agrupaciones identificadas



Nota. De Clustering con Python, por J. Amat Rodrigo, 2020 (https://cienciadedatos.net/documentos/py20-clustering-con-python#Algoritmo).

Este tipo de visualización es extremadamente útil y proporciona una gran cantidad de información, aunque su aplicación se limita a situaciones donde se trabaja con dos dimensiones. Cuando los datos contienen más de dos variables (o dimensiones), una alternativa apropiada consiste en utilizar las

dos primeras componentes principales derivadas de un análisis de componentes principales (PCA).

El rendimiento del modelo puede evaluarse mediante una matriz de confusión, la cual muestra los aciertos y errores en la clasificación. Al interpretar estas matrices, es fundamental tener en cuenta que el proceso de clustering asigna observaciones a clusters, cuyos identificadores no necesariamente coinciden con la nomenclatura de los grupos reales. Por ejemplo, en este caso, el grupo 1 se ha asignado al cluster 3. Por consiguiente, en cada fila de la matriz se espera observar un valor elevado que representa las coincidencias correctas en una posición determinada, y valores bajos en las demás posiciones, los cuales reflejan errores de clasificación; sin embargo, los nombres no necesariamente deben coincidir con los ubicados en la diagonal.

K-medoids

Es un método de agrupamiento estrechamente relacionado con k-means, ya que ambos dividen las observaciones en k clústeres, siendo k un valor definido previamente por el analista. La diferencia fundamental radica en que, en k-medoids, cada clúster se representa mediante una observación real del conjunto de datos (denominada *medoid*), mientras que en k-means, la representación corresponde al centroide, calculado como el promedio de todas las observaciones del clúster, sin necesariamente coincidir con una observación específica.

Un medoid se define de manera más precisa como el elemento dentro de un cluster cuya distancia promedio a todas las demás observaciones del mismo cluster es la menor posible, lo que lo convierte en el punto más central y

representativo del cluster. El uso de medoides en lugar de centroides convierte a k-medoids en un método más robusto que k-means, ya que presenta una menor sensibilidad frente a valores atípicos o ruido en los datos. De forma intuitiva, esta diferencia puede entenderse de manera análoga a la distinción existente entre la media y la mediana**.**

El algoritmo más utilizado para la implementación de k-medoids es PAM (Partitioning Around Medoids), el cual se desarrolla a través de las siguientes etapas:

1. Seleccionar de forma aleatoria k observaciones del conjunto de datos como medoides iniciales, aunque también es posible elegirlas de manera determinada según algún criterio previo.
2. Calcular la matriz de distancias entre todas las observaciones del conjunto de datos, en caso de que aún no se haya generado.
3. Asignar cada observación al medoide más cercano, de acuerdo con la distancia calculada entre las observaciones y los medoides.
4. Para cada clúster formado, evaluar si reemplazar el medoide actual por otra observación reduce la distancia promedio interna del clúster; en caso afirmativo, actualizar el medoide seleccionando dicha observación como el nuevo representante del grupo.
5. Si en el paso 4 se modifica al menos un medoide, repetir el paso 3; de lo contrario, el algoritmo concluye al haberse alcanzado la convergencia.

A diferencia del algoritmo k-means, que busca minimizar la suma de los cuadrados de las distancias de cada observación respecto a su centroide dentro de los clústeres, el algoritmo k-medoids minimiza la suma de las

distancias absolutas (o diferencias) entre cada observación y su respectivo medoide.

El método k-medoids es especialmente útil cuando se sospecha o se conoce la presencia de outliers. En tales casos, se recomienda utilizar la distancia de Manhattan como medida de similitud, ya que es menos sensible a outliers en comparación con la distancia euclidiana.

Hierarchical clustering

El agrupamiento jerárquico (hierarchical clustering) constituye una alternativa a los métodos de agrupamiento por partición, ya que no requiere definir previamente el número de clústeres. Este enfoque se clasifica en dos tipos principales, según la estrategia empleada para conformar los grupos:

- Aglomerativo (agglomerative clustering *o* bottom-up): este enfoque inicia el proceso de agrupamiento considerando cada observación como un clúster individual. Posteriormente, los clústeres se van fusionando de manera progresiva en función de su similitud, hasta que finalmente todas las observaciones se integran en un único grupo.
- Divisivo (divisive clustering *o* top-down): representa la estrategia opuesta al enfoque aglomerativo. En este caso, el proceso comienza con todas las observaciones agrupadas en un único clúster, el cual se divide de manera sucesiva hasta que cada observación pasa a constituir un clúster individual**.**

Ambas estrategias pueden representarse visualmente mediante una estructura arbórea llamada dendrograma.

Aglomerativo

El procedimiento del agrupamiento aglomerativo se desarrolla siguiendo las siguientes etapas:

1. Inicio: considerar cada una de las n observaciones como un clúster individual, constituyendo de esta manera la base del dendrograma (sus hojas).
2. Proceso iterativo: continuar el procedimiento hasta que todas las observaciones se integren en un único clúster**.**
    2.1. Calcular la distancia entre cada par de clústeres posibles. El tipo de medida empleada para cuantificar la similitud o disimilitud entre las observaciones o grupos (tanto la métrica de distancia como el método de enlace) debe ser determinado por el investigador.
    2.2. Fusionar los dos clústeres más similares, es decir, aquellos con la menor distancia entre sí, reduciendo así el número total de clústeres a $n - 1$.
3. Finalización: realizar un corte en el dendrograma a una altura determinada**,** con el fin de definir el número final de clústeres resultantes del proceso de agrupamiento.

Para que el proceso de agrupamiento se lleve a cabo conforme a este algoritmo, es fundamental establecer el criterio de similitud entre clústeres, lo cual requiere ampliar el concepto de distancia originalmente aplicado a pares de observaciones para hacerlo extensivo a pares de grupos, cada uno conformado por múltiples observaciones. Este procedimiento se conoce como linkage (método de enlace). A continuación, se presentan los cinco tipos más comunes de linkage junto con sus respectivas definiciones:

- **Complete o Maximum**: La distancia entre dos clústeres se obtiene calculando todas las distancias posibles entre cada observación del clúster A y cada observación del clúster B; de estas, se toma la mayor distancia, que se considera como la distancia entre ambos clústeres. Este enfoque es el más conservador, pues representa la máxima disimilitud interclúster.
- **Single o Minimum**: De manera análoga al método anterior, se calculan todas las distancias posibles entre una observación del clúster A y una del clúster B; sin embargo, en este caso se selecciona la distancia más pequeña entre todos los pares. Este enfoque es el menos conservador, ya que representa la mínima disimilitud interclúster.
- **Average**: Se calcula la distancia entre todos los pares posibles de observaciones entre los clusters A y B, y se selecciona el valor promedio como la distancia entre los dos clusters (mean intercluster dissimilarity).
- **Centroid**: Se calcula el centroide de cada clúster y la distancia entre estos centroides se considera como la distancia representativa entre los dos clústeres.
- **Ward**: Se trata de un método general en el que la elección del par de clústeres a fusionar en cada paso del agrupamiento jerárquico aglomerativo se basa en la optimización de una función objetivo, la cual puede ser definida libremente por el analista. Un caso particular de este enfoque es el método de mínima varianza de Ward, cuyo objetivo es minimizar la suma total de la varianza intra-clúster. En cada iteración, se seleccionan los dos clústeres cuya fusión implique el menor incremento en la varianza total intra-clúster, una métrica que guarda similitud con la utilizada en k-means.

Los métodos de complete linkage**,** average linkage y la mínima varianza de Ward suelen ser preferidos por los analistas, ya que tienden a generar dendrogramas más equilibrados. No obstante, no es posible establecer una jerarquía de superioridad entre ellos, ya que la elección depende del contexto del estudio; por ejemplo, en genómica es frecuente el uso del método de centroides. Al presentar los resultados de un agrupamiento jerárquico, es indispensable especificar de forma explícita tanto la medida de distancia como el método de linkage utilizados, ya que las soluciones obtenidas pueden variar de manera significativa en función de estos parámetros.

Divisivo

El algoritmo más representativo del agrupamiento jerárquico divisivo es DIANA (Divisive Analysis Clustering). Este método inicia considerando un único clúster que contiene todas las observaciones y realiza divisiones sucesivas hasta que cada observación constituye un clúster independiente. En cada iteración, se selecciona el clúster con el mayor diámetro, definido como la máxima distancia entre dos de sus observaciones. Dentro de este clúster, se identifica la observación más disímil, es decir, aquella cuya distancia promedio respecto al resto de las observaciones es mayor. Esta observación se convierte en el inicio de un nuevo clúster, mientras que las observaciones restantes se reasignan según su proximidad al nuevo clúster o al conjunto restante, dividiendo así el clúster original en dos nuevos clústeres.

1. Todas las n observaciones se agrupan inicialmente en un solo clúster.
2. El proceso se repite de manera continua hasta que se formen n clústeres:

2.1. Se calcula el diámetro de cada clúster, es decir, la distancia más grande entre cualquier par de observaciones dentro del mismo.
2.2. Se elige el clúster que presenta el diámetro más grande.
2.3. Se calcula la distancia promedio de cada observación respecto al resto de las observaciones del clúster.
2.4. La observación que presenta la mayor distancia promedio se utiliza para iniciar un nuevo clúster.
2.5. Las observaciones restantes se reasignan al nuevo clúster o al original, según cuál de los dos se encuentre más cercano.

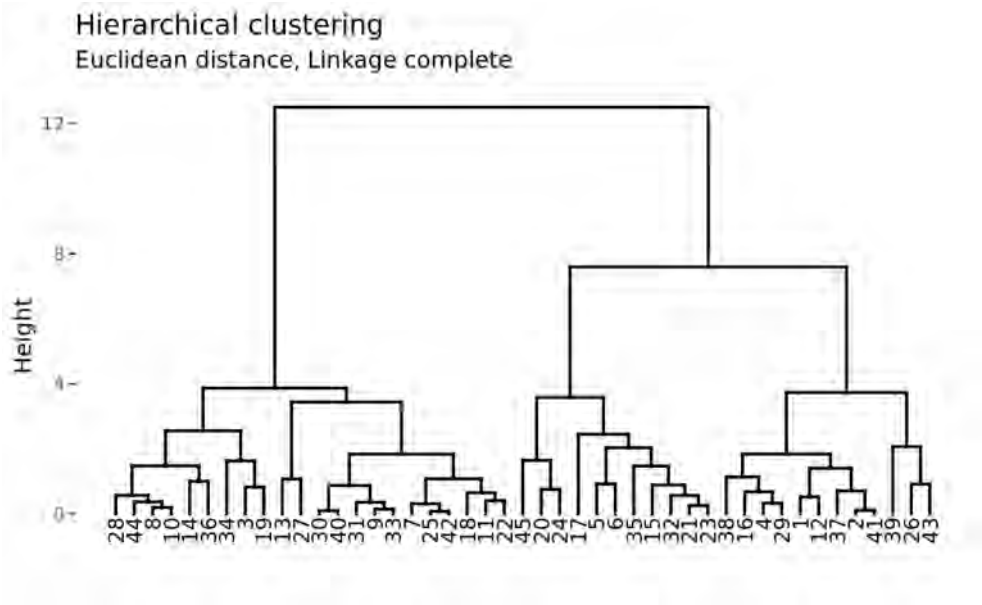
A diferencia del agrupamiento aglomerativo, que requiere definir tanto una medida de distancia como un método de enlace, en el clustering divisivo únicamente es necesario seleccionar la distancia, sin precisar un tipo de linkage.

Dendrograma

Los resultados obtenidos mediante hierarchical clustering pueden visualizarse en forma de un árbol, donde las ramas reflejan la jerarquía de las fusiones de clusters a lo largo del proceso. Imaginemos que contamos con 45 observaciones en un espacio bidimensional y se aplica un agrupamiento jerárquico para identificar posibles grupos. El dendrograma obtenido refleja las fusiones sucesivas de los clústeres a lo largo del proceso.

Figura 3

Dendrograma



Nota. De Clustering con Python, por J. Amat Rodrigo, 2020 (https://cienciadedatos.net/documentos/py20-clustering-con-python#Algoritmo).

En la base del dendrograma, cada observación se representa como una hoja individual del árbol. Al avanzar hacia niveles superiores, estas hojas se van fusionando en pares, formando las primeras ramas, lo que refleja que dichas observaciones son las más similares entre sí. Estas ramas pueden luego fusionarse con otras ramas o con nuevas hojas. Cuanto más próxima a la base del dendrograma ocurre una fusión, mayor es la similitud entre las observaciones que se están agrupando.

Para cualquier par de observaciones, se puede determinar el nivel del dendrograma en el que se unen las ramas que las contienen. La altura a la que ocurre esta fusión en el eje vertical indica el grado de similitud o diferencia entre las dos observaciones. Así, los dendrogramas deben interpretarse considerando únicamente el eje vertical, ya que las posiciones

en el eje horizontal son meramente estéticas y pueden variar según el programa utilizado.

Por ejemplo, al analizar la observación 8, se observa que su par más similar es la observación 10, ya que representan la primera fusión que involucra a la 10 (y viceversa). Aunque la observación 14, ubicada justo a la derecha de la 10, podría parecer la siguiente más similar, en realidad las observaciones 28 y 44 presentan una mayor similitud con la 10, a pesar de estar más alejadas en el eje horizontal. De manera análoga, no es correcto asumir que la observación 14 es más similar a la 10 que la 36 únicamente por su proximidad horizontal. La única interpretación válida, basada en la altura a la que se unen las ramas, es que la similitud entre los pares 10-14 y 10-36 es equivalente.

Análisis de Correspondencias Múltiples (ACM)

El Análisis de Correspondencias Múltiples (ACM) es una técnica estadística multivariada de carácter exploratorio diseñada específicamente para analizar tablas de contingencia y estudiar la relación de asociación conjunta entre múltiples variables cualitativas o categóricas (Greenacre, 2017). Conceptualmente, constituye una extensión del Análisis de Correspondencias Simple y es equivalente a aplicar un Análisis de Componentes Principales (ACP) a datos cualitativos (Husson, Le, & Pagès, 2017).

Clustering Jerárquico sobre Componentes Principales (HCPC)

El Clustering Jerárquico sobre Componentes Principales (HCPC, por sus siglas en inglés Hierarchical Clustering on Principal Components) es un método avanzado de aprendizaje no supervisado que combina de manera

secuencial la reducción de dimensionalidad con las técnicas de clasificación y agrupamiento (Husson, Le, & Pagès, 2017).

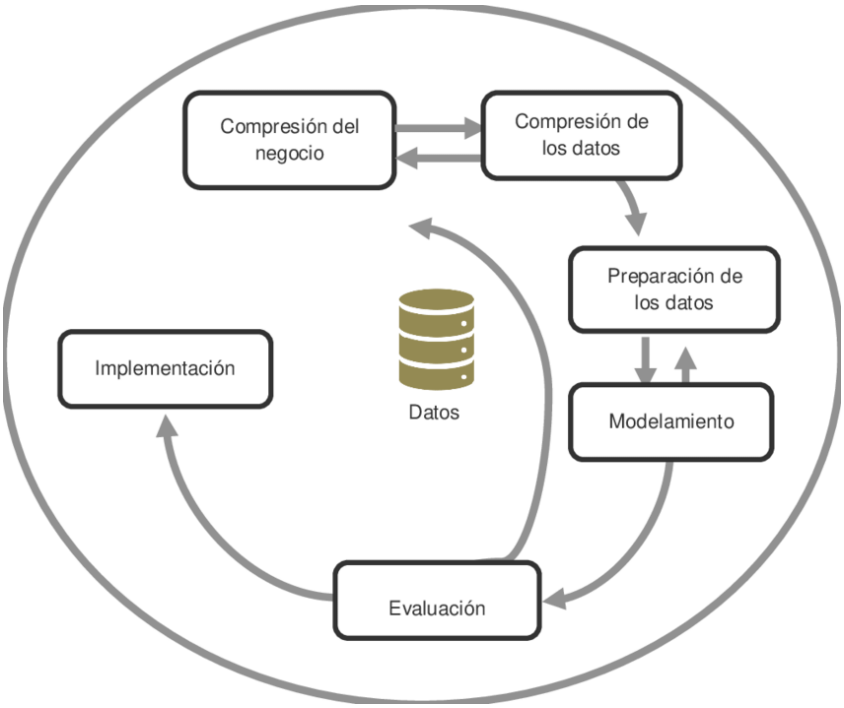
A diferencia del clustering convencional (como el método de k-means clásico) que se aplica de manera directa sobre los datos categóricos originales, el HCPC se ejecuta sobre las coordenadas de las dimensiones continúas obtenidas previamente mediante un método factorial, como el ACM en el caso de datos categóricos (Le, Josse, & Husson, 2008). Metodológicamente, esta secuencia ofrece una ventaja crítica: las primeras dimensiones del ACM retienen la señal o variabilidad más importante de los datos, mientras que las dimensiones posteriores se consideran "ruido estadístico".

Metodología CRISP-DM

El CRISP-DM (Cross Industry Standard Process for Data Mining) es un modelo de proceso que describe cómo los profesionales de minería de datos abordan y resuelven problemas dentro de su área de trabajo.

Figura 4

El proceso de la Metodología CRISP-DM



Nota. Adaptado de Ciclo de vida de minería de datos, por IBM, 2021 (https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview).

A continuación, se detalla cada una de las fases:

1. **Comprensión del negocio.**

   Esta fase inicial es probablemente la más determinante, ya que comprende las actividades orientadas a entender los objetivos y requerimientos del proyecto desde la perspectiva del negocio, con el fin de traducirlos en metas técnicas y en un plan de acción concreto. Sin una comprensión clara de estos objetivos, incluso el algoritmo más avanzado no podrá generar resultados confiables. Para maximizar los beneficios de la minería de datos, es esencial tener un entendimiento completo del problema que se busca resolver, lo que permitirá tanto la correcta

recolección de datos como la adecuada interpretación de los resultados.

En esta etapa resulta clave convertir lo aprendido sobre el funcionamiento del negocio en un problema concreto de minería de datos, junto con un plan preliminar que permita avanzar hacia los objetivos planteados por la organización.

2. **Comprensión de los datos.**

   Esta segunda fase consiste en realizar la primera recopilación de datos para explorar el problema, familiarizarse con la información disponible, evaluar su calidad y detectar patrones o relaciones preliminares que ayuden a formular las primeras hipótesis. Esta etapa, junto con las dos siguientes, suele ser la que más tiempo y esfuerzo demanda dentro de un proyecto de minería de datos.

   Cuando la organización cuenta con una base de datos corporativa, es aconsejable crear una base independiente destinada exclusivamente al proyecto. Esto se debe a que, durante el desarrollo del análisis, es común realizar consultas frecuentes y aplicar modificaciones, lo que podría generar problemas si se trabaja directamente sobre la base principal.

3. **Preparación de los datos.**

   En esta etapa, una vez concluida la recolección inicial, los datos se preparan para poder aplicarles las técnicas de minería de datos que se utilizarán más adelante. Estas técnicas pueden incluir visualizaciones, análisis de relaciones entre variables u otros métodos de explotación de datos. La preparación comprende actividades como seleccionar la información que será usada en un modelo específico, limpiar los datos, generar nuevas variables, integrar distintas fuentes y ajustar formatos.

Esta fase está muy relacionada con la fase de modelado, ya que el tipo de procesamiento requerido depende de la técnica de modelado elegida, lo que hace que ambas etapas interactúen de forma constante.

4. **Modelado.**

   En esta fase del proceso CRISP-DM se eligen las técnicas de modelado más adecuadas para el proyecto de minería de datos. La selección depende de varios aspectos: que la técnica se ajuste al tipo de problema, que los datos disponibles sean suficientes y apropiados, que cumpla con los requisitos planteados, que el tiempo necesario para construir el modelo sea razonable y, además, que se cuente con el conocimiento necesario para aplicarla correctamente.

   Antes de comenzar a modelar, es importante definir cómo se evaluará cada modelo, es decir, qué método permitirá medir qué tan bien se ajusta a los objetivos establecidos. Una vez definidas estas consideraciones, se construyen y evalúan los modelos. Los parámetros utilizados durante su creación varían según las características de los datos y el nivel de precisión que se busca alcanzar.

5. **Evaluación.**

   En esta fase se evalúa el modelo desarrollado con el propósito de verificar si cumple efectivamente con los criterios de éxito definidos para el problema de estudio. No obstante, es importante considerar que el nivel de fiabilidad alcanzado se limita a los datos utilizados durante el análisis, por lo que sus resultados no deben generalizarse de manera automática. Asimismo, se realiza una revisión integral de todo el proceso, a la luz de los resultados obtenidos, con el fin de identificar y corregir

posibles errores cometidos en etapas previas. De forma complementaria, pueden emplearse diversas herramientas que faciliten la interpretación de los resultados.

Una vez concluida esta revisión, si el modelo satisface los criterios previamente establecidos, se procede a la fase final, correspondiente a su implementación y aplicación práctica.

6. **Despliegue o implementación.**

   En esta fase, una vez construido y validado el modelo, el conocimiento generado se traduce en acciones concretas dentro del proceso organizacional. Esta implementación puede materializarse mediante la formulación de recomendaciones derivadas de la interpretación del modelo o a través de su aplicación sobre nuevos conjuntos de datos como parte de las operaciones habituales, por ejemplo, en la evaluación del riesgo crediticio o en la detección de fraudes.

   En términos generales, un proyecto de minería de datos no concluye con la implementación del modelo. También es indispensable documentar y comunicar los resultados de manera clara y comprensible para los usuarios finales, con el propósito de facilitar su interpretación y respaldar la toma de decisiones. Asimismo, durante la fase de explotación, resulta fundamental garantizar el mantenimiento de la aplicación y, cuando corresponda, promover la difusión de los hallazgos obtenidos.

2.2. Marco conceptual

Evaluación Censal de Estudiantes: La Evaluación Censal de Estudiantes (ECE) es una evaluación estandarizada de gran escala implementada por el Ministerio de Educación del Perú con el propósito de medir el logro de los

aprendizajes de los estudiantes en áreas fundamentales del currículo nacional. Su aplicación censal en determinados grados del sistema educativo permite generar información representativa a nivel nacional, regional, local e institucional. Desde el punto de vista metodológico, la ECE se fundamenta en la Teoría de Respuesta al Ítem (TRI), un enfoque psicométrico que posibilita estimar el nivel de competencia de los estudiantes a partir de sus patrones de respuesta a los ítems, independientemente del conjunto específico de preguntas administrado. Este enfoque facilita la comparación de los resultados entre estudiantes y grupos, así como la construcción de escalas de logro con propiedades métricas apropiadas (Ministerio de Educación del Perú, 2019).

Algoritmo: Un algoritmo es un conjunto finito de instrucciones o reglas definidas, precisas y secuenciales, diseñado para resolver un problema o ejecutar una tarea específica de manera sistemática.

Modelo de machine learning: Un modelo de machine learning es un programa computacional capaz de aprender a partir de datos y ejemplos previos para identificar patrones, establecer relaciones y generar predicciones o apoyar la toma de decisiones de manera automatizada.

Análisis de Correspondencias Múltiples (ACM): El Análisis de Correspondencias Múltiples (ACM) es una técnica estadística multivariada de carácter exploratorio, orientada al análisis de relaciones entre variables categóricas. Constituye una extensión del Análisis de Correspondencias Simple y permite representar en un espacio de baja dimensión la estructura de asociaciones entre modalidades de varias variables cualitativas, facilitando la identificación de patrones y perfiles en conjuntos de datos

complejos. Diversos estudios han señalado que la combinación del ACM con técnicas de agrupamiento o clustering permite mejorar la identificación de perfiles homogéneos, ya que el ACM reduce la dimensionalidad y elimina redundancias antes del proceso de clasificación. Esta integración metodológica favorece la interpretación de los resultados y mejora la estabilidad de los grupos identificados, especialmente en estudios basados en datos censales y encuestas educativas de gran tamaño.

Nivel socioeconómico: El nivel socioeconómico se refiere al conjunto de condiciones económicas, sociales y culturales que caracterizan el entorno del estudiante y su familia, e incluye dimensiones como ingresos, nivel educativo de los padres, ocupación y condiciones de vivienda. En el ámbito educativo, el nivel socioeconómico es considerado uno de los factores más influyentes en el rendimiento académico, al incidir en el acceso a recursos educativos, el apoyo familiar y las oportunidades de aprendizaje.

Gestión educativa: La gestión educativa comprende el conjunto de procesos orientados a la planificación, organización, dirección y evaluación de las instituciones educativas, con el propósito de garantizar el logro de los objetivos pedagógicos y la mejora continua de la calidad educativa. Incluye dimensiones relacionadas con la gestión institucional, el liderazgo directivo, la organización escolar y el uso eficiente de los recursos.

Ruralidad: La ruralidad hace referencia a la ubicación geográfica de la institución educativa en áreas rurales o urbanas, según la clasificación oficial del Ministerio de Educación. Esta variable está asociada a diferencias en el acceso a recursos educativos, infraestructura, conectividad y servicios

complementarios, factores que influyen directamente en el desempeño académico de los estudiantes.

Legua materna: La lengua materna se define como el primer idioma aprendido y utilizado de manera habitual por el estudiante en su entorno familiar. En la ECE, esta variable es particularmente relevante en contextos interculturales, como el departamento de Cusco, donde una proporción significativa de estudiantes tiene como lengua materna el quechua.

Control de calidad educativa: El control de calidad en educación se refiere al conjunto de mecanismos, procesos y herramientas orientados a asegurar que los servicios educativos cumplan con estándares establecidos de desempeño y aprendizaje. En el contexto de los sistemas educativos, el control de calidad se apoya en evaluaciones estandarizadas, indicadores de logro y sistemas de monitoreo que permiten evaluar el desempeño de estudiantes e instituciones educativas.

Técnicas de agrupamiento (Clustering): El clustering o análisis de agrupamiento es un conjunto de técnicas de aprendizaje no supervisado cuyo objetivo es agrupar observaciones en función de su similitud, de modo que los elementos dentro de un grupo sean más similares entre sí que respecto a los de otros grupos. En el análisis educativo, estas técnicas se utilizan para identificar perfiles de estudiantes con características académicas y contextuales similares, superando los enfoques basados en promedios o clasificaciones unidimensionales.

2.3. Antecedentes de la investigación

Antecedentes nacionales

Cueto, León y Muñoz (2016) realizaron un estudio exhaustivo sobre los factores asociados al rendimiento académico de los estudiantes peruanos a partir del análisis de los resultados de la Evaluación Censal de Estudiantes (ECE), utilizando información correspondiente a la educación primaria. Mediante la aplicación de modelos estadísticos multivariados y enfoques de análisis contextual, los autores evidenciaron que el desempeño en comprensión lectora está estrechamente influido por factores de naturaleza socioeconómica, lingüística y territorial. En particular, encontraron que los estudiantes provenientes de hogares con menor nivel socioeconómico, aquellos cuya lengua materna es distinta del castellano principalmente lenguas originarias y quienes asisten a instituciones educativas ubicadas en zonas rurales presentan, de manera sistemática, menores niveles de logro en las evaluaciones censales.

Asimismo, el estudio pone de manifiesto que estas brechas trascienden las diferencias individuales y responden a desigualdades estructurales del sistema educativo peruano, vinculadas con la distribución desigual de recursos, las oportunidades de aprendizaje y el capital cultural. Los autores concluyen que los resultados de la ECE constituyen una herramienta fundamental para visibilizar dichas desigualdades; sin embargo, advierten que los análisis sustentados exclusivamente en promedios y porcentajes agregados son insuficientes para comprender la complejidad de los patrones de rendimiento académico. En consecuencia, plantean la necesidad de incorporar enfoques analíticos más integrales y desagregados que permitan

identificar grupos diferenciados de estudiantes con características y trayectorias educativas similares.

Antecedentes internacionales

Guichard, Chimienti, Bolzman y Le Goff (2024), en la investigación titulada When National Origins Equal Socio-economic Background: The Effect of the Ethno-class Parental Background on the Education of Children, examinaron las barreras que enfrentan los jóvenes para alcanzar el éxito en la educación superior mediante una estrategia metodológica secuencial que integró el Análisis de Correspondencias Múltiples (ACM) y el Clustering Jerárquico sobre Componentes Principales (HCPC) en R, con el propósito de construir una tipología de logro académico basada en variables como el nivel socioeconómico de los padres, el idioma de origen y el área geográfica; los resultados mostraron que el ACM permitió reducir la alta dimensionalidad de las variables de origen, mientras que el HCPC posibilitó clasificar a los estudiantes en clusters claramente diferenciados, evidenciando que la combinación de un idioma materno no oficial de la región junto con un bajo nivel socioeconómico configuraba un perfil de alta vulnerabilidad educativa y mayor riesgo de deserción escolar; en consecuencia, el estudio concluyó que la integración metodológica ACM + HCPC constituye una estrategia altamente eficiente para resumir bases de datos poblacionales complejas, transformando variables abstractas en perfiles sociales claramente identificables y útiles para la formulación de políticas de equidad educativa.

Timarán Pereira, Caicedo Zambrano, y Hidalgo Troya (2019) en su artículo sobre la identificación de factores vinculados al desempeño en matemáticas en las pruebas Saber 11 mediante técnicas de minería de datos, señalan

que variables como el estrato socioeconómico, la jornada escolar, el índice TIC, la edad y el sexo fueron las que más influyeron en la generación de patrones relacionados con un rendimiento alto o bajo en dicha prueba. Estos hallazgos se obtuvieron empleando un modelo de clasificación basado en árboles de decisión.

En la investigación de Holgado Apaza (2018), titulado: “Detección de patrones de bajo rendimiento académico mediante técnicas de minería de datos de los estudiantes de la Universidad Nacional Amazónica de Madre de Dios 2018”, se identificó que las variables con mayor impacto en la predicción del rendimiento académico fueron la cantidad de asignaturas matriculadas, el acceso al comedor universitario, la carrera profesional y la existencia de deudas con la universidad. Entre los tres algoritmos evaluados (Random Forest, C5.0 y CART), el modelo C5.0 obtuvo el mejor desempeño en la clasificación del bajo rendimiento académico, alcanzando una exactitud del 77.80% y un coeficiente Kappa de 0.56.

Acosta, La Red Martínez, y Primorac (2018) señalan en su estudio que, para analizar el desempeño académico estudiantil mediante técnicas de minería de datos, tomaron en cuenta tanto las calificaciones como diversas variables socioeconómicas, demográficas y actitudinales. Los perfiles de rendimiento académico identificados correspondieron a la asignatura de Álgebra de la carrera de Licenciatura en Sistemas de Información y a la asignatura Matemática I de la carrera de Ingeniería Agronómica en la Universidad Nacional del Nordeste (UNNE). Para este propósito, emplearon un Data Warehouse, lo que permitió caracterizar con mayor precisión dichos perfiles.

En el artículo de La Red Martínez, Karanik, Giovannini, Báez, y Torre (2016) los autores buscan identificar patrones relacionados con las características de los estudiantes para definir perfiles que expliquen el éxito o el fracaso académico. Para establecer estos perfiles en la asignatura Algoritmos y Estructura de Datos de la carrera de Ingeniería en Sistemas de Información, aplicaron un modelo basado en técnicas de minería de datos. Entre sus conclusiones destacan que los estudiantes que dedican menos horas al estudio, pero asignan mayor importancia a esta actividad tienden a obtener mejores resultados. Asimismo, señalan que los alumnos que no trabajan presentan un rendimiento superior, al igual que aquellos cuyo empleo está vinculado a sus estudios.

Fernández Cañete y Martínez Espinal (2014) en su estudio titulado "Determinación de factores que mejoran el rendimiento académico aplicando minería de datos en el colegio Convenio Andrés Bello – Huancayo", señalan que la aplicación de técnicas de minería de datos permite identificar los elementos que contribuyen a mejorar el rendimiento académico y, con ello, elevar la calidad del servicio educativo. Los autores concluyen que la convivencia con los padres, el lugar de origen del estudiante, la situación socioeconómica y el nivel educativo de los padres son factores que influyen significativamente en el desempeño académico. Asimismo, reportan que los árboles de decisión mostraron un mejor desempeño predictivo en comparación con los algoritmos de clustering y las redes neuronales.

III. HIPÓTESIS Y VARIABLES

3.1. Hipótesis

3.1.1. Hipótesis general

Existen perfiles diferenciados de rendimiento académico en lectura en la Evaluación Censal de Estudiantes en el departamento de Cusco usando minería de datos.

3.1.2. Hipótesis especificas

- El nivel socioeconómico, la lengua materna, el área y la gestión de la institución educativa se relacionan de forma diferenciada con el rendimiento en lectura de los estudiantes.
- Los estudiantes con un perfil de rendimiento bajo en lectura en el departamento de Cusco presentan un nivel socioeconómico predominante muy bajo, lengua materna quechua, área rural y gestión estatal.

3.2. Operacionalización de variables

Tabla 1

Operacionalización de variables

Variable	Indicador	Valor final	Tipo de variable
Código modular	Código modular	0207373, 0207381, 0207407, 0207449, …	Categórica nominal
Nombre de la institución educativa	Nombre de la institución educativa	Técnico Agropecuario, Almirante Miguel Grau, …	Categórica nominal
Sección	Sección	A, B, C, D, E, F, G, H, I, J, K, L, M, N	Categórica nominal
Código de la UGEL	Código de la UGEL	080001, 080002, 080003, 080004, …	Categórica nominal
Nombre de la UGEL	Nombre de la UGEL	Acomayo, Anta, Calca, Canas, …	Categórica nominal
Código geográfico	Código geográfico	080101, 080102, 080103, 080104, …	Categórica nominal
Provincia	Provincia	Acomayo, Anta, Calca, Canas, Canchis, …	Categórica nominal
Distrito	Distrito	Accha, Acomayo, Acopia, Acos, …	Categórica nominal
Gestión	Gestión	Estatal, No estatal	Categórica nominal
Área	Área	Rural, Urbano	Categórica nominal
Sexo	Sexo	Hombre, Mujer	Categórica nominal
Lengua materna	Lengua materna	Aimara, Castellano, Lengua extranjera, …	Categórica nominal
Nivel socioeconómico	Nivel socioeconómico	Alto, Medio, Bajo, Muy bajo	Categórica ordinal
Medida en lectura	Medida en lectura	Puntaje	Numérica continua
Nivel de desempeño en lectura	Nivel de desempeño en lectura	Previo al inicio, En inicio, En proceso, Satisfactorio	Categórica ordinal
Ajuste por no respuesta en lectura	Ajuste por no respuesta en lectura	Peso	Numérica continua
Medida en matemática	Medida en matemática	Puntaje	Numérica continua
Nivel de desempeño en matemática	Nivel de desempeño en matemática	Previo al inicio, En inicio, En proceso, Satisfactorio	Categórica ordinal
Ajuste por no respuesta en matemática	Ajuste por no respuesta en matemática	Peso	Numérica continua
Medida en ciencia y tecnología	Medida en ciencia y tecnología	Puntaje	Numérica continua
Nivel de desempeño en ciencia y tecnología	Nivel de desempeño en ciencia y tecnología	Previo al inicio, En inicio, En proceso, Satisfactorio	Categórica ordinal
Ajuste por no respuesta en ciencia y tecnología	Ajuste por no respuesta en ciencia y tecnología	Peso	Numérica continua

IV. METODOLOGÍA

4.1. Ámbito de estudio

La presente investigación se desarrolló en el departamento de Cusco, ubicado en la región sur del Perú, el cual presenta una marcada diversidad geográfica, social, cultural y lingüística. Estas características convierten a Cusco en un escenario particularmente relevante para el análisis del rendimiento académico y las desigualdades educativas, especialmente en el nivel de educación secundaria.

Desde el punto de vista de la gestión educativa, el departamento de Cusco se encuentra organizado administrativamente en catorce Unidades de Gestión Educativa Local (UGEL), las cuales corresponden a sus trece provincias: Cusco, Calca, Canas, Canchis, Chumbivilcas, Anta, Acomayo, Paruro, Paucartambo, Quispicanchi, Urubamba, La Convención y Espinar. Cada UGEL es responsable de la gestión pedagógica, administrativa e institucional de las instituciones educativas públicas y privadas de su ámbito territorial, lo que genera contextos educativos diferenciados en función de las condiciones locales.

En términos de distribución territorial, el sistema educativo cusqueño se caracteriza por una alta proporción de instituciones educativas ubicadas en zonas rurales, especialmente en provincias como Chumbivilcas, Canas, Paucartambo, Paruro y Acomayo. Estas zonas presentan mayores dificultades de acceso geográfico, menor disponibilidad de recursos educativos y una alta prevalencia de estudiantes cuya lengua materna es el quechua. En contraste, UGEL como Cusco y, en menor medida, Urubamba y La Convención, concentran una mayor proporción de instituciones

educativas urbanas, con mejores condiciones de infraestructura, mayor acceso a servicios educativos complementarios y una población estudiantil predominantemente castellanohablante.

4.2. Tipo y nivel de investigación

Tipo de investigación

La presente investigación adopta un enfoque cuantitativo, dado que se sustenta en el análisis estadístico de datos numéricos provenientes de la Evaluación Censal de Estudiantes (ECE) 2019, aplicada por el Ministerio de Educación del Perú. El análisis se realiza a partir de información secundaria oficialmente recolectada, lo que permite trabajar con una base de datos censal y de cobertura regional completa.

El diseño de la presente investigación es no experimental, dado que las variables de estudio no son objeto de manipulación deliberada, sino que se analizan tal como se manifiestan en el contexto educativo. Asimismo, corresponde a un estudio de corte transversal, debido a que utiliza información recolectada en un único momento temporal, y de carácter retrospectivo, ya que se basa en registros previamente recopilados y disponibles en fuentes oficiales.

En cuanto a su alcance metodológico, la investigación adopta un enfoque exploratorio-descriptivo, cuyo propósito es identificar y caracterizar perfiles de rendimiento académico en la competencia de lectura mediante el análisis integrado de variables socioeconómicas, académicas e institucionales. Con este fin, se emplean técnicas de minería de datos y análisis multivariado, como el Análisis de Correspondencias Múltiples y los métodos de agrupamiento, que permiten explorar la estructura subyacente de los datos

e identificar patrones de asociación sin establecer relaciones causales ni desarrollar modelos explicativos.

Nivel de investigación

El nivel de investigación es exploratorio y descriptivo**.**

Es exploratorio, porque el estudio busca identificar patrones, estructuras y agrupamientos subyacentes en los datos de la Evaluación Censal de Estudiantes, sin partir de hipótesis causales previamente definidas. Este nivel es pertinente cuando el fenómeno de estudio los perfiles de rendimiento académico en la competencia de lectura no han sido analizado previamente con técnicas de minería de datos en el contexto regional del departamento de Cusco.

Es descriptivo, ya que permite caracterizar los perfiles de rendimiento académico identificados, describiendo sus principales atributos en términos de nivel socioeconómico, lengua materna, tipo de gestión de la institución educativa, ámbito geográfico y otros factores contextuales incluidos en la ECE 2019. El análisis se centra en la descripción sistemática de las configuraciones observadas, sin pretender establecer relaciones de causalidad.

Asimismo, la investigación es de nivel aplicado, dado que los resultados obtenidos tienen una finalidad práctica y están orientados a generar evidencia empírica útil para la gestión educativa, la toma de decisiones institucionales y el diseño de estrategias de intervención focalizadas por parte de las instituciones educativas, las Unidades de Gestión Educativa Local y la Dirección Regional de Educación del departamento de Cusco.

4.3. Unidad de análisis

La unidad de análisis de la presente investigación es el estudiante**,** específicamente cada estudiante de 2.º grado de educación secundaria del nivel de Educación Básica Regular (EBR) del departamento de Cusco que participó en la Evaluación Censal de Estudiantes (ECE) 2019.

4.4. Población de estudio

La población de estudio está constituida por todos los estudiantes de 2.º grado de educación secundaria del nivel de Educación Básica Regular (EBR) del departamento de Cusco que participaron en la Evaluación Censal de Estudiantes (ECE) 2019, aplicada por el Ministerio de Educación del Perú. Esta población comprende a 23071 estudiantes, matriculados en instituciones educativas de gestión pública y privada, ubicadas en las trece provincias del departamento de Cusco: Cusco, Acomayo, Anta, Calca, Canas, Canchis, Chumbivilcas, Espinar, La Convención, Paruro, Paucartambo, Quispicanchi y Urubamba. Al tratarse de una evaluación de carácter censal, la población incluye la totalidad de los estudiantes evaluados en dicho grado y año en el ámbito geográfico definido.

4.5. Tamaño de muestra

Dado que la presente investigación se basa en datos provenientes de una evaluación censal**,** no se realizó un proceso de muestreo, ya sea probabilístico o no probabilístico. En este sentido, la muestra coincide conceptualmente con la población de estudio.

No obstante, para efectos del análisis estadístico y de la aplicación de técnicas de minería de datos, se trabajó con una muestra analítica, conformada por los registros de estudiantes que contaban con información

completa en todas las variables consideradas en el estudio. En consecuencia, se excluyeron aquellos registros que presentaban datos faltantes en alguna de las variables socioeconómicas, académicas o institucionales utilizadas.

La exclusión de estos registros representó un porcentaje mínimo del total de estudiantes evaluados y se realizó con el objetivo de garantizar la consistencia y validez de los análisis multivariados aplicados, sin que ello implique la introducción de sesgos de selección ni la pérdida del carácter censal del estudio.

4.6. Técnicas de recolección de información

La información electrónica de la Evaluación Censal de Estudiantes (ECE) del año 2019 fue obtenida de la página web (**http://umc.minedu.gob.pe/resultadosnacionales2019/**) de la Oficina de Medición de la Calidad de los Aprendizajes. El archivo digital descargado están en formato SAV, relacionado al software SPSS (Statistical Package for the Social Sciences).

4.7. Técnicas de análisis e interpretación de la información

El análisis de la información se realizó mediante técnicas de análisis multivariado para variables categóricas, específicamente el Análisis de Correspondencias Múltiples (Multiple Correspondence Analysis, MCA) y el análisis de conglomerados jerárquicos sobre componentes principales (Hierarchical Clustering on Principal Components, HCPC). Todo el procesamiento estadístico se llevó a cabo en el software R, utilizando principalmente los paquetes FactoMineR y factoextra, los cuales permiten realizar análisis multivariados bajo un enfoque reproducible.

En primer lugar, se efectuó la preparación de los datos mediante la selección de variables categóricas relevantes para el estudio: área de residencia, nombre de la institución educativa, tipo de gestión institucional, sexo, lengua materna, nivel socioeconómico y nivel de desempeño en lectura. Posteriormente, se realizó la depuración de la base de datos mediante la identificación de registros incompletos, la verificación de la consistencia de los niveles de cada variable y la recodificación de categorías cuando fue necesario, con el fin de garantizar la calidad de la información antes de la aplicación de las técnicas multivariadas.

Debido al carácter censal de la base de datos de la ECE 2019 (N = 23071 estudiantes), los registros excluidos representaron una proporción mínima e insignificante del universo estudiado. De acuerdo con la teoría de datos perdidos (Little & Rubin, 2019), cuando la tasa de omisión es mínima, el Análisis de Casos Completos es el método más seguro y robusto, ya que evita introducir sesgos de selección, preserva el poder estadístico de la muestra y no requiere la aplicación de algoritmos de imputación artificiales que podrían alterar la varianza real de las categorías analizadas.

El Análisis de Correspondencias Múltiples se llevó a cabo mediante la función MCA() del paquete FactoMineR. Esta técnica se fundamenta en la construcción de la matriz disyuntiva completa y en su descomposición en valores singulares (Singular Value Decomposition, SVD), procedimiento que permite reducir la dimensionalidad del conjunto de datos y representar, de forma simultánea, a los individuos y a las categorías de las variables en un espacio factorial de menor dimensión. De manera equivalente, el MCA puede interpretarse como un análisis factorial aplicado a la matriz de Burt, la

cual reúne todas las tablas de contingencia posibles entre las variables categóricas incluidas en el estudio.

Para la estimación del modelo se definieron como variables activas Gestión, Área, Sexo, Lengua materna y Nivel socioeconómico, mientras que Nombre de la institución educativa y Nivel de desempeño en lectura se incorporaron como variables suplementarias. Esta estrategia permitió interpretar la estructura factorial sin influir en la construcción de los ejes principales. En consecuencia, las variables activas fueron las responsables de la formación de las dimensiones factoriales, en tanto que las variables suplementarias se proyectaron posteriormente sobre el espacio factorial con fines exclusivamente interpretativos.

La interpretación del modelo factorial se efectuó a partir de los valores propios (eigenvalues), el porcentaje de inercia explicado por cada dimensión y la inercia total del sistema. En el contexto del MCA, la inercia representa la variabilidad total contenida en la matriz disyuntiva completa; por ello, las primeras dimensiones concentran la mayor proporción de la información estadística del conjunto de datos. En consecuencia, para el análisis se conservaron únicamente aquellas dimensiones con mayor capacidad discriminante y mayor contribución de las variables.

De forma complementaria, la interpretación de las dimensiones se apoyó en indicadores de calidad de representación, específicamente el $\cos^2$ (coseno cuadrado) de las categorías y de los individuos, así como en la contribución absoluta y relativa de cada variable a la conformación de los ejes factoriales.

Estos indicadores permiten identificar qué variables explican cada dimensión y qué categorías presentan mayor capacidad de diferenciación dentro del modelo.

A partir de las coordenadas factoriales obtenidas en el MCA se aplicó posteriormente un análisis de conglomerados jerárquicos utilizando la función HCPC() del paquete FactoMineR, procedimiento conocido como Hierarchical Clustering on Principal Components (HCPC). Este método permite combinar reducción de dimensionalidad y clasificación de observaciones, mejorando la estabilidad del análisis de clustering en comparación con los métodos aplicados directamente sobre datos categóricos.

El agrupamiento se realizó utilizando la distancia euclidiana calculada sobre las coordenadas factoriales y el método de enlace de Ward, el cual minimiza la varianza intra-cluster en cada etapa del proceso de agrupamiento. Este método permite obtener conglomerados internamente homogéneos y externamente bien diferenciados, lo que resulta especialmente adecuado en estudios socioeducativos con alta heterogeneidad entre individuos.

El número óptimo de clusters no se fijó previamente, sino que fue determinado automáticamente por el algoritmo (nb.clust = -1) mediante el criterio de ganancia de inercia entre particiones sucesivas. Este criterio permite seleccionar el número de grupos que maximiza la diferenciación entre clusters y minimiza la variabilidad interna dentro de cada grupo. Asimismo, el punto de corte del dendrograma fue verificado visualmente con el fin de garantizar la coherencia estadística de los grupos obtenidos.

Finalmente, los clusters obtenidos se proyectaron sobre el plano factorial mediante las funciones fviz_mca_var() y fviz_cluster() del paquete factoextra, lo que facilitó la interpretación visual de la estructura socioeducativa de los estudiantes y el análisis de los patrones identificados en cada UGEL en función del contexto territorial, la lengua materna, el nivel socioeconómico y el nivel de logro académico. Esta representación gráfica permitió evidenciar las relaciones estructurales entre las variables educativas, socioculturales y territoriales, proporcionando una interpretación integral, coherente y estadísticamente consistente de los resultados obtenidos en la investigación.

V. RESULTADOS Y DISCUSIÓN

5.1. Procesamiento de los datos

Para verificar los objetivos se aplicaron técnicas estadísticas de Análisis de Correspondencias Múltiples (ACM) en combinación con el Clúster Jerárquico. Las variables de análisis incluyeron: el área de la institución educativa (rural o urbana), el tipo de gestión de la IE (estatal o no estatal), el sexo del estudiante (hombre o mujer), la lengua materna del estudiante considerando las categorías más frecuentes: castellano y quechua, el nivel socioeconómico del estudiante (alto, medio, bajo y muy bajo) y, como referencia adicional, el nivel de desempeño en lectura del estudiante (previo al inicio, en inicio, en proceso y satisfactorio) y el nombre de la institución educativa. Los análisis se realizaron de manera diferenciada por Unidad de Gestión Local (UGEL)**,** abarcando un total de 14 unidades pertenecientes a la Dirección Regional de Educación (DRE) de Cusco.

5.2. Análisis multivariado y perfiles de rendimiento académico en lectura por UGEL.

5.2.1. Inercia explicada del Análisis de Correspondencia Múltiples (MCA)

Tabla 2

Inercia explicada por las dos primeras dimensiones del MCA en las UGEL

UGEL	Dim. 1 (%)	Dim. 2 (%)	% Acumulado
Cusco	23.34	16.32	39.66
Canchis	24.54	15.00	39.55
La Convención	24.63	15.58	40.21
Quispicanchi	23.72	16.22	39.94
Calca	23.84	15.28	39.13
Urubamba	26.71	15.47	42.19
Acomayo	19.46	16.27	35.73
Paruro	22.96	18.08	41.04
Paucartambo	23.82	17.79	41.61
Espinar	25.89	16.06	41.95
Canas	26.60	15.34	41.95
Chumbivilcas	27.63	16.88	44.52
Pichari-Kimbiri	22.71	15.40	38.11
Anta	26.66	16.39	43.05

Los resultados del Análisis de Correspondencias Múltiples muestran que, en las 14 UGEL evaluadas, las dos primeras dimensiones concentran entre 35.73% y 44.52% de la inercia total, lo que indica que el plano bidimensional permite representar adecuadamente la estructura de las variables sociodemográficas y del rendimiento académico en lectura.

En la mayoría de las UGEL, la varianza acumulada se sitúa alrededor del 40%, lo que evidencia una estructura relativamente estable entre los contextos educativos analizados. UGEL como Chumbivilcas

(44.52%), Anta (43.05%) y Urubamba (42.19%) presentan los valores más altos de inercia acumulada, lo que sugiere que en estas jurisdicciones las variables sociodemográficas y académicas muestran una mayor capacidad de diferenciación entre los estudiantes. Es decir, las desigualdades socioeducativas se encuentran más claramente estructuradas en estas UGEL.

Por el contrario, UGEL como Acomayo (35.73%), Pichari-Kimbiri (38.11%) y Calca (39.13%) presentan los valores más bajos de varianza acumulada. Esto indica que, en estas UGEL, la estructura de las variables es menos marcada, por lo que las diferencias entre los estudiantes tienden a ser más homogéneas o menos definidas en el espacio factorial.

En cuanto a la contribución individual de las dimensiones, la Dimensión 1 explica entre 19.46% y 27.63% de la inercia total, siendo generalmente la dimensión con mayor poder explicativo en todas las UGEL. Destacan especialmente Chumbivilcas (27.63%), Urubamba (26.71%), Anta (26.66%) y Canas (26.60%), donde la primera dimensión representa un eje de diferenciación particularmente fuerte. En contraste, Acomayo (19.46%) presenta el valor más bajo, lo que confirma que la estructura de las variables es menos marcada en esta UGEL.

Por su parte, la Dimensión 2 presenta valores más estables, con una varianza explicada que oscila entre 15.00% y 18.08%, lo que indica que esta dimensión cumple un papel complementario en la interpretación de la estructura socioeducativa. UGEL como Paruro

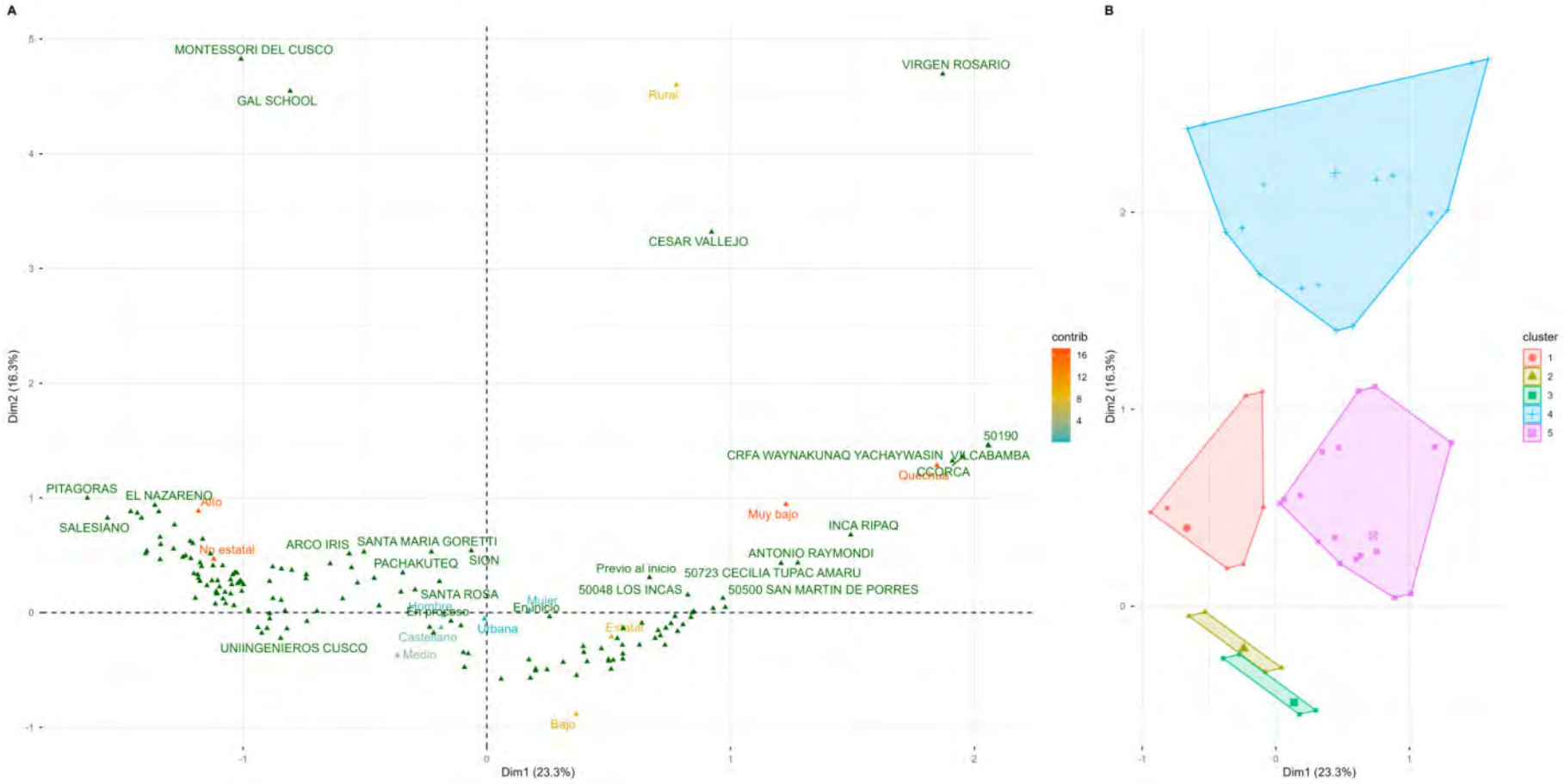
(18.08%) y Paucartambo (17.79%) muestran una mayor contribución de esta segunda dimensión, lo que sugiere la presencia de factores adicionales de diferenciación relacionados probablemente con características territoriales o socioculturales.

En conjunto, los resultados evidencian que, aunque existen ligeras variaciones entre las UGEL, la estructura general del MCA es consistente, ya que en todas ellas las dos primeras dimensiones concentran la mayor proporción de la inercia total. Esto justifica el uso del plano bidimensional para la interpretación de los biplots y el análisis de los perfiles de estudiantes en cada UGEL.

5.2.2. Análisis factorial y formación de conglomerados en las UGEL del Departamento de Cusco

Figura 5

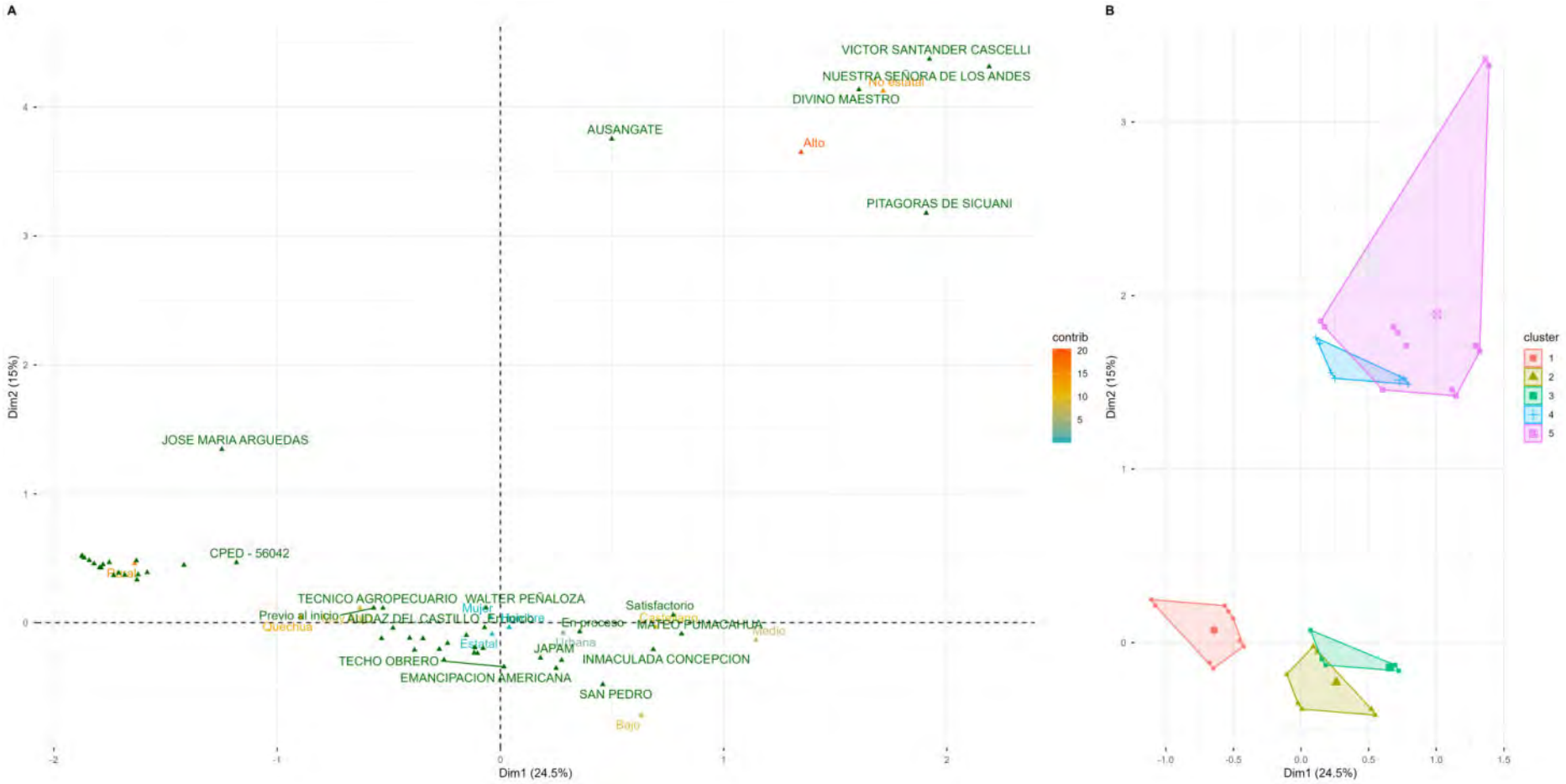
Biplot y clusters formados en la UGEL Cusco



El biplot de la UGEL Cusco evidencia que la primera dimensión está principalmente vinculada al tipo de gestión de la institución educativa, al idioma materno de los estudiantes y al nivel socioeconómico, mientras que la segunda dimensión se asocia sobre todo al área de residencia y a los niveles socioeconómicos intermedios. En el plano factorial se aprecia que los estudiantes cuya lengua materna es el quechua, que pertenecen a niveles socioeconómicos bajos y que provienen de zonas rurales tienden a concentrarse en un mismo sector, mientras que las instituciones (Salesiano, Pitagoras, El Nazareno, etc.) no estatales y los estudiantes con mejores condiciones socioeconómicas se ubican en una posición opuesta, lo que pone en evidencia la presencia de desigualdades socioeducativas dentro de la UGEL Cusco (ver figura 5).

Figura 6

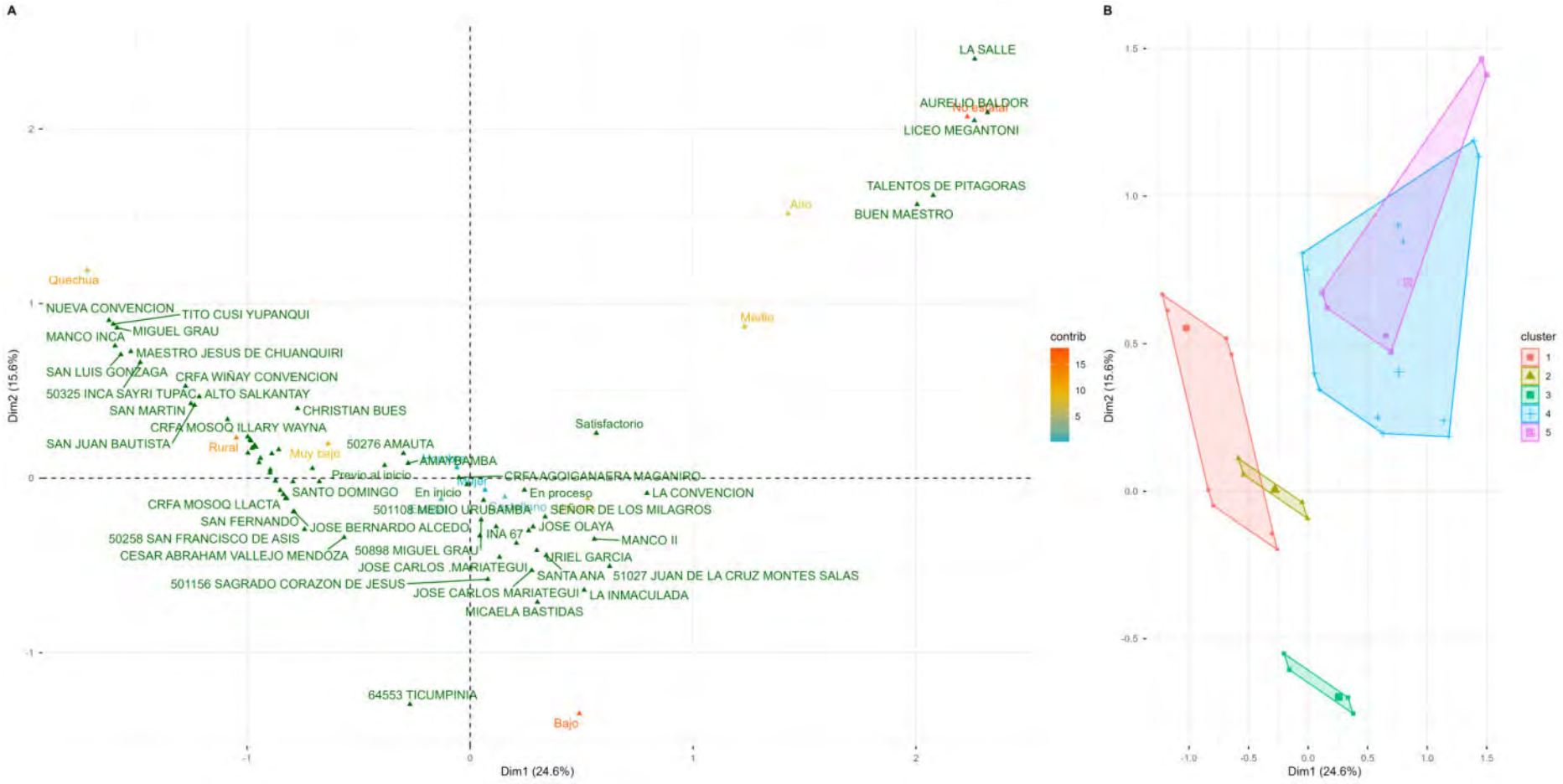
Biplot y clusters formados en la UGEL Canchis



En la figura 6, el biplot correspondiente a la UGEL Canchis evidencia que la primera dimensión está principalmente influenciada por el área de residencia, la lengua materna y el nivel socioeconómico de los estudiantes, lo que indica que este eje refleja principalmente diferencias de carácter territorial y sociocultural. En cambio, la segunda dimensión se relaciona de manera más marcada con el tipo de gestión institucional y con los niveles socioeconómicos más altos, lo que sugiere la presencia de un eje de diferenciación institucional y socioeconómica. En el plano factorial se observa que los estudiantes provenientes de zonas rurales, con lengua materna quechua y niveles socioeconómicos bajos tienden a concentrarse en un mismo sector del gráfico, mientras que las instituciones no estatales y los niveles socioeconómicos más favorables se ubican en una zona opuesta, lo que pone en evidencia la existencia de desigualdades socioeducativas claramente definidas en la UGEL Canchis.

Figura 7

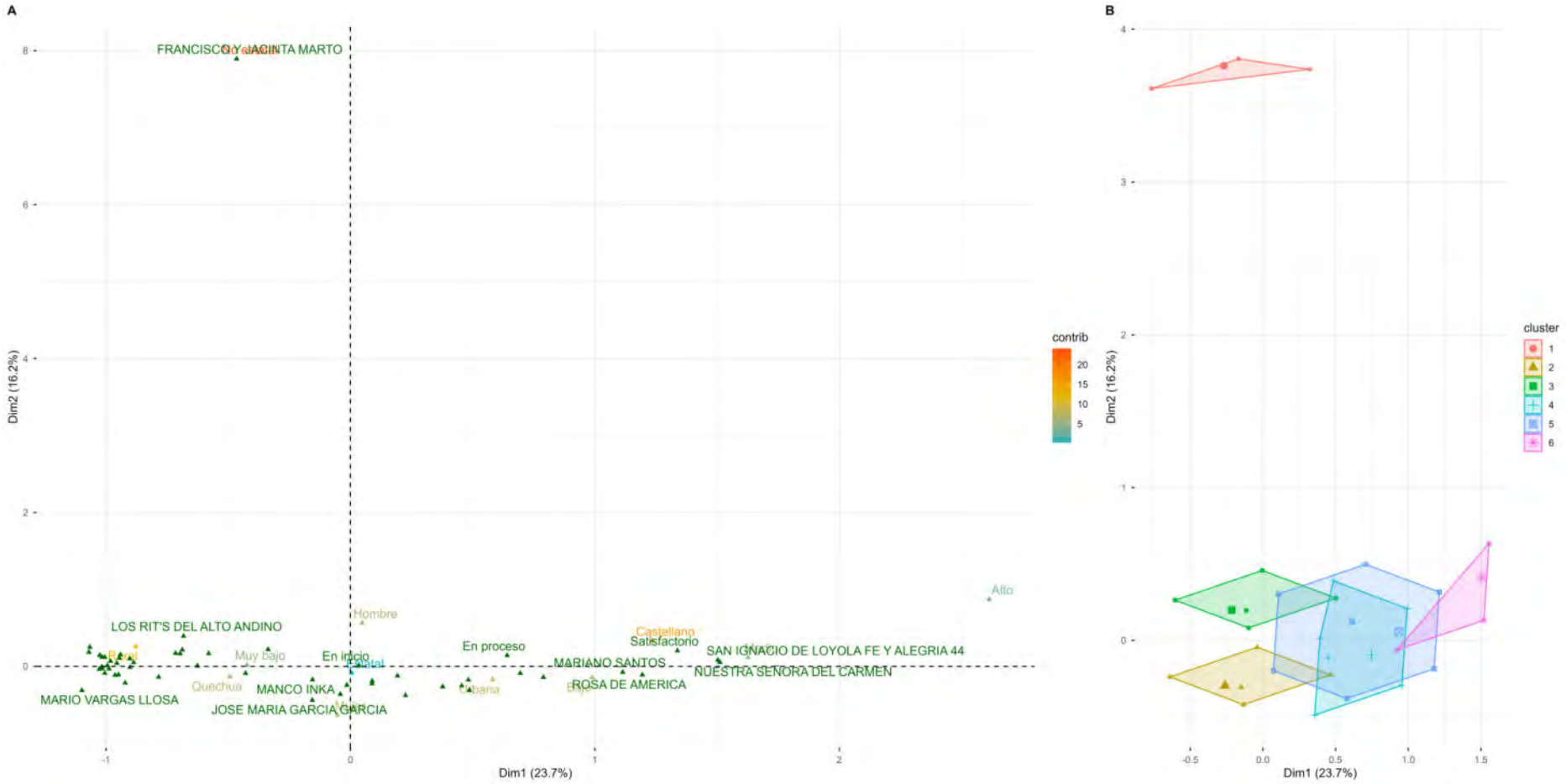
Biplot y clusters formados en la UGEL La Convención



El biplot de la UGEL La Convención evidencia que la primera dimensión está estrechamente asociada al contexto territorial de los estudiantes, al nivel socioeconómico y al tipo de gestión de la institución educativa. Este eje permite distinguir con claridad a los estudiantes de instituciones educativas como Nueva Convención, Manco Inca, Miguel Grau, entre otras, que provienen de zonas rurales, con lengua materna quechua y con menores condiciones socioeconómicas, respecto a aquellos que pertenecen a contextos más favorables. En cambio, la segunda dimensión refleja diferencias internas dentro de estos grupos, principalmente relacionadas con el nivel socioeconómico y el tipo de institución educativa. En el plano factorial se observa que los estudiantes en situación de mayor vulnerabilidad tienden a concentrarse en un mismo sector del gráfico, mientras que las instituciones no estatales y los estudiantes con mejores condiciones socioeconómicas se sitúan en una posición opuesta, lo que evidencia la presencia de patrones socioeducativos diferenciados dentro de la UGEL La Convención (ver figura 7).

Figura 8

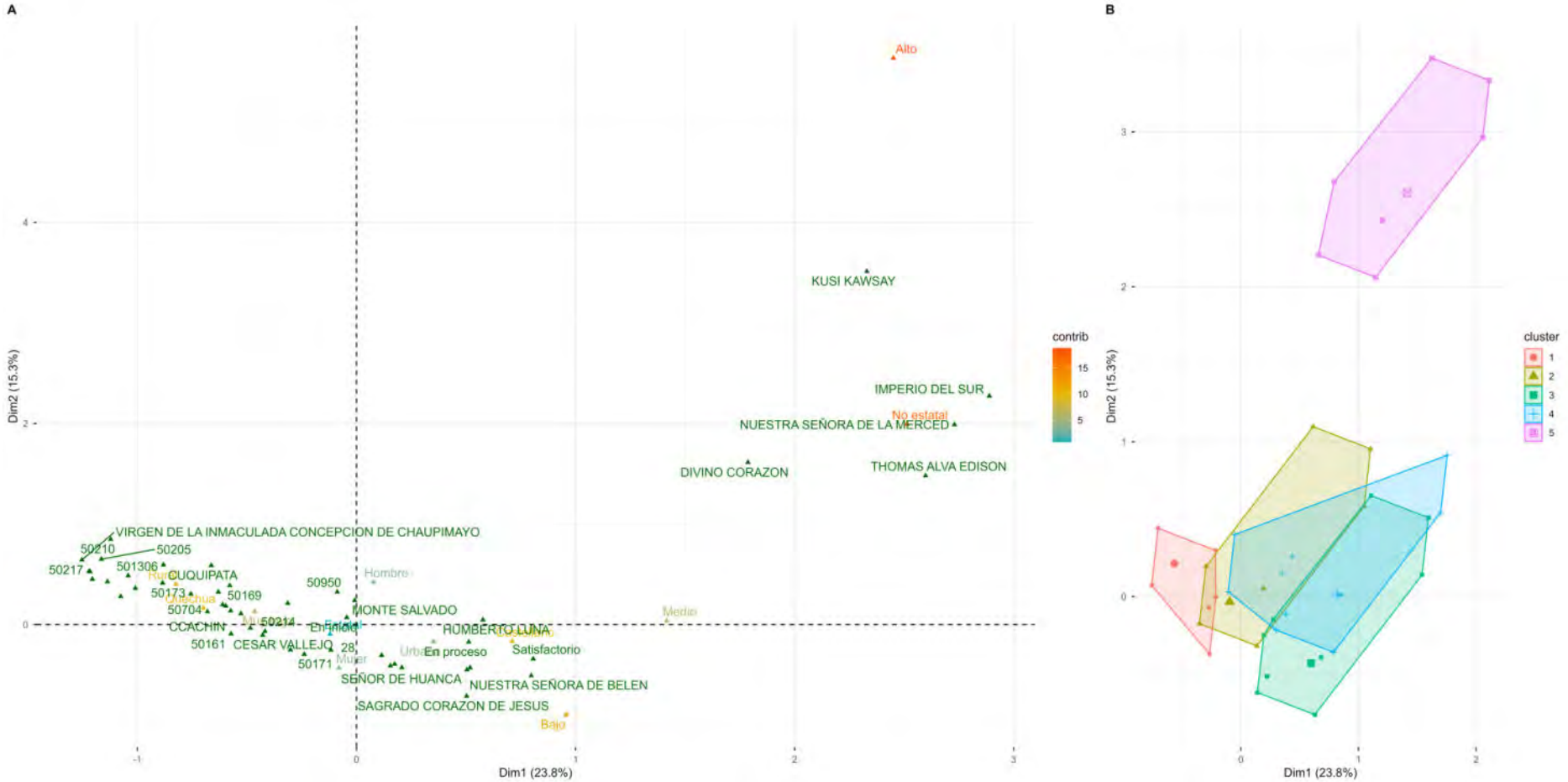
Biplot y clusters formados en la UGEL Quispicanchi



El biplot de la UGEL Quispicanchi evidencia que la primera dimensión está principalmente vinculada al idioma materno y al contexto territorial de los estudiantes, lo que permite diferenciar claramente a los estudiantes de lengua materna castellana de aquellos cuya lengua materna es el quechua, así como distinguir entre estudiantes provenientes de zonas urbanas y rurales. Por su parte, la segunda dimensión está fuertemente relacionada con el tipo de gestión institucional y el sexo de los estudiantes, evidenciándose una mayor contribución de las instituciones no estatales y de las categorías hombre y mujer en la diferenciación del espacio factorial. En el plano factorial se observa que los estudiantes de contextos rurales y con lengua materna quechua tienden a concentrarse en un mismo sector del gráfico, mientras que los estudiantes con lengua materna castellana y aquellos vinculados a instituciones no estatales se ubican en una posición diferente, lo que evidencia la existencia de patrones socioeducativos diferenciados dentro de la UGEL Quispicanchi (ver figura 8).

Figura 9

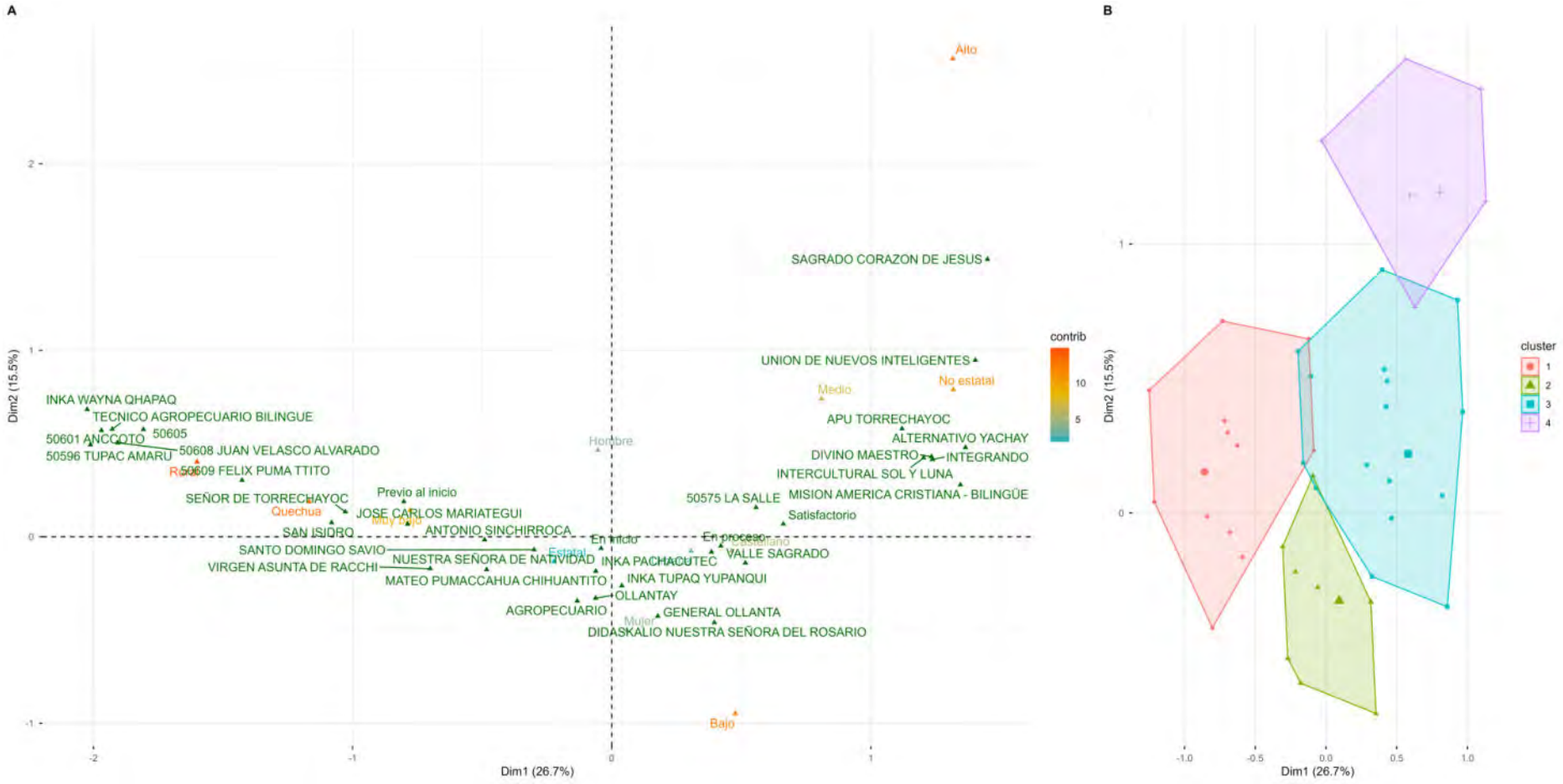
Biplot y clusters formados en la UGEL Calca



En la figura 9, el biplot de la UGEL Calca evidencia que la estructura del espacio factorial está determinada principalmente por diferencias vinculadas al contexto territorial de los estudiantes, al idioma materno y al tipo de institución educativa a la que pertenecen. En la primera dimensión se aprecia una separación vinculada al ámbito rural y urbano, junto con la distribución de los estudiantes según lengua materna, lo que evidencia la influencia de factores socioculturales en la organización del plano. Por su parte, la segunda dimensión permite distinguir con mayor claridad los niveles de logro académico, especialmente en los estudiantes que alcanzan niveles más altos de desempeño, los cuales se ubican en una zona claramente diferenciada del gráfico. Asimismo, se observa que las instituciones no estatales tienden a posicionarse en un sector distinto respecto a las instituciones estatales, lo que sugiere la existencia de patrones educativos diferenciados dentro de la UGEL Calca. En conjunto, el biplot muestra que los estudiantes no se distribuyen de manera uniforme, sino que se agrupan según características territoriales, lingüísticas y académicas.

Figura 10

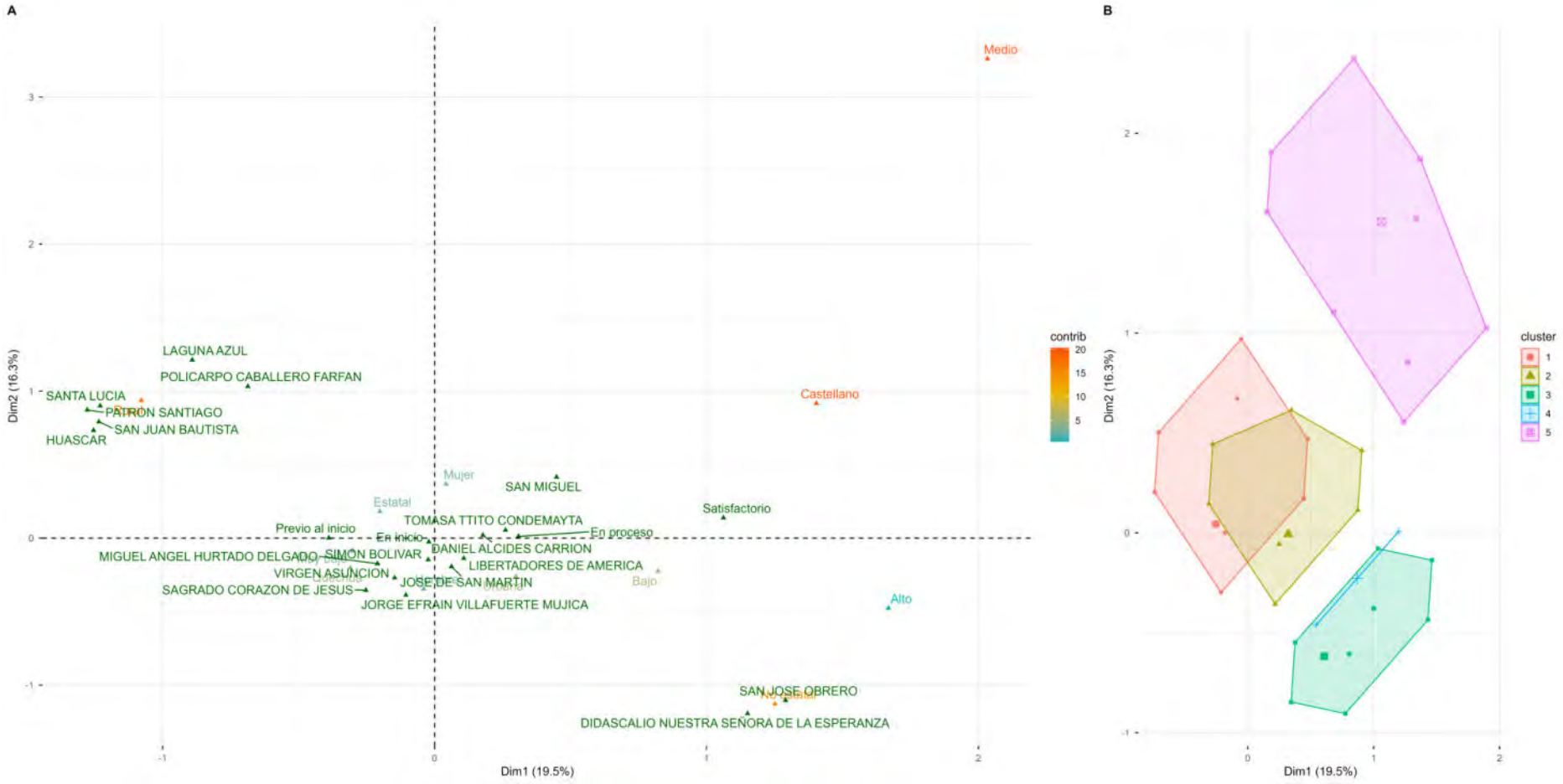
Biplot y clusters formados en la UGEL Urubamba



El biplot de la UGEL Urubamba muestra que la estructura del plano factorial está principalmente determinada por el contexto sociolingüístico y por el nivel de desempeño académico de los estudiantes. En la primera dimensión se observa una clara separación asociada principalmente al entorno rural y a la lengua materna quechua, lo que sugiere que los factores culturales y territoriales tienen un peso importante en la estructura del gráfico. En contraste, la segunda dimensión permite distinguir con mayor precisión los niveles de logro, especialmente los niveles bajo y alto, que se ubican en posiciones opuestas dentro del plano. Asimismo, las instituciones no estatales tienden a posicionarse en un sector distinto respecto a las instituciones estatales, lo que indica la presencia de diferencias educativas dentro de la UGEL. En conjunto, el biplot muestra que los estudiantes se agrupan en función de características socioculturales y académicas, evidenciando una estructura claramente diferenciada en el espacio factorial (ver figura 10).

Figura 11

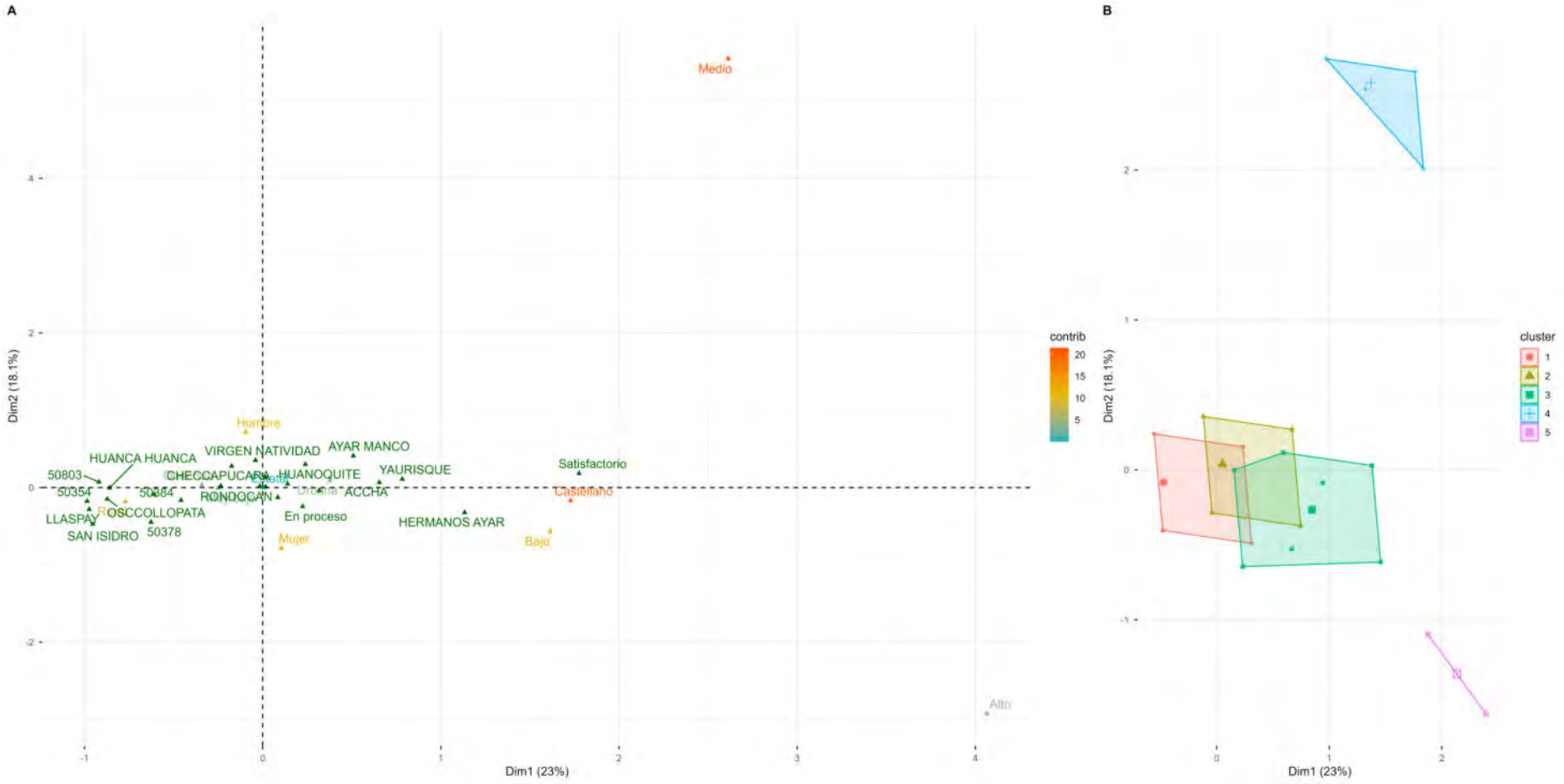
Biplot y clusters formados en la UGEL Acomayo



En la UGEL Acomayo, el biplot revela un patrón diferente al identificado en las demás UGEL, debido a que la configuración del plano factorial se explica principalmente por el idioma materno y por los niveles intermedios de logro académico, más que por el contexto territorial de los estudiantes. En la primera dimensión se aprecia una diferenciación marcada asociada principalmente a los estudiantes de lengua materna castellana y a las instituciones no estatales, mientras que la segunda dimensión está más vinculada al nivel de logro medio y a la distribución entre contextos rurales y urbanos. En el gráfico se observa que los estudiantes no se concentran únicamente en función del entorno rural, sino que se agrupan principalmente según el nivel de desempeño alcanzado, lo que indica que el factor académico tiene un mayor peso en la configuración del espacio factorial en esta UGEL. En conjunto, el biplot evidencia una estructura donde las diferencias educativas se explican más por el rendimiento académico y el idioma materno que por el contexto territorial (ver figura 11).

Figura 12

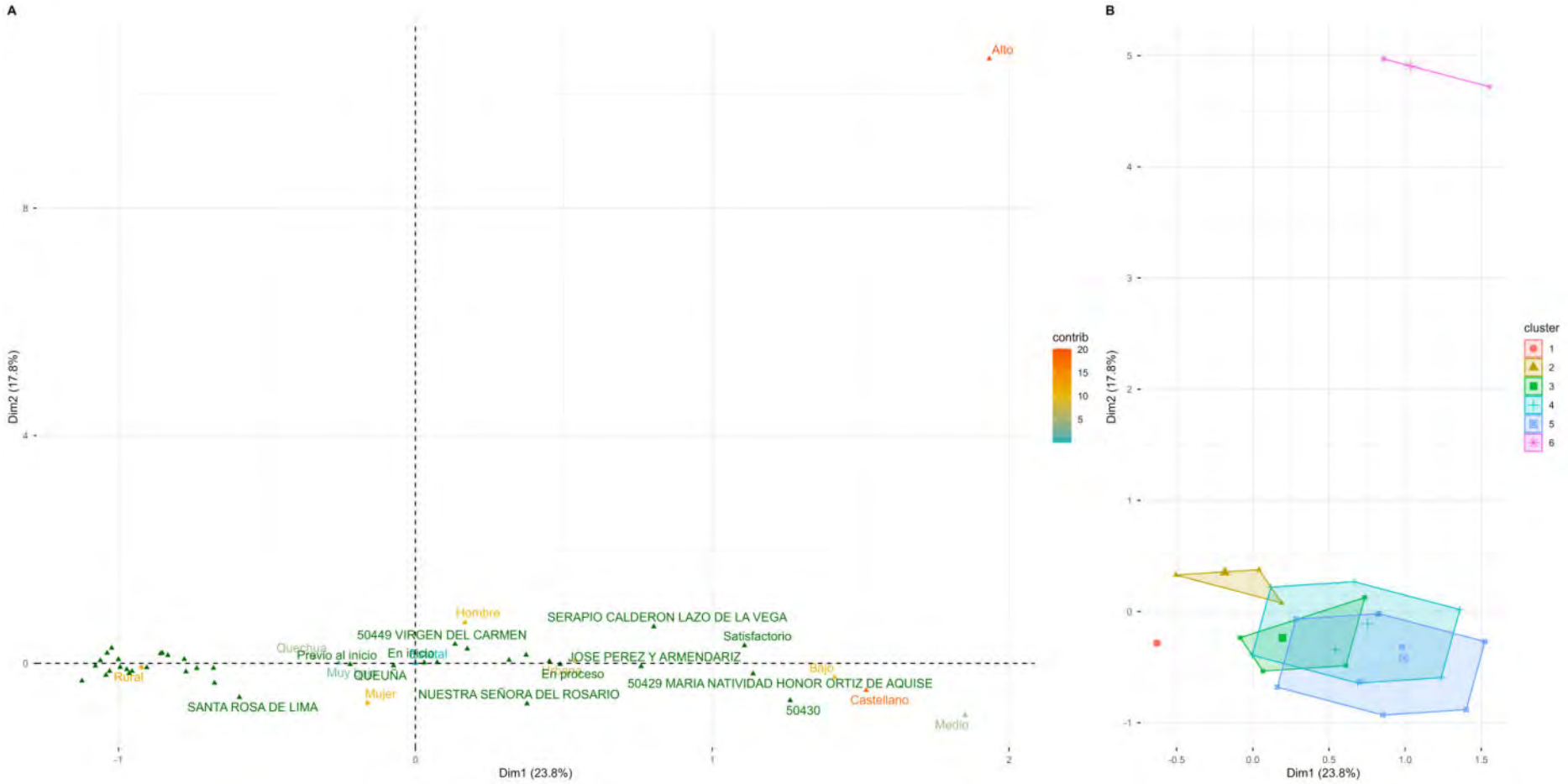
Biplot y clusters formados en la UGEL Paruro



En la UGEL Paruro, el biplot evidencia una estructura específica en la que la diferenciación entre los estudiantes se explica principalmente por el idioma materno y por los niveles de logro académico, más que por el tipo de institución educativa a la que pertenecen. En la primera dimensión se observa una clara separación asociada principalmente a los estudiantes de lengua materna castellana, mientras que la segunda dimensión está más relacionada con el nivel de logro medio y con la distribución según sexo, lo que genera una organización distinta del plano factorial. Asimismo, en el gráfico se aprecia que los estudiantes no se agrupan únicamente en función del contexto rural, sino que la diferenciación se produce principalmente según el rendimiento académico, donde los niveles bajo, medio y alto ocupan posiciones claramente separadas dentro del espacio factorial. En conjunto, el biplot evidencia que la estructura de la UGEL Paruro está determinada principalmente por el desempeño académico y el idioma materno, mostrando una organización diferente a la observada en las demás UGEL (ver figura 12).

Figura 13

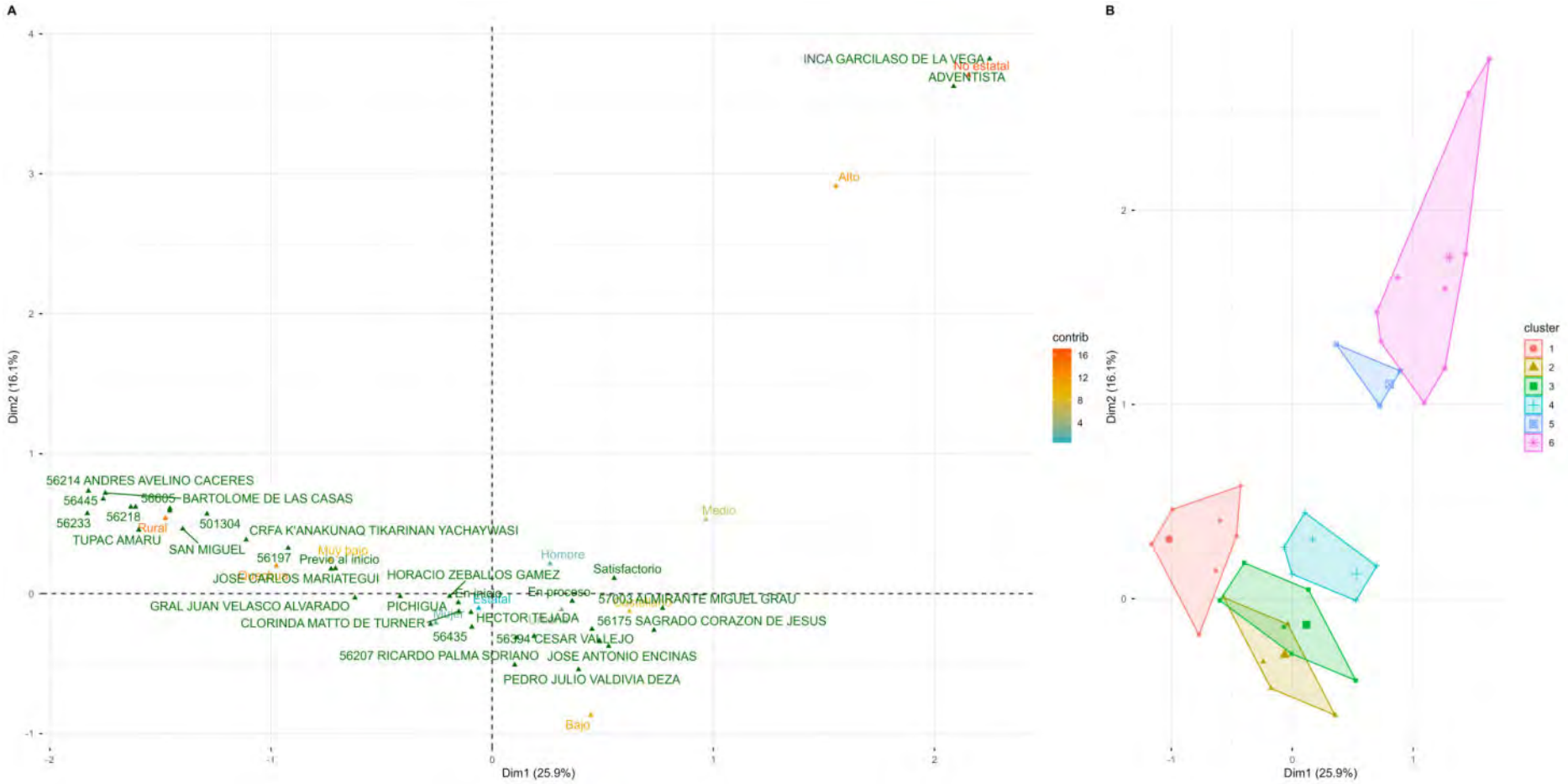
Biplot y clusters formados en la UGEL Paucartambo



En la figura 13, en la UGEL Paucartambo, el biplot muestra que la diferenciación principal de los estudiantes no se debe principalmente al tipo de institución educativa, sino a la interacción entre el idioma materno, el contexto territorial y el nivel de logro académico. En la primera dimensión se observa una marcada separación asociada a los estudiantes de lengua materna castellana y a quienes pertenecen a contextos rurales y urbanos, lo que evidencia la influencia de factores socioculturales en la configuración del plano factorial. Por otro lado, la segunda dimensión está más relacionada con los niveles de logro, especialmente el nivel alto, que se ubica en una posición claramente diferenciada dentro del gráfico, generando una organización distinta respecto a los demás niveles. Asimismo, se aprecia que los estudiantes no se concentran en un solo sector, sino que se distribuyen en función de su desempeño académico y del contexto sociolingüístico, lo que refleja una estructura socioeducativa heterogénea dentro de la UGEL Paucartambo.

Figura 14

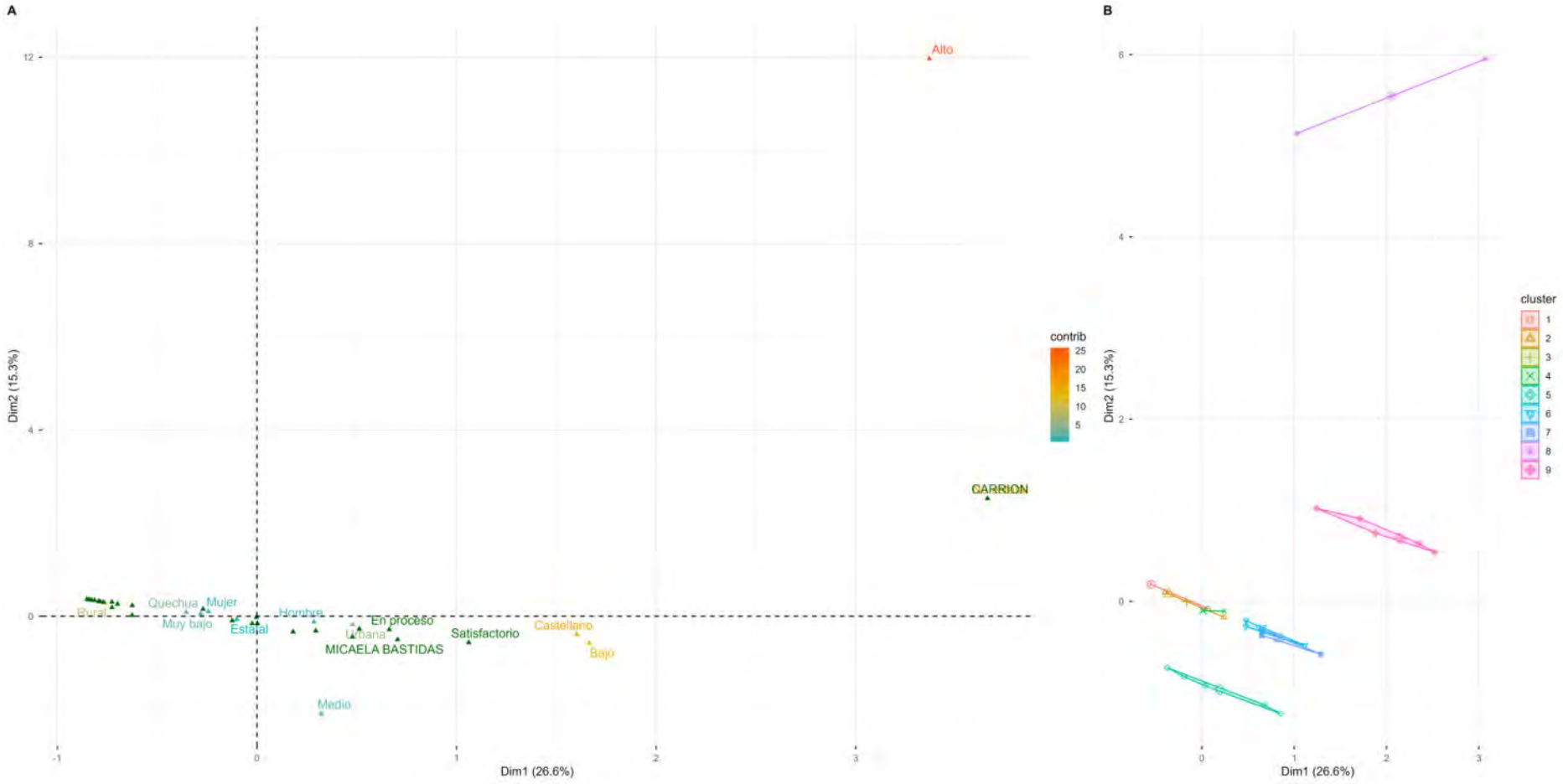
Biplot y clusters formados en la UGEL Espinar



En la UGEL Espinar, el biplot evidencia una estructura en la que la diferenciación entre los estudiantes está determinada principalmente por el contexto sociocultural y por el nivel de desempeño académico, más que por el tipo de gestión institucional. En la primera dimensión se identifica una separación marcada vinculada al ámbito rural y a la lengua materna quechua, lo que refleja la influencia predominante de factores culturales y territoriales en la configuración del plano factorial. Por su parte, la segunda dimensión permite distinguir con mayor claridad los niveles de logro académico, especialmente los niveles bajo y alto, que se ubican en posiciones claramente diferenciadas dentro del gráfico. Además, se aprecia que las instituciones no estatales tienden a situarse en un sector distinto respecto a las instituciones estatales, aunque su influencia es menor en comparación con las variables socioculturales. En conjunto, el biplot refleja que los estudiantes se agrupan principalmente según su contexto sociolingüístico y su nivel de rendimiento, mostrando una estructura socioeducativa diferenciada dentro de la UGEL Espinar (ver figura 14).

Figura 15

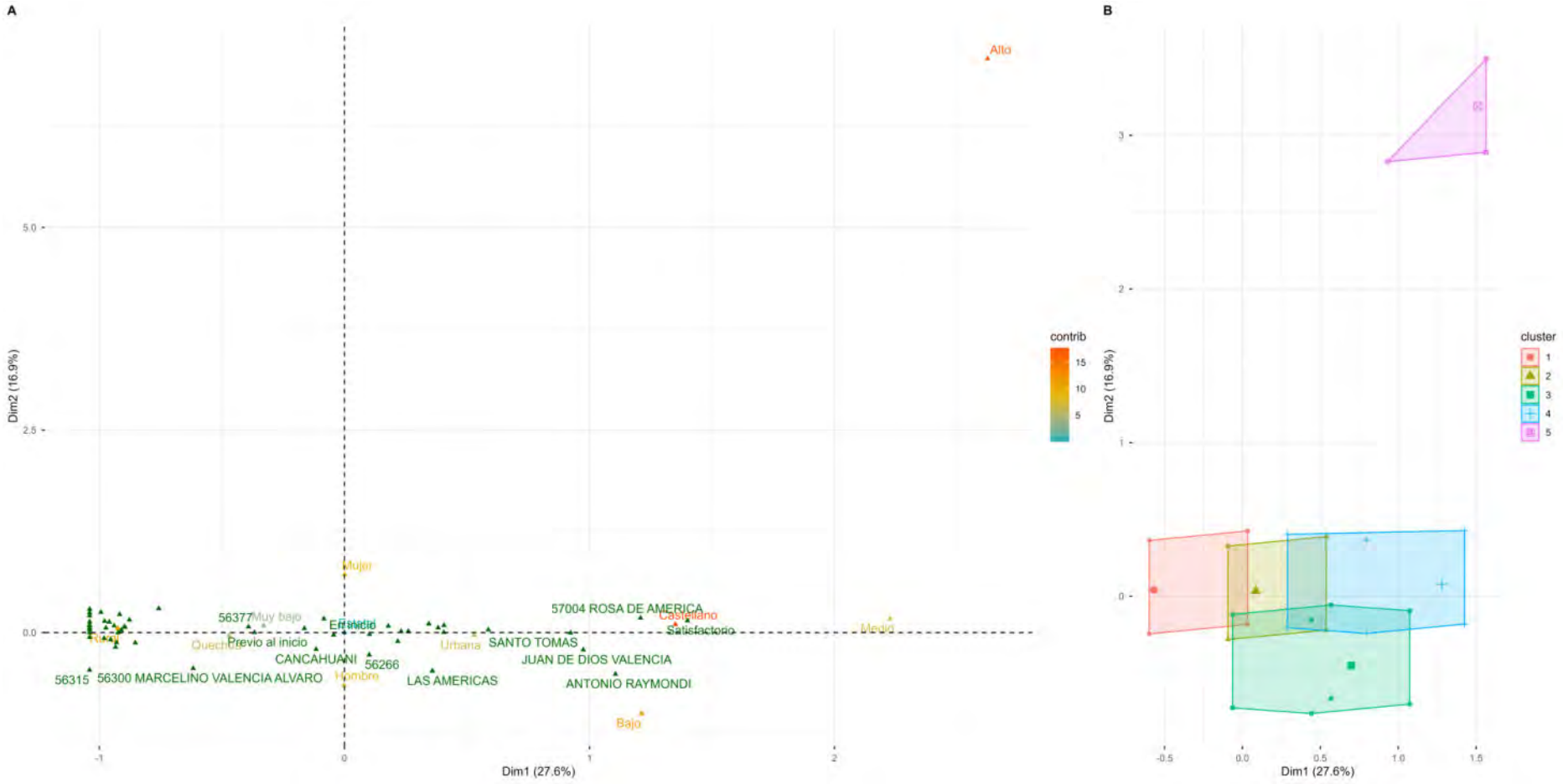
Biplot y clusters formados en la UGEL Canas



Para la UGEL Canas, el biplot evidencia que la Dimensión 1 está principalmente asociada a las categorías Castellano, No estatal, Bajo y Rural, reflejando un eje que diferencia según tipo de gestión, lengua materna y nivel socioeconómico. Por su parte, la Dimensión 2 está dominada por la categoría Alto, lo que indica que este eje discrimina principalmente a los estudiantes con niveles socioeconómicos más elevados. En cambio, las categorías Medio y el sexo (Hombre/Mujer) presentan mayor relevancia en dimensiones posteriores, por lo que su aporte en el plano principal es limitado. En conjunto, el biplot muestra una diferenciación entre perfiles asociados a contextos socioeconómicos bajos y características rurales frente a aquellos vinculados a niveles socioeconómicos altos (ver figura 15).

Figura 16

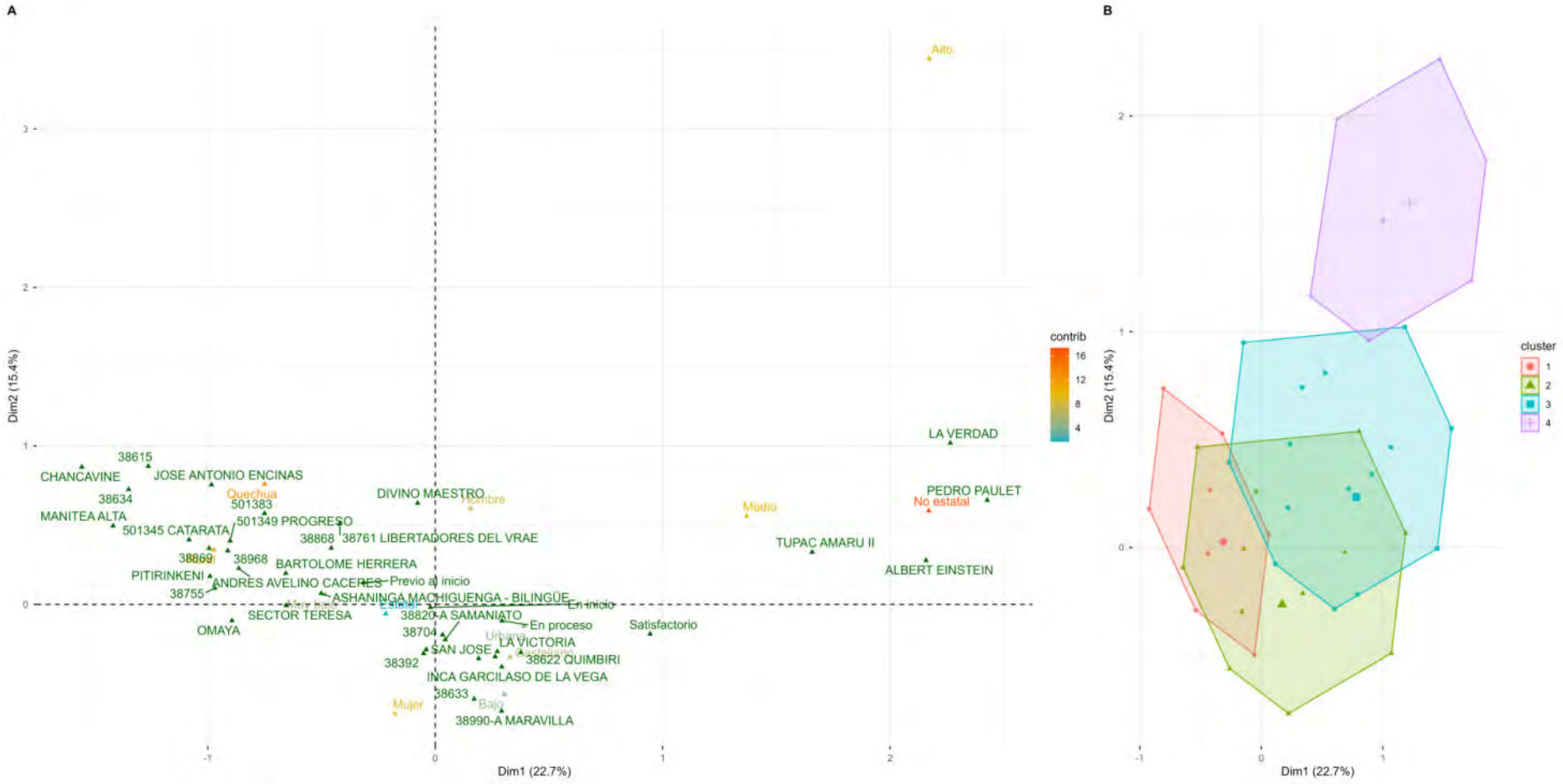
Biplot y clusters formados en la UGEL Chumbivilcas



En la figura 16, en la UGEL Chumbivilcas, el biplot sugiere que el primer eje organiza principalmente a las categorías vinculadas con Castellano, ámbitos Rural/Urbana y el nivel Bajo, reflejando un gradiente asociado a condiciones territoriales y socioeconómicas menos favorables. En contraste, el segundo eje está definido sobre todo por la categoría Alto y por el sexo (Hombre/Mujer), lo que indica una diferenciación relacionada con niveles socioeconómicos más altos y cierta separación por género. Por otro lado, la categoría Medio adquiere mayor protagonismo en dimensiones adicionales, por lo que su incidencia en el plano principal es reducida. En conjunto, la configuración del biplot evidencia una oposición entre contextos predominantemente rurales y de menor nivel socioeconómico frente a aquellos con mejores condiciones.

Figura 17

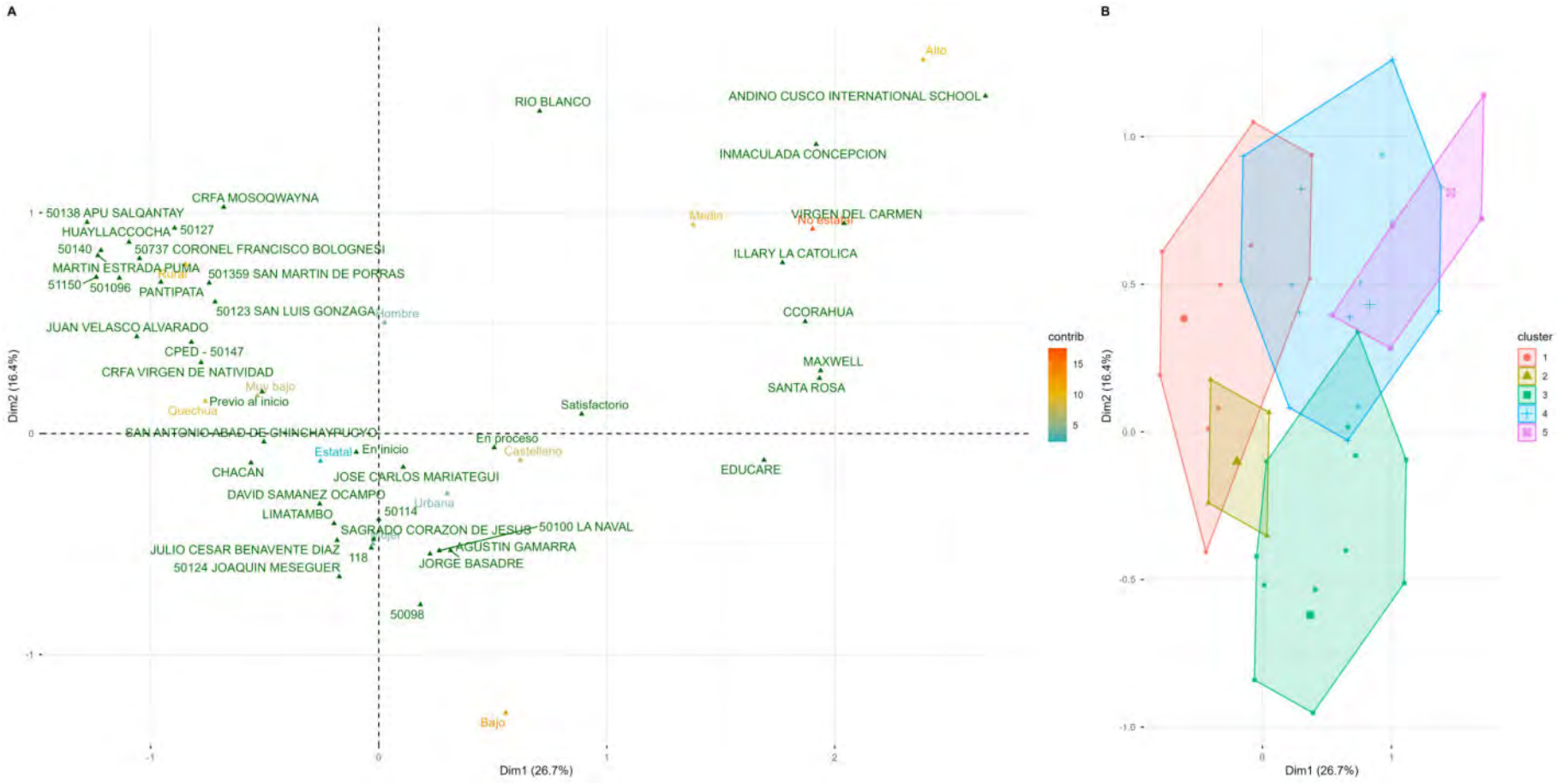
Biplot y clusters formados en la UGEL Pichari-Kimbiri



En la UGEL Pichari-Kimbiri, el biplot muestra que el primer eje está determinado principalmente por las categorías No estatal, Rural, Medio y Quechua, lo que evidencia una diferenciación asociada tanto al tipo de gestión como a características territoriales y lingüísticas, junto con ciertos niveles socioeconómicos intermedios. En cambio, el segundo eje se encuentra más influenciado por el sexo (Hombre/Mujer), la categoría Alto y en menor medida Quechua, indicando una separación vinculada a niveles socioeconómicos más elevados y diferencias de género. Asimismo, la categoría Bajo adquiere mayor relevancia en dimensiones posteriores, especialmente fuera del plano principal, por lo que su aporte en este es limitado. En conjunto, el biplot refleja una estructura donde se contraponen perfiles ligados a contextos rurales, gestión no estatal y lengua originaria frente a aquellos con mejores condiciones socioeconómicas y diferenciación por sexo (ver figura 17).

Figura 18

Biplot y clusters formados en la UGEL Anta



En la UGEL Anta, el biplot evidencia que la primera dimensión está principalmente definida por la categoría No estatal, en combinación con Quechua, Castellano, Rural y el nivel Alto, configurando un eje que integra características de gestión institucional, contexto territorial y diversidad lingüística. Por otro lado, la segunda dimensión está dominada por la categoría Bajo, lo que señala una diferenciación marcada por condiciones socioeconómicas menos favorables. A su vez, las categorías Medio y Alto, junto con el sexo (Hombre/Mujer), muestran mayor incidencia en dimensiones posteriores, limitando su aporte en el plano principal. En conjunto, el biplot sugiere una estructura donde interactúan factores institucionales, territoriales y socioeconómicos, generando contrastes entre distintos niveles de bienestar (ver figura 18).

5.3. Perfiles de rendimiento académico por UGEL

Tabla 3

Porcentaje de Características Sociodemográficas y de Rendimiento en Lectura por Cluster en la UGEL Cusco

		GESTIÓN		ÁREA		SEXO		LENGUA MATERNA		ISE				LECTURA			
Cluster	N	Estatal	No estatal	Urbana	Rural	Mujer	Hombre	Castellano	Quechua	Muy bajo	Bajo	Medio	Alto	Previo inicio	En inicio	En proceso	Satisfactorio
1	1535	32.64	67.36	100	0	46.51	53.49	99.61	0.39	2.87	0	0	97	2.67	23.84	36.94	36.55
2	1963	61.59	38.41	100	0	47.94	52.06	100	0	0	0	100	0	5.15	31.58	36.73	26.54
3	1964	82.28	17.72	100	0	50.1	49.9	100	0	0	100	0	0	7.23	40.02	35.34	17.41
4	85	64.71	35.29	0	100	43.53	56.47	72.94	27.06	41.18	16.5	18.82	24	22.35	34.12	23.53	20
5	1628	94.9	5.1	100	0	57.06	42.94	61.12	38.88	83.05	11.8	4.67	0.5	17.01	48.77	25.31	8.91

Como se presenta en la Tabla 3, los clusters 1 y 5, que concentran la mayor cantidad de estudiantes, muestran un predominio de instituciones no estatales (cluster 1: 67.36%) y estatales (cluster 5: 94.9%), con niveles socioeconómicos mayoritariamente altos (cluster 1: 97.13%) y muy bajos (cluster 5: 83,05%), respectivamente. Mientras que el cluster 1 evidencia un rendimiento aceptable (En proceso + Satisfactorio = 73.49%), el cluster 5 presenta un resultado significativamente inferior (34.22%), destacando la fuerte influencia del nivel socioeconómico. Por otro lado, los clusters 2 y 3, con mayoría de gestión estatal y niveles socioeconómicos medio y bajo, alcanzan un rendimiento aceptable del 63.27% y 52.75%, respectivamente, mostrando que, aun en contextos menos favorables, existen patrones de desempeño moderado. El cluster 4, aunque pequeño y rural, presenta un perfil mixto con un 43.53% de rendimiento aceptable. En conjunto, la UGEL Cusco refleja una brecha educativa asociada a la

desigualdad socioeconómica, donde los estudiantes de mayores recursos y de gestión no estatal tienden a alcanzar mejores resultados, mientras que los de entornos estatales y socioeconómicamente vulnerables enfrentan mayores desafíos para lograr un rendimiento satisfactorio.

Tabla 4

Porcentaje de Características Sociodemográficas y de Rendimiento en Lectura por Cluster en la UGEL Canchis

		GESTIÓN		ÁREA		SEXO		LENGUA MATERNA		ISE				LECTURA			
Cluster	N	Estatal	No estatal	Urbana	Rural	Mujer	Hombre	Castellano	Quechua	Muy bajo	Bajo	Medio	Alto	Previo inicio	En inicio	En proceso	Satisfactorio
1	813	100	0	62.12	37.88	47.6	52.4	5.17	94.83	96.8	2.83	0.37	0	37.39	47.36	12.79	2.46
2	948	100	0	99.89	0.11	47.36	52.64	86.08	13.92	46.1	53.9	0	0	14.87	45.89	27.22	12.03
3	267	100	0	99.25	0.75	44.94	55.06	91.01	8.99	0	0	100	0	9.74	37.45	27.72	25.09
4	73	100	0	95.89	4.11	41.1	58.9	95.89	4.11	0	0	0	100	9.59	26.03	28.77	35.62
5	46	0	100	100	0	54.35	45.65	80.43	19.57	23.91	34.8	26.09	15	26.09	39.13	23.91	10.87

En los datos tabulados (Tabla 4), la UGEL Canchis muestra una estructura de clusters que refleja una transición progresiva desde condiciones de alta vulnerabilidad hacia perfiles con mayor ventaja socioeducativa, observándose una relación directa entre el nivel socioeconómico y el rendimiento aceptable. El cluster 1, mayoritariamente rural, con lengua materna quechua (94.83%) y nivel socioeconómico muy bajo (96.80%), presenta un rendimiento aceptable mínimo (15.25%), evidenciando los mayores desafíos de aprendizaje. El cluster 2, urbano y con nivel bajo, mejora ligeramente este indicador a 39.25%. Un salto significativo se da en el cluster 3, con nivel socioeconómico medio (100%), donde el rendimiento aceptable alcanza el 52.81%.

Los clusters 4 y 5, con niveles socioeconómicos alto y mixto (alto, medio, bajo), respectivamente, muestran los mejores resultados: 64.39% y 34.78% de rendimiento aceptable. Este patrón indica que, en Canchis, la superación del nivel socioeconómico bajo es un factor crítico para mejorar los resultados académicos, siendo el dominio del castellano y el contexto urbano factores asociados a una mayor probabilidad de alcanzar un desempeño en proceso o satisfactorio.

Tabla 5

Porcentaje de Características Sociodemográficas y de Rendimiento en Lectura por Cluster en la UGEL La Convención

Cluster	N	GESTIÓN		ÁREA		SEXO		LENGUA MATERNA		ISE				LECTURA			
		Estatal	No estatal	Urbana	Rural	Mujer	Hombre	Castellano	Quechua	Muy bajo	Bajo	Medio	Alto	Previo inicio	En inicio	En proceso	Satisfactorio
1	135	100	0	28.89	71.11	40	60	0	100	92.59	7.41	0	0	36.3	45.19	14.81	3.7
2	809	100	0	53.15	46.85	49.44	50.56	100	0	100	0	0	0	26.33	53.15	17.43	3.09
3	355	100	0	81.97	18.03	45.63	54.37	100	0	0	100	0	0	15.77	52.39	25.35	6.48
4	257	72.37	27.63	96.11	3.89	45.14	54.86	99.22	0.78	3.11	8.56	88.33	0	9.34	48.64	28.4	13.62
5	92	78.26	21.74	97.83	2.17	47.83	52.17	100	0	0	0	0	100	11.96	36.96	28.26	22.83

De acuerdo a los valores consignados en la Tabla 5, la UGEL La Convención evidencia perfiles de rendimiento académico fuertemente segmentados según las características culturales y socioeconómicas de los estudiantes. Los clusters 1 y 2, definidos por una lengua materna exclusivamente quechua y un nivel socioeconómico mayoritariamente muy bajo, presentan los porcentajes más bajos de estudiantes en los niveles de logro En proceso y Satisfactorio (18.51% y 20.52%, respectivamente), identificando así un patrón claro de bajo rendimiento académico asociado a condiciones de ruralidad y contexto cultural quechua.

En contraste, se observa una transición positiva en los clusters 3, 4 y 5, donde la lengua materna predominante es el castellano. Esta relación se hace más evidente en el cluster 5, que, aun siendo de gestión principalmente estatal, combina un nivel socioeconómico alto con el mayor porcentaje de rendimiento aceptable (51.09%), lo que sugiere que las variables institucionales y socioeconómicas actúan de forma interrelacionada. En conjunto, el análisis identifica que en esta UGEL la variable lengua materna y el nivel socioeconómico están estrechamente relacionados con el rendimiento en la evaluación censal, señalando que los estudiantes quechua hablantes de bajo nivel socioeconómico constituyen el perfil prioritario para intervenciones educativas focalizadas.

Tabla 6

Porcentaje de Características Sociodemográficas y de Rendimiento en Lectura por Cluster en la UGEL Quispicanchi

		GESTIÓN		ÁREA		SEXO		LENGUA MATERNA		ISE				LECTURA			
Cluster	N	Estatal	No estatal	Urbana	Rural	Mujer	Hombre	Castellano	Quechua	Muy bajo	Bajo	Medio	Alto	Previo inicio	En inicio	En proceso	Satisfactorio
1	19	0	100	0	100	0	100	78.95	21.05	94.74	5.26	0	0	10.53	57.89	26.32	5.26
2	655	100	0	49.31	50.69	100	0	17.71	82.29	95.42	4.58	0	0	44.73	45.95	7.94	1.37
3	698	100	0	51.58	48.42	0	100	16.62	83.38	94.13	5.87	0	0	39.54	46.13	12.75	1.58
4	283	100	0	98.59	1.41	46.29	53.71	62.9	37.1	0	100	0	0	18.37	45.23	24.38	12.01
5	95	100	0	88.42	11.58	43.16	56.84	66.32	33.68	0	0	100	0	27.37	34.74	28.42	9.47
6	17	100	0	100	0	41.18	58.82	94.12	5.88	0	0	0	100	0	41.18	29.41	29.41

El análisis de correspondencias múltiples y clusters jerárquicos aplicado a los datos de la UGEL Quispicanchi (Tabla 6) evidencia una asociación significativa entre las características contextuales de los estudiantes y su rendimiento académico en la Evaluación Censal de Estudiantes. Los resultados muestran que tanto la ubicación geográfica como el nivel socioeconómico desempeñan un papel determinante en el logro de los aprendizajes. En particular, los clusters 2 y 3, conformados mayoritariamente por estudiantes de zonas rurales (50.69% y 48.42%, respectivamente) y con un nivel socioeconómico clasificado como muy bajo (≥94%), registran las menores proporciones de estudiantes ubicados en los niveles de logro En Proceso y Satisfactorio (9.31% y 14.33%, respectivamente). Estos hallazgos sugieren que la convergencia entre la ruralidad y la vulnerabilidad socioeconómica configura un perfil con mayor riesgo de bajo rendimiento académico.

En contraste, se observa un gradiente favorable en los clusters con mayor proporción de estudiantes provenientes de áreas urbanas (clusters 4, 5 y 6, con 98.59%, 88.42% y 100%, respectivamente) y con niveles socioeconómicos progresivamente más altos (bajo, medio y alto). En estos grupos se aprecia un incremento sostenido en la proporción de estudiantes con rendimiento aceptable, alcanzando valores de 36.39%, 37.89% y 58.82%, respectivamente. La transición desde un nivel socioeconómico muy bajo hacia categorías superiores aparece consistentemente asociada a una mejora en los resultados de aprendizaje. En conclusión, los datos analizados apoyan la existencia de una brecha educativa vinculada a disparidades territoriales y socioeconómicas en la UGEL Quispicanchi.

Tabla 7

Porcentaje de Características Sociodemográficas y de Rendimiento en Lectura por Cluster en la UGEL Chumbivilcas

		GESTIÓN		ÁREA		SEXO		LENGUA MATERNA		ISE				LECTURA			
Cluster	N	Estatal	No estatal	Urbana	Rural	Mujer	Hombre	Castellano	Quechua	Muy bajo	Bajo	Medio	Alto	Previo inicio	En inicio	En proceso	Satisfactorio
1	518	100	0	0	100	46.53	53.47	4.83	95.17	100	0	0	0	50.39	40.93	8.11	0.58
2	714	100	0	100	0	49.16	50.84	28.29	71.71	100	0	0	0	31.93	49.44	15.13	3.5
3	200	100	0	87	13	45.5	54.5	51	49	0	100	0	0	26.5	37	21.5	15
4	59	100	0	94.92	5.08	44.07	55.93	81.36	18.64	0	0	100	0	13.56	28.81	37.29	20.34
5	12	100	0	100	0	50	50	91.67	8.33	0	0	0	100	8.33	33.33	33.33	25

Como se puede observar en la tabla correspondiente (Tabla 7), la aplicación de técnicas de correspondencias múltiples y agrupamiento jerárquico en la UGEL Chumbivilcas permite identificar perfiles estudiantiles con distintos niveles de desempeño académico, los cuales se encuentran vinculados a condiciones contextuales y socioeconómicas específicas. El análisis evidencia notorias diferencias en el porcentaje de estudiantes que alcanzan los niveles de logro En proceso y Satisfactorio entre los distintos grupos identificados. El Cluster 1, caracterizado por población 100% rural, lengua materna quechua (95.17%) y nivel socioeconómico muy bajo (100%), presenta el porcentaje más bajo de rendimiento aceptable (8.69%: 8.11% En proceso + 0.58% Satisfactorio), indicando condiciones educativas críticas asociadas al aislamiento geográfico y la vulnerabilidad socioeconómica. El cluster 2, aunque mantiene un nivel Muy bajo (100%), muestra una composición 100% urbana y un incremento en el rendimiento aceptable (18.63%: 15.13% En proceso + 3.50% Satisfactorio), sugiriendo que el contexto urbano mitiga

parcialmente los efectos del bajo nivel socioeconómico. Una transición significativa se observa en el cluster 3, con nivel socioeconómico bajo (100%) y mayor proporción urbana (87%), donde el rendimiento aceptable alcanza el 36.50% (21.50% En proceso + 15.00% Satisfactorio). Este patrón se consolida en los clusters 4 y 5, que presentan niveles socioeconómicos medio (100%) y alto (100%) respectivamente, con rendimientos aceptables de 57.63% (37.29% En proceso + 20.34% Satisfactorio) y 58.33% (33.33% En proceso + 25.00% Satisfactorio). Esta progresión evidencia una asociación positiva consistente entre la mejora en el nivel socioeconómico y el logro académico.

Tabla 8

Porcentaje de Características Sociodemográficas y de Rendimiento en Lectura por Cluster en la UGEL Calca

		GESTIÓN		ÁREA		SEXO		LENGUA MATERNA		ISE				LECTURA			
Cluster	N	Estatal	No estatal	Urbana	Rural	Mujer	Hombre	Castellano	Quechua	Muy bajo	Bajo	Medio	Alto	Previo inicio	En inicio	En proceso	Satisfactorio
1	333	100	0	0	100	45.65	54.35	29.43	70.57	96.1	3.9	0	0	41.44	47.45	8.71	2.4
2	568	97.01	2.99	100	0	52.99	47.01	41.73	58.27	100	0	0	0	29.4	48.42	16.55	5.63
3	227	90.75	9.25	90.31	9.69	47.58	52.42	78.41	21.59	0	100	0	0	11.89	40.09	28.63	19.38
4	100	87	13	84	16	45	55	89	11	0	0	100	0	4	42	32	22
5	17	64.71	35.29	94.12	5.88	47.06	52.94	88.24	11.76	0	0	0	100	5.88	47.06	29.41	17.65

A través del análisis de correspondencias múltiples y agrupamiento jerárquico en la UGEL Calca (Tabla 8), se identifica una relación significativa entre el contexto socioeducativo de los estudiantes y sus resultados en la evaluación censal. Se observa un incremento gradual en el logro de los niveles En proceso y Satisfactorio a medida que se presentan mejores condiciones

socioeconómicas y entornos educativos más favorables. El cluster 1, caracterizado por población 100% rural, lengua materna quechua (70.57%) y nivel socioeconómico muy bajo (96.10%), presenta el porcentaje más bajo de rendimiento aceptable (11.11%: 8.71% En proceso + 2.40% Satisfactorio), identificando condiciones educativas críticas asociadas al aislamiento geográfico y la vulnerabilidad socioeconómica. El cluster 2, pese a estar conformado en su totalidad por estudiantes con un nivel socioeconómico muy bajo (100%), presenta una composición completamente urbana (100%) y un incremento moderado en el rendimiento aceptable, que alcanza el 22.18% (16.55% en En Proceso y 5.63% en Satisfactorio). Una mejora más marcada se observa en el cluster 3, integrado por estudiantes con nivel socioeconómico Bajo (100%) y predominio de población urbana (90.31%), donde el porcentaje de rendimiento aceptable asciende al 48.01% (28.63% en En Proceso y 19.38% en Satisfactorio). Esta tendencia se mantiene en los clusters 4 y 5, caracterizados por niveles socioeconómicos Medio (100%) y Alto (100%), respectivamente. En estos grupos, la proporción de estudiantes con rendimiento aceptable alcanza el 54.00% (32.00% en En Proceso y 22.00% en Satisfactorio) en el cluster 4, y el 47.06% (29.41% en En Proceso y 17.65% en Satisfactorio) en el cluster 5. En conjunto, los resultados de la UGEL Calca evidencian un gradiente en el rendimiento académico, donde la transición desde un nivel socioeconómico muy bajo hacia categorías socioeconómicas más favorables se asocia con mejoras sustanciales en los resultados de aprendizaje, particularmente cuando estas condiciones se combinan con contextos predominantemente urbanos.

Tabla 9

Porcentaje de Características Sociodemográficas y de Rendimiento en Lectura por Cluster en la UGEL Urubamba

		GESTIÓN		ÁREA		SEXO		LENGUA MATERNA		ISE				LECTURA			
Cluster	N	Estatal	No estatal	Urbana	Rural	Mujer	Hombre	Castellano	Quechua	Muy bajo	Bajo	Medio	Alto	Previo inicio	En inicio	En proceso	Satisfactorio
1	305	98.69	1.31	44.26	55.74	45.9	54.1	13.44	86.56	95.41	3.28	1.31	0	42.62	47.21	8.52	1.64
2	539	100	0	97.22	2.78	50.65	49.35	90.35	9.65	42.3	57.7	0	0	10.76	52.5	26.53	10.2
3	289	50.52	49.48	98.27	1.73	46.71	53.29	93.43	6.57	7.96	23.2	68.86	0	8.65	42.21	33.56	15.57
4	55	54.55	45.45	98.18	1.82	41.82	58.18	96.36	3.64	0	0	0	100	9.09	34.55	29.09	27.27

Mediante la aplicación del Análisis de Correspondencias Múltiples y del agrupamiento jerárquico a los datos de la UGEL Urubamba (Tabla 9), se identificaron diversos perfiles de desempeño académico derivados de la interacción entre el nivel socioeconómico, la ubicación geográfica y el tipo de gestión educativa. Los resultados evidencian diferencias significativas en la proporción de estudiantes que alcanzan los niveles En Proceso y Satisfactorio entre los distintos perfiles identificados.

El cluster 1, conformado mayoritariamente por estudiantes de zonas rurales (55.74%), pertenecientes a instituciones de gestión estatal (98.69%) y con un nivel socioeconómico muy bajo (95.41%), registra el menor porcentaje de rendimiento aceptable, con un 10.16% (8.52% en En Proceso y 1.64% en Satisfactorio). Este perfil representa las condiciones de mayor vulnerabilidad educativa dentro de la UGEL.

Por su parte, el cluster 2, aunque integrado exclusivamente por instituciones de gestión estatal (100%), presenta un predominio de estudiantes de contextos urbanos (97.22%) y un nivel socioeconómico bajo (57.70%), alcanzando un rendimiento aceptable del 36.73% (26.53% en En Proceso y 10.20% en Satisfactorio), lo que evidencia una mejora sustancial respecto al perfil anterior.

Una evolución más favorable se observa en el cluster 3, caracterizado por una distribución equilibrada entre instituciones de gestión estatal y no estatal (50.52% y 49.48%, respectivamente), un contexto predominantemente urbano (98.27%) y un nivel socioeconómico medio (68.86%). En este grupo, el porcentaje de estudiantes con rendimiento aceptable asciende al 49.13%, compuesto por un 33.56% en En Proceso y un 15.57% en Satisfactorio.

Finalmente, el cluster 4, integrado por estudiantes con nivel socioeconómico alto (100%) y una composición similar entre instituciones de gestión estatal (54.55%) y no estatal (45.45%), presenta el mejor desempeño académico de la UGEL, con un 56.36% de rendimiento aceptable, distribuido entre un 29.09% en En Proceso y un 27.27% en Satisfactorio.

En conjunto, los resultados de la UGEL Urubamba evidencian que la mejora del nivel socioeconómico, acompañada de una mayor diversidad en la oferta educativa, se asocia con incrementos significativos en los niveles de logro académico, reflejándose en una mayor proporción de estudiantes que alcanzan desempeños satisfactorios.

Tabla 10

Porcentaje de Características Sociodemográficas y de Rendimiento en Lectura por Cluster en la UGEL Espinar

		GESTIÓN		ÁREA		SEXO		LENGUA MATERNA		ISE				LECTURA			
Cluster	N	Estatal	No estatal	Urbana	Rural	Mujer	Hombre	Castellano	Quechua	Muy bajo	Bajo	Medio	Alto	Previo inicio	En inicio	En proceso	Satisfactorio
1	154	100	0	0	100	63.64	36.36	9.74	90.26	92.86	5.84	1.3	0	33.12	50	14.29	2.6
2	359	100	0	98.33	1.67	100	0	63.79	36.21	54.6	45.4	0	0	15.32	44.57	27.3	12.81
3	339	100	0	94.99	5.01	0	100	64.01	35.99	47.79	52.2	0	0	11.21	49.85	26.55	12.39
4	147	100	0	94.56	5.44	39.46	60.54	87.76	12.24	0	0	100	0	6.8	32.65	36.73	23.81
5	29	100	0	100	0	41.38	58.62	96.55	3.45	0	0	0	100	6.9	24.14	34.48	34.48
6	29	0	100	100	0	31.03	68.97	96.55	3.45	10.34	27.6	41.38	21	10.34	17.24	41.38	31.03

El análisis de correspondencias múltiples y clusters jerárquicos aplicado a la UGEL Espinar (Tabla 10) evidencia un marcado gradiente socioeducativo en el rendimiento académico de los estudiantes. Los clusters con nivel socioeconómico muy bajo (clusters 1, 2 y 3) registran las menores proporciones de rendimiento aceptable, con valores que oscilan entre 16.89% y 40.11%. En contraste, se observa una mejora considerable en los clusters con condiciones socioeconómicas más favorables. En particular, el cluster 4, correspondiente al nivel socioeconómico medio, alcanza un rendimiento aceptable de 60.54%, mientras que los clusters 5 y 6, asociados a niveles socioeconómicos alto y mixto, respectivamente, superan el 68%.

La distribución geográfica también muestra una asociación importante con el desempeño académico. El cluster 1, caracterizado por un predominio de población rural, presenta los resultados más desfavorables, mientras que los clusters con mayor presencia

de estudiantes de zonas urbanas exhiben mejoras progresivas en el rendimiento a medida que aumentan las condiciones socioeconómicas favorables. De igual forma, la variable de gestión educativa evidencia un patrón relevante, ya que los clusters con participación de instituciones de gestión no estatal (clusters 5 y 6) registran los niveles más altos de desempeño académico. En conjunto, los resultados de la UGEL Espinar muestran que la transición desde un nivel socioeconómico muy bajo hacia condiciones más favorables constituye un punto de inflexión para la mejora de los resultados de aprendizaje, especialmente cuando esta se combina con contextos urbanos y una oferta educativa más diversificada.

Tabla 11

Porcentaje de Características Sociodemográficas y de Rendimiento en Lectura por Cluster en la UGEL Pichari Kimbiri

		GESTIÓN		ÁREA		SEXO		LENGUA MATERNA		ISE				LECTURA			
Cluster	N	Estatal	No estatal	Urbana	Rural	Mujer	Hombre	Castellano	Quechua	Muy bajo	Bajo	Medio	Alto	Previo inicio	En inicio	En proceso	Satisfactorio
1	507	100	0	63.71	36.29	49.9	50.1	63.71	36.29	100	0	0	0	31.56	54.24	11.64	2.56
2	253	88.54	11.46	77.08	22.92	45.85	54.15	73.52	26.48	0	100	0	0	17.79	51.38	23.72	7.11
3	124	62.9	37.1	88.71	11.29	37.1	62.9	84.68	15.32	14.52	0	85.48	0	23.39	50	16.13	10.48
4	15	53.33	46.67	93.33	6.67	33.33	66.67	73.33	26.67	0	0	0	100	6.67	66.67	20	6.67

La aplicación del Análisis de Correspondencias Múltiples y del agrupamiento jerárquico en la UGEL Pichari Kimbiri (Tabla 11) permitió identificar diferentes perfiles de rendimiento académico asociados al nivel socioeconómico y al tipo de gestión educativa. Los resultados evidencian que los clusters correspondientes a los niveles socioeconómicos Muy bajo (cluster 1) y Bajo (cluster

2) presentan las menores proporciones de rendimiento aceptable, con 14.20% (11.64% en En Proceso y 2.56% en Satisfactorio) y 30.83% (23.72% en En Proceso y 7.11% en Satisfactorio), respectivamente.

Posteriormente, se observa una mejora en el cluster 3, correspondiente al nivel socioeconómico Medio, que alcanza un rendimiento aceptable de 26.61%, distribuido en 16.13% de estudiantes en En Proceso y 10.48% en Satisfactorio. Por su parte, el cluster 4, integrado por estudiantes con nivel socioeconómico Alto, registra el mejor desempeño de la UGEL, con un rendimiento aceptable de 26.67% (20.00% en En Proceso y 6.67% en Satisfactorio).

En cuanto al tipo de gestión educativa, se identifica un patrón diferenciado entre los perfiles analizados. Los clusters con predominio de instituciones de gestión estatal (clusters 1 y 2) concentran las menores proporciones de estudiantes con rendimiento aceptable, mientras que aquellos con una mayor participación de instituciones de gestión no estatal (clusters 3 y 4) presentan resultados relativamente más favorables.

En conjunto, los resultados de la UGEL Pichari Kimbiri indican que la mejora del nivel socioeconómico, acompañada de una mayor diversificación de la oferta educativa, se asocia con avances en los logros de aprendizaje. No obstante, la magnitud de esta mejora es menor en comparación con la observada en las demás UGEL analizadas.

Tabla 12

Porcentaje de Características Sociodemográficas y de Rendimiento en Lectura por Cluster en la UGEL Anta

		GESTIÓN		ÁREA		SEXO		LENGUA MATERNA		ISE				LECTURA			
Cluster	N	Estatal	No estatal	Urbana	Rural	Mujer	Hombre	Castellano	Quechua	Muy bajo	Bajo	Medio	Alto	Previo inicio	En inicio	En proceso	Satisfactorio
1	226	95.58	4.42	0	100	46.02	53.98	29.2	70.8	96.46	3.54	0	0	44.25	45.58	7.52	2.65
2	397	100	0	100	0	53.9	46.1	47.1	52.9	100	0	0	0	22.17	54.91	18.39	4.53
3	197	83.76	16.24	91.88	8.12	52.79	47.21	75.13	24.87	0	100	0	0	14.72	50.25	25.38	9.64
4	107	55.14	44.86	90.65	9.35	42.06	57.94	87.85	12.15	10.28	0	89.72	0	14.02	39.25	27.1	19.63
5	36	33.33	66.67	100	0	44.44	55.56	97.22	2.78	0	0	0	100	0	11.11	47.22	41.67

El análisis de la UGEL Anta (Tabla 12), realizado mediante técnicas multivariadas, evidencia una marcada segmentación del rendimiento académico asociada a la estratificación socioeconómica y a las características geográficas de los estudiantes. Los resultados muestran diferencias importantes en los niveles de logro entre los clusters identificados.

Los clusters que concentran las condiciones de mayor vulnerabilidad (clusters 1 y 2), caracterizados por un nivel socioeconómico Muy bajo (96.46% y 100%, respectivamente) y por un predominio de estudiantes de zonas rurales (100% y 0% rural, respectivamente), registran los porcentajes más bajos de rendimiento aceptable. El cluster 1 alcanza únicamente un 10.17% (7.52% en En Proceso y 2.65% en Satisfactorio), mientras que el cluster 2 presenta un 22.92% (18.39% en En Proceso y 4.53% en Satisfactorio). Esta diferencia sugiere que el contexto urbano podría ejercer un efecto favorable sobre el rendimiento académico, incluso en escenarios de elevada vulnerabilidad socioeconómica.

Una mejora importante se observa en el cluster 3, correspondiente al nivel socioeconómico Bajo, donde el rendimiento aceptable alcanza el 35.02%, distribuido entre un 25.38% en En Proceso y un 9.64% en Satisfactorio. Esta tendencia continúa en los clusters con condiciones socioeconómicas más favorables. El cluster 4, asociado al nivel socioeconómico Medio, registra un rendimiento aceptable de 46.73% (27.10% en En Proceso y 19.63% en Satisfactorio), mientras que el cluster 5, correspondiente al nivel socioeconómico Alto, presenta el mejor desempeño de la UGEL, con un 88.89% de rendimiento aceptable (47.22% en En Proceso y 41.67% en Satisfactorio).

En relación con la gestión educativa, se observa un patrón diferenciado entre los perfiles identificados. El cluster 5, que exhibe el mayor nivel de rendimiento académico, está conformado en un 66.67% por instituciones de gestión no estatal, mientras que los clusters con menor desempeño corresponden predominantemente a instituciones de gestión estatal (95.58% a 100%). Este comportamiento sugiere la existencia de una interacción entre el tipo de gestión educativa y el nivel socioeconómico en la configuración de los resultados de aprendizaje.

En conjunto, los hallazgos de la UGEL Anta ponen de manifiesto la necesidad de implementar políticas educativas focalizadas en los clusters que presentan mayores condiciones de vulnerabilidad, con el propósito de contribuir a la reducción de las brechas educativas y favorecer una mayor equidad en los logros de aprendizaje.

Tabla 13

Porcentaje de Características Sociodemográficas y de Rendimiento en Lectura por Cluster en la UGEL Paucartambo

		GESTIÓN		ÁREA		SEXO		LENGUA MATERNA		ISE				LECTURA			
Cluster	N	Estatal	No estatal	Urbana	Rural	Mujer	Hombre	Castellano	Quechua	Muy bajo	Bajo	Medio	Alto	Previo inicio	En inicio	En proceso	Satisfactorio
1	166	100	0	0	100	100	0	0	100	100	0	0	0	49.4	47.59	3.01	0
2	334	100	0	56.89	43.11	0	100	1.8	98.2	100	0	0	0	40.42	48.8	9.88	0.9
3	290	100	0	97.59	2.41	83.79	16.21	38.97	61.03	100	0	0	0	35.17	48.62	14.14	2.07
4	110	100	0	81.82	18.18	46.36	53.64	35.45	64.55	0	100	0	0	32.73	47.27	16.36	3.64
5	22	100	0	81.82	18.18	45.45	54.55	45.45	54.55	0	0	100	0	31.82	45.45	18.18	4.55
6	4	100	0	100	0	0	100	25	75	0	0	0	100	0	50	25	25

La UGEL Paucartambo (Tabla 13) presenta un perfil de rendimiento académico predominantemente bajo, con la mayoría de sus clusters (1, 2, 3 y 4) mostrando un 100% de estudiantes en el nivel socioeconómico muy bajo y un alto porcentaje de lengua materna Quechua (entre 61.03% y 98.2%). En estos clusters, los niveles de logro En proceso y Satisfactorio son muy reducidos (entre 3.01% y 20.45% combinados), lo que indica que la gran mayoría de los estudiantes se encuentra en niveles previos al inicio o en inicio de aprendizaje. Solo los clusters 5 y 6, aunque con muy pocos estudiantes (22 y 4 respectivamente), muestran un mejor desempeño, con niveles socioeconómicos medio y alto, y porcentajes más altos en “En proceso” y “Satisfactorio” (22.73% y 50% respectivamente). Esto sugiere que el bajo rendimiento en Paucartambo está fuertemente asociado a

condiciones socioeconómicas desfavorables y a la predominancia del Quechua como lengua materna, reflejando la necesidad de intervenciones focalizadas en estos factores para mejorar los resultados académicos.

Tabla 14

Porcentaje de Características Sociodemográficas y de Rendimiento en Lectura por Cluster en la UGEL Paruro

		GESTIÓN		ÁREA		SEXO		LENGUA MATERNA		ISE				LECTURA			
Cluster	N	Estatal	No estatal	Urbana	Rural	Mujer	Hombre	Castellano	Quechua	Muy bajo	Bajo	Medio	Alto	Previo inicio	En inicio	En proceso	Satisfactorio
1	152	100	0	0	100	49.34	50.66	5.92	94.08	100	0	0	0	38.82	48.03	12.5	0.66
2	284	100	0	100	0	47.18	52.82	16.9	83.1	100	0	0	0	42.25	48.24	7.04	2.46
3	51	100	0	82.35	17.65	50.98	49.02	37.25	62.75	0	100	0	0	21.57	41.18	17.65	19.61
4	7	100	0	85.71	14.29	14.29	85.71	57.14	42.86	0	0	100	0	0	57.14	14.29	28.57
5	2	100	0	50	50	50	50	100	0	0	0	0	100	0	0	50	50

De acuerdo con los datos analizados mediante Análisis de Correspondencias Múltiples (MCA) y Clustering Jerárquico (HCPC), la UGEL Paruro (Tabla 14) evidencia un perfil de rendimiento académico crítico, predominantemente asociado a condiciones de alta vulnerabilidad socioeducativa. Los clusters 1 y 2, que concentran la mayor población estudiantil (N = 152 y N = 284, respectivamente), muestran que el 100% de los estudiantes se ubican en el nivel socioeconómico muy bajo, con una predominancia del Quechua como lengua materna (94.08% y 83.10%) y una gestión estatal del 100%. En estos grupos, los niveles de logro (En proceso y Satisfactorio) son extremadamente bajos (13.16% y 9.50%, respectivamente), indicando que la mayoría del estudiantado se encuentra en fases previas al aprendizaje esperado. Aunque los clusters 3, 4 y 5 presentan mejoras

en el nivel socioeconómico y logro académico, su representatividad es mínima (N = 60), lo que no modifica la tendencia general. Estos hallazgos sugieren una relación significativa entre el bajo rendimiento en la evaluación censal y variables estructurales como la pobreza y la barrera lingüística, destacando la urgencia de políticas educativas interculturales y de equidad en este contexto.

Tabla 15

Porcentaje de Características Sociodemográficas y de Rendimiento en Lectura por Cluster en la UGEL Acomayo

		GESTIÓN		ÁREA		SEXO		LENGUA MATERNA		ISE				LECTURA			
Cluster	N	Estatal	No estatal	Urbana	Rural	Mujer	Hombre	Castellano	Quechua	Muy bajo	Bajo	Medio	Alto	Previo inicio	En inicio	En proceso	Satisfactorio
1	335	100	0	73.73	26.27	50.75	49.25	12.54	87.46	100	0	0	0	35.22	48.36	13.13	3.28
2	79	100	0	81.01	18.99	39.24	60.76	29.11	70.89	0	100	0	0	18.99	49.37	21.52	10.13
3	65	0	100	100	0	47.69	52.31	21.54	78.46	75.38	24.6	0	0	15.38	44.62	23.08	16.92
4	2	100	0	100	0	0	100	50	50	0	0	0	100	0	50	50	0
5	17	76.47	23.53	70.59	29.41	52.94	47.06	58.82	41.18	0	0	100	0	11.76	35.29	29.41	23.53

El análisis mediante MCA y HCPC revela que el perfil académico de la UGEL Acomayo (Tabla 15) se sitúa mayoritariamente en niveles iniciales de rendimiento, mostrando una correlación evidente con factores de tipo sociolingüístico y de gestión institucional.

Los clusters 1 (N = 335) y 3 (N = 65), que representan la mayor parte de la población estudiantil, muestran una composición socioeconómica en los niveles muy bajo (100% y 75.38%, respectivamente) y una abrumadora predominancia del Quechua

como lengua materna (87.46% y 78.46%). En estos grupos, los niveles de logro aceptable combinados (En proceso y Satisfactorio) son notablemente bajos (16.41% y 40.00%, respectivamente), indicando que la mayoría de los estudiantes se encuentran en las fases Previo inicio o En inicio. Solo el cluster 5 (N = 17), aunque de tamaño reducido, presenta un nivel socioeconómico medio y un rendimiento aceptable más alto (52.94%), mientras que el cluster 4 (N = 2) es atípico y no representativo. Estos hallazgos sugieren que en la UGEL Acomayo el bajo rendimiento en la evaluación censal está fuertemente relacionado con condiciones de vulnerabilidad socioeconómica y la prevalencia del Quechua, señalando la necesidad de implementar estrategias pedagógicas diferenciadas e interculturales para mejorar los resultados de aprendizaje.

Tabla 16

Porcentaje de Características Sociodemográficas y de Rendimiento en Lectura por Cluster en la UGEL Canas

		GESTIÓN		ÁREA		SEXO		LENGUA MATERNA		ISE				LECTURA			
Cluster	N	Estatal	No estatal	Urbana	Rural	Mujer	Hombre	Castellano	Quechua	Muy bajo	Bajo	Medio	Alto	Previo inicio	En inicio	En proceso	Satisfactorio
1	92	100	0	0	100	100	0	0	100	96.74	3.26	0	0	60.87	34.78	3.26	1.09
2	63	100	0	0	100	0	100	0	100	93.65	6.35	0	0	49.21	44.44	6.35	0
3	84	100	0	100	0	100	0	0	100	100	0	0	0	32.14	55.95	10.71	1.19
4	73	100	0	98.63	1.37	0	100	1.37	98.63	100	0	0	0	35.62	52.05	8.22	4.11
5	5	100	0	60	40	60	40	40	60	0	0	100	0	40	20	20	20
6	38	100	0	100	0	100	0	78.95	21.05	52.63	47.4	0	0	15.79	50	26.32	7.89
7	42	100	0	100	0	0	100	76.19	23.81	47.62	52.4	0	0	21.43	47.62	23.81	7.14
8	2	50	50	100	0	50	50	50	50	0	0	0	100	50	50	0	0
9	10	0	100	100	0	30	70	80	20	30	70	0	0	40	20	30	10

El análisis de la UGEL Canas (Tabla 16), realizado mediante el Análisis de Correspondencias Múltiples (MCA) y el Clustering Jerárquico sobre Componentes Principales (HCPC), evidencia una marcada heterogeneidad en los perfiles de rendimiento académico. La mayor parte de los clusters (1, 2, 3, 4, 6 y 7) presenta un perfil caracterizado por condiciones de alta vulnerabilidad, integrado exclusivamente por estudiantes de instituciones educativas de gestión estatal, pertenecientes principalmente a zonas rurales, con predominio de la lengua materna Quechua (entre 70.89% y 100%) y con un nivel socioeconómico muy bajo o bajo. Estas características se asocian con bajos niveles de rendimiento académico, reflejados en porcentajes combinados de los niveles En Proceso y Satisfactorio que oscilan entre 4.35% y 34.21%.

En contraste, se identifican perfiles diferenciados en los clusters 5, 8 y 9, los cuales, aunque cuentan con un número reducido de estudiantes (N entre 2 y 10), presentan condiciones socioeconómicas más favorables, correspondientes a los niveles medio o alto, así como una participación importante de instituciones de gestión no estatal, que alcanza hasta el 100%. Estos clusters registran porcentajes de rendimiento aceptable considerablemente superiores, con valores que varían entre 40% y 100%.

En conjunto, los resultados de la UGEL Canas indican que el nivel socioeconómico y el tipo de gestión institucional constituyen factores estrechamente asociados con el rendimiento académico. La coexistencia de perfiles con condiciones estructurales marcadamente distintas pone de manifiesto la heterogeneidad educativa presente en esta UGEL y evidencia la importancia de considerar estas diferencias para comprender los patrones de desempeño académico.

5.4. Discusión de Resultados

5.4.1. Hallazgo principal y contraste con estudios previos

El principal hallazgo de esta investigación no se limita a evidenciar diferencias en el rendimiento académico, sino que consiste en la identificación de perfiles estructurales claramente diferenciados entre los estudiantes que participaron en la Evaluación Censal de Estudiantes (ECE) del departamento de Cusco. La aplicación del Análisis de Correspondencias Múltiples (MCA) permitió demostrar que el bajo rendimiento académico no puede atribuirse a un único factor, sino que emerge de la interacción simultánea entre variables de carácter socioeconómico, lingüístico, territorial e institucional. En consecuencia, los resultados evidencian que el rendimiento académico se configura en estructuras latentes bien definidas, lo que sugiere que las desigualdades educativas responden, principalmente, a factores de naturaleza estructural más que a características individuales.

Estos resultados son consistentes con los reportados por Cueto, León y Muñoz (2016), quienes demostraron que el rendimiento en comprensión lectora en el Perú está estrechamente asociado con el nivel socioeconómico, la lengua materna y el contexto territorial. No obstante, la presente investigación incorpora un aporte metodológico adicional. Mientras el estudio de Cueto et al. se fundamenta en modelos estadísticos tradicionales, este trabajo identifica perfiles diferenciados mediante técnicas multivariadas no supervisadas, lo que permite comprender la heterogeneidad del rendimiento

académico desde una perspectiva estructural, en lugar de limitarse a establecer relaciones de carácter explicativo.

De igual manera, los hallazgos guardan concordancia con lo señalado por Timarán Pereira, Caicedo Zambrano y Hidalgo Troya (2019), quienes identificaron patrones de rendimiento académico mediante la aplicación de técnicas de minería de datos, evidenciando la influencia de variables socioeconómicas y contextuales en la conformación de perfiles académicos. En este contexto, los resultados del presente estudio corroboran que el empleo de técnicas multivariadas permite identificar patrones y estructuras subyacentes que difícilmente pueden ser detectados mediante análisis descriptivos convencionales.

5.4.2. El rol del nivel socioeconómico en la estructuración del rendimiento académico

Uno de los hallazgos más consistentes de la investigación es el papel determinante del nivel socioeconómico en la configuración de los perfiles de rendimiento académico. El Análisis de Correspondencias Múltiples (MCA) evidencia que el Índice Socioeconómico (ISE) no solo mantiene una estrecha asociación con el rendimiento académico, sino que constituye uno de los ejes principales que estructuran el espacio factorial. Este resultado indica que las diferencias observadas en el desempeño académico no pueden atribuirse exclusivamente a las capacidades individuales de los estudiantes, sino que reflejan desigualdades estructurales en las condiciones de aprendizaje.

Este hallazgo es consistente con la evidencia reportada en la literatura internacional, que reconoce al nivel socioeconómico como uno de los predictores más sólidos del rendimiento académico. Berkowitz et al. (2017) sostienen que los estudiantes provenientes de contextos socioeconómicos desfavorecidos alcanzan sistemáticamente menores niveles de logro debido a las limitaciones en el acceso a recursos educativos, al apoyo familiar y a las oportunidades de aprendizaje. De manera similar, en el contexto peruano, Cueto et al. (2016) evidencian que las desigualdades socioeconómicas se traducen en brechas educativas persistentes.

No obstante, la presente investigación aporta un elemento adicional de relevancia al demostrar que el nivel socioeconómico no opera de manera independiente, sino en estrecha interacción con variables territoriales y lingüísticas. En consecuencia, los resultados sugieren que el bajo rendimiento académico debe comprenderse como la expresión de múltiples desventajas acumuladas, lo que refuerza la necesidad de diseñar políticas educativas con un enfoque multidimensional e integral.

5.4.3. La brecha lingüística y el contexto de la Educación Intercultural Bilingüe

La lengua materna se configura como una de las variables con mayor capacidad para diferenciar los perfiles de rendimiento académico identificados en la investigación. Los resultados del Análisis de Correspondencias Múltiples (MCA) evidencian que los estudiantes cuya lengua materna es el quechua se concentran principalmente en

los clusters con menor rendimiento académico, mientras que aquellos cuya lengua materna es el castellano presentan una mayor representación en los clusters de mejor desempeño.

Este hallazgo trasciende una explicación basada exclusivamente en las diferencias lingüísticas y pone de manifiesto la existencia de desigualdades estructurales dentro del sistema educativo. La literatura sobre educación intercultural bilingüe ha documentado de manera consistente que los estudiantes cuya lengua materna difiere del castellano enfrentan barreras de carácter pedagógico, cultural y lingüístico que inciden directamente en su desempeño académico. En esta línea, UNESCO (2020) señala que los sistemas educativos que no incorporan de manera adecuada la diversidad lingüística tienden a reproducir y profundizar las desigualdades educativas.

En el contexto del departamento de Cusco, donde una proporción importante de estudiantes tiene como lengua materna el quechua, el presente estudio aporta evidencia empírica que respalda esta problemática. Los resultados indican que la condición lingüística no solo se asocia con el rendimiento académico, sino que también interactúa con el nivel socioeconómico y el contexto territorial en la configuración de los perfiles estudiantiles, dando lugar a escenarios de mayor vulnerabilidad educativa.

5.4.4. Diferencia urbano-rurales en el rendimiento académico

El análisis realizado evidencia una marcada diferenciación entre los estudiantes de zonas urbanas y rurales en cuanto a sus perfiles de rendimiento académico. Los resultados muestran que los

estudiantes pertenecientes a instituciones educativas ubicadas en áreas rurales se concentran principalmente en los clusters con menor desempeño, mientras que aquellos que asisten a instituciones de zonas urbanas presentan una mayor representación en los clusters con mejores niveles de rendimiento.

Este hallazgo es consistente con la evidencia reportada en la literatura, que reconoce al contexto territorial como un factor determinante del rendimiento académico. En general, las zonas rurales enfrentan mayores limitaciones en el acceso a recursos educativos, presentan deficiencias en la infraestructura escolar, disponen de una menor oferta de docentes especializados y concentran niveles más elevados de pobreza. En consecuencia, el rendimiento académico debe entenderse no solo como el resultado de las capacidades individuales de los estudiantes, sino también como una expresión de las condiciones estructurales asociadas al territorio.

La presente investigación aporta evidencia adicional al demostrar que las diferencias entre los contextos urbano y rural no se manifiestan únicamente en los niveles promedio de rendimiento, sino también en la conformación de perfiles académicos claramente diferenciados. Este resultado refuerza la importancia de considerar la dimensión territorial como un componente fundamental para comprender las desigualdades educativas y orientar el diseño de políticas e intervenciones más focalizadas.

5.4.5. Heterogeneidad entre UGELs

Uno de los aportes más relevantes de la presente investigación es la identificación de una marcada heterogeneidad entre las UGEL del departamento de Cusco. Aunque los perfiles de bajo rendimiento académico comparten características similares en términos socioeconómicos y lingüísticos, la magnitud y configuración de estos patrones difieren de manera significativa entre las distintas UGEL analizadas.

Esta variabilidad evidencia que el sistema educativo no presenta un comportamiento homogéneo en el territorio, sino que responde a dinámicas locales específicas que inciden en el rendimiento académico de los estudiantes. Entre los factores que podrían explicar estas diferencias se encuentran la dispersión geográfica, el acceso a los servicios educativos, la infraestructura escolar y las condiciones socioeconómicas propias de cada población.

Asimismo, la aplicación del Análisis de Correspondencias Múltiples (MCA) permitió evidenciar que cada UGEL posee una estructura particular en el espacio factorial, lo que refleja configuraciones diferenciadas de las variables asociadas al rendimiento académico. Este hallazgo sugiere que las políticas educativas no deberían diseñarse bajo un enfoque uniforme, sino adaptarse a las características y necesidades específicas de cada contexto territorial, con el fin de responder de manera más efectiva a las desigualdades identificadas.

5.4.6. Limitaciones del estudio

A pesar de las contribuciones del estudio, resulta necesario reconocer algunas limitaciones metodológicas que deben considerarse al interpretar los resultados. En primer lugar, el análisis se fundamenta en datos secundarios procedentes de la Evaluación Censal de Estudiantes, por lo que las variables analizadas se encuentran restringidas a la información disponible en dicha base de datos. En este sentido, no fue posible incorporar variables asociadas al contexto familiar, al nivel educativo de los padres o a las condiciones pedagógicas del aula, factores que podrían tener una influencia relevante sobre el rendimiento académico.

En segundo lugar, el enfoque analítico empleado posee un carácter exploratorio y no causal. Si bien el Análisis de Correspondencias Múltiples (MCA) y el clustering jerárquico permiten identificar patrones de asociación y estructuras subyacentes en los datos, estas técnicas no posibilitan establecer relaciones de causalidad entre las variables estudiadas. Por ello, los hallazgos deben interpretarse como configuraciones o asociaciones estructurales presentes en la información analizada, y no como evidencia de efectos causales directos.

Asimismo, la interpretación de los clusters identificados puede estar sujeta a cierto grado de subjetividad, dado que la caracterización de los perfiles académicos requiere un proceso de interpretación por parte del investigador. Sin embargo, esta limitación fue mitigada mediante la aplicación de criterios estadísticos objetivos para la

selección de las dimensiones relevantes y la determinación del número óptimo de clusters.

5.4.7. Fortalezas metodológicas del estudio

Una de las fortalezas principales del estudio radica en la aplicación de técnicas multivariadas avanzadas para el análisis del rendimiento académico. A diferencia de investigaciones previas centradas exclusivamente en análisis descriptivos o modelos de regresión, este estudio emplea el Análisis de Correspondencias Múltiples y el análisis de conglomerados jerárquicos con el propósito de identificar estructuras latentes presentes en los datos.

Estas metodologías permiten examinar de manera simultánea múltiples variables categóricas y reconocer patrones complejos que difícilmente podrían ser detectados mediante enfoques estadísticos convencionales. Asimismo, la adopción de una perspectiva multivariada contribuye a comprender el rendimiento académico como un fenómeno de naturaleza estructural, determinado por la interacción de diversos factores relacionados entre sí.

Otra fortaleza relevante del estudio corresponde al uso de datos censales, lo que proporciona una elevada representatividad de los resultados y posibilita una caracterización más precisa del rendimiento académico en el contexto del departamento de Cusco.

5.4.8. Implicaciones para la política educativa

Los resultados obtenidos en el estudio presentan importantes implicancias para el diseño y la implementación de políticas educativas. En primer lugar, la identificación de perfiles asociados al

bajo rendimiento académico permite orientar las intervenciones hacia los grupos con mayores niveles de vulnerabilidad, superando la aplicación de estrategias generales que podrían presentar una menor efectividad al no considerar las particularidades de los estudiantes.

En segundo lugar, los hallazgos evidencian que las políticas educativas deben abordar de manera simultánea los factores socioeconómicos, lingüísticos y territoriales. El bajo rendimiento académico no puede ser enfrentado únicamente mediante acciones orientadas a mejorar los procesos de enseñanza dentro del aula, sino que requiere la implementación de estrategias integrales destinadas a disminuir las desigualdades estructurales que condicionan las oportunidades educativas de los estudiantes.

Finalmente, el estudio pone de manifiesto la importancia de fortalecer la educación intercultural bilingüe, particularmente en regiones como Cusco, donde una proporción significativa de estudiantes tiene al quechua como lengua materna. La reducción de las brechas lingüísticas constituye un componente fundamental para favorecer la mejora del rendimiento académico y contribuir a la disminución de las desigualdades existentes en el ámbito educativo.

CONCLUSIONES

Se determinó que el rendimiento en lectura de los estudiantes de segundo de secundaria en Cusco no responde a un comportamiento aleatorio; por el contrario, los perfiles de desempeño académico se encuentran estrechamente asociados con factores como el nivel socioeconómico, la lengua materna, el área geográfica de residencia y el tipo de gestión de la institución educativa. Mientras que quienes pertenecen a un nivel socioeconómico medio o alto logran un perfil de rendimiento aceptable, quienes enfrentan desventajas en estas variables se quedan en un perfil de rendimiento bajo.

Se concluye que el rendimiento en lectura de los estudiantes depende de la interacción conjunta de su nivel socioeconómico, su lengua materna, el área donde viven y el tipo de gestión de su colegio. Estas variables no actúan solas: el análisis demuestra que estudiar en la ciudad (área urbana) no es suficiente para garantizar el éxito escolar si el alumno pertenece a un nivel socioeconómico muy bajo y habla quechua.

Se identificó que el perfil de rendimiento bajo en lectura (que afecta a casi una cuarta parte de la población urbana y a la mayoría de la rural) está directamente asociado a la acumulación de desventajas, específicamente a estudiantes con un nivel socioeconómico: muy bajo, de lengua materna: quechua y que asisten a instituciones de gestión: estatal. Esto evidencia que la brecha más crítica en el Cusco es de origen estructural y cultural, no pedagógico.

RECOMENDACIONES

Se recomienda a la Dirección Regional de Educación (DRE) y a las UGEL del Cusco abandonar las políticas educativas uniformes y diseñar intervenciones diferenciadas. Los recursos de apoyo pedagógico, materiales y acompañamiento deben priorizarse en las escuelas de gestión: estatal que atienden al perfil más crítico: estudiantes con nivel socioeconómico: muy bajo y de lengua materna: quechua.

Se recomienda fortalecer los programas de Educación Intercultural Bilingüe (EIB), expandiendo su cobertura no solo en las zonas rurales, sino también en las periferias urbanas de la ciudad del Cusco, donde se concentra un porcentaje significativo de estudiantes quechua-hablantes en situación de pobreza que actualmente no reciben una enseñanza adaptada a su lengua materna.

Se recomienda capacitar a los especialistas pedagógicos y directivos de las UGEL en el uso de herramientas de análisis de datos de evaluaciones nacionales. Esto permitirá identificar tempranamente los perfiles de riesgo en cada colegio y planificar estrategias preventivas antes de que los estudiantes finalicen su trayectoria escolar.

BIBLIOGRAFÍA

Acosta, J. C., La Red Martínez, D., & Primorac, C. (2018). *Determinación de perfiles de rendimiento académico en la UNNE con Minería de Datos Educacional.* Corrientes: RedUNCI - UNNE.

Berkowitz, R., Moore, H., Astor, R. A., & Benbenishty, R. (2017). A research synthesis of the associations between socioeconomic background, inequality, school climate, and academic achievement. *Review of Educational Research, 87*(2), 425-469.

Bourdieu, P. (1986). The forms of capital. En J. Richardson (Ed.). *Handbook of theory and research for the sociology of education*, 241-258.

Cueto, S., Leon, J., & Muñoz, I. (2016). *Brechas de rendimiento académico en el Perú: Un análisis a partir de los resultados de la Evaluación Censal de Estudiantes.* Lima: Grupo de Análisis para el Desarrollo (GRADE).

Everitt, B., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis (5th ed.).* Wiley.

Fernández Cañete, G. Y., & Martínez Espinal, G. (2014). *Determinación de factores que mejoran el rendimiento académico aplicando minería de datos en el colegio Convenio Andrés Bello - Huancayo.* Huancavelica: Universidad Nacional de Huancavelica.

Greenacre, M. (2017). *Correspondence analysis in practice* (Tercera ed.). Chapman & Hall/CRC.

Guichard, E., Chimienti, M., Bolzman, C., & Le Goff, J.-M. (2024). When national origins equal socio-economic background: The effect of the ethno-class parental background on the education of children coming of

age in Switzerland. *Journal of International Migration and Integration*. doi:https://doi.org/10.1007/s12134-024-01129-w

Holgado Apaza, L. (2018). *Detección de patrones de bajo rendimiento académico mediante técnicas de minería de datos de los estudiantes de la Universidad Nacional Amazónica de Madre de Dios 2018.* Puno: Universidad Nacional del Altiplano.

Husson, F., Le, S., & Pagès, J. (2017). *Exploratory multivariate analysis by example using R* (Segunda ed.). Chapman & Hall/CRC.

La Red Martínez, D. L., Karanik, M., Giovannini, M., Báez, M. E., & Torre, J. (2016). *Descubrimiento de perfiles de rendimiento estudiantil: un modelo de integración de datos académicos y socioeconómicos.* Campus Virtuales.

Le, S., Josse, J., & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software, 25*(1), págs. 1-18.

Lewis, N. (2017). *Machine Learning Made Easy with R: An Intuitive Step by Step Blueprint for Beginners.* CreateSpace Independent Publishing Platform.

Little, R., & Rubin, D. (2019). *Statistical analysis with missing data* (Tercera ed.). John Wiley & Sons.

Ministerio de Educación. (2019). *Resultados de las evaluaciones nacionales de logros de aprendizaje 2019.* Lima: Oficina de Medición de la Calidad de los Aprendizajes.

Ministerio de Educación del Perú. (2019). *Evaluación Censal de Estudiantes 2019: Marco de evaluación y fundamentos técnicos.* Ministerio de

Educación del Perú – Oficina de Medición de la Calidad de los Aprendizajes (UMC), Lima, Perú.

Montero Rojas, E., Villalobos Palma, J., & Valverde Bermúdez, A. (2007). FACTORES INSTITUCIONALES, PEDAGÓGICOS, PSICOSOCIALES Y SOCIODEMOGRÁFICOS ASOCIADOS AL RENDIMIENTO ACADÉMICO EN LA UNIVERSIDAD DE COSTA RICA: UN ANÁLISIS MULTINIVEL. *Revista Electrónica de Investigación y Evaluación Educativa*, 215-234.

OECD. (2018). *Equity in Education: Breaking Down Barriers to Social Mobility. PISA.* Paris: OECD Publishing.

OECD. (2019). *PISA 2018 Results (Volume I): What Students Know and Can Do.* OECD Publishing.

Pérez Lopez, C., & Santin Gonzáles, D. (2007). *MINERÏA DE DATOS Técnicas Y Herramientas.* Madrid: PARANINFO.

R Core Team. (2023). R: A Language and Environment for Statistical Computing. Vienna. Obtenido de https://www.R-project.org/

Sen, A. (1999). *Development as freedom.* Oxford University Press.

Supo, J. (2015). *CÓMO EMPEZAR UNA TESIS.* Arequipa: BIOESTADISTICO EIRL.

Timarán Pereira, R., Caicedo Zambrano, J., & Hidalgo Troya, A. (2019). Identificación de factores asociados al desempeño académico de matemáticas en las pruebas Saber 11 aplicando minería de datos. *17th LACCEI International Multi-Conference for Engineering, Education, and Technology*, (págs. 24-26). Jamaica.

UNESCO. (2020). Global Education Monitoring Report 2020: Inclusion and education: All means all.

ANEXOS

CÓDIGO EN R

```r
# CARGAR PAQUETES
library(tidyverse)
library(FactoMineR)
library(factoextra)
library(cowplot)
```

```r
# IMPORTAR DATOS
load("datos/ece_2s_2019.RData")
```

```r
# PROCESAR DATOS
datos <- ece_2s_2019 |>
  filter(nom_UGEL == "Cusco") |>
  select(
    area, Nombre_IE, gestion2, sexo, lengua_materna, n_ise, grupo_L
  ) |>
  mutate(
    n_ise = factor(
      n_ise,
      levels = c("Muy bajo", "Bajo", "Medio", "Alto")
    ),
    grupo_L = factor(
      grupo_L,
      levels = c("Previo al inicio", "En inicio", "En proceso", "Satisfactorio")
    ),
    lengua_materna =
      case_when(
        lengua_materna %in% c("Castellano", "Quechua")
        ~ as.character(lengua_materna)
      ),
    across(where(is.character), factor)
  ) |>
  drop_na()
```

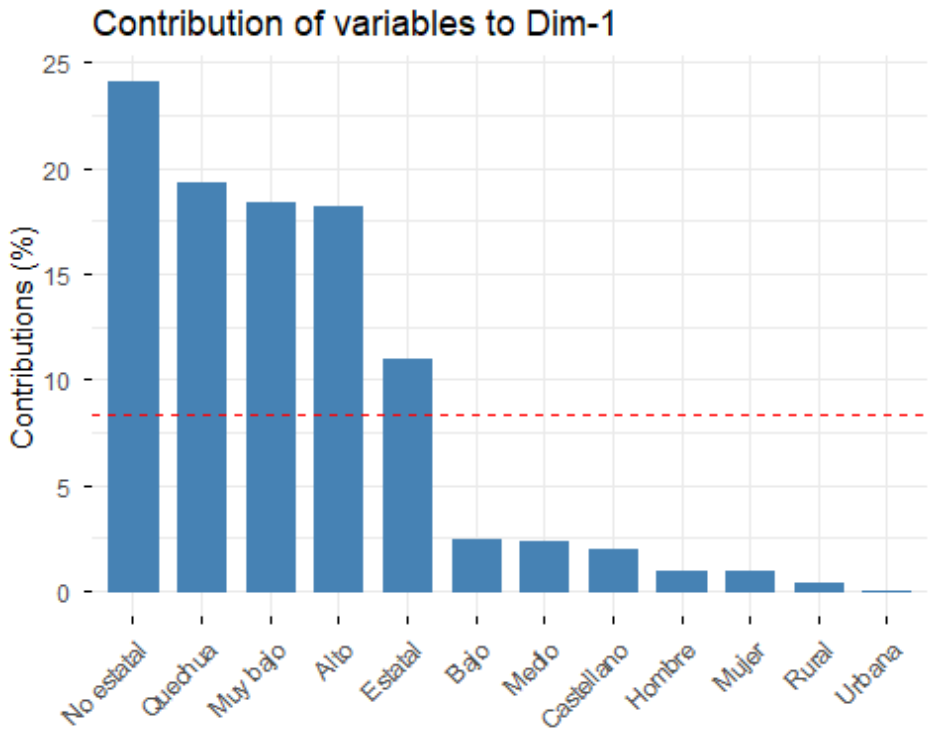
```r
# DATOS
## # A tibble: 7,175 × 7
##    area   Nombre_IE   gestion2 sexo  lengua_materna n_ise    grupo_L
##    <fct>  <fct>       <fct>    <fct> <fct>          <fct>    <fct>
##  1 Urbana COMERCIO 41 Estatal  Mujer Castellano     Medio    Satisfactorio
##  2 Urbana COMERCIO 41 Estatal  Mujer Castellano     Bajo     Satisfactorio
##  3 Urbana COMERCIO 41 Estatal  Mujer Castellano     Medio    En proceso
##  4 Urbana COMERCIO 41 Estatal  Mujer Castellano     Medio    En inicio
##  5 Urbana COMERCIO 41 Estatal  Mujer Castellano     Bajo     Previo al inicio
##  6 Urbana COMERCIO 41 Estatal  Mujer Castellano     Muy bajo Satisfactorio
##  7 Urbana COMERCIO 41 Estatal  Mujer Castellano     Bajo     En inicio
##  8 Urbana COMERCIO 41 Estatal  Mujer Castellano     Bajo     En inicio
##  9 Urbana COMERCIO 41 Estatal  Mujer Castellano     Bajo     En inicio
## 10 Urbana COMERCIO 41 Estatal  Mujer Castellano     Alto     En inicio
## # ℹ 7,165 more rows
```

```r
# 2. ANÁLISIS DE CORRESPONDENCIAS MÚLTIPLES (MCA)
mca_result <- MCA(datos,
                  quali.sup = c(2, 7),
                  graph = FALSE)
```

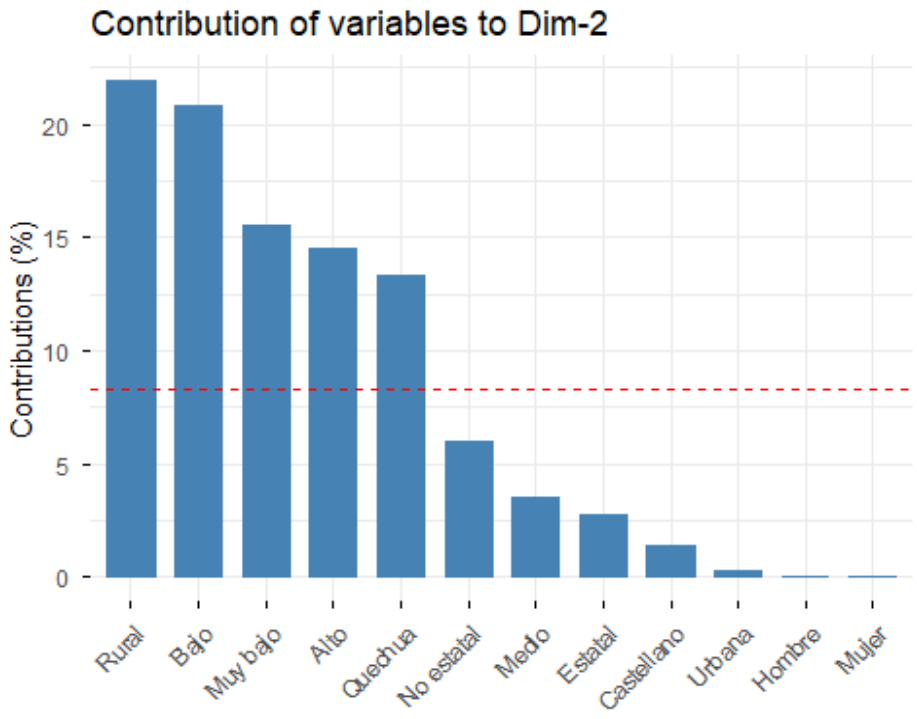
```r
## 2.1. Inercia
mca_result$eig
```

```
##        eigenvalue percentage of variance cumulative percentage of variance
## dim 1  0.3267632               23.34023                           23.34023
## dim 2  0.2285033               16.32166                           39.66189
## dim 3  0.2020765               14.43403                           54.09593
## dim 4  0.1997252               14.26608                           68.36201
## dim 5  0.1880915               13.43510                           81.79712
## dim 6  0.1491428               10.65305                           92.45017
## dim 7  0.1056976                7.54983                          100.00000
```

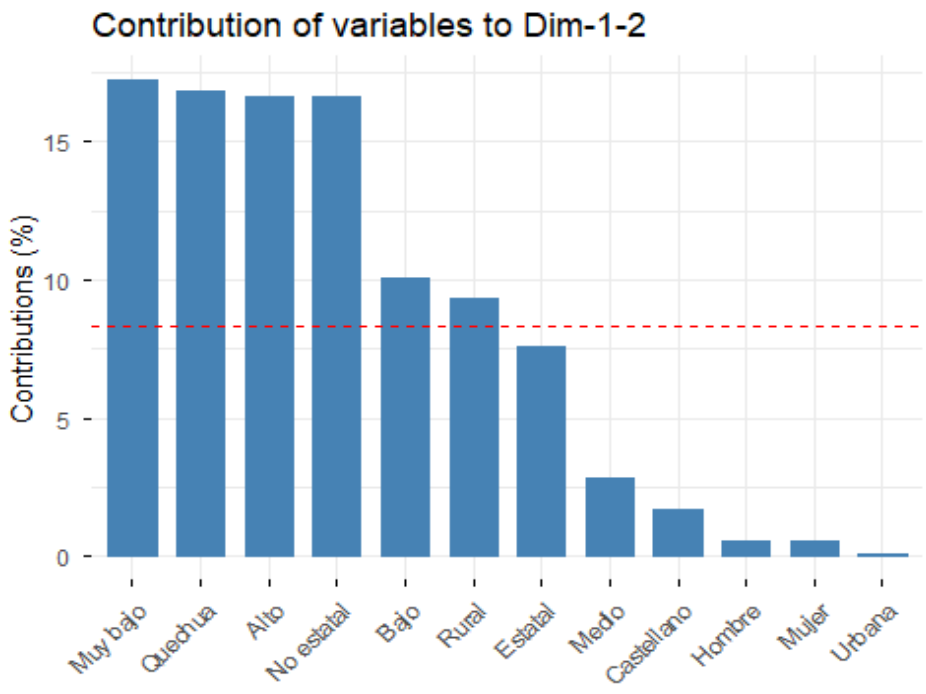
```r
## 2.2. Contribución de las Categorías de Variables Cualitativas a las Dimensiones 1 y 2
##      del MCA en Estudiantes de la UGEL Canas
fviz_contrib(mca_result, choice = "var", axes = 1)
```



```r
fviz_contrib(mca_result, choice = "var", axes = 2)
```

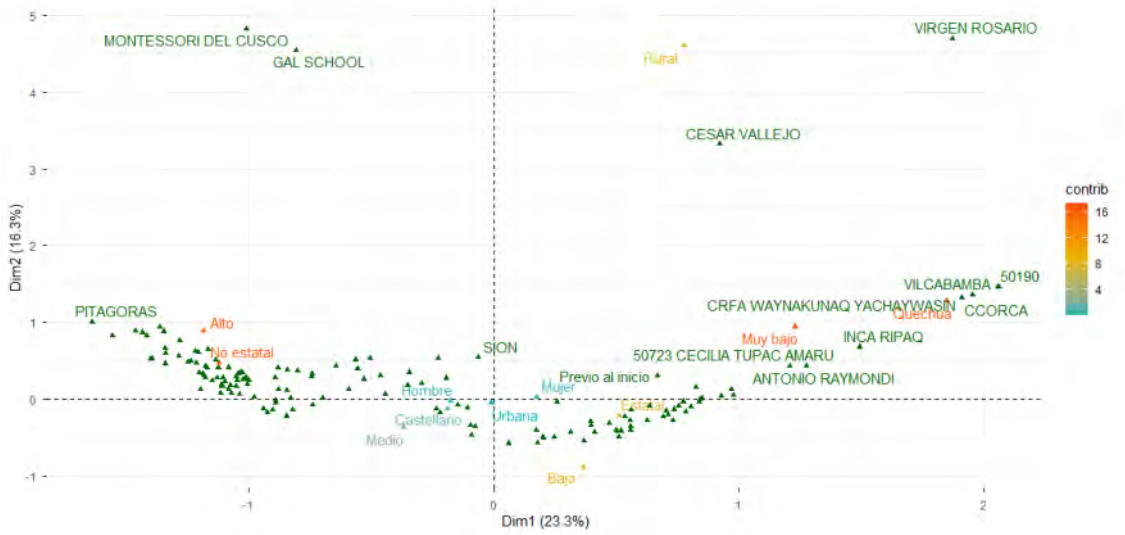



```r
fviz_contrib(mca_result, choice = "var", axes = 1:2)
```




```r
## 2.3. Plano Factorial del MCA de Variables Cualitativas en Estudiantes de la UGEL Canas
mca_plot <- fviz_mca_var(mca_result,
                         col.var = "contrib",
                         gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
                         repel = TRUE,
                         title = "",
                         ggtheme = theme_minimal())
```

```
mca_plot
```

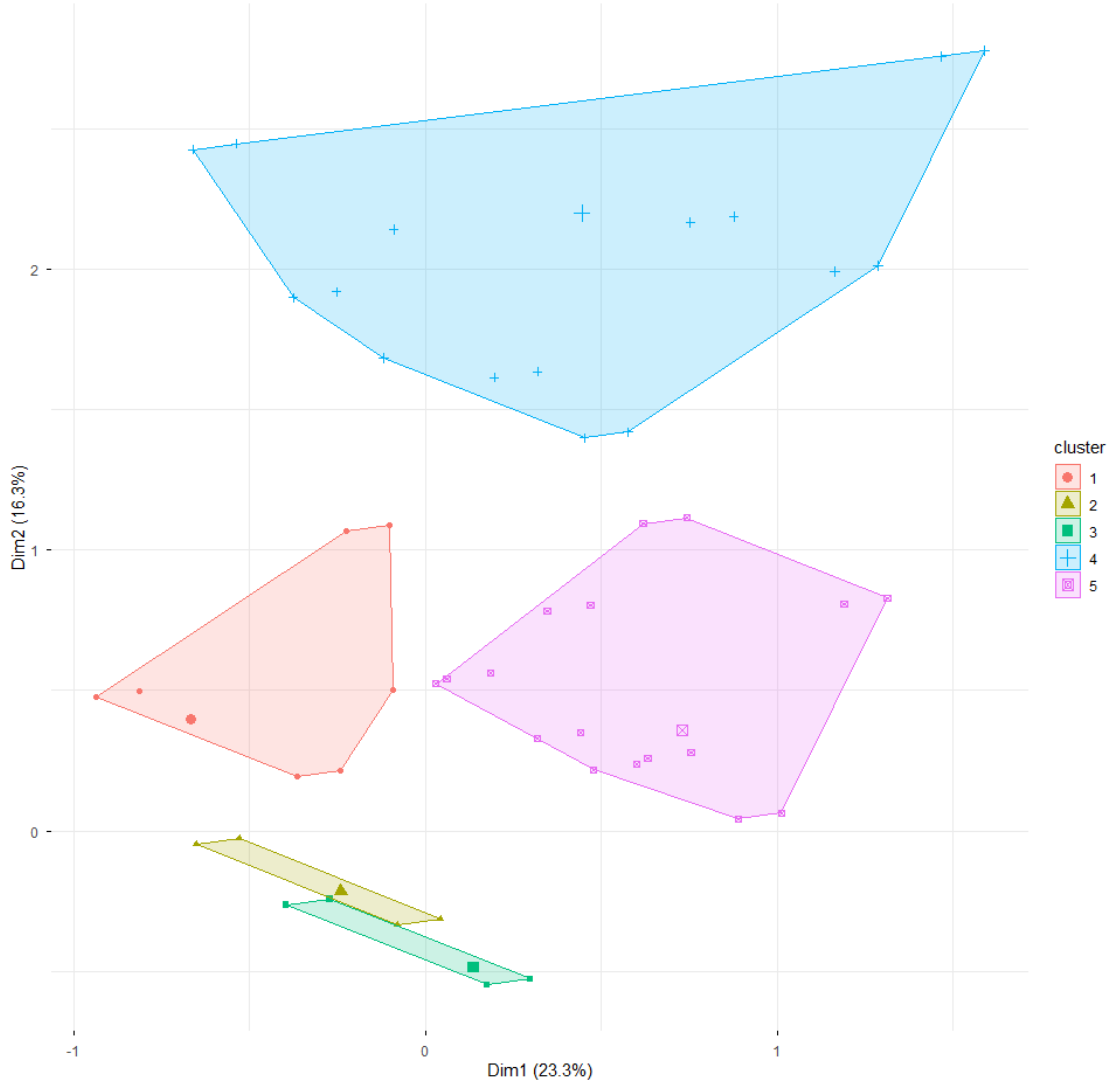



```
# 3. ANÁLISIS DE CLUSTERS JERÁRQUICO (HCPC)
hcpc_result <- HCPC(mca_result, nb.clust = -1, graph = FALSE)
```

```
# 4. AGREGAR LOS CLUSTERS A LOS DATOS ORIGINALES
datos_cluster <- cbind(datos, cluster = hcpc_result$data.clust$clust)
```

```
# 5. ANÁLISIS DE CLUSTERS
## 5.1. Distribución de clusters estudiantiles en el plano factorial (MCA) de la UGEL Canas
clusters_plot <- fviz_cluster(hcpc_result,
                              geom = "point",
                              ggtheme = theme_minimal(),
                              main = "")
```

```
clusters_plot
```




```
## 5.2. Pruebas de Chi-Cuadrado para la Asociación entre Variables Cualitativas
##      y Perfiles de Clusters Estudiantiles
hcpc_result$desc.var$test.chi2
```

```
##                     p.value  df
## area           0.000000e+00   4
## Nombre_IE      0.000000e+00 564
## gestion2       0.000000e+00   4
## lengua_materna 0.000000e+00   4
## n_ise          0.000000e+00  12
## grupo_L       4.433408e-158  12
## sexo           4.915427e-09   4
```

```r
## 6. Perfiles de rendimiento academico en lectura de la UGEL Cusco
perfiles <- datos_cluster |>
  group_by(cluster) |>
  summarise(
    n = n(),
    # Porcentaje por gestión
    Estatal = mean(gestion2 == "Estatal") * 100,
    No_estatal = mean(gestion2 == "No estatal") * 100,
    # Porcentaje por área
    Urbana = mean(area == "Urbana") * 100,
    Rural = mean(area == "Rural") * 100,
    # Porcentaje por sexo
    Mujer = mean(sexo == "Mujer") * 100,
    Hombre = mean(sexo == "Hombre") * 100,
    # Porcentaje por lengua materna
    Castellano = mean(lengua_materna == "Castellano") * 100,
    Quechua = mean(lengua_materna == "Quechua") * 100,
    # Distribución de n_ise
    Muy_bajo = mean(n_ise == "Muy bajo") * 100,
    Bajo = mean(n_ise == "Bajo") * 100,
    Medio = mean(n_ise == "Medio") * 100,
    Alto = mean(n_ise == "Alto") * 100,
    # Distribución de grupo_L
    Previo_inicio = mean(grupo_L == "Previo al inicio") * 100,
    En_inicio = mean(grupo_L == "En inicio") * 100,
    En_proceso = mean(grupo_L == "En proceso") * 100,
    Satisfactorio = mean(grupo_L == "Satisfactorio") * 100
  )
```

```r
# PERFILES
## # A tibble: 5 × 18
##   cluster     n Estatal No_estatal Urbana Rural Mujer Hombre Castellano Quechua
##   <fct>   <int>   <dbl>      <dbl>  <dbl> <dbl> <dbl>  <dbl>      <dbl>   <dbl>
## 1 1        1535    32.6       67.4     100     0  46.5   53.5       99.6   0.391
## 2 2        1963    61.6       38.4     100     0  47.9   52.1      100     0
## 3 3        1964    82.3       17.7     100     0  50.1   49.9      100     0
## 4 4          85    64.7       35.3       0   100  43.5   56.5       72.9  27.1
## 5 5        1628    94.9        5.10    100     0  57.1   42.9       61.1  38.9
##   Muy_bajo  Bajo  Medio   Alto Previo_inicio En_inicio En_proceso Satisfactorio
##      <dbl> <dbl>  <dbl>  <dbl>         <dbl>     <dbl>      <dbl>         <dbl>
## 1     2.87   0      0    97.1            2.67      23.8       36.9          36.5
## 2     0      0    100     0              5.15      31.6       36.7          26.5
## 3     0    100      0     0              7.23      40.0       35.3          17.4
## 4    41.2   16.5   18.8  23.5           22.4       34.1       23.5          20
## 5    83.0   11.8    4.67  0.491         17.0       48.8       25.3           8.91
```