



UNIVERSIDAD NACIONAL DE SAN ANTONIO ABAD DEL CUSCO

ESCUELA DE POSGRADO

MAESTRÍA EN ESTADÍSTICA

TESIS

**ANÁLISIS FACTORIAL DE CORRESPONDENCIAS APLICADO AL
ESTUDIO DE PATRONES EN TESIS DE PREGRADO DE LA
UNSAAC, CUSCO – 2025**

**PARA OPTAR AL GRADO ACADÉMICO DE MAESTRO EN
ESTADÍSTICA**

AUTOR

Br. MIJAEL ANDRE LIMA RODRIGUEZ

ASESOR

Dr. GUILLERMO PAUCAR CARLOS

CÓDIGO ORCID:

0000-0001-5476-1188

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 GUILLERMO DAUCAR CARLOS quien aplica el software de detección de similitud al trabajo de investigación/tesis titulada: ANÁLISIS FACTORIAL DE CORRESPONDENCIAS APLICADO AL ESTUDIO DE PATRONES EN TESIS DE PREGRADO DE LA UNSAAC, CUSCO-2025

Presentado por: MIGUEL ANDRÉ LIMA RODRÍGUEZ DNI N° 73825752;
presentado por: DNI N°:
Para optar el título Profesional/Grado Académico de MAESTRO EN ESTADÍSTICA

Informo que el trabajo de investigación ha sido sometido a revisión por 3 veces, mediante el Software de Similitud, conforme al Art. 6° del **Reglamento para Uso del Sistema Detección de Similitud en la UNSAAC** y de la evaluación de originalidad se tiene un porcentaje de 03%.

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.	☒
Del 11 al 30 %	Devolver al usuario para las subsanaciones.	☐
Mayor 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 las primeras páginas del reporte del Sistema de Detección de Similitud.

Cusco, 18 de AGOSTO de 2026


Firma

Post firma GUILLERMO DAUCAR CARLOS

Nro. de DNI 23892621

ORCID del Asesor 0000-0001-5476-1188

Se adjunta:

1. Reporte generado por el Sistema Antiplagio.
2. Enlace del Reporte Generado por el Sistema de Detección de Similitud: oid: 272592615211003

MIJAEL ANDRE LIMA RODRIGUEZ

ANÁLISIS FACTORIAL DE CORRESPONDENCIAS APLICADO AL ESTUDIO DE PATRONES EN TESIS DE PREGRADO DE LA U...

 Universidad Nacional San Antonio Abad del Cusco

Detalles del documento

Identificador de la entrega

trn:oid::27259:615211003

129 páginas

Fecha de entrega

18 ago 2026, 9:37 a.m. GMT-5

41.191 palabras

Fecha de descarga

18 ago 2026, 9:50 a.m. GMT-5

227.461 caracteres

Nombre del archivo

ANÁLISIS FACTORIAL DE CORRESPONDENCIAS APLICADO AL ESTUDIO DE PATRONES EN TESIS DE pdf

Tamaño del archivo

6.2 MB

3% Similitud general

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

Filtrado desde el informe

- Bibliografía
- Texto citado

Fuentes principales

- 2%  Fuentes de Internet
- 1%  Publicaciones
- 3%  Trabajos entregados (trabajos del estudiante)

Marcas de integridad

N.º de alertas de integridad para revisión

No se han detectado manipulaciones de texto sospechosas.

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

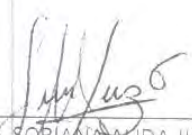
Dr. LEONCIO SOLIS QUISPE, Director de la Escuela de Posgrado, nos dirigimos a usted en condición de integrantes del jurado evaluador de la tesis intitulada **ANÁLISIS FACTORIAL DE CORRESPONDENCIAS APLICADO AL ESTUDIO DE PATRONES EN TESIS DE PREGRADO DE LA UNSAAC, CUSCO - 2025** del / de la Br. MIJAEL ANDRE LIMA RODRIGUEZ. Hacemos de su conocimiento que el (la) sustentante ha cumplido con el levantamiento de las observaciones realizadas por el Jurado el día CINCO DE AGOSTO DE 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, 17 de agosto 2026


Mgt. CARLA PATRICIA ZUÑIGA VILCA
Primer Replicante


Mgt. YETSABEL AUCCAILLE QUISPE
Segundo Replicante


Mtro. SORIANA ALIDA JIMENEZ COA
Primer Dictaminante


Mgt. RINA MARICELA ZAMALLOA CORNEJO
Segundo Dictaminante

DEDICATORIA

Quiero dar las gracias a quienes caminaron a mi lado en el sendero. A los que abrazaron mis grietas, me acompañaron en el silencio y me enseñaron que no se trata de poseer el mosaico completo de una vez, sino de hallar la armonía en lo incompleto y la paz en lo transitorio. Me recordaron que la verdadera riqueza es saber beber del agua que tenemos en el instante exacto en que el río pasa frente a nosotros.

Esta obra es la huella de un esfuerzo, el fruto de la sabiduría de actuar sin forzar, de sembrar y confiar en que el brote no se apresura. Es el testimonio de que la paciencia y la constancia son como el agua que, sin violencia, termina por moldear la piedra.

Porque la hoja cae cuando el viento lo indica, y el fruto madura en su propia grieta. Todo llega, en su perfecta e imperfecta armonía, cuando debe llegar.

AGRADECIMIENTOS

A la Universidad, por ser un refugio al que se llega con la esperanza de la semilla y se parte con la fortaleza del árbol. Un sendero con sus propias curvas y desafíos, donde el aprendizaje floreció en cada paso.

Al Dr. Guillermo Paucar Carlos, asesor de la investigación. Por actuar como guía serena y constante, por sus sugerencias que, dieron forma, claridad y profundidad a esta investigación, y por su genuino compromiso con el saber.

Por actuar con amor silencioso, a mi familia por su paciencia infinita y por ser el suelo firme que me sostuvo en los momentos en que el camino se sintió pesado. Su presencia discreta, pero inquebrantable, ha sido el verdadero cimiento de este logro.

ÍNDICE GENERAL

DEDICATORIA	i
ÍNDICE GENERAL	iii
LISTA DE TABLAS	vi
LISTA DE FIGURAS.....	vii
RESUMEN	viii
ABSTRACT.....	ix
INTRODUCCIÓN	x
CAPÍTULO I.	1
PLANTEAMIENTO DEL PROBLEMA	1
1.1 Situación problemática.....	1
1.2 Problema general y específicos	3
a. Problema general	3
b. Problemas específicos.....	3
1.3 Justificación de la investigación.....	3
1.4 Objetivos de la investigación	5
a. Objetivo general.....	5
b. Objetivos específicos	5
CAPÍTULO II.	7
MARCO TEÓRICO CONCEPTUAL	7
2.1. Bases teóricas	7
2.1.1. Producción de Tesis en la UNSAAC y Contexto nacional	7
2.1.2. Principios Teóricos de la Metodología de Investigación.....	8
2.1.3. Metodologías de Investigación en Tesis de Pregrado	9
2.1.4. Herramientas estadísticas aplicadas a tesis de pregrado.....	13
2.1.5. Análisis factorial de correspondencias (AFC).....	14

2.1.6. Prueba Chi-cuadrado de independencia	15
2.1.7. Análisis de correspondencia simple	19
2.1.8. Análisis de correspondencia múltiple	23
2.1.9. Procesamiento de Lenguaje Natural (NLP).....	25
2.1.10. Inteligencia Artificial Generativa	26
2.2. Marco conceptual	28
2.3. Antecedentes empíricos de la investigación.....	32
2.3.1. Antecedentes internacionales	32
2.3.2. Antecedentes nacionales.....	36
2.3.3. Antecedentes locales.....	38
CAPÍTULO III.....	43
HIPÓTESIS Y VARIABLES	43
3.1. Hipótesis.....	43
a. Hipótesis general.....	43
b. Hipótesis específicas.....	43
3.2. Identificación de variables	45
3.3. Operacionalización de variables.....	47
CAPÍTULO IV.....	49
METODOLOGÍA	49
4.1 Ámbito del estudio: localización política y geográfica.....	49
4.2 Tipo y nivel de investigación	49
4.3 Unidad de análisis	49
4.4. Población y muestra de estudio.....	49
4.5. Técnicas de recolección de información e instrumento de recolección de datos.....	51
4.6. Técnicas de análisis e interpretación de la información.....	53
CAPÍTULO V.....	56
RESULTADOS Y DISCUSION	56

5.1. Análisis descriptivo	56
5.2. Análisis bivariado de las variables de estudio.....	66
5.3. Análisis multivariado	78
DISCUSIÓN	94
CONCLUSIONES	97
RECOMENDACIONES	99
REFERENCIAS BIBLIOGRAFÍA	100
ANEXOS	103
Anexo 1. Ficha de observación de tesis universitarias.....	103
Anexo 2. Matriz de consistencia	106
Anexo 3. Código	107
Anexo 3. Imágenes.....	119

LISTA DE TABLAS

Tabla 1 Ejemplo de tabla de contingencia	16
Tabla 1 Composición de la población de tesis de pregrado UNSAAC	50
Tabla 2 Distribución de palabras en título variable contenido temático.....	56
Tabla 3 Análisis de variables nominales.....	57
Tabla 4 Análisis de variables nominales.....	58
Tabla 5 Análisis de variables nominales.....	60
Tabla 6 Análisis descriptivo de variables dicotómicas	62
Tabla 7 Descripción de variables numéricas y conversión a variables ordinales.	64
Tabla 8 Medida de asociación V de Cramer y frecuencia de pares y variables significativas	73
Tabla 9 Informe del primer análisis factorial de correspondencia.....	82
Tabla 10 Informe del segundo grafo análisis factorial de correspondencia.....	85
Tabla 11 Informe del tercer grafo mediante análisis factorial de correspondencia	88
Tabla 12 Informe del cuarto grafo mediante análisis factorial de correspondencia	91

LISTA DE FIGURAS

Figura 1 Nube de palabras del título de tesis.	56
Figura 2 Matriz del valor de significancia (sig.) de la prueba Chi cuadrado a pares de las 62 variables.	66
Figura 3 Matriz del valor de coeficiente de contingencia a pares de las 62 variables.....	68
Figura 4 Matriz del valor de V de cramer a pares de las 62 variables.....	70
Figura 5 Valor V de cramer y número de combinaciones significativas	73
Figura 6 Valor V de cramer y número de variables significativas.....	74
Figura 7 Matriz del valor de V de cramer a pares de las variables con correlación moderada alta 32 variables.....	77
Figura 8 Grafo 1 de la matriz de correlación	78
Figura 9 Grafo 2 de la matriz de correlación	79
Figura 10 Grafo 3 de la matriz de correlación.....	80
Figura 11 Grafo 4 de la matriz de correlación.....	81
Figura 12 Mapa factorial del Análisis de Correspondencias Múltiples (AFCM) Grafo 1	83
Figura 13 Mapa factorial del Análisis de Correspondencias Múltiples (AFCM) Grafo 2	86
Figura 14 Mapa factorial del Análisis de Correspondencias Múltiples (AFCM) Grafo 3	89
Figura 15 Mapa factorial del Análisis de Correspondencias Múltiples (AFCM) Grafo 4	92

RESUMEN

Esta investigación determinó las asociaciones y patrones estructurales en las tesis de pregrado de la Universidad Nacional de San Antonio Abad del Cusco (UNSAAC) mediante un enfoque cuantitativo, correlacional y transversal. Se analizaron 8,324 tesis del repositorio institucional (2017-2025), extrayendo datos de forma automatizada con técnicas de Procesamiento de Lenguaje Natural. Se aplicaron la prueba Chi-cuadrado, el coeficiente V de Cramer y el Análisis Factorial de Correspondencias Múltiples (AFCM).

Los resultados revelaron que el 84,3 % de las 1,891 combinaciones evaluadas presentaron dependencia estadística significativa ($p < 0,01$). El AFCM identificó a la Escuela Profesional como el nodo articulador principal del sistema, condicionando la arquitectura investigativa. Se caracterizaron cuatro tipologías: Determinismo Disciplinar-Metodológico, Gestión y Asesoría Desconectada, Gradiente de Complejidad Metodológica e Integridad Estructural y Calidad Documental. Predominaron el enfoque cuantitativo (81,1 %), los diseños descriptivos (50,1 %) y el autofinanciamiento (90,1 %).

Se concluye que la producción académica constituye un sistema estructurado donde la identidad disciplinar dicta las decisiones metodológicas. El AFCM demostró ser una herramienta robusta para revelar tipologías ocultas, proporcionando insumos empíricos para la planificación estratégica, la diversificación metodológica y la mejora de las políticas de financiamiento y asesoría en la investigación de pregrado de la UNSAAC.

Palabras clave: Análisis de Correspondencias Múltiples, Chi-cuadrado, Tesis de pregrado, Variables categóricas, Patrones estructurales.

ABSTRACT

This research determined the associations and structural patterns in the undergraduate theses of the Universidad Nacional de San Antonio Abad del Cusco (UNSAAC) using a quantitative, correlational, and cross-sectional approach. A total of 8,324 theses from the institutional repository (2017–2025) were analyzed, with data automatically extracted using Natural Language Processing (NLP) techniques. The Chi-square test, Cramer's V coefficient, and Multiple Correspondence Analysis (MCA) were applied.

The results revealed that 84.3% of the 1,891 evaluated combinations showed significant statistical dependence ($p < 0.01$). MCA identified the Professional School as the main articulating node of the system, conditioning the overall research architecture. Four thesis typologies were characterized: Disciplinary-Methodological Determinism, Disconnected Management and Advising, Methodological Complexity Gradient, and Structural Integrity and Documentary Quality. A clear predominance of the quantitative approach (81.1%), descriptive designs (50.1%), and self-financing (90.1%) was observed.

It is concluded that academic production constitutes a structured system where disciplinary identity dictates methodological decisions. MCA proved to be a robust tool for revealing hidden typologies, providing key empirical inputs for strategic planning, methodological diversification, and the improvement of funding and advising policies in UNSAAC's undergraduate research.

Keywords: Multiple Correspondence Analysis, Chi-square, Undergraduate theses, Categorical variables, Structural patterns

INTRODUCCIÓN

La investigación científica es un indicador importante que marca como será el desarrollo de los países. Las tesis de pregrado son importantes porque muestran el nivel formativo de los estudiantes y las tendencias en las diferentes áreas del conocimiento.

En el Perú, la ley universitaria exige que los estudiantes realicen una tesis para obtener su título profesional. Esto ha aumentado la producción de trabajos de investigación, aunque no siempre se han realizado estudios detallados para analizar estas investigaciones.

En la UNSAAC, hay muchas tesis disponibles, pero falta un análisis estructurado que permita entender las relaciones entre variables que estas aportan. Esto hace que sea difícil identificar tendencias y tomar decisiones informadas.

Para solucionar este problema, se propone utilizar la técnica estadística de Análisis Factorial de Correspondencias Múltiples (AFCM). Esta técnica permite explorar y visualizar las relaciones entre variables categóricas de manera simultánea. De esta forma, lo que permite identificar estructuras y tipos de tesis.

El presente estudio tiene como propósito identificar y analizar los patrones estructurales presentes en las tesis de pregrado de la UNSAAC. Para ello, se propone describir la distribución de la producción académica, determinar la existencia de asociaciones significativas entre variables categóricas y revelar las dimensiones factoriales subyacentes.

La investigación adopta un enfoque cuantitativo, de alcance correlacional y diseño no experimental de tipo transversal. La información se obtiene a partir de las tesis registradas en el repositorio institucional. El documento se estructura en cinco capítulos:

El Capítulo I plantea el problema de investigación y formula los objetivos.

El Capítulo II desarrolla el marco teórico y conceptual.

El Capítulo III presenta las hipótesis y la operacionalización de las variables.

El Capítulo IV describe la metodología empleada.

El Capítulo V expone los resultados, su discusión, las conclusiones y las recomendaciones derivadas de la investigación.

CAPÍTULO I.

PLANTEAMIENTO DEL PROBLEMA

1.1 Situación problemática

Las tesis de pregrado son una parte importante del proceso educativo universitario que muestra cómo ha sido el proceso de aprendizaje. A través de ellas, podemos ver qué métodos estadísticos se emplean frecuentemente y cuáles son los temas que son tendencia en diferentes áreas de estudio. Estar al tanto de estas tendencias es fundamental para poder brindar una mejor investigación y enseñanza.

A nivel mundial, las universidades tienen que mejorar la calidad en dos áreas importantes: la enseñanza y la investigación. Esto se nota en las tesis que realizan los estudiantes, donde la digitalización de tesis de pregrado en los repositorios universitarios que promueven la transparencia en el mundo académico ha llevado a que se utilicen diferentes metodologías, software especializado y enfoques interdisciplinarios para su sistematización y análisis. La importancia científica de las tesis cada vez está más relacionada con su publicación en revistas reconocidas y con la revisión por expertos en el área. (Hidalgo León, 2022). En el entorno académico, Debusho et al. (2025) indican que este se encuentra cada vez más enfocado en evaluar en profundidad la producción investigativa con fines de toma de decisiones estratégicas, influenciado por rankings y organismos de acreditación.

En el Perú, Cardaña Quispe (2021) indica que actualmente el reto principal de las universidades es elevar sus niveles de calidad y esto implica también que se deba mejorar la preparación de los estudiantes principalmente en la solidez de sus investigaciones y con esta finalidad, la Ley Universitaria N° 30220 especifica que realizar el trabajo de investigación, no es una opción sino un pilar fundamental que deben cumplir todas las instituciones superiores y es a partir de esta norma que la prioridad tanto, para docentes, alumnos y egresados colaboren estrechamente para generar nuevo conocimiento y desarrollar tecnologías innovadoras. (Cárdenas & Gutierrez, 2024).

Las tesis representan una fuente muy valiosa de información sobre las metodologías que se usan, los temas que se tratan y las prácticas de investigación más comunes en cada programa académico. En estudios como el desarrollado por Hidalgo León (2022) ha detectado algunos problemas, como "Documentación insuficiente (bajo evidencia) de políticas claras y actualizadas acerca de cómo atender las necesidades del entorno social y científico. Estudios a nivel nacional han revelado problemáticas recurrentes en la producción de tesis de pregrado, tales como la escasa publicación en revistas científicas indexadas, las limitaciones económicas para financiar las investigaciones, la falta de apoyo institucional sostenido y la concentración

temática en áreas ya saturadas. Estudios en áreas como Comunicación, Psicología y Medicina han identificado tendencias en la metodología, como el enfoque mayoritario en métodos cuantitativos. Asimismo, estos estudios señalan la existencia de temas que se repiten con frecuencia, así como desafíos persistentes tales como la escasa publicación en revistas científicas, las limitaciones económicas y la falta de apoyo institucional (Cárdenas Ortega & Gutierrez Blanco, 2024). De igual manera, se destaca un incremento en la elaboración de tesis con dos o tres autores a partir de la promulgación de la Ley Universitaria N.º 30220.

La función de SUNEDU, orientada a supervisar la calidad de la educación universitaria en el Perú, genera una presión constante, aunque indirecta, para que las instituciones eleven la calidad de sus investigaciones, incluidas las tesis de pregrado (SUNEDU, 2020).

En la UNSAAC, se ha detectado que falta un estudio completo y bien organizado que ayude a entender patrones o relaciones en las tesis de pregrado. Esta falta de información cubre aspectos importantes como los temas de investigación, las metodologías usadas, el año de sustentación, el género de los estudiantes que realizaron las tesis, y si el trabajo fue individual o en grupo. Esta situación hace difícil comprender cómo ha ido cambiando la producción académica en los distintos programas de estudio, además de complicar la identificación en los métodos usados. En un momento donde la calidad y la innovación en la investigación universitaria son cada vez más valoradas, es muy importante saber si hay diferencias notables en las formas en que investigan los estudiantes según las carreras de la UNSAAC.

La UNSAAC, consciente de su compromiso con la investigación, tiene su Reglamento General de Grados y Título y reglamentos específicos para cada facultad. Estos guían todo el proceso de investigación, estableciendo qué requisitos hay que cumplir, cómo asesorar, los criterios para evaluar, los formatos a usar y los plazos a seguir. La asesoría para tesis, tanto de pregrado como de posgrado, es una parte obligatoria en el trabajo adicional que hacen los docentes fuera de sus clases. (UNSAAC, 2015). La universidad también promueve la investigación científica a través de programas como “Yachayninchis Wiñarinanpaq”. El objetivo es despertar interés en la ciencia, fortalecer las habilidades de investigación y crear alianzas estratégicas que respondan a las necesidades tanto regionales como nacionales.

La posibilidad de trabajar en este problema en la UNSAAC es factible, ya que cuentan con una base de datos en su repositorio oficial de tesis que todavía no ha sido analizada por medios estadístico. Esto hace que el estudio sea, pionero, útil y tenga un impacto social, ya que permite a los estudiantes definir temas de tesis y explorar diferentes líneas de investigación. Para encontrar una solución, nos preguntamos:

1.2 Problema general y específicos

a. Problema general

- ¿Qué asociaciones y patrones estructurales existen entre las características de las tesis de pregrado de la UNSAAC, considerando variables identidad académica, contenido temático, enfoque metodológico, herramientas analíticas, características estructurales, calidad, producción y gestión académicas?

b. Problemas específicos

- ¿Cómo se distribuyen las tesis de pregrado registradas en el repositorio institucional de la UNSAAC, según variables identidad académica, contenido temático, enfoque metodológico, herramientas analíticas, características estructurales, calidad, producción y gestión académicas?
- ¿Existe asociación estadísticamente significativa entre las variables categóricas que caracterizan las tesis de pregrado de la UNSAAC, evaluadas mediante la prueba de Chi-cuadrado?
- ¿Qué dimensiones factoriales subyacentes explican la estructura de asociación entre las variables categóricas presentes en las tesis de pregrado de la UNSAAC, a través del Análisis Factorial de Correspondencias Múltiples (AFCM)?
- ¿Qué patrones estructurales o tipologías de tesis pueden identificarse a partir de la proyección en el espacio factorial obtenido mediante el Análisis Factorial de Correspondencias Múltiples (AFCM) en las tesis de pregrado de la UNSAAC?

1.3 Justificación de la investigación

La presente investigación posee una relevancia multidimensional, abarcando áreas académicas, institucionales, metodológicas, sociales y técnicas. A continuación, se detallan las justificaciones que sustentan la importancia de este estudio:

Justificación Académica

Esta investigación se realiza porque, hasta la fecha, no se ha llevado a cabo un análisis integral y sistemático que abarque la totalidad de la producción reciente de tesis de pregrado en la UNSAAC (8 324 trabajos entre 2017 y 2025). La ausencia de estudios cuantitativos y bibliométricos de esta magnitud impide comprender los patrones, tendencias y relaciones sistémicas que configuran la producción del conocimiento a nivel de pregrado. En consecuencia, el estudio se justifica porque permite transitar de la observación anecdótica a la evidencia estadística, revelando la existencia de un determinismo disciplinar-metodológico en

el que la identidad de la escuela profesional condiciona la arquitectura completa de la investigación. La identificación de estas tipologías estructurales contribuye al cierre de una brecha significativa en la cienciometría y la bibliometría a nivel local y nacional.

Justificación Institucional

La investigación se lleva a cabo porque las autoridades académicas, la Vicerrectoría de Investigación, los comités de ética, las comisiones de tesis y los docentes asesores carecen actualmente de un diagnóstico empírico que fundamente la toma de decisiones. Dado que los resultados evidencian que el 90,1 % de las tesis se autofinancia, que solo el 5,1 % cuenta con co-asesor y que existe una carga desigual en la distribución de la asesoría, el estudio se justifica porque proporciona la base de evidencia necesaria para que la universidad rediseñe sus políticas de incentivos, reoriente los concursos de fondos internos (como el programa Yachayninchis Wiñarinanpaq) y reformule las normativas de acompañamiento académico.

Justificación Metodológica

El estudio se desarrolla porque los análisis bibliométricos tradicionales, basados en metadatos superficiales o en codificación manual, resultan insuficientes para abordar volúmenes documentales de esta escala. Por ello, la investigación se justifica porque introduce un pipeline automatizado que combina técnicas de Procesamiento de Lenguaje Natural (NLP) y modelos de Inteligencia Artificial para la extracción estructurada de variables desde el texto completo de miles de documentos PDF. Asimismo, se incorpora el Análisis Factorial de Correspondencias Múltiples (AFCM) articulado con el filtro de ruido mediante el coeficiente V de Cramer, técnicas que no suelen emplearse de manera conjunta en el análisis de repositorios académicos en el Perú, lo que constituye una innovación metodológica significativa.

Justificación Social

La investigación se realiza porque las tesis de pregrado no constituyen únicamente un requisito académico, sino que representan el reflejo de las problemáticas regionales y el mecanismo principal mediante el cual se construye y transfiere conocimiento en las universidades públicas del sur del Perú. Al mapear las tendencias temáticas (con un fuerte enfoque en salud, educación, minería y el contexto geográfico del Cusco), el estudio se justifica porque proporciona una línea base que permitirá a los futuros tesisistas evitar la saturación de temas repetitivos, explorar vacíos de conocimiento y orientar sus esfuerzos hacia problemáticas sociales y productivas reales de la región.

Justificación Práctica y Técnica

El estudio se ejecuta porque resulta factible, conveniente y altamente escalable. Dado que se utilizan datos públicos disponibles en el repositorio institucional de la UNSAAC mediante el protocolo OAI-PMH (Open Archives Initiative – Protocol for Metadata Harvesting), el proceso es transparente, de bajo costo y libre de sesgos de selección humana. La metodología de extracción y análisis desarrollada en este estudio (scripts en Python, limpieza de datos y modelado estadístico) sienta un precedente técnico que puede ser replicado por otras universidades del país para auditar su propia producción científica. En un momento donde la transformación digital y la analítica de datos son indispensables, esta investigación demuestra cómo la estadística multivariada y la inteligencia artificial pueden aliarse para optimizar la gestión del conocimiento académico.

1.4 Objetivos de la investigación

a. Objetivo general

- Determinar las asociaciones y patrones estructurales en las tesis de pregrado de la UNSAAC mediante la prueba Chi-cuadrado y el Análisis Factorial de Correspondencias Múltiples (AFCM) de las variables identidad académica, contenido temático, enfoque metodológico, herramientas analíticas, características estructurales, calidad, producción académica y gestión académica

b. Objetivos específicos

- Describir la distribución de las tesis de pregrado registradas en el repositorio institucional de la UNSAAC, considerando variables categóricas identidad académica, contenido temático, enfoque metodológico, herramientas analíticas, características estructurales, calidad y producción académica y gestión académica.
- Establecer la asociación estadísticamente significativa entre las variables categóricas que caracterizan las tesis de pregrado de la UNSAAC, mediante la aplicación de la prueba de Chi-cuadrado, a fin de identificar relaciones relevantes entre dichas variables.
- Determinar las dimensiones factoriales subyacentes que explican la estructura de asociación entre las variables categóricas de las tesis de pregrado de la UNSAAC, mediante la aplicación del Análisis Factorial de Correspondencias Múltiples (AFCM), con el fin de reducir la dimensionalidad de los datos y sintetizar la información.

- Caracterizar patrones estructurales o tipologías de tesis de pregrado en la UNSAAC, a partir de la interpretación del espacio factorial generado por el Análisis Factorial de Correspondencias Múltiples (AFCM), con el propósito de reconocer perfiles diferenciados de producción académica.

CAPÍTULO II.

MARCO TEÓRICO CONCEPTUAL

2.1. Bases teóricas

2.1.1. Producción de Tesis en la UNSAAC y Contexto nacional

Presentar un trabajo de tesis es requisito para obtener el título profesional. Para graduarte como Bachiller, necesitas aprobar todos tus estudios de pregrado y entregar un trabajo de investigación. Si quieres obtener el Título Profesional, debes tener el grado de Bachiller y también elaborar una tesis.

En la UNSAAC, el proceso de presentar la tesis cuenta con respaldo institucional. Todo el apoyo, como la asesoría, evaluación y las actividades del jurado tanto para tesis de pregrado como de posgrado son parte de las responsabilidades que los docentes deben cumplir fuera de sus horas de clase y son requisitos obligatorios. La universidad asegura que la asesoría, el desarrollo y sustentación de la tesis sean gratuitas. La UNSAAC fomenta la publicación y difusión de los trabajos académicos producidos por los estudiantes, siempre respetando los derechos de autor. (UNSAAC, 2015)

A nivel nacional, el Registro Nacional de Trabajos de investigación RENATI (Registro Nacional de Trabajos de Investigación), administrado por SUNEDU (Superintendencia Nacional de Educación Superior Universitaria), recopila toda la información sobre los trabajos realizados para obtener grados de Bachiller, Maestría y Doctorado, así como los títulos profesionales de Licenciatura y Segundas Especialidades. Es obligatorio inscribir estos trabajos en el RENATI, lo que ayuda a llevar un control de la producción de tesis en el país.

El Segundo Informe Bienal sobre la Realidad Universitaria en el Perú de SUNEDU no da datos específicos sobre cómo ha cambiado la cantidad de tesis producidas. Pero sí comparte información sobre la producción científica en general.

Las publicaciones de universidades nacionales en revistas indexadas han aumentado con el tiempo. Estas publicaciones incluyen artículos y libros. El crecimiento ha sido constante. Se han publicado más artículos y libros. La cantidad de publicaciones universitarias en estas revistas sigue subiendo. En 2009, había alrededor de 500 publicaciones, y esa cifra aumentó a 1643 en 2017. Los artículos son la forma más común de publicación, representando aproximadamente el 58.5% del total en 2017. Los investigadores universitarios peruanos casi triplicaron su producción de artículos entre 2014 y 2017, pasando de 572 a 961 publicaciones en ese período. (SUNEDU, 2020)

En esta producción se notó una fuerte concentración tanto en las instituciones como en los temas abordados. En 2017, tres universidades juntas representaron más de la mitad, específicamente el 56%, de toda la producción nacional. La mayoría de las publicaciones, aproximadamente el 68%, en campos de salud y ciencias naturales. Cuando miramos el tipo de universidades, las privadas asociativas publicaron más estudios (alrededor de 1100 en 2017) en comparación con las públicas, que sumaron 394, y las privadas societarias, con 151. En regiones como la sierra y la selva, las universidades públicas fueron las que más produjeron en revistas indexadas. En la sierra, las universidades públicas produjeron el 54.2 % del total. En la selva, las universidades públicas produjeron el 75.9 % del total. Esto muestra que las universidades públicas en estas regiones tienen un papel importante en la producción de investigaciones y publicaciones. (SUNEDU, 2020)

2.1.2. Principios Teóricos de la Metodología de Investigación

Estructura y características de una tesis universitaria

La tesis, como producto final del proceso formativo en pregrado, debe adherirse a una estructura clara y lógica (Hernández Sampieri et al., 2014). Esta organización estandarizada, guiada por estilos de publicación aceptados como APA, Vancouver, Harvard, Chicago o MLA (Hernández Sampieri et al., 2014), lo que asegura la calidad, coherencia y validez del trabajo, permitiendo una presentación clara y ordenada del problema de investigación, su marco teórico o perspectiva, la metodología usada, resultados, conclusiones y recomendaciones.

Esta estructura no solo ayuda a que el jurado pueda evaluar de manera más objetiva y facilita a otros investigadores entender mejor, sino que también fomenta una cultura académica basada en la ética y el orden.

Un reporte de índole académico considera aspectos estructurales como la caratula o portada, tablas de contenido, resumen y el contenido central o cuerpo que incluye una parte introductoria, la revisión de contenidos literarios o las bases conceptuales en las que se basa la investigación, los aspectos o pasos metodológicos, el reporte de resultados y la comparación crítica con otros autores o discusión de resultados, además incluye una lista de referencias y la sección de apéndices (Hernández et al., 2014) en tanto, la formulación de las hipótesis se integra, si corresponde, en el planteamiento del problema (Kerlinger & Lee, 2002). Cada parte cumple una función específica, y es muy importante que todas estén alineadas. Lo que ayuda a que la investigación realmente aporte conocimientos, solucione problemas y tenga un impacto positivo. Identificamos las partes más importantes de una investigación:

- **Temática:** Es el núcleo del estudio y se refiere al tema, fenómeno o problema específico que queremos explorar. Elegir un tema relevante significa reconocer una necesidad de aprender más o identificar una brecha que aún no ha sido cubierta por la literatura científica. (Hernández Sampieri et al., 2014). Es fundamental que la investigación tenga un impacto positivo en la sociedad, lo académico o en la ciencia. Esto nos ayuda a crear nuevos conocimientos, solucionar problemas y hacer la vida de las personas mejor.
- **Palabras clave:** Son conceptos clave que resumen los temas, variables o aspectos principales del estudio. Tienen dos funciones ayudan a identificar y organizar el contenido en bases de datos académicas, y también guían al lector sobre los temas principales que se tratan.
- **Carrera profesional:** Disciplina académica que orienta el tipo de conocimiento y metodología empleada.
- **Año de sustentación:** Periodo cronológico en que se aprueba o defiende la tesis puede influir en el enfoque adoptado.
- **Género del tesista:** Variable sociodemográfica que puede estar relacionada con el enfoque temático o estilo investigativo.
- **Tipo de colaboración:** Forma en la que se organiza el trabajo de tesis: individual o colectiva.
- **Asesor:** Función técnica y pedagógica del docente que orienta y garantiza la calidad del trabajo de tesis.
- **Unidad de estudio:** Elemento analizado en la investigación.

2.1.3. Metodologías de Investigación en Tesis de Pregrado

Enfoques metodológicos.

Los principales enfoques que usamos son el método cuantitativo, cualitativo y la combinación de ambos llamados mixto.

- **Enfoque Cuantitativo:** Hernández Sampieri et al. (2014) indica que este enfoque sostiene que podemos aprender cosas válidas sobre el mundo a través de la observación directa, mediciones objetivas y poniendo a prueba nuestras hipótesis. Ve la realidad como algo que puede ser medido, objetivo y aplicable en diferentes situaciones. Recopila datos numéricos usando instrumentos estandarizados y usa estadísticas para analizarlos.

Intenta descubrir relaciones causales o asociaciones entre variables, así como establecer patrones de comportamiento, tendencias y regularidades que puedan generalizarse a poblaciones más amplias. Los resultados obtenidos pueden ser verificados por otros estudios y contribuir a la acumulación progresiva de evidencia científica dentro de un campo disciplinar. Los resultados que obtiene suelen aplicarse a poblaciones más grandes, usando muestras representativas. Propone pruebas de hipótesis, comparar diferentes grupos o medir cambios en variables específicas.

- **Enfoque Cualitativo:** Este enfoque se basa en ideas de la fenomenología, el interpretativismo y el constructivismo, que defienden que la realidad la construimos socialmente y que hay que entenderla a partir de las experiencias, significados y contextos de las personas. No busca hacer generalizaciones, sino comprender en profundidad un fenómeno o una experiencia específica. Trabaja con datos que no son numéricos, y el investigador es quien recoge la mayor parte de los datos. Para recopilar información, se usan técnicas como entrevistas, grupos de discusión, observación directa, análisis de contenido y estudios de caso. El análisis que se realiza es inductivo y va surgiendo a medida que se trabaja con los datos, en lugar de seguir un esquema rígido.(Kerlinger & Lee, 2002)
- **Enfoque Mixto:** Combinar métodos cuantitativos y cualitativos ayuda a superar sus propias limitaciones: los datos numéricos pueden faltar de contexto, mientras que los datos cualitativos a veces no se pueden aplicar a toda una población. Basándose en un enfoque pragmático, este método se enfoca en la utilidad y en qué funciona mejor según el problema de investigación. Permite recolectar y analizar tanto números como textos, lo que ayuda a corroborar, complementar y entender mejor los resultados.

Diseño de investigación

Cada uno de estos enfoques posee sus propios tipos de diseños de investigación:

❖ Enfoque cuantitativo.

- **Diseños experimentales:** El investigador modifica una o varias variables independientes se hacen para ver cómo afectan a las variables dependientes. El objetivo es entender si existe una relación de causa y efecto entre las variables independientes y las variables dependientes bajo condiciones controladas. Este tipo de estudios se clasifican en diferentes categorías.(Hernández Sampieri et al., 2014)
 - **Preexperimental:** Constituye el nivel más básico de control dentro de la familia experimental. Se caracteriza por la ausencia de asignación aleatoria de los sujetos a

los grupos y, en muchos casos, por la inexistencia de un grupo de control comparable. El investigador administra un tratamiento o estímulo a un único grupo y observa sus efectos mediante una posprueba, sin contar con una condición de comparación que permita descartar amenazas a la validez interna como la maduración, la historia o la selección. (Hernández Sampieri et al., 2014).

- **Cuasiexperimentales:** Se diferencia del diseño experimental propiamente dicho precisamente en el método de asignación de los sujetos a los grupos. Mientras que en el experimento verdadero los participantes son asignados al azar, en el cuasiexperimento los sujetos ya se encuentran organizados en grupos intactos o preformados, cuya conformación obedece a criterios naturales, administrativos o institucionales ajenos al investigador. En consecuencia, no existe asignación aleatoria, lo que introduce el riesgo de que los grupos no sean equivalentes en sus características iniciales y compromete parcialmente la validez interna. (Hernández Sampieri et al., 2014; Kerlinger & Lee, 2002).
- **Experimental:** Representa el estándar de mayor rigor dentro de los diseños experimentales. Su rasgo definitorio es la asignación aleatoria de los participantes a los grupos de tratamiento y de control, procedimiento que garantiza que las características iniciales de los sujetos se distribuyan de manera equitativa entre las condiciones experimentales y neutraliza los sesgos de selección. Esta equivalencia inicial entre los grupos, sumada a la manipulación deliberada de la variable independiente y al control de variables extrañas, asegura tanto el control como la validez interna del estudio. Estos diseños pueden incorporar prepruebas y pospruebas, así como una o varias variables independientes y dependientes. (Kerlinger & Lee, 2002)
- **Diseños no experimentales:** Este tipo de investigación se realiza observando fenómenos en su entorno natural sin manipular nada deliberadamente. Se analizan los fenómenos tal como ocurren, para entender cómo funcionan. Según la duración del proceso, estos estudios se dividen en diferentes categorías.
 - **Diseños Transversales:** Recolectar datos en un solo momento, en lugar de hacerlo, puede clasificar en diferentes tipos:
 - **Exploratorios:** El objetivo es empezar a comprender algo, como una sola variable, un conjunto de variables, una comunidad, un contexto o una situación. Esto es especialmente importante cuando nos enfrentamos a problemas que son nuevos o que no conocemos bien. (Hernández Sampieri et al., 2014)

- **Descriptivos:** Tiene como propósito comprender las cualidades, rasgos y características de individuos, grupos, comunidades, procesos u otros fenómenos.
- **Correlacional:** Esta describen cómo dos o más variables se relacionan, o están conectadas en un momento dado, ya sea en términos de que una causa afecta a otra o simplemente están relacionadas. Las causas y los efectos ya sucedieron o están sucediendo mientras se realiza el estudio. El investigador observa las causas y los efectos. En este tipo de investigación, es común encontrar el uso de encuestas de opinión suelen ser un tipo de estudio que no busca experimentar, sino describir y encontrar relaciones o causas y efectos en un momento específico.(Hernández Sampieri et al., 2014)
- **Diseños Longitudinales:** Se recopilan datos en diferentes momentos para entender cómo evoluciona el problema, sus causas y efectos. Esto se dividen en:
 - **Tendencia:** Analizan cambios en una población a lo largo del tiempo.
 - **Cohortes:** Son grupos o subpoblaciones específicos que se analizan para detectar cambios significativos a lo largo del tiempo.
 - **Panel:** Los mismos participantes o casos se analizan en diferentes momentos o en distintos puntos del estudio.

❖ Diseños de Investigación Cualitativa

En el enfoque cualitativo, el concepto de diseño hace referencia al marco general que se empleará durante el proceso de investigación, caracterizándose por su flexibilidad y apertura a las condiciones específicas del contexto. Los principales diseños que suelen considerarse son:

- **Diseños de Teoría Fundamentada:** El objetivo es crear una teoría basada en hechos reales que se obtienen directamente del contexto estudiado y siendo una de las personas involucradas en el problema.
- **Diseños Etnográficos:** Tienen como finalidad principal estudiar a las personas dentro de su propio entorno natural, de modo que el investigador se sumerja en el sentir, el actuar y la manera de pensar de los distintos grupos humanos, culturas o comunidades. Para lograrlo, participa activamente en el proceso mediante la observación participante, lo que le permite comprender de manera más integral y contextualizada la realidad estudiada (Kerlinger & Lee, 2002)
- **Diseños Narrativos:** Las historias bien contadas nos ayudan a entender mejor un problema o fenómeno, ya que nos ofrecen diferentes puntos de vista que enriquecen la

perspectiva. Estas narrativas pueden variar en forma y estilo, adaptándose a lo que quieren mostrar o transmitir.(Hernández Sampieri et al., 2014)

- **Diseños Fenomenológicos:** El objetivo es conocer y entender las experiencias que tienen las personas sobre un fenómeno en particular, considerando diferentes opiniones al describir y analizar cómo viven esa experiencia desde su propia perspectiva.
- **Diseños de Investigación-Acción:** La idea principal es que la investigación debe generar un cambio y que ese cambio se integre en el propio proceso de investigación. Esto implica investigar mientras se interviene, en paralelo. Una modalidad muy conocida es la investigación-acción participativa, donde tanto la comunidad como los investigadores trabajan juntos de manera durante todo el proceso.

❖ **Diseños de Investigación Mixta**

Los métodos mixtos son una forma de investigación que combina datos numéricos y datos descriptivos en un solo estudio. Esto permite recopilar, analizar e integrar los datos de ambas maneras. Los métodos mixtos han ganado popularidad porque ofrecen una visión completa al combinar lo cuantitativo y lo cualitativo. Esta estrategia busca aprovechar las fortalezas inherentes de ambos enfoques, al tiempo que minimiza sus posibles limitaciones. Un estudio de carácter mixto inicia con una definición del problema que requiere la utilización conjunta de ambos enfoques. Los modelos generales que rigen los diseños mixtos son los siguientes:

- **Diseños Concurrentes:** El proceso consiste en recopilar y analizar juntos tanto datos numéricos como cualitativos para obtener una visión más completa.
- **Diseños Secuenciales:** Consiste en una fase inicial que emplea un método, ya sea cuantitativo o cualitativo, seguida de una etapa en la que se utiliza el otro método.
- **Diseño Exploratorio Secuencial (DEXPLOS):** El proceso comienza con una fase cualitativa que incluye una exploración inicial y el desarrollo de una teoría emergente, seguida de una etapa cuantitativa para probar y generalizar dicha teoría.
- **Diseño Explicativo Secuencial (DEXPLIS):** El proceso empieza con una etapa cuantitativa, donde se recopilan y analizan datos, para luego pasar a una fase cualitativa que ayuda a entender o profundizar en los resultados que se obtuvieron antes.

2.1.4. Herramientas estadísticas aplicadas a tesis de pregrado

Las herramientas estadísticas son de importancia en las tesis de pregrado que adoptan un enfoque cuantitativo facilitan el análisis de las mediciones numéricas y el análisis

sistemático de los datos con el propósito de probar hipótesis, detectar patrones y construir conocimientos fundamentados en evidencia empírica.

Las herramientas estadísticas son importantes porque ayudan a respaldar las ideas que tenemos antes de recopilar los datos. Esas ideas se prueban con análisis estadísticos que nos dicen si los resultados son significativos o no. Sirven para describir fenómenos, investigar cómo se relacionan diferentes variables, ya sea en forma de correlación o para entender si una causa lleva a un efecto.(Cuadras, 2014)

La estadística se divide en dos grandes líneas que cumplen diferentes funciones, pero que se complementan mutuamente; la primera, descriptiva y se ocupa de ordenar y simplificar los datos recolectados con la finalidad de que estos sean más fáciles de entender y para lograrlo, recurre al uso de tablas de frecuencias, a valores que son representativos, tales como la media, la mediana o la moda, así como a indicadores de variabilidad como el rango o la desviación estándar; la segunda línea inferencial, avanza en un paso más adelante y es que no solo describe sino que aprovecha la información para lograr conclusiones más amplias, además presenta un gran aporte, y es que puede predecir comportamientos o características de todo un grupo o de la población y esto basándose únicamente en el estudio de una parte de estos y para llevar a cabo este proceso, dispone de diferentes herramientas, como el coeficiente de Pearson que es de utilidad para medir la fuerza de la correlación entre dos cantidades, la prueba t que verifica las diferencias entre dos grupos, el ANOVA que amplía la comparación hacia tres o más grupos y la regresión lineal que ayuda a la estimación del valor de una variable de acuerdo con otra, cuando los datos no cumplen ciertos supuestos o son de naturaleza categórica, se recurre a pruebas no paramétricas como la Chi-cuadrado (χ^2), útil para examinar la asociación entre variables.(Parra, 2019)

2.1.5. Análisis factorial de correspondencias (AFC)

Nos ayuda a explorar datos con muchas variables y entender cómo se relacionan entre sí. Es especialmente útil cuando se trabaja con variables cualitativas o categóricas. A diferencia del análisis factorial clásico, en el Análisis de Correspondencias se usa la estadística Chi-Cuadrado (χ^2) para medir las diferencias entre categorías, en lugar de calcular distancias euclidianas como en otros métodos. (Greenacre, 2008). El objetivo principal es descubrir cómo diferentes categorías en una tabla de datos se relacionan entre sí. Esto nos ayuda a entender y analizar cómo se conectan los diferentes atributos de esas variables. Tiene como finalidad representar dichas relaciones de forma visual, facilitando la comprensión de los datos, la

identificación de patrones relevantes y el entendimiento de la estructura global del conjunto de información. (Greenacre, 2008).

El AFC (Análisis Factorial de Correspondencias) ayuda a identificar variables que no podemos ver directamente, como algunos aspectos o factores que influyen en los datos. Gracias a la inercia, podemos entender cuánto del cambio o la dispersión en los datos se explica por estos factores que elegimos.

El Análisis de Correspondencias Múltiples (AFCM) es una extensión del AFC (Greenacre, 2008). Esta técnica fue creada para entender cómo se relacionan más de dos variables categóricas entre sí. Su desarrollo es gracias al trabajo de Jean-Paul Benzécri y su equipo en Francia, especialmente con su publicación en 1973. Se puede realizar usando una matriz binaria, también llamada tabla disyuntiva completa, o mediante una matriz de Burt.

Base estadística

Para entender cómo funciona el análisis de correspondencia y la correspondencia múltiple, es importante tener en cuenta conceptos como la prueba de chi cuadrado para la independencia de variables.

2.1.6. Prueba Chi-cuadrado de independencia

Antes de aplicar el análisis de correspondencias, se realiza la prueba de Chi-cuadrado de independencia:

- H_0 : Las variables son independientes
- H_1 : Las variables no son independientes

Si $p < 0.05$, se rechaza la independencia de variables y el análisis de correspondencias es pertinente. Para poder realizar la prueba se debe conocer algunos conceptos. (Cuadras, 2014)

Tabla de contingencia

Sea una tablade tamaño $I \times J$, donde:

- n_{ij} : frecuencia observada en la celda (i, j) .
- $n_{i+} = \sum_{j=1}^J n_{ij}$: total marginal de la fila i .
- $n_{+j} = \sum_{i=1}^I n_{ij}$: total marginal de la columna j .
- $n_{++} = \sum_{i=1}^I \sum_{j=1}^J n_{ij}$: total general.

queda expresada como.

Tabla 1

Ejemplo de tabla de contingencia

	C_1	C_2	$\dots$	C_J	Total, fila
F_1	n_{11}	n_{12}	$\dots$	n_{1J}	n_{1+}
F_2	n_{21}	n_{22}	$\dots$	n_{2J}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
F_I	n_{I1}	n_{I2}	$\dots$	n_{IJ}	n_{I+}
Total col.	n_{+1}	n_{+2}	$\dots$	n_{+J}	n_{++}

Con la tabla de contingencia se puede generar perfil fila y perfil columna los cuales permiten una interpretación no influenciada por la cantidad de datos. Los perfiles son distribuciones de frecuencias relativas que permiten comparar categorías. (Sarmiento, s. f.)

Perfil fila: Distribución condicional de columnas dada una fila:

$$\text{Perfil fila } i: \left(\frac{n_{i1}}{n_{i+}}, \frac{n_{i2}}{n_{i+}}, \dots, \frac{n_{iJ}}{n_{i+}} \right)$$

Perfil columna: Distribución condicional de filas dada una columna:

$$\text{Perfil columna } j: \left(\frac{n_{1j}}{n_{+j}}, \frac{n_{2j}}{n_{+j}}, \dots, \frac{n_{Ij}}{n_{+j}} \right)$$

Hipótesis de Independencia

Para la verificación de independencia de las variables categóricas de estudio se plantea la hipótesis nula

- H_0 : Las variables fila y columna son independientes.
- H_1 : Las variables fila y columna no son independientes.

Si se cumple H_0 , la frecuencia esperada en la celda (i, j) esta dada por:

$$E_{ij} = \frac{n_{i+} \cdot n_{+j}}{n_{++}}$$

Esta fórmula que define el valor esperado en la celda (i, j) para cual la distribución seria independiente que proviene de las definiciones de probabilidad conjunta e independencia:

$$P(F_i \cap C_j) = P(F_i) \cdot P(C_j) = \frac{n_{i+}}{n_{++}} \cdot \frac{n_{+j}}{n_{++}} \Rightarrow E_{ij} = n_{++} \cdot P(F_i \cap C_j)$$

El estadístico χ^2 mide las variaciones entre frecuencias observadas y esperadas, entre el valor real observado y el valor esperado o valor en el cual la frecuencia seria cumplirían el

supuesto de independencia. Al sumar estos errores tiende a distribuirse de forma chi cuadrado. (Sarmiento, s. f.)

$$\chi^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

También se observa que el valor se calcula por celda donde cada celda contribuye al total del χ^2 con:

$$\chi_{ij}^2 = \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

Estas contribuciones permiten identificar celdas con desviaciones más fuertes respecto a la independencia.

Si se cumple H_0 , el estadístico χ^2 sigue aproximadamente una distribución chi-cuadrado con:

$$gl = (I - 1)(J - 1)$$

Valor que sirve para realizar la comparación con la distribución chi cuadrado, la cual es calculada para cada término mediante $\frac{(n_{ij} - E_{ij})^2}{E_{ij}}$ que es el cuadrado de una desviación estandarizada.

- Bajo H_0 , $(n_{ij} - E_{ij})/\sqrt{E_{ij}}$ se aproxima a una normal estándar.
- La suma de cuadrados de k normales estándar independientes sigue una χ_k^2 .
- Debido a las restricciones de los totales marginales, hay $(I - 1)(J - 1)$ grados de libertad. La comparación de los valores chi cuadrado calculado y de tabla son los que definen si las dos variables categóricas son independientes o no.

Coefficiente de contingencia de Pearson

El coeficiente de contingencia C , es una medida derivada del estadístico χ^2 que resume en un solo valor la intensidad de la asociación entre las variables de una tabla de contingencia. Se define como:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

donde χ^2 es el estadístico de la prueba de independencia y n el tamaño de la muestra. Dado que C es una función monótona creciente de χ^2 , valores próximos a 0 indican independencia entre las variables, en tanto que valores próximos a 1 reflejan asociaciones muy intensas. Una

limitación importante de C es que su valor máximo teórico no es 1, sino que depende de las dimensiones de la tabla.

Esta dependencia de las dimensiones de la tabla es, precisamente, la razón por la que en el presente estudio el coeficiente de contingencia se emplea como medida complementaria para visualizar la magnitud de las asociaciones, mientras que para la selección de relaciones relevantes se privilegia el coeficiente V de Cramer, cuya escala $[0,1]$ es independiente de r y c .

Coeficiente V de Cramer para tablas de contingencia

El coeficiente V de Cramer es un indicador de asociación entre variables categóricas que se fundamenta en el estadístico Chi-cuadrado. A diferencia de este último, el V de Cramer permite medir la fuerza de la relación de manera estandarizada, con valores comprendidos entre 0 y 1, no depende del tamaño de la tabla de contingencia.(Sarmiento, s. f.)

Definición del coeficiente V de Cramer

El coeficiente V de Cramer se define como:

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(I - 1, J - 1)}}$$

donde:

- χ^2 : estadístico de la prueba de independencia
- n : tamaño de la muestra
- $\min(I - 1, J - 1)$: número mínimo de grados de libertad.

Interpretación: El coeficiente V de Cramer toma valores en el intervalo:

$$0 \leq V \leq 1$$

donde:

$V = 0$: independencia entre variables

$V \approx 1$: asociación fuerte

Se utilizan criterios de interpretación aproximados:

Valor de V	Interpretación
0.00 – 0.10	Asociación muy débil
0.10 – 0.30	Asociación débil
0.30 – 0.50	Asociación moderada
> 0.50	Asociación fuerte

2.1.7. Análisis de correspondencia simple

Permite examinar la relación entre dos variables de tipo categórico, utilizando una tabla de contingencia que refleja la frecuencia con la que se presentan conjuntamente sus distintas categorías. La idea principal de esta técnica es visualizar esas relaciones en un mapa, que puede ser en 2 dimensiones o 3 dimensiones. En estos mapas, las categorías que están más cercanas entre sí representan relaciones similares. No intenta predecir una variable, sino, explicar y resumir cómo están relacionadas las categorías en los datos.(Greenacre, 2008)

Una tabla de contingencia con I filas y J columnas. Cada elemento n_{ij} representa el número de individuos que pertenecen a la categoría i de la primera variable y a la categoría j de la segunda.

Definimos a N como la matriz de la tabla de contingencia que contiene n_{ij} de frecuencias de las categorías. Definimos la matriz de correspondencias (P) como las frecuencias n_{ij} divididas entre el total(Cuadras, 2014)

$$P = \frac{1}{n_{++}} N$$

Los porcentajes totales parciales se denominan masas donde:

- Vector de masas fila: $r = (r_1, \dots, r_I)^\top$, donde:

$$r_i = \sum_{j=1}^J p_{ij} = n_{i+}/n_{++}$$

- Vector de masas columna: $c = (c_1, \dots, c_J)^\top$, donde:

$$c_j = \sum_{i=1}^I p_{ij} = n_{+j}/n_{++}$$

Dado los vectores de masa se puede definir las matrices diagonales de estas como

$$D_r = \text{diag}(r), \quad D_c = \text{diag}(c)$$

Para el cálculo de AFC se requiere el perfil fila y columna los cuales se pueden calcular de manera matricial de las siguientes maneras.

$$\text{Perfil fila } R = D_r^{-1}P$$

$$\text{Perfil columna } C = PD_c^{-1}$$

Distancia entre perfiles fila

La distancia chi-cuadrado entre dos perfiles fila i y i' es:

$$d_x^2(i, i') = \sum_{j=1}^J \frac{1}{c_j} \left(\frac{p_{ij}}{r_i} - \frac{p_{i'j}}{r_{i'}} \right)^2$$

Lo que genera una distancia entre dos características de la variable categórica.

Inercia total

Mide la dispersión global de los perfiles fila respecto al perfil medio c . Se define como:

$$\text{Inercia} = \sum_{i=1}^I r_i \cdot d_{\chi}^2(i, c) = \frac{\chi^2}{n_{++}}$$

donde χ^2 es el estadístico de la prueba de independencia.

Matriz estandarizada

Definimos la matriz:

$$S = D_r^{-1/2}(P - rc^T)D_c^{-1/2}$$

Esta matriz:

- Resta el modelo de independencia (rc^T) a P .
- Estandariza con $D_r^{-1/2}$ y $D_c^{-1/2}$ para tener en cuenta las masas.

Cada elemento s_{ij} mide la desviación estandarizada de p_{ij} respecto a $r_i c_j$.

Descomposición en Valores Singulares (DVS)

La DVS de la matriz S es:

$$S = U\Delta V^T$$

Esta descomposición expresa S como el producto de tres matrices con propiedades geométricas y algebraicas clave. Donde U y V son ortogonales y Δ es diagonal y sus elementos cumplen que $v_1 \geq v_2 \geq \dots$

Componentes de la DVS. Matriz $U \in \mathbb{R}^{I \times K}$: Matriz ortogonal cuyas columnas son los vectores singulares izquierdos. Cumple: $U^T U = I_K$, donde I_K es la matriz identidad de tamaño K . Cada columna u_k representa una dirección en el espacio de las filas.

Matriz $V \in \mathbb{R}^{J \times K}$: Matriz ortogonal cuyas columnas son los vectores singulares derechos. Cumple: $V^T V = I_K$. Cada columna v_k representa una dirección en el espacio de las columnas.

Matriz $\Delta \in \mathbb{R}^{K \times K}$: Matriz diagonal con los valores singulares en orden decreciente:

$$\Delta = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_K), \quad \sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_K \geq 0$$

Estos valores miden la importancia o magnitud de cada componente. Dimensión K : Número máximo de componentes no triviales:

$$K = \min(I - 1, J - 1)$$

Esta restricción se debe a que, bajo el modelo de independencia, hay una dependencia lineal entre las filas y columnas (los totales marginales imponen restricciones).

La DVS no solo descompone S , sino que también permite extraer una estructura geométrica útil para visualizar los datos. (Sarmiento, s. f.)

Método de Cálculo de la DVS

La DVS no se calcula con una fórmula directa, sino mediante un algoritmo numérico basado en álgebra lineal computacional. A continuación, se describen los pasos teóricos:

a) Calcular $S^T S$ y SS^T

$S^T S \in \mathbb{R}^{J \times J}$: matriz de covarianzas entre columnas

$SS^T \in \mathbb{R}^{I \times I}$: matriz de covarianzas entre filas

Ambas son simétricas y semi definida positivas.

b) Calcular valores y vectores propios

Los valores propios de $S^T S$ son $\lambda_k = \sigma_k^2$ - Los vectores propios de $S^T S$ forman las columnas de V - Los vectores propios de SS^T forman las columnas de U .

Relación entre U , V y Δ

Para $k = 1, \dots, K$ y $\sigma_k > 0$:

$$u_k = \frac{1}{\sigma_k} S v_k, \quad v_k = \frac{1}{\sigma_k} S^T u_k$$

Esto garantiza que $S = U \Delta V^T$.

Relación con la inercia

En el AC, la inercia es una medida de la asociación entre filas y columnas. Está directamente relacionada con los valores singulares:

La inercia explicada por el eje k (o dimensión k) es:

$$\sigma_k^2$$

Cuanto mayor sea σ_k , más estructura capta ese eje.

La inercia total del análisis es la suma de todas las inercias por eje:

$$\text{Inercia total} = \sum_{k=1}^K \sigma_k^2$$

Dada las matrices U, Δ, V se puede calcular las coordenadas para el grafico de ACS

- Coordenadas principales de filas:

$$F = D_r^{-1/2} U \Delta$$

- Coordenadas principales de columnas:

$$G = D_c^{-1/2} V \Delta$$

- Coordenadas estándar (o recíprocas)

Estas coordenadas se usan para interpretar la inercia y las contribuciones.

- Para las filas:

$$\Phi = D_r^{-1/2} U \in \mathbb{R}^{I \times K}$$

- Para las columnas:

$$\Gamma = D_c^{-1/2} V \in \mathbb{R}^{J \times K}$$

Las coordenadas estándar tienen la propiedad de que:

$$\sum_{i=1}^I r_i \phi_{ik}^2 = 1, \quad \sum_{j=1}^J c_j \gamma_{jk}^2 = 1$$

es decir, están normalizadas por las masas.

Mapas del Análisis de Correspondencias

Hay dos tipos principales de mapas:

a) Mapa simétrico

Se representan filas y columnas con sus coordenadas principales (F y G). Es útil para comparar la posición relativa de filas y columnas, pero no permite interpretar distancias entre filas y columnas directamente.

b) Mapa asimétrico

- Las filas se representan con coordenadas principales (F).
- Las columnas se representan con coordenadas estándar (Γ) o con Γ ponderadas.
- Permite interpretar cómo cada columna explica la posición de las filas.

Inercia explicada por un eje

Tras aplicar DVS a la matriz estandarizada S , se obtienen los valores singulares $\sigma_1, \sigma_2, \dots, \sigma_K$. La inercia explicada por el eje k es:

$$\lambda_k = \sigma_k^2$$

Porcentaje de inercia explicado

El porcentaje de inercia explicado por el eje k es:

$$\%_k = \frac{\lambda_k}{\sum_{k=1}^K \lambda_k} \times 100 = \frac{\sigma_k^2}{\sum_{k=1}^K \sigma_k^2} \times 100$$

Este valor indica la importancia relativa del eje. Por ejemplo, si $\%_1 = 60\%$, el primer eje explica el 60% de la estructura de asociación.

Relación de AFCM con V de Cramer

La inercia total está dada por la expresión $\frac{\chi^2}{n}$ lo que permite establecer que el coeficiente V de Cramer es una transformación normalizada de dicha inercia. Esta relación se expresa como:

$$V = \sqrt{\frac{\text{Inercia total}}{\min(I - 1, J - 1)}}$$

El coeficiente V de Cramer no solo permite detectar la existencia de asociación, sino que también puede interpretarse como una medida estandarizada de la desviación respecto a la independencia entre variables categóricas.

Una de sus principales ventajas es que cuantifica la intensidad de la relación en una escala normalizada entre 0 y 1, lo que facilita su interpretación. Este coeficiente es independiente del tamaño de la tabla de contingencia, a diferencia de otros indicadores como el coeficiente de contingencia, cuya interpretación depende de las dimensiones de la tabla. Gracias a estas características, el coeficiente V de Cramer permite identificar asociaciones relevantes que posteriormente pueden ser analizadas con mayor profundidad mediante técnicas multivariadas como el Análisis Factorial de Correspondencias Múltiples (AFCM). (Sarmiento, s. f.)

2.1.8. Análisis de correspondencia múltiple

Pensado como una ampliación natural que se requería para el Análisis de Correspondencias Simples (AC), el nuevo Análisis de Correspondencias Múltiple (AFCM) surge con la finalidad de responder una necesidad concreta y es la de manejar escenarios donde aparecen más de dos variables categóricas, ello mientras que en el AFC que circunscrito al estudio de una única tabla de contingencia entre dos factores y a ello, el AFCM da un salto cualitativo ya que abarca un conjunto más amplio de variables de manera simultánea y es gracias a esta flexibilidad que, la técnica no se limita a contrastar pares, sino que invita a explorar la trama de relaciones que se establecen entre todas las categorías implicadas, de este modo, ofrece una visión más orgánica e integral de los datos (Greenacre, 2008).

Matriz Binaria (Disyuntiva Completa)

La primera representación de los datos es la matriz binaria o disyuntiva completa $Z \in \mathbb{R}^{n \times M}$, donde:

- n : número de individuos (filas)

- $M = \sum_{q=1}^Q J_q$: número total de categorías (columnas), siendo J_q el número de categorías de la variable q

Cada elemento z_{ij} es:

$$z_{ij} = \begin{cases} 1 & \text{si el individuo } i \text{ pertenece a la categoría } j \\ 0 & \text{en caso contrario} \end{cases}$$

Cada fila de Z tiene exactamente un 1 por cada variable (ya que cada individuo elige una categoría por variable).

Matriz de Burt

La matriz de Burt es una generalización de la tabla de contingencia al caso múltiple. Se define como:

$$B = Z^T Z \in \mathbb{R}^{M \times M}$$

- B es una matriz simétrica por bloques.
- El bloque B_{qq} es una matriz diagonal con las frecuencias de las categorías de la variable q .
- El bloque B_{qr} ($q \neq r$) es la tabla de contingencia entre las variables q y r .

$$B = \begin{bmatrix} Z_1^T Z_1 & Z_1^T Z_2 & \cdots & Z_1^T Z_Q \\ Z_2^T Z_1 & Z_2^T Z_2 & \cdots & Z_2^T Z_Q \\ \vdots & \vdots & \ddots & \vdots \\ Z_Q^T Z_1 & Z_Q^T Z_2 & \cdots & Z_Q^T Z_Q \end{bmatrix}$$

La matriz de Burt es la base para una de las definiciones del AFCM.

Definiciones del AFCM

Existen dos enfoques principales para definir el AFCM:

a) AFCM sobre la matriz binaria

Se aplica el AFC al perfil fila de la matriz binaria Z , ponderando filas y columnas por sus masas.

- Masa de los individuos: $r = \frac{1}{n} \mathbf{1}_n$
- Masa de las categorías: $c = \frac{1}{n} Z^T \mathbf{1}_n = \frac{1}{n} \text{diag}(B)$

La matriz de correspondencias es:

$$P = \frac{1}{n} Z$$

Luego se procede como en el AC, calculando la matriz estandarizada y aplicando la DVS.

b) AFCM sobre la matriz de Burt

Se aplica el AFC a la matriz de Burt B , tratándola como una tabla de contingencia donde:

- Las filas y columnas son las categorías de todas las variables. - La masa de cada categoría es $c_j = b_{jj}/n_{++}$, con $n_{++} = nQ$ (ya que cada individuo contribuye a Q celdas).

El AFCM se define como la DVS de:

$$S = D_c^{-1/2} \left(\frac{B}{n_{++}} - cc^T \right) D_c^{-1/2}$$

2.1.9. Procesamiento de Lenguaje Natural (NLP)

El Procesamiento de Lenguaje Natural (NLP, por sus siglas en inglés: Natural Language Processing) es un subcampo interdisciplinario de la inteligencia artificial y la lingüística computacional que se ocupa de desarrollar métodos, algoritmos y herramientas para que los sistemas computacionales puedan analizar, comprender, interpretar y generar lenguaje humano de manera automática (Jurafsky & Martin, 2024).

El NLP se sitúa en la intersección de tres disciplinas: la lingüística (que aporta el conocimiento sobre la estructura del lenguaje), la informática (que proporciona los algoritmos y la capacidad de cómputo) y la estadística (que sustenta los modelos probabilísticos de aprendizaje automático). Esta convergencia permite transformar texto no estructurado en representaciones cuantificables susceptibles de análisis estadístico posterior.

El procesamiento de lenguaje natural sigue una secuencia de etapas que transforman progresivamente el texto crudo en información estructurada:

- Tokenización: Segmentación del texto en unidades significativas (palabras, frases o símbolos), que constituyen los elementos básicos de análisis.
- Normalización: Estandarización del texto mediante la conversión a minúsculas, eliminación de caracteres especiales, corrección de codificación y unificación de formatos.
- Eliminación de palabras vacías (stopwords): Remoción de términos de alta frecuencia y baja carga semántica (artículos, preposiciones, conjunciones) que no aportan información relevante para el análisis.
- Lematización y stemming: Reducción de las palabras a su forma base o raíz, permitiendo agrupar variantes morfológicas bajo un mismo término.
- Extracción de características: Identificación de elementos relevantes como palabras clave, entidades nombradas, frecuencias de términos y patrones sintácticos.

- Clasificación y etiquetado: Asignación automática de categorías o etiquetas a los documentos o segmentos de texto según criterios predefinidos.

Para que el lenguaje natural pueda ser procesado computacionalmente, es necesario convertirlo en representaciones numéricas o vectoriales. Entre las técnicas más utilizadas se encuentran:

- Bolsa de palabras (Bag of Words): Representación de un documento como un vector de frecuencias de términos, ignorando el orden sintáctico.
- TF-IDF (Term Frequency – Inverse Document Frequency): Ponderación de términos que combina la frecuencia local dentro de un documento con la rareza global en un corpus, permitiendo identificar los términos más distintivos de cada texto.
- Embeddings y modelos de lenguaje: Representaciones vectoriales densas que capturan relaciones semánticas entre palabras. (Zúñiga Rojas, 2024).

2.1.10. Inteligencia Artificial Generativa

La Inteligencia Artificial Generativa (IAG) constituye una rama de la inteligencia artificial orientada a la creación, interpretación y transformación de contenido a partir de patrones aprendidos de grandes volúmenes de datos. A diferencia de los sistemas de IA tradicionales, diseñados para tareas específicas de clasificación o predicción, los modelos generativos son capaces de comprender, sintetizar y producir lenguaje natural con un grado de coherencia y contextualización comparable al lenguaje humano (Zúñiga Rojas, 2024).

Dentro de esta categoría se encuentran los Modelos de Lenguaje de Gran Escala (LLM, por sus siglas en inglés: Large Language Models), arquitecturas basadas en redes neuronales profundas de tipo Transformer que se entrenan con corpus textuales de miles de millones de parámetros. Estos modelos operan mediante mecanismos de atención que les permiten capturar dependencias contextuales de largo alcance, lo que les otorga la capacidad de comprender instrucciones complejas, inferir información implícita y generar respuestas estructuradas a partir de texto no organizado (Vaswani et al., 2017).

El Modelo Gemini

Gemini es una familia de modelos de inteligencia artificial multimodal desarrollada por Google DeepMind, presentada inicialmente en diciembre de 2023. A diferencia de modelos exclusivamente textuales, Gemini fue diseñado para procesar de manera nativa múltiples modalidades de entrada texto, imagen, audio, video y código, lo que amplía significativamente

su capacidad de comprensión contextual (Google DeepMind, 2023). Las características principales que distinguen a Gemini son:

- **Arquitectura multimodal nativa:** El modelo procesa simultáneamente texto e información visual, lo que permite analizar documentos PDF con tablas, figuras y texto continuo sin necesidad de conversiones intermedias que generen pérdida de información.
- **Ventana de contexto extendida:** Gemini admite ventanas de contexto de hasta 1 000 000 de tokens, lo que posibilita el procesamiento de documentos extensos como tesis de más de 200 páginas en una sola inferencia, evitando la fragmentación del texto y la pérdida de coherencia entre secciones.
- **Capacidad de razonamiento estructurado:** El modelo puede seguir instrucciones complejas (prompts) que especifican formatos de salida, criterios de clasificación y reglas de inferencia, lo que lo hace apto para tareas de extracción de información con esquema predefinido.
- **Acceso mediante API programática:** Gemini ofrece interfaces de programación (API) que permiten su integración en pipelines automatizados desarrollados en lenguajes como Python, facilitando el procesamiento por lotes de miles de documentos de manera reproducible y escalable.

La incorporación de Gemini como herramienta de extracción responde a la justificación metodológica del estudio (sección 1.3), que plantea la necesidad de superar el paradigma tradicional de los análisis bibliométricos manuales. Las ventajas principales son:

- **Escalabilidad:** El procesamiento automatizado de 8 324 documentos habría sido inviable mediante codificación manual en un tiempo razonable. La API de Gemini permitió el procesamiento por lotes en un periodo acotado.
- **Reducción del sesgo subjetivo:** La codificación manual de variables metodológicas (como el nivel estadístico o el tipo de diseño) está sujeta a la interpretación del investigador. El uso de un modelo de IA con criterios de clasificación predefinidos minimiza la variabilidad Inter codificador y garantiza la aplicación uniforme de las reglas de categorización.
- **Reproducibilidad:** Al estar implementado en scripts de Python con prompts versionados, el proceso de extracción puede ser replicado exactamente por otros investigadores, cumpliendo con los principios de transparencia y verificabilidad de la investigación científica.

- **Profundidad de análisis:** A diferencia de la minería de metadatos superficiales (título, autor, año), Gemini permite acceder al texto completo de la tesis y extraer información que no se encuentra en los campos bibliográficos estándar del repositorio, como el enfoque metodológico declarado, las técnicas estadísticas empleadas o la calidad redaccional del documento.

2.2. Marco conceptual

- **Análisis documental:** Es un método de recopilación de información que implica el análisis sistemático y crítico de documentos escritos, específicamente de tesis de pregrado en este contexto.
- **Enfoque mixto:** Estrategia que combina métodos cuantitativos y cualitativos para enriquecer el análisis y superar las limitaciones de cada enfoque por separado.
- **Análisis Factorial de Correspondencias Múltiples (AFCM):** Es un método estadístico multivariado que facilita el análisis y la representación de las relaciones entre varias variables categóricas. (Sarmiento, s. f.)
- **Calidad de la tesis:** Valoración integral del rigor académico, la solidez metodológica y la pertinencia científica de un trabajo de investigación de pregrado, expresada en una escala numérica de 0 a 20 puntos. En el presente estudio, la calidad se evaluó mediante un modelo de inteligencia artificial aplicado al texto completo de las tesis, considerando criterios como la coherencia argumentativa, la adecuación del diseño metodológico, la consistencia entre objetivos y resultados, y la profundidad del análisis.
- **Coefficiente V de Cramer:** Medida estandarizada de asociación entre variables categóricas, derivada del estadístico Chi-cuadrado de independencia, que cuantifica la intensidad de la relación en una escala normalizada entre 0 (independencia total) y 1 (asociación perfecta). A diferencia del coeficiente de contingencia de Pearson, su valor máximo es independiente de las dimensiones de la tabla de contingencia, lo que permite comparar de manera objetiva la fuerza de asociación entre pares de variables con distinto número de categorías.
- **Enfoque cualitativo:** Es un método que busca entender cómo funcionan las cosas en profundidad, interpretando sus significados, experiencias y contextos, sin usar mediciones numéricas.

- **Enfoque cuantitativo:** Es un enfoque de investigación que utiliza datos cuantitativos y herramientas estadísticas para contrastar hipótesis y analizar las relaciones entre distintas variables.
- **Escuela profesional:** Unidad académica específica dentro de una facultad que imparte una carrera universitaria.
- **Grafo de red (Análisis de redes):** Representación gráfica compuesta por un conjunto de nodos (vértices) y aristas (arcos), donde los nodos representan entidades del sistema en este caso, variables categóricas y las aristas representan las relaciones o asociaciones significativas entre ellas. El análisis de redes permite identificar la topología del sistema, detectar nodos centrales o articuladores (vértices con mayor grado de conectividad), reconocer subgrafos aislados y visualizar agrupamientos naturales de variables que comparten vínculos intensos.
- **Inercia:** En el análisis de correspondencias, es la medida de dispersión total de los perfiles (filas o columnas) respecto al perfil medio. Se calcula como el estadístico Chi-cuadrado dividido entre el tamaño muestral (χ^2/N) y se descompone en inercias parciales asociadas a cada eje factorial. La inercia explicada por un eje equivale al cuadrado del valor singular correspondiente (λ_k^2), y su porcentaje respecto a la inercia total indica la proporción de estructura de asociación capturada por dicha dimensión (Greenacre, 2008; Cuadras, 2014).
- **Inteligencia Artificial generativa:** Rama de la inteligencia artificial que emplea modelos de lenguaje de gran escala (LLM) capaces de comprender, interpretar y generar texto natural a partir de instrucciones estructuradas (prompts). En el contexto de la presente investigación, se utilizó el modelo Gemini como motor de extracción estructurada, capaz de identificar, inferir y registrar variables metodológicas, temáticas y de calidad definidas en la operacionalización (enfoque de investigación, tipo de diseño, nivel estadístico, riqueza gramatical, calidad, entre otras) a partir del título, resumen y texto completo de cada tesis. Esta aplicación constituye un componente innovador del pipeline metodológico, al sustituir la codificación manual por un proceso automatizado, transparente y reproducible, reduciendo significativamente el sesgo subjetivo del investigador y el tiempo de procesamiento. (Zúñiga Rojas, 2024)
- **Mapa factorial:** Representación gráfica bidimensional (o tridimensional) generada por el AFC o AFCM, donde cada categoría se proyecta mediante sus coordenadas principales o estándar. La proximidad entre dos puntos indica asociación o similitud de

perfiles, mientras que la distancia al origen refleja la contribución a la inercia. Se distinguen el mapa simétrico (filas y columnas en coordenadas principales) y el mapa asimétrico (un conjunto en coordenadas principales y el otro en estándar), cada uno con una interpretación diferente de las distancias (Greenacre, 2008).

- **Metodología de investigación:** Enfoque sistemático utilizado para recoger y analizar datos. Puede ser cuantitativa, cualitativa o mixta, y define la estrategia científica del estudio.
- **Minería de texto:** Conjunto de técnicas computacionales orientadas a la extracción automática de información, patrones y conocimiento a partir de grandes volúmenes de texto no estructurado. A diferencia del análisis documental tradicional, la minería de texto permite procesar simultáneamente miles de documentos mediante algoritmos de tokenización, eliminación de palabras vacías, normalización, análisis de frecuencia y clasificación automática. (Zúñiga Rojas, 2024)
- **Nivel de inferencia:** Grado de profundidad analítica y capacidad interpretativa que alcanza una investigación, clasificado en tres categorías ordinales: Bajo (estudios que se limitan a describir o tabular datos), Medio (estudios que establecen relaciones, comparaciones o asociaciones entre variables) y Alto (estudios que formulan explicaciones causales, modelos predictivos o generalizaciones teóricas).
- **OAI-PMH (Open Archives Initiative – Protocol for Metadata Harvesting):** Es un protocolo de interoperabilidad que estandariza la cosecha automatizada de metadatos desde repositorios digitales de acceso abierto. Mediante sus seis verbos (Identify, ListMetadataFormats, ListSets, ListIdentifiers, ListRecords y GetRecord) permite recuperar de forma estructurada los registros bibliográficos de las tesis (título, autor, fecha, resumen, facultad, palabras clave e identificadores), garantizando transparencia y reproducibilidad en la recolección de datos.
- **Patrones:** Configuraciones recurrentes de asociación entre variables categóricas que se manifiestan de forma consistente en un conjunto de datos. En el ámbito de la producción académica, representan perfiles tipológicos de tesis que comparten características metodológicas, temáticas o estructurales, identificables mediante técnicas multivariadas como el AFCM (Greenacre, 2008).
- **Procesamiento de Lenguaje Natural (NLP):** Es un subcampo de la inteligencia artificial y de la lingüística computacional que desarrolla métodos y herramientas para que los sistemas computacionales analicen, comprendan y transformen el lenguaje

humano. En el estudio agrupa técnicas como la tokenización, normalización, extracción de palabras clave y clasificación de textos, que permiten convertir el contenido no estructurado de las tesis (título, resumen y texto completo) en variables categóricas susceptibles de análisis estadístico.

- **Prueba Chi-cuadrado (χ^2):** Prueba estadística no paramétrica que evalúa si existe asociación significativa entre dos variables categóricas a partir de una tabla de contingencia. Compara las frecuencias observadas con las frecuencias esperadas bajo el supuesto de independencia (H_0). El estadístico se distribuye aproximadamente como χ^2 con $(r-1)(c-1)$ grados de libertad. En el presente estudio se aplica con un nivel de significancia $\alpha = 0,01$ (Cuadras, 2014; Greenacre, 2008)
- **Repositorio institucional:** Plataforma digital donde se almacenan, organizan y difunden los trabajos académicos de una universidad, como las tesis de pregrado y posgrado. En la UNSAAC opera bajo el protocolo OAI-PMH y que contiene 8,324 tesis de pregrado (2017-2025), vinculándolo con la población del estudio.
- **Riqueza gramatical:** Indicador de calidad lingüística que evalúa el nivel de sofisticación, variedad y corrección del lenguaje empleado en la redacción de un texto académico. Se clasifica en tres niveles ordinales Alta, Media y Baja a partir del análisis automatizado de la diversidad léxica, la complejidad sintáctica, la correcta aplicación de normas ortográficas y la coherencia textual.
- **Tabla de contingencia:** Matriz bidimensional de frecuencias que cruza dos variables categóricas, donde cada celda (n_{ij}) registra el número de observaciones que pertenecen simultáneamente a la categoría i de la primera variable y a la categoría j de la segunda. A partir de esta matriz se derivan los totales marginales por fila y por columna, las frecuencias esperadas bajo el supuesto de independencia, los perfiles fila y columna (distribuciones condicionales de frecuencias relativas) y el estadístico Chi-cuadrado.
- **Tesis de pregrado:** Es un estudio científico original desarrollado por estudiantes universitarios como requisito para obtener un grado académico o título profesional, en el que se emplean métodos sistemáticos para examinar un problema y producir conocimiento nuevo o aplicado.
- **Tipo de colaboración:** Modalidad de elaboración del trabajo de investigación, clasificada como individual o grupal (con dos o más autores).

- **Variables categóricas:** Características o atributos que se clasifican en categorías no numéricas, como género, metodología, área temática o tipo de colaboración. Son el material principal para el AFCM. (Sarmiento, s. f.)

2.3. Antecedentes empíricos de la investigación

2.3.1. Antecedentes internacionales

Simahate et al. (2022) en el artículo que titula **Bibliometric Analysis of Social Sciences** planteó como hilo conductor la necesidad de tener comprensión sola forma en cómo ha evolucionado la producción científica en las ciencias sociales a través del tiempo y para dar respuesta a ello, los autores recurrieron a realizar un análisis de la bibliografía disponible eligiendo principalmente trabajos indexados dentro de la base de datos del gestor Scopus y de este modo, la investigación no solo se dedica a escrutar pautas y frecuencias de publicación, sino también centró su búsqueda en autores e instituciones de mayor peso académico, esto sin dejar de lado el papel que tienen los diferentes países a la hora de marcar una pauta en esta disciplina.

Los resultados señalan un crecimiento sostenido de la producción en ciencias sociales durante los últimos años, siendo Estados Unidos el país con mayor volumen de publicaciones, seguido de varias naciones europeas y asiáticas. Se evidencia que las instituciones con más alto nivel de influencia se encuentran principalmente en países desarrollados, lo que refleja una concentración del conocimiento científico en determinadas regiones. En cuanto a las áreas temáticas, se observa un predominio de estudios vinculados a la educación, el desarrollo sostenible y las políticas sociales, los cuales han despertado gran interés en la comunidad académica internacional.

El estudio también resalta la consolidación de una red de colaboración internacional entre investigadores, aunque persisten notorias desigualdades geográficas en la producción de conocimiento, especialmente en comparación con regiones menos representadas. En conclusión, esta investigación demuestra que el análisis bibliométrico es una herramienta útil para identificar patrones de investigación y orientar políticas que promuevan una mayor equidad en la producción científica a nivel global.

Çakır & Cönk (2025) en su artículo de investigación titulado **A Bibliometric Analysis of Theses on Marketing Communication in the National Thesis Center of Turkey**, elaborado por Elif Akben, tuvo como finalidad analizar de manera bibliométrica las tesis que incluyen el concepto comunicación de marketing en su título y que se encuentran registradas

en el Centro Nacional de Tesis de Turquía CoHE (Consejo de Educación Superior de Turquía). La investigación se centró en identificar la evolución del campo y las principales tendencias de estudio. Los datos se recopilaron a partir de las páginas de índice de las tesis para su análisis de contenido y estadística descriptiva. Los resultados muestran que los trabajos académicos sobre comunicación de marketing en Turquía se han incrementado con el tiempo y que muchos de ellos abordan la influencia de las nuevas tecnologías y de los medios digitales en el ámbito de la comunicación. También se evidencian vacíos de investigación en áreas emergentes que aún no reciben suficiente atención académica. El estudio concluye que los hallazgos pueden servir de guía para futuros investigadores, ya que ayudan a orientar nuevas líneas de investigación y contribuyen a mejorar la producción académica en este campo. Aporta insumos valiosos para que los programas de posgrado diseñen propuestas de investigación que respondan a las necesidades actuales del entorno social y empresarial. Este antecedente pone de relieve cómo un análisis bibliométrico es una herramienta útil para mapear el desarrollo de un área de estudio, detectar tendencias, señalar vacíos y orientar investigaciones futuras, en especial en disciplinas que evolucionan con rapidez como la comunicación de marketing.

Viloria et al. (2019) en el estudio, desarrollado en el marco del **International Workshop on Applying Data Mining Techniques to E-Learning and Pedagogical Approaches**, la finalidad de los autores fue identificar patrones de conocimiento, ello a través de la revisión de los trabajos de investigación llevados a cabo por los estudiantes de la Facultad de Ciencias de la Universidad de Mumbai y para lograr ello, recurrieron a comparaciones del desempeño de cinco algoritmos de clasificación, estos fueron, Máquinas de Vectores de Soporte, Redes Neuronales, Árboles de Decisión, Bosques Aleatorios y métodos de potenciación, ello considerando las áreas de Diseño Experimental y Análisis Multivariado; posterior al análisis, los hallazgos mostraron que, en la temática de Diseño Experimental, el algoritmo de Bosques Aleatorios pudo lograr el mejor desempeño al alcanzar una precisión de 71.48%; en cambio, para Análisis Multivariado, todos los modelos registraron una precisión próxima al 97%, sin presentar diferencias importantes entre ellos; asimismo, las Redes Neuronales mostraron una capacidad más elevada para identificar correctamente los casos positivos, mientras que los Bosques Aleatorios mostraron mayor eficacia para reconocer los casos negativos; se llega a la conclusión de que la aplicación de técnicas de minería de datos hace más fácil el descubrimiento de patrones y tendencias en las investigaciones académicas, lo que genera contribución a la mejora en la toma de decisiones en el ámbito académico; además, destacaron el uso del software Rattle, esto principalmente porque es una herramienta sencilla y eficiente para implementar estos modelos.

Sousa et al. (2021) en el artículo titulado **NewResearch and Trends in Higher Education** realizaron un estudio con el objetivo de identificar y analizar las nuevas tendencias y líneas de investigación en la educación superior, abordando innovaciones metodológicas, tecnológicas y contextuales, así como su impacto en los procesos de enseñanza-aprendizaje. La metodología consistió en un análisis de 26 artículos incluidos en un número especial de Education Sciences, abarcando investigaciones empíricas, estudios de caso y desarrollos conceptuales. Se aplicaron técnicas bibliométricas y de análisis de redes mediante el software VOSviewer para mapear asociaciones entre autores, palabras clave y atributos de investigación presentes en títulos y resúmenes. Se analizaron redes de coautoría, clústeres temáticos y vínculos de palabras clave para identificar patrones de colaboración y tendencias emergentes. Los resultados revelaron que las principales áreas de innovación giran en torno a: Uso de tecnologías emergentes (realidad virtual, aprendizaje móvil, automatización), metodologías centradas en el estudiante, desarrollo de competencias transversales, analítica del aprendizaje, adaptación de la enseñanza ante el COVID-19 y fortalecimiento de las habilidades blandas. El análisis de coautoría mostró baja colaboración internacional y grupos de trabajo pequeños y poco estables. En cuanto a palabras clave, educación superior fue el término central, aunque también destacaron conceptos como realidad virtual, pensamiento creativo, evaluación, inteligencia emocional y comunicación efectiva. Los clústeres temáticos identificaron líneas de trabajo sobre clima motivacional, integración de tareas, calidad del aprendizaje, actitudes hacia la educación superior, y prácticas docentes en contextos de pandemia. Los autores concluyeron que, aunque existen avances significativos en la incorporación de tecnologías y metodologías innovadoras, persisten desafíos en la colaboración académica y en el desarrollo consolidado de ciertas áreas, lo que abre oportunidades para investigaciones más integradas y cooperativas a nivel global.

Chen et al. (2016) llevaron a cabo un estudio **Mapping the research trends by co-word analysis based on keywords from funded project** con la principal finalidad de reconocer las principales líneas de estudio dentro del campo de las Ciencias de la Gestión en Ingeniería en el país de China y para ello, los autores recurrieron al análisis de los proyectos que fueron financiados por la Fundación Nacional de Ciencias Naturales de China durante el periodo comprendido entre los años 2011 y 2015 y a partir de esta información, se centraron en identificar los temas con mayor presencia en la investigación científica y con ello evaluaron el nivel de desarrollo y consolidación por cada uno de ellos. Se basó en el análisis de co-palabras, aplicando técnicas de clúster y análisis de redes sociales a las palabras clave extraídas de 2,054 proyectos (con un total de 6,153 términos y 9,348 ocurrencias). Se seleccionaron los

100 términos de mayor frecuencia (≥ 8 menciones) y se construyó una matriz de co-ocurrencia, procesada mediante herramientas como Ucinet y VOSviewer para la visualización de redes, densidad y estructuras núcleo-periferia (k-core). Los resultados mostraron que las áreas de investigación más frecuentes y conectadas fueron teoría de juegos, gestión de la cadena de suministro, minería de datos, redes complejas, optimización, gestión de riesgos y análisis envolvente de datos. El análisis de clúster identificó ocho grupos temáticos, destacando el de redes complejas y big data como el más centralizado. La visualización de densidad reveló que, aunque existen temas bien desarrollados, la mayoría de los campos aún se encuentran en etapas tempranas de madurez. El análisis k-core confirmó que teoría de juegos, gestión de la cadena de suministro y minería de datos son los núcleos temáticos principales. Los autores concluyeron que, pese a la diversidad de líneas de investigación, son pocas las que alcanzan un nivel de desarrollo consolidado, lo que plantea la necesidad de ampliar y profundizar en campos emergentes dentro de la disciplina.

Debushe et al. (2025) desarrollaron un estudio denominado **Evaluation of statistical methods applied in theses and dissertations in an Open, Distance and e-Learning University** cuyo objetivo fue identificar y analizar las tendencias de investigación en el campo de la ingeniería de software a través de un mapeo sistemático de la literatura. La metodología consistió en la aplicación de un mapeo sistemático que incluyó la búsqueda, selección y categorización de artículos científicos publicados entre 2008 y 2018 para clasificar los estudios según sus temas, enfoques metodológicos y áreas de aplicación. Los resultados mostraron que las principales áreas de interés se centraron en el desarrollo ágil, la ingeniería de requisitos, la calidad del software y la gestión de proyectos.

La técnica estadística más empleada fue la estadística descriptiva (80,6%), seguida del ANOVA (32,4%), el análisis de correlación (30,9%), el análisis de regresión (24,5%) y la prueba de chi-cuadrado (21,6%). Se observó un aumento gradual en las publicaciones vinculadas a la automatización de pruebas y al uso de la inteligencia artificial en el desarrollo de software. El análisis mostró que la mayoría de las investigaciones se sustentaron en diseños empíricos, principalmente estudios de caso y encuestas, mientras que los experimentos controlados fueron menos comunes. Se concluye que las tendencias actuales apuntan a una creciente integración de enfoques interdisciplinarios y al uso de herramientas automatizadas, lo que plantea nuevas oportunidades para futuras investigaciones en este campo.

2.3.2. Antecedentes nacionales

Cárdenas Ortega & Gutierrez Blanco (2024) en el trabajo de tesis con título **Tesis de Comunicación en las universidades públicas licenciadas del Perú (2015-2022). Un estudio sobre las tesis de pregrado** Autores: Kelly Cárdenas Ortega y Natalia Alessandra Gutiérrez Blanco. La finalidad principal de los autores fue realizar la descripción de las principales características de las investigaciones que se realizaron para obtener el grado de licenciados en universidades públicas dentro del territorio peruano entre los años 2015 y 2022 y para ello, consideraron seguir un enfoque cuantitativo, un alcance descriptivo y basaron su investigación en un diseño no experimental de tipo transversal. La población estuvo constituida por 1173 tesis registradas en los repositorios institucionales de 19 universidades públicas. Para la recopilación de información se utilizó una ficha de observación elaborada a partir de diversos aportes teóricos. El mayor número de tesis se produjo en el año 2019. Respecto a la autoría, el 55,9% correspondió a trabajos individuales y el 40,8% a trabajos en pareja. El área de investigación más predominante fue la comunicación organizacional. En términos metodológicos, destacó el enfoque cuantitativo (44,2%), y el 65,3% de los estudios se ubicaron en un nivel descriptivo, siendo los sujetos de estudio la principal unidad de análisis (79,8%). Las autoras concluyen que, tras la promulgación de la Ley Universitaria N.º 30220, se evidenciaron cambios relevantes en la elaboración de tesis, especialmente el incremento de investigaciones desarrolladas por dos o tres autores.

Caceres Luna (2021) en el trabajo Titulado **Calidad de las tesis de pregrado y rendimiento académico en estudiantes de Psicología** autora: Gabriela Ladid Cáceres Luna; la finalidad de los autores fue determinar si existía una asociación entre la calidad de las tesis de pregrado en Psicología y el rendimiento académico alcanzado por los estudiantes en los cursos de Metodología de la Investigación Científica y Seminario de Tesis y para lo cual, efectuaron un estudio de tipo asociativo, con enfoque predictivo y diseño transversal, tomando como muestra 64 tesis sustentadas entre 2013 y 2019 en la Escuela Profesional de Psicología de la Universidad Católica San Pablo; además, la información fue obtenida a través de la aplicación de una ficha de evaluación que fue diseñada y validada específicamente para la presente investigación, Los resultados reflejaron un vínculo positivo y significativo entre la calidad de las tesis y las calificaciones obtenidas en los cursos metodológicos, evidenciando además que el curso de Metodología de la Investigación Científica constituye un factor importante para predecir la calidad de los trabajos de titulación. Se observó que la asesoría de docentes con perfil investigador influye favorablemente, al asociarse con tesis de mayor calidad

y mejores evaluaciones. Se identificó una mejora progresiva en la calidad de las tesis a lo largo del tiempo, y que aquellas desarrolladas en coautoría (dos autores) presentan niveles superiores de calidad en comparación con las elaboradas de manera individual.

Arce Rivera (2023) en el trabajo de tesis cuyo título **Tendencias temáticas de las tesis de maestría de docencia universitaria de una universidad peruana, periodo 2015-2022** autor Franz Rosendo Arce Rivera. Con el propósito identificar las principales tendencias temáticas abordadas en las tesis de Maestría en Docencia Universitaria, planteando la pregunta central: ¿qué investigan los investigadores en este campo? La investigación adoptó un enfoque cuantitativo, utilizando muestreo probabilístico, distribución de frecuencias y análisis de correspondencia. La muestra estuvo conformada por 342 tesis seleccionadas aleatoriamente del repositorio institucional de una universidad peruana, comprendiendo el periodo 2015 de octubre de 2022. Los resultados revelaron que los temas más recurrentes fueron: la aplicación de enfoques psicológicos (23%), el uso de TICs y recursos tecnológicos (21%), el aprendizaje del estudiante (11%), el rendimiento académico (8%) y el desempeño docente (7%). Un 22% de las investigaciones se orientó hacia otras áreas relevantes, como programas o talleres específicos, estrategias y modelos de enseñanza, métodos de evaluación, estilos y hábitos de aprendizaje. La investigación concluye que estas tendencias temáticas permiten trazar un panorama de los principales intereses investigativos en el área de docencia universitaria y constituyen una base sólida para futuros estudios vinculados al posgrado en este campo.

Pinto Ariza (2023) en el trabajo de tesis con el Título **Revisión sistemática de las tesis de Ingeniería en Higiene y Seguridad Industrial en una Universidad Peruana, 2000-2022** con el Autor: Pablo José Pinto Ariza; El objetivo principal de este estudio fue determinar los ejes temáticos y la evolución de la investigación en higiene y seguridad industrial en una universidad peruana entre 2000 y 2022. Mediante un enfoque bibliométrico, se examinaron 50 tesis de maestría, sistematizando variables como autoría, año, temática, género y metodología. Los hallazgos revelaron una predominancia del diseño de sistemas de gestión de seguridad industrial (48%) y la prevención de riesgos laborales (18%), acompañada de un crecimiento sostenido en la producción académica. El estudio concluye que es imperativo diversificar las metodologías, fomentando el uso de enfoques cualitativos y mixtos para robustecer el campo investigativo.

Soncco Chullunquia (2017) presentan el trabajo con título **Factores que influyen en la elaboración de la tesis para optar el título profesional de Cirujano Dentista en los estudiantes de internado hospitalario y taller de tesis de la Universidad Alas Peruanas, Lima, 2017** autor: Bach. Jorge Luis Soncco Chullunquia La presente investigación tuvo como

propósito determinar las variables que influyen en el proceso de elaboración de la tesis para optar el título de cirujano dentista en alumnos de la Universidad Alas Peruanas, filial Lima, durante el año 2017. El trabajo se ejecutó bajo un diseño observacional, con un enfoque prospectivo, de corte transversal y nivel descriptivo, abarcando una muestra de 104 educandos del décimo ciclo, matriculados en el periodo 2017-I y egresados del taller de tesis. Para la obtención de la información se aplicó un cuestionario cerrado, sometido a validación por juicio de expertos y estructurado en 25 ítems de respuesta dicotómica. Los hallazgos demostraron que la población estuvo mayoritariamente representada por mujeres (70,2%) y por jóvenes de 20 a 25 años (51%). Respecto a las variables de índole personal, estas mostraron una tendencia ampliamente favorable, resaltando una actitud positiva (73,1%), un alto nivel de motivación intrínseca (64,4%) y una adecuada disponibilidad de tiempo (84,6%). más de la mitad de los estudiantes (54,8%) no contaba con recursos económicos suficientes para financiar su investigación. Por otro lado, los factores institucionales fueron considerados mayormente desfavorables (57,7%), debido a limitaciones en tres de sus cuatro dimensiones, a excepción de la disponibilidad de asesores de tesis. En conclusión, se determinó que los factores institucionales constituyen la principal dificultad en el desarrollo de la tesis, mientras que los factores personales, en su mayoría positivos, no representan un obstáculo significativo.

2.3.3. Antecedentes locales

Catunta Guillen (2024) presentan la investigación con el título **Aplicaciones de los métodos de análisis de clúster y correspondencia en el estudio de resultados de examen de admisión de la UNSAAC, 2022** La finalidad de los autores fue analizar los resultados logrados en los procesos de admisión o ingreso a la UNSAAC durante el año 2022, esto a través del uso de diferentes técnicas de análisis estadístico y para ello, la investigación se realizó con un enfoque cuantitativo, de alcance descriptivo y bajo un diseño no experimental de corte transversal; asimismo, el estudio consideró a los postulantes egresados de la educación básica regular que participaron en los procesos de admisión de ese año; además, para clasificar a los participantes en grupos con características similares, fue necesario aplicar tres algoritmos de agrupamiento: PAM (Partitioning Around Medoids), CLARA (Clustering Large Applications) y el método bietápico. La matriz de distancias se construyó utilizando el coeficiente de Gower para datos mixtos. Tras la comparación de los algoritmos, se determinó que el método bietápico presentaba un mejor desempeño en términos de cohesión y separación, medido a través del índice de silueta, identificándose cuatro clústeres como la solución más adecuada para describir los perfiles de los postulantes. Finalmente, los resultados permitieron caracterizar dichos

clústeres mediante el análisis de correspondencias, evidenciando diferencias en los perfiles obtenidos:

Clúster 1: Estudiantes que no lograron ingresar, en su mayoría mujeres del Cusco, procedentes de colegios nacionales, con promedio de edad de 18.49 años y nota promedio de 7.44, postulantes principalmente al grupo D.

Clúster 2: Postulantes que sí obtuvieron vacante, mayoritariamente hombres del Cusco, egresados de colegios nacionales, con edad promedio de 20.7 años y nota promedio de 12.58.

Clúster 3: Jóvenes que no ingresaron, en su mayoría hombres del Cusco, provenientes de colegios particulares, con promedio de edad de 18.49 años, nota de 7.67, postulantes al grupo A.

Clúster 4: Postulantes no ingresantes, en su mayoría mujeres del Cusco, egresadas de colegios particulares, con edad promedio de 18.50 años y nota promedio de 7.75, postulantes principalmente al grupo A.

La investigación evidenció la utilidad del análisis de clúster para identificar perfiles y análisis de correspondencia para la caracterización de los clústeres de ingreso y no ingreso en la UNSAAC, aportando información valiosa para comprender las características académicas y sociodemográficas de los postulantes.

Zuñiga Rojas (2024) en el trabajo de tesis con título: **Modelo de extracción automática de glosario de términos utilizando técnicas de procesamiento de lenguaje natural y clustering** del autor: Br. Gabriela Zuñiga Rojas. El propósito de este estudio fue desarrollar un método automatizado para extraer y clasificar términos de glosario en proyectos de software de gran envergadura, simplificando así una actividad que normalmente requiere un complejo trabajo manual. El enfoque planteado combina técnicas de preprocesamiento y reglas heurísticas para identificar los términos relevantes, junto con el uso de embeddings generados con FastText, que permiten medir la similitud semántica entre los términos. En el proceso de agrupamiento se aplicaron distintos algoritmos: K-means, Expectation Maximization (EM) y clusterización jerárquica. La técnica fue probada sobre un conjunto de 2966 requisitos funcionales, logrando formar 318 grupos semánticos. Para evaluar la eficacia del modelo, se empleó la métrica Word Mover's Distance (WMD), obteniendo un valor de 0.0113, lo cual evidenció que los resultados automáticos presentan menor distancia respecto a los agrupamientos manuales. Los experimentos mostraron que la combinación de FastText con el algoritmo EM permite una agrupación semántica más precisa y consistente, lo que valida la viabilidad del modelo en contextos reales de documentación y desarrollo de software. Constituye un aporte al campo del procesamiento de lenguaje natural aplicado a ingeniería de

software, al demostrar cómo el uso de técnicas automatizadas puede optimizar la generación de glosarios y mejorar la gestión de requisitos en proyectos complejos.

Guzman Mamani & Huamani Muñoz (2024) en el trabajo de tesis cuyo título **Tildación en la redacción de las tesis de licenciatura en Educación: Estudio realizado desde la Universidad Nacional de San Antonio Abad del Cusco, periodo 2019-2022** con autores: Br. Ronald Guzmán Mamani y Br. Katerine Esther Huamani Muñoz. Con el objetivo de determinar el nivel de cumplimiento de las reglas ortográficas de tildación en las tesis de licenciatura en Educación de la UNSAAC, se desarrolló un estudio de diseño no experimental, transversal y básico. La muestra estuvo compuesta por 24 tesis sustentadas entre 2019 y 2022, seleccionadas de manera no probabilística de las especialidades de Lengua y Literatura, Educación Primaria, Matemática y Física, Ciencias Naturales, Educación Física y Ciencias Sociales. Para la recolección de información se empleó la técnica de análisis documental, apoyándose en una ficha de registro diseñada ad hoc. Los resultados evidenciaron que los errores de tildación se atribuyen tanto a los tesisistas como al personal encargado de la revisión, debido a la falta de un control riguroso en este aspecto. Ninguna especialidad estuvo exenta de errores; sin embargo, Matemática y Física, junto con Lengua y Literatura, presentaron menor cantidad de fallas. Se concluye que existe un alto nivel de incumplimiento de las normas generales de tildación, siendo este el tipo de error más recurrente, mientras que las reglas de tildación diacrítica, enfática y diacrítica mostraron menor incidencia de errores.

Enriquez Apaza (2020) realizó el estudio bibliométrico titulado **Características bibliométricas de las tesis de pregrado de la Escuela Profesional de Medicina Humana registradas en el Repositorio Institucional Digital de la Universidad Nacional de San Antonio Abad del Cusco (UNSAAC), 2000–2022**. La finalidad fue analizar las características de producción científica de las tesis de pregrado, considerando variables como año de registro, diseño metodológico, población de estudio, lugar de ejecución, especialidad médica, área de investigación, vinculación a líneas de investigación institucionales y publicación en revistas científicas.

El estudio tuvo un enfoque observacional, descriptivo, transversal y retrospectivo, revisando 187 tesis registradas en el repositorio digital de la UNSAAC. Entre los hallazgos principales, se identificó que el año 2019 concentró el mayor número de tesis (16 %), el 90 % contó con un solo autor, y el diseño transversal fue el más frecuente (56,2 %). La población más estudiada fue la de adultos y adultos mayores (29,9 %), el entorno hospitalario fue el lugar preferido para la ejecución (85,1 %), y la especialidad más abordada fue Ginecología y Obstetricia (25,3 %). El área clínico-quirúrgica fue la más investigada (82 %).

Las líneas de investigación, solo el 2,7 % de las tesis se enmarcó en una línea de investigación nacional y el 2,1 % en una regional. únicamente 3 tesis (1,6 %) fueron publicadas en revistas científicas indexadas. Se concluyen que existe una baja articulación de las tesis con las líneas de investigación institucionales y una escasa producción científica publicada.

Cardeña Quispe (2021)Cardeña Quispe, Andre (2021) en la investigación titulada **Factores que limitan la realización de tesis en internos y egresados de la Escuela Profesional de Odontología de la UNSAAC en el periodo 2019-2020**, el estudio tuvo como objetivo identificar las principales dificultades que enfrentan los estudiantes en la elaboración de sus trabajos de investigación. Los autores realizaron la investigación considerando un alcance descriptivo, empleando un diseño no experimental y trabajando con una muestra censal de 54 participantes; además, la información se recolectó a través de la aplicación de un instrumento conformado por 24 interrogantes, estructuradas de acuerdo a dos grandes grupos de factores: Personales e institucionales; además, estos incluyeron aspectos relacionados con la motivación e interés, la disponibilidad de tiempo, la situación económica, la experiencia en investigación, el apoyo brindado por la institución, la asesoría recibida y el plan de estudios; finalmente, posterior al análisis, los hallazgos revelaron que la gran parte de los participantes percibía que ambos tipos de factores generan efectos en el desarrollo de la tesis; específicamente, el 92,6 % indicó que los factores personales representan la principal dificultad, mientras que el 85,2 % consideró que las condiciones institucionales también influyen de manera significativa en este proceso. Dentro de los factores personales, la falta de motivación e interés (90.7%) y la escasa experiencia en investigación (66.7%) fueron los más relevantes, además de las limitaciones económicas. En cuanto a los factores institucionales, el plan de estudios fue señalado por el 94.6% y la falta de apoyo adecuado de los asesores por el 75.9% como los mayores problemas. Se concluye que la realización de tesis en los estudiantes de odontología se ve restringida tanto por factores individuales como la motivación, el interés y la experiencia investigativa como por factores institucionales, principalmente el diseño del plan de estudios y la labor de los asesores.

Hidalgo León (2022) en la investigación titulada **Relevancia social y científica de las tesis de pregrado a través de una propuesta metodológica en las escuelas de ingeniería de sistemas e informática en la región sur de Perú** autora Mg. Pilar Vanessa Hidalgo León, se planteó como propósito analizar hasta qué punto los trabajos de investigación cumplen con criterios de relevancia académica y social. Basándose en la propuesta de Karl Popper, se consideró que la relevancia científica está ligada a la creatividad, la posibilidad de reproducir, compartir y debatir los resultados, así como a indicadores bibliométricos reconocidos

internacionalmente (versiones, citas, líneas de investigación). La relevancia social se asocia con el impacto positivo que los estudios pueden generar en la comunidad, respondiendo a necesidades de forma sostenible y con evidencia verificable. El estudio se aplicó a tesis de pregrado de las Escuelas Profesionales de Ingeniería de Sistemas e Informática en universidades del sur del Perú, las tesis de pregrado en este campo no alcanzan el nivel mínimo exigido en los indicadores de relevancia científica y social planteados. Esto refleja limitaciones tanto en la producción académica como en su capacidad de generar aportes significativos a la sociedad.

Zúñiga Huamán (2019) en la investigación titulada **Calidad de tesis de los egresados de la Escuela Profesional de Medicina Humana de la Universidad Andina del Cusco**, La investigación, desarrollada por el bachiller Karen Evelyn Zúñiga Huamán, tuvo como objetivo principal evaluar la calidad de las tesis elaboradas por los egresados de la Escuela Profesional de Medicina Humana. Se trató de un estudio observacional, descriptivo, retrospectivo y de corte transversal, que consideró como muestra un total de 61 tesis sustentadas. Para el análisis se empleó una ficha de recopilación de datos validada institucionalmente. Los resultados mostraron que el 83,6% de las tesis alcanzaron un nivel de calidad aceptable, mientras que el 16,4% fueron clasificadas como no aceptables, sin registrarse trabajos con nivel de excelencia. El índice de similitud presentó un promedio de 35,6%, con valores que oscilaron entre 20% y 63%. Asimismo, se observó que 13 tesis no fueron incorporadas en el repositorio institucional. En cuanto al tipo de investigación, predominó el enfoque descriptivo (57,4%), seguido de estudios analíticos (37,7%), experimentales (3,3%) y correlacionales (1,6%). Por áreas temáticas, las tesis se concentraron principalmente en Medicina (37,7%), Ginecología (27,9%), Pediatría (18%), Cirugía (11,5%) y Salud Pública (4,9%). Se concluye que, si bien la mayoría de los trabajos presenta un nivel aceptable de calidad, ninguno alcanza estándares de excelencia; además, se evidencia una predominancia de diseños descriptivos y la presencia de niveles considerables de similitud en varios trabajos.

CAPÍTULO III

HIPÓTESIS Y VARIABLES

3.1. Hipótesis

a. Hipótesis general

- **H₀ (Hipótesis nula):** No existe asociación estadísticamente significativa entre las variables categóricas de identidad académica, contenido temático, enfoque metodológico, herramientas analíticas, características estructurales, calidad, producción académica y gestión académica que caracterizan las tesis de pregrado de la Universidad Nacional de San Antonio Abad del Cusco (UNSAAC), por lo que no se identifican patrones estructurales relevantes.
- **H₁ (Hipótesis alterna):** Existe asociación estadísticamente significativa entre las variables categóricas de identidad académica, contenido temático, enfoque metodológico, herramientas analíticas, características estructurales, calidad, producción académica y gestión académica que caracterizan las tesis de pregrado de la Universidad Nacional de San Antonio Abad del Cusco (UNSAAC), lo que permite identificar patrones estructurales mediante el Análisis Factorial de Correspondencias Múltiples (AFCM).

b. Hipótesis específicas

Hipótesis específica 1

- La distribución de las tesis de pregrado de la UNSAAC presenta predominancia en determinadas categorías, tales como enfoque metodológico, tipo de diseño y modalidad de autoría.

Hipótesis específica 2

- **H₀:** No existe asociación estadísticamente significativa entre las variables categóricas de identidad académica, contenido temático, enfoque metodológico, herramientas analíticas, características estructurales, calidad, producción académica y gestión académica de las tesis de pregrado de la UNSAAC, evaluada mediante la prueba de Chi-cuadrado ($p \geq 0,01$).
- **H₁:** Existe asociación estadísticamente significativa entre las variables categóricas de identidad académica, contenido temático, enfoque metodológico, herramientas analíticas, características estructurales, calidad, producción académica y gestión

académica de las tesis de pregrado de la UNSAAC, evaluada mediante la prueba de Chi-cuadrado ($p < 0,01$).

Hipótesis específica 3

- **H₀:** Las dimensiones factoriales obtenidas mediante el Análisis Factorial de Correspondencias Múltiples (AFCM) no explican una proporción significativa de la variabilidad de las variables categóricas de identidad académica, contenido temático, enfoque metodológico, herramientas analíticas, características estructurales, calidad, producción académica y gestión académica de las tesis de pregrado de la UNSAAC.
- **H₁:** Las dimensiones factoriales obtenidas mediante el Análisis Factorial de Correspondencias Múltiples (AFCM) explican una proporción significativa de la variabilidad de las variables categóricas de identidad académica, contenido temático, enfoque metodológico, herramientas analíticas, características estructurales, calidad, producción académica y gestión académica de las tesis de pregrado de la UNSAAC, permitiendo interpretar la estructura subyacente de los datos.

Hipótesis específica 4

- **H₀:** No se identifican patrones estructurales ni tipologías diferenciadas en las tesis de pregrado de la UNSAAC a partir de la interpretación del espacio factorial generado por el Análisis Factorial de Correspondencias Múltiples (AFCM) de las variables categóricas de identidad académica, contenido temático, enfoque metodológico, herramientas analíticas, características estructurales, calidad, producción y gestión académicas.
- **H₁:** Se identifican patrones estructurales y tipologías diferenciadas en las tesis de pregrado de la UNSAAC a partir de la interpretación del espacio factorial generado por el Análisis Factorial de Correspondencias Múltiples (AFCM) de las variables categóricas de identidad académica, contenido temático, enfoque metodológico, herramientas analíticas, características estructurales, calidad, producción y gestión académicas.

3.2. Identificación de variables

Se definieron las variables basándose en el objetivo principal de explorar los patrones estructurales que se encuentran en las tesis de pregrado de la UNSAAC. utilizamos el Análisis Factorial de Correspondencias (AFC), técnica estadística multivariada para datos categóricos. Razón por la cual las variables son de tipo cualitativas o categóricas. Las variables se agrupan en seis:

- **Identidad Académica:** Comprende el conjunto de variables que permiten ubicar cada tesis dentro de un contexto temporal, institucional y disciplinario. Incluye el año de sustentación, la facultad y la escuela profesional, permite ver la evolución de la producción académica a lo largo del tiempo, así como la identificación de diferencias y patrones entre las distintas áreas de conocimiento.
- **Contenido Temático:** Agrupa las variables relacionadas con el contenido conceptual de la investigación, tales como el título de la tesis, el resumen, las palabras clave y la temática principal. Este factor permite clasificar las investigaciones según sus líneas de estudio predominantes (salud, educación, tecnología, medio ambiente, entre otras), facilitando la identificación de tendencias temáticas y áreas de mayor concentración investigativa.
- **Enfoque Metodológico:** Integra las variables que describen el proceso metodológico del estudio, incluyendo el tipo de enfoque (cuantitativo, cualitativo o mixto), la dimensión temporal (transversal o longitudinal), el diseño de investigación (descriptivo, correlacional, experimental o preexperimental), las fuentes de información utilizadas (documental, encuestas, entre otros) y las herramientas estadísticas empleadas. Este factor permite evaluar el nivel de rigurosidad y las características metodológicas de las tesis.
- **Composición del Trabajo:** Incluye las variables asociadas a la autoría y estructura de elaboración de la tesis, tales como el tipo de colaboración (individual o grupal), el género de los tesistas (tesista 1 y tesista 2), la presencia de co-asesor y la carga de asesoría del asesor principal (clasificada según el número de tesis asesoradas). Este factor permite analizar las dinámicas académicas y organizativas en la producción de las investigaciones.
- **Recursos Gráficos y Datos:** Comprende variables de naturaleza cuantitativa que son categorizadas en escalas ordinales, como el número de tablas, figuras, tamaño de muestra y número de páginas. Estas variables permiten aproximar el nivel de

complejidad estructural, el grado de desarrollo analítico y la profundidad del tratamiento de los datos en las tesis.

- **Calidad y Originalidad:** Evalúa aspectos relacionados con el rigor académico y la calidad de la redacción científica, mediante variables como la cantidad de citas, el porcentaje de originalidad, la valoración de la calidad del trabajo, la riqueza gramatical, los verbos más utilizados y la cantidad de errores ortográficos. Este factor permite analizar la calidad global de las tesis desde una perspectiva tanto técnica como lingüística.

3.3. Operacionalización de variables

Variables	Definición operacional	Definición teórica	Nº	Indicador	Expresión
Identidad Académica	Datos institucionales extraídos automáticamente de la tesis	Conjunto de variables que permiten ubicar cada tesis dentro de un contexto temporal, institucional y disciplinario, determinando la procedencia académica del trabajo (Hernández Sampieri et al., 2014).	1	Año de publicación	Año
			2	Facultad	Categoría nominal
			3	Escuela profesional	Categoría nominal
Contenido Temático	Información temática extraída mediante NLP desde título y resumen	Agrupar los elementos que describen el objeto de estudio, su categoría temática y palabras clave, constituyendo el núcleo conceptual que orienta la investigación hacia un fenómeno o problema específico (Kerlinger & Lee, 2002).	4	Título de la tesis	Texto
			5	Resumen	Texto
			6	Palabras clave	Lista de términos relevantes
			7	Temática principal	Categoría (salud, educación, ingeniería, etc.)
			8	Mención de COVID-19	Sí / No
Enfoque Metodológico	Variables metodológicas identificadas mediante NLP	Describe el enfoque científico, diseño y estructura metodológica de la investigación, definiendo la estrategia sistemática utilizada para recoger y analizar datos (Hernández Sampieri et al., 2014)	9	Enfoque de investigación	Cuantitativo / Cualitativo / Mixto
			10	Tipo de investigación	Descriptivo / Correlacional / Explicativo / Experimental
			11	Diseño de investigación	No experimental / Cuasiexperimental / Experimental
			12	Dimensión temporal	Transversal / Longitudinal
			13	Tipo de relación	Asociación / Causalidad / Comparación
			14	Nivel estadístico	Bajo / Medio / Alto
			15	Tipo de estadística utilizada	Descriptiva / Inferencial / Multivariada
			16	Presencia de muestra	Sí / No
			17	Tipo de muestra	Individuo / Organización / Documento
			18	Tamaño de muestra	Número
			19	Fuente de información	Encuesta / Entrevista / Experimento / Documental / Observación

Herramientas Analíticas	Uso de software y técnicas estadísticas detectadas automáticamente	Representa los recursos técnicos y software estadísticos utilizados para el análisis de datos, facilitando la descripción de fenómenos y la verificación de relaciones entre variables (Cuadras, 2014).	20	Software estadístico	SPSS / R / Excel / Stata / Minitab / Python / MATLAB / SAS
			21	Técnicas estadísticas utilizadas	Frecuencias / Media / Mediana / Chi-cuadrado / t de Student / ANOVA / Pearson / Spearman / Regresión / Alfa de Cronbach / Shapiro-Wilk / Kolmogorov-Smirnov / Análisis factorial / Componentes principales
Características Estructurales	Propiedades físicas y cuantitativas del documento extraídas automáticamente	Describe la composición y extensión del documento académico, incluyendo la organización estandarizada que asegura la calidad, coherencia y validez del trabajo (Hernández Sampieri et al., 2014).	22	Condición del documento en repositorio	Tesis completa / Resumen
			23	Número de páginas	Número
			24	Número de palabras del título	Número
			25	Número de palabras del resumen	Número
			26	Número total de palabras del documento	Número
			27	Número de palabras clave	Número
			28	Peso del archivo PDF	MB
			29	Número de tablas	Número
			30	Número de figuras	Número
			31	Número de enlaces	Número
Calidad y Producción Académica	Indicadores cuantitativos derivados del contenido del documento	Evalúa el nivel de calidad, producción científica y rigor académico del trabajo, considerando la solidez metodológica, la pertinencia científica y la profundidad del análisis (SUNEDU, 2020).	32	Número de citas	Número
			33	Calidad de la tesis	Escala 0 – 20
			34	Riqueza gramatical	Alta / Media / Baja
			35	Nivel de inferencia	Bajo / Medio / Alto
Gestión Académica	Información relacionada con asesoría y producción académica	Describe la estructura de dirección y acompañamiento académico de la tesis, donde la asesoría constituye una función técnica y pedagógica que orienta y garantiza la calidad del trabajo (UNSAAC, 2015).	36	Género del asesor	Masculino / Femenino
			37	Número de tesis por asesor	Número
			38	Presencia de co-asesor	Sí / No
			39	Tipo de tesis según participación	Individual / Grupal
			40	Tipo de financiamiento	Institucional / Externo / Propio

CAPÍTULO IV

METODOLOGÍA

4.1 Ámbito del estudio: localización política y geográfica

La presente investigación se desarrolla en la UNSAAC, con un alcance de carácter institucional, al centrarse en el análisis del repositorio digital desde su implementación en el año 2017 hasta diciembre de 2025, periodo en el que se efectuó la recolección de datos.

4.2 Tipo y nivel de investigación

Este estudio adopta un enfoque cuantitativo, de tipo correlacional, para analizar las tesis de pregrado de la UNSAAC. Mediante un diseño no experimental y transversal, se examinarán los documentos en su estado natural en un único momento, con el propósito de establecer y explicar las relaciones entre variables metodológicas (enfoque, tipo de investigación, nivel estadístico y técnicas analíticas) y su influencia en la calidad y complejidad de los trabajos. Se emplearán técnicas de Procesamiento de Lenguaje Natural (NLP) y minería de texto, que permitirán automatizar el análisis, garantizar objetividad, convertir la información textual en variables categóricas y cuantificar patrones y tendencias mediante frecuencias y porcentajes, reduciendo así la intervención subjetiva del investigador.

4.3 Unidad de análisis

Este estudio analiza cada tesis de pregrado que ha sido presentada y publicada en el repositorio institucional de la UNSAAC. Cada tesis se considera como una unidad independiente de observación y análisis, de la cual se extraen y revisan diferentes características tanto cualitativas como cuantitativas.

4.4. Población y muestra de estudio.

La población de estudio estuvo conformada por la totalidad de las tesis de pregrado registradas en el repositorio digital institucional de la Universidad Nacional de San Antonio Abad del Cusco (UNSAAC). Este universo asciende a un total de 8324 documentos, abarcando desde el inicio oficial de funcionamiento del sistema en el año 2017 para la recolección de datos. La muestra de tipo censal con el propósito de detallar la distribución de esta población y muestra según facultad carrera profesional se presenta la Tabla 1, donde se desglosan las frecuencias absolutas y relativas correspondientes.

Tabla 2*Composición de la población de tesis de pregrado UNSAAC*

		Frecuencia (n=8324)	Porcentaje
Arquitectura e Ing Civil 595 (7,1%)	Arquitectura	180	2,2%
	Ingeniería civil	415	5,0%
Ciencias Administrativas, Contables, Económicas y Turismo 1400 (16,8%)	Ciencias administrativas	255	3,1%
	Contabilidad	439	5,3%
	Economía	462	5,6%
	Ecoturismo	6	0,1%
	Turismo	238	2,9%
Ciencias Agrarias 921 (11,1%)	Agronomía	305	3,7%
	Agronomía tropical	173	2,1%
	Ingeniería agropecuaria	101	1,2%
	Ingeniería forestal	22	0,3%
	Medicina veterinaria	48	0,6%
	Zootecnia	272	3,3%
Ciencias de la Salud 1453 (17,5%)	Enfermería	270	3,2%
	Farmacia y Bioquímica	284	3,4%
	Medicina Humana	561	6,7%
	Obstetricia	105	1,3%
	Odontología	233	2,8%
Ciencias 476 (5,7%)	Biología	361	4,3%
	Física	24	0,3%
	Matemática	60	0,7%
	Química	31	0,4%
Derecho y Ciencias Sociales 831 (10%)	Antropología	191	2,3%
	Arqueología	159	1,9%
	Derecho	227	2,7%
	Historia	79	0,9%
	Psicología	175	2,1%
Educación y Ciencias de la Comunicación 1141 (13,7%)	Ciencias de la Comunicación	254	3,1%
	Educación	887	10,7%
Ingeniería de Procesos 200 (2,4%)	Ingeniería Agroindustrial	57	0,7%
	Ingeniería de Petroquímica	15	0,2%
	Ingeniería en Industrias Alimentarias	24	0,3%
	Ingeniería Química	104	1,2%
	Ingeniería Eléctrica	253	3,0%
Ingeniería Eléctrica, Electrónica, Informática y Mecánica 574 (6,9%)	Ingeniería Electrónica	99	1,2%
	Ingeniería Informática y de Sistemas	106	1,3%
	Ingeniería Mecánica	116	1,4%
Ingeniería Geológica, Minas y Metalúrgica 733 (8,8%)	Ingeniería de Minas	310	3,7%
	Ingeniería Geológica	301	3,6%
	Ingeniería Metalúrgica	122	1,5%

La Tabla 1 muestra que, de 8324 tesis, hay una distribución irregular en áreas académicas. La mayoría de los estudiantes eligen Ciencias de la Salud, que es el 17,5 %, y Ciencias Administrativas, Contables, Económicas y Turismo, que es el 16,8 %. También hay muchos estudiantes en Educación y Ciencias de la Comunicación, que es el 13,7 %, y en Ciencias Agrarias, que es el 11,1 %.

Hay algunas áreas que tienen muy pocos estudiantes. Ingeniería de Procesos solo tiene el 2,4 % y Ciencias tiene el 5,7 %. En cuanto a las escuelas, Medicina Humana tiene el 6,7 %, Ingeniería Civil tiene el 5 %, Economía tiene el 5,6 % y Contabilidad tiene el 5,3 %.

Algunas carreras tienen muy pocos estudiantes, como Ecoturismo, que tiene el 0,1 %, Física, que tiene el 0,3 %, y Ingeniería de Petroquímica, que tiene el 0,2 %. La mayoría de los estudiantes eligen carreras de salud, administración y educación. Esto nos dice qué es lo que los estudiantes prefieren estudiar en la universidad.

Criterios de inclusión:

- Tesis de pregrado registradas, aprobadas y disponibles en el repositorio digital institucional de la UNSAAC.
- Documentos correspondientes al periodo comprendido entre el año 2017 (inicio oficial del repositorio digital) y diciembre de 2025.
- Trabajos de investigación que hayan culminado el proceso de sustentación para la obtención del grado de bachiller o título profesional.
- Documentos en formato digital (PDF) o con metadatos completos que permitan la extracción automatizada de variables mediante técnicas de NLP (Procesamiento de Lenguaje Natural) y minería de texto, tal como se describe en el instrumento de recolección de datos.

Criterios de exclusión:

- Trabajos de investigación que no correspondan al nivel de pregrado (tesis de maestría, doctorado, trabajos de suficiencia profesional, informes de prácticas preprofesionales o proyectos de investigación no sustentados).
- Documentos pertenecientes a otras instituciones educativas diferentes a la UNSAAC.
- Registros duplicados dentro del repositorio institucional.
- Archivos con metadatos incompletos, formatos corruptos o documentos PDF (Portable Document Format) no legibles (escaneos de solo imagen sin texto reconocible) que impidan el procesamiento automatizado y la extracción estructurada de las variables de estudio.
- Tesis registradas con anterioridad al año 2017, por no formar parte del periodo de estudio del repositorio digital analizado.

4.5. Técnicas de recolección de información e instrumento de recolección de datos

Técnica: La recopilación de datos se efectuó a través de la técnica de análisis documental automatizado, aplicada a las tesis de pregrado disponibles en el repositorio institucional de la UNSAAC. Esta técnica permitió revisar, recuperar, transformar y estructurar de manera sistemática la información contenida tanto en los metadatos del repositorio como en el texto completo de los documentos analizados.

Instrumento: El instrumento de recopilación de datos estuvo conformado por una ficha estructurada de extracción automatizada, implementada mediante scripts desarrollados en lenguaje Python. Dicha ficha fue diseñada en función de la operacionalización de variables del estudio. La automatización permitió registrar de forma consistente variables identidad académica, contenido temático, enfoque metodológico, herramientas analíticas, características estructurales, calidad y producción académica y gestión académica. El proceso de recolección y tratamiento de la información se desarrolló en etapas secuenciales y complementarias:

- **Cosecha de metadatos del repositorio institucional**

En una primera etapa se realizó la recuperación automatizada de registros bibliográficos desde el repositorio institucional mediante el protocolo OAI-PMH, lo que permitió cosechar metadatos básicos de las tesis, tales como título, autor, fecha, resumen, facultad, palabras clave, identificadores y enlaces institucionales. Esta etapa constituyó la base del proceso de construcción de la base documental del estudio.

- **Extracción del enlace real al documento PDF**

Posteriormente, a partir de los identificadores tipo handle obtenidos en los metadatos, se accedió a la página individual de cada tesis con el fin de identificar y recuperar el enlace real de descarga del archivo PDF.

- **Descarga masiva de documentos**

Una vez identificados los enlaces válidos, se efectuó la descarga masiva automatizada de las tesis en formato PDF, generando además un registro de control con el estado de descarga de cada archivo. Este procedimiento permitió conformar una colección local de documentos homogénea y organizada, apta para su posterior procesamiento computacional.

- **Conversión de PDF a texto y extracción de metadatos documentales**

Los archivos PDF descargados fueron convertidos a formato de texto plano (.txt) Archivo de texto plano mediante herramientas automatizadas. Paralelamente, se extrajeron metadatos documentales complementarios, tales como número de páginas, tamaño del archivo, cantidad de enlaces detectados, total de palabras, total de caracteres

y presencia aproximada de tablas o recursos gráficos. Esta fase resultó esencial para facilitar el análisis textual posterior y para incorporar variables documentales a la base de datos del estudio.

- **Limpieza, normalización y preprocesamiento**

Se aplicaron rigurosamente los criterios de inclusión y exclusión establecidos para el estudio, lo que permitió filtrar y descartar documentos con metadatos incompletos, formatos corruptos o aquellos que no correspondían a tesis de pregrado completas. El contenido textual de los documentos seleccionados fue sometido a una etapa de limpieza y preprocesamiento, que incluyó la normalización de caracteres, corrección de codificación, eliminación de ruido textual, depuración de saltos de línea innecesarios y estandarización de formatos.

- **Extracción estructurada de variables mediante inteligencia artificial**

En la siguiente etapa se empleó un sistema de extracción estructurada asistida por inteligencia artificial, utilizando el modelo Gemini para identificar, inferir y registrar variables definidas en la operacionalización. A partir del título, resumen y texto completo de cada tesis, el sistema recuperó información relacionada con el enfoque de investigación, tipo y diseño de estudio, dimensión temporal, técnicas estadísticas, software utilizado, temática principal, calidad redaccional y otros elementos metodológicos y académicos. Esta extracción se realizó siguiendo un esquema estructurado de campos, de modo que cada salida pudiera incorporarse directamente a una matriz de análisis.

- **Estructuración final de la base de datos**

Toda la información recuperada fue integrada en una base de datos en formato Excel, donde cada fila representó una tesis y cada columna una variable de estudio. En el caso de variables originalmente numéricas, estas fueron categorizadas para hacerlas compatibles con las técnicas de análisis propuestas, el AFCM, que requiere variables cualitativas o categóricas. Se emplearon criterios de agrupación como intervalos y categorías ordinales, según la naturaleza de cada indicador.

4.6. Técnicas de análisis e interpretación de la información

Análisis descriptivo

En un primer nivel se realizará un análisis descriptivo mediante tablas de frecuencias absolutas, frecuencias relativas, porcentajes y distribuciones categóricas, con el propósito de

caracterizar las tesis de pregrado según los factores definidos en la operacionalización. Este nivel permitirá describir la distribución de variables como escuela profesional, año de sustentación, temática principal, enfoque metodológico, modalidad de autoría, presencia de co-asesor, uso de técnicas estadísticas, recursos gráficos, originalidad y calidad documental.

Este análisis servirá para ofrecer una visión general de la producción de tesis en la UNSAAC y para identificar tendencias preliminares en los distintos componentes del estudio.

Análisis bivariado

Se enfoca en analizar la relación entre pares de variables categóricas mediante el uso de tablas de contingencia. Para determinar si existe una asociación estadísticamente significativa entre ellas, se empleará la prueba de independencia Chi-cuadrado (χ^2)

Complementariamente, se calculará el coeficiente V de Cramer, con la finalidad de estimar la intensidad de asociación entre las variables analizadas. Esta medida permitirá no solo identificar si la relación existe, sino también valorar su magnitud relativa, facilitando la priorización de asociaciones más relevantes dentro del conjunto de variables consideradas.

La construcción de matrices simétricas de valores p y V de Cramer permitirá disponer de una visión global de las asociaciones entre variables, semejante a una matriz de correlaciones para datos categóricos. Este nivel de análisis permitirá responder preguntas específicas, por ejemplo, si existe relación entre la temática principal y el enfoque metodológico, entre el año de sustentación y el diseño de investigación, o entre la modalidad de autoría y determinadas características estructurales de las tesis.

Selección analítica de relaciones relevantes

Dado el número elevado de variables incluidas en el estudio, el análisis bivariado generará múltiples cruces posibles. Por ello, se tomará como criterio complementario de selección la magnitud del coeficiente V de Cramer, a fin de identificar relaciones de mayor relevancia analítica. Este criterio permitirá distinguir asociaciones débiles de aquellas que reflejan patrones más consistentes, facilitando una lectura más sintética y útil de la estructura relacional de los datos. De forma complementaria, estas asociaciones podrán representarse mediante esquemas o grafos exploratorios de relaciones, con el fin de visualizar agrupamientos de variables que presenten vínculos más intensos.

Análisis exploratorio multivariado

Se empleará el Análisis Factorial de Correspondencias Múltiples (AFCM), considerado la técnica principal del estudio, ya que resulta apropiada para el análisis simultáneo de varias variables categóricas.

La aplicación del AFCM en esta investigación permitirá:

- Visualizar simultáneamente las asociaciones entre categorías de múltiples variables.
- Identificar perfiles o tipologías de tesis con características semejantes.
- Reducir la dimensionalidad del conjunto de datos, sintetizando estructuras subyacentes que no se evidencian con claridad en análisis univariados o bivariados.

Los resultados del AFCM permitirán tanto el análisis tabular de coordenadas e inercias como la inspección gráfica interactiva de los mapas factoriales.

Técnicas para demostrar las hipótesis planteadas

Para comprobar las hipótesis específicas que analizan la asociación entre variables categóricas, se utilizará la prueba de independencia Chi-cuadrado mediante tablas de contingencia; además, el análisis se realizará con un nivel de significancia de 0,01, en ese sentido, cuando el valor de p sea inferior a 0,01, se rechazará la hipótesis nula, concluyendo que existe una asociación estadísticamente significativa entre las variables.

De modo complementario, se calculará el coeficiente V de Cramer como medida de tamaño del efecto, a fin de valorar la intensidad de la asociación observada. Este criterio permitió reforzar la interpretación de los resultados, diferenciando asociaciones significativas pero débiles de aquellas que presentan una mayor relevancia estructural dentro del sistema de variables estudiado.

Para verificación de la hipótesis general referida a la existencia de patrones estructurales en las tesis de pregrado de la UNSAAC, se utilizará el Análisis Factorial de Correspondencias Múltiples. La confirmación de esta hipótesis se sustentará en la identificación de configuraciones interpretables en el espacio factorial, la proximidad entre categorías asociadas y el porcentaje de inercia explicado por los primeros ejes factoriales, los cuales permitirán evidenciar estructuras subyacentes, perfiles recurrentes y relaciones multivariadas significativas entre las características de las tesis.

CAPÍTULO V. RESULTADOS Y DISCUSION

5.1. Análisis descriptivo

Figura 1

Nube de palabras del título de tesis.



Tabla 3

Distribución de palabras en título variable contenido temático

Palabra	Cusco	Distrito	Educación	Empresa	Estudiante	Efecto	Provincia	Rendimiento	San	Evaluación
Ocurrencia	6,2%	2,0%	1,7%	1,5%	1,5%	1,4%	1,2%	1,1%	1,1%	1,0%
Palabra	Hospital	Periodo	Centro	Asociación	Factor	Secundario	Sistema	Gestión	Comunidad	Ciudad
Ocurrencia	0,9%	0,9%	0,8%	0,8%	0,7%	0,7%	0,7%	0,6%	0,6%	0,6%
Palabra	Grado	Servicio	Antonio	Paciente	Región	Nacional	Ano	Minero	Salud	Convención
Ocurrencia	0,6%	0,5%	0,5%	0,5%	0,5%	0,5%	0,5%	0,5%	0,5%	0,4%
Palabra	Regional	Diseño	Producción	Habilidad	Calidad	Social	Municipalidad	Cultivo	Determinación	Jerónimo
Ocurrencia	0,4%	0,4%	0,4%	0,4%	0,4%	0,4%	0,4%	0,3%	0,3%	0,3%
Palabra	Perú	Control	Departamento	Implementación	Sector	Proyecto	Apurímac	Riesgo	Profesional	Mercado
Ocurrencia	0,3%	0,3%	0,3%	0,3%	0,3%	0,3%	0,3%	0,3%	0,3%	0,3%
Palabra	Primario	Niño	Fortunato	Actividad	Económico	Herrera	Área	Tratamiento	Lorena	Turístico
Ocurrencia	0,3%	0,3%	0,3%	0,3%	0,3%	0,3%	0,3%	0,3%	0,3%	0,3%

En la Figura 1 y la Tabla 2 se observa que los títulos de las tesis se enfocan predominantemente en temáticas vinculadas a la educación y al territorio. La palabra Cusco presenta la mayor frecuencia de ocurrencia (6,2 %), lo cual resulta coherente con el ámbito geográfico de la universidad y evidencia que las investigaciones se circunscriben mayoritariamente a esta región. Destacan términos como distrito, provincia y ciudad, lo que

refleja una clara orientación hacia la delimitación geográfica de los estudios, en cumplimiento de los lineamientos establecidos en el reglamento interno de formato de tesis de la institución.

En cuanto a las líneas temáticas, sobresalen vocablos como educación, estudiante, rendimiento y evaluación, lo que sugiere un fuerte interés de los tesisistas por analizar el desempeño estudiantil y los procesos de desarrollo educativo. Se identifica un bloque importante orientado al sector salud y servicios, representado por palabras como hospital, paciente, salud y tratamiento. A esto se suma un conjunto de investigaciones enfocadas en la gestión organizacional y productiva, evidenciado por términos como empresa, gestión, sistema y producción.

La presencia de vocablos como factor, diseño, implementación y control determina que los estudios buscan no solo analizar fenómenos, sino también aplicar soluciones prácticas. La producción académica es heterogénea y se orienta a resolver problemáticas locales en los ámbitos de la educación, la salud y el desarrollo socioeconómico de la región. Se evidencia un marcado predominio de las temáticas educativas y de gestión, lo que refleja el enfoque por impulsar la mejora continua en estos sectores a través de la investigación de pregrado.

Tabla 4
Análisis de variables nominales

Variable	Categorías	Frecuencia	Porcentaje
Identidad Académica – Año de sustentación	2016	356	4,3%
	2017	1267	15,2%
	2018	145	1,7%
	2019	1224	14,7%
	2020	471	5,7%
	2021	512	6,2%
	2022	641	7,7%
	2023	805	9,7%
	2024	1727	20,7%
	2025	1176	14,1%
Contenido Temático – Temática de la investigación principal	Salud	1919	23,1%
	Ingeniería	1525	18,3%
	Economía	1213	14,6%
	Educación	1016	12,2%
	Ambiental	951	11,4%
	Sociales	878	10,5%
	Tecnología	479	5,8%
	Cultura	299	3,6%
	Otras	44	0,5%

La Tabla 3 presenta la distribución temporal y temática de las tesis sustentadas, evidenciando una concentración significativa de la producción académica en años específicos. Destaca el año 2024 con el mayor volumen de investigaciones (20,7 %), seguido por 2017 (15,2 %), 2019 (14,7 %) y 2025 (14,1 %). Esta tendencia refleja un consolidado crecimiento en la actividad investigativa de pregrado en la institución a lo largo del tiempo. Se observan fluctuaciones marcadas, como el drástico descenso registrado en el año 2018 (1,7 %), lo cual podría estar asociado a factores coyunturales, administrativos o de promoción de la titulación durante dicho periodo.

En lo concerniente a la distribución temática, el área de Salud concentra el mayor porcentaje de investigaciones (23,1 %), seguida por Ingeniería (18,3 %) y Economía (14,6 %). Esta jerarquía temática sugiere que la producción científica de los tesisistas está fuertemente orientada hacia disciplinas que atienden necesidades sanitarias, productivas y socioeconómicas prioritarias. Las áreas de Educación (12,2 %), Ambiental (11,4 %) y Sociales (10,5 %) mantienen una participación relevante en el conjunto de la producción. Por el contrario, las líneas de Tecnología (5,8 %), Cultura (3,6 %) y Otras (0,5 %) registran una menor densidad investigativa, lo que evidencia oportunidades de expansión y diversificación para futuros trabajos de grado.

Tabla 5
Análisis de variables nominales

variable	Categorías	Frecuencia	Porcentaje
Enfoque Metodológico - Enfoque de investigación	Cuantitativo	6748	81,1%
	Cualitativo	921	11,1%
	Mixto	655	7,9%
Enfoque Metodológico - Tipo de investigación	Descriptivo	4168	50,1%
	Correlacional	1984	23,8%
	Explicativo	2172	26,10%
Enfoque Metodológico - Diseño de investigación	No experimental	7447	89,5%
	Experimental	446	5,4%
	Cuasiexperimental	355	4,3%
	Casos	63	0,8%
	Casos y control	13	0,2%
Enfoque Metodológico – Tipo relación	Descriptivo	4044	48,6%
	Relación	1857	22,3%
	Influencia	1552	18,6%
	Evaluación	808	9,7%
	Predicción	63	0,8%

Se evidencia un marcado predominio del enfoque cuantitativo en la producción científica de la UNSAAC, el cual concentra el 81,1 % de las investigaciones, en detrimento de los enfoques cualitativo (11,1 %) y mixto (7,9 %). Esta tendencia refleja una clara preferencia por la medición objetiva de variables y el análisis estadístico, en concordancia con los paradigmas positivistas predominantes en la formación de pregrado.

En lo concerniente al tipo de investigación, los estudios de carácter descriptivo constituyen la modalidad predominante (50,1 %), seguidos por los correlacionales (23,8 %) y explicativos (16,0 %). Esta distribución evidencia que la mayoría de los tesisistas optan por caracterizar fenómenos y establecer relaciones entre variables, en lugar de desarrollar diseños que permitan explicar causas o intervenir directamente sobre la realidad estudiada. Los diseños experimentales y cuasiexperimentales representan apenas un 8,9 % y 1,2 % respectivamente, lo que confirma una orientación metodológica predominantemente observacional.

Respecto al diseño de investigación, el enfoque no experimental alcanza el 89,5 %, lo que ratifica que la gran mayoría de los estudios se desarrollan mediante la observación de fenómenos en su contexto natural, sin manipulación deliberada de variables. Por el contrario, los diseños experimentales (5,4 %) y cuasiexperimentales (4,3 %) registran una presencia marginal, lo que limita el potencial de inferencia causal en la producción académica institucional.

En cuanto al tipo de relación analizada, la descripción predomina con un 48,6 %, seguida por los estudios de relación (22,3 %) e influencia (18,6 %). Los enfoques más avanzados, como la predicción, presentan una incidencia mínima (0,8 %). Estos resultados permiten concluir que la investigación de pregrado en la UNSAAC se circunscribe mayoritariamente a niveles descriptivos y correlacionales, con una limitada aplicación de diseños experimentales y modelos predictivos, lo que sugiere oportunidades de fortalecimiento en la formación metodológica de los tesisistas.

Tabla 6
Análisis de variables nominales

Variable	Categorías	Frecuencia	Porcentaje
Enfoque Metodológico- Nivel estadístico	Avanzado	1287	15,50%
	Intermedio	4608	55,40%
	Básico	2429	29,20%
Enfoque Metodológico - Tipo de unidad de análisis	Individuo	4438	53,30%
	Organización	1844	22,20%
	Objeto	1342	16,10%
	Documento	517	6,20%
	Grupo	183	2,20%
Calidad y Producción Académica – Calidad de tesis (0 a 20)	10-11.	1612	19,40%
	12-13.	24	0,30%
	14-15.	1618	19,40%
	16-17.	5021	60,30%
	18	49	0,60%
Calidad y Producción Académica – Riqueza gramatical	Alta	6385	76,70%
	Media	1939	23,30%
	Baja	0	0,00%
Calidad y Producción Académica – Nivel inferencia	Alto	471	5,70%
	Medio	6563	78,80%
	Bajo	1290	15,50%
Gestión Académica – Genero asesor	Masculino	6230	74,80%
	Femenino	1890	22,70%
	Sin registro	204	2,50%
Gestión Académica – Tesis grupal	Individual varón	2453	29,50%
	Individual mujer	1880	22,60%
	Ambas mujeres	1495	18,00%
	Varón y mujer	1432	17,20%
	Ambos varones	1064	12,80%
Gestión Académica – Financiamiento	Propio	7498	90,10%
	Institucional (UNSAAC)	679	8,20%
	Externo	147	1,80%

A partir de la Tabla 5, se evidencia que la mayoría de las tesis de pregrado de la UNSAAC presentan un nivel estadístico intermedio (55,4 %), seguido por el nivel básico (29,2 %), mientras que el nivel avanzado registra una incidencia marginal (15,5 %). Esta distribución revela un empleo predominantemente austero de las técnicas estadísticas en la producción académica institucional, lo que limita el alcance analítico de las investigaciones. En cuanto a la unidad de análisis, predomina el estudio centrado en el individuo (53,3 %), seguido por las organizaciones (22,2 %) y los objetos (16,1 %), mientras que las unidades de tipo documento (6,2 %) y grupo (2,2 %) presentan una presencia residual.

Respecto a los indicadores de calidad, se observa que el 60,3 % de las tesis obtiene una calificación entre 16 y 17, mientras que apenas el 0,6 % alcanza la calificación máxima de 18, lo que evidencia un techo en la excelencia académica. La riqueza gramatical se concentra mayoritariamente en el nivel alto (76,7 %), sin registrarse casos clasificados en el nivel bajo, mientras que el nivel de inferencia se ubica predominantemente en la categoría media (78,8 %), con una escasa representación del nivel alto (5,7 %). Esta configuración sugiere limitaciones en la capacidad de los tesisistas para desarrollar análisis de mayor profundidad y complejidad interpretativa.

En el ámbito de la gestión académica, destaca una marcada predominancia de asesores de género masculino (74,8 %), frente a una participación femenina significativamente menor (22,7 %). La modalidad de autoría muestra una distribución heterogénea, con predominio de las tesis individuales elaboradas por varones (29,5 %), seguidas por las individuales de mujeres (22,6 %), las coautorías femeninas (18,0 %) y los equipos mixtos (17,2 %). El financiamiento de las investigaciones es mayoritariamente propio (90,1 %), con una reducida participación del apoyo institucional (8,2 %) y externo (1,8 %), lo que refleja una limitada inversión institucional y externa en la investigación de pregrado, condicionando así el alcance y la rigurosidad de los estudios desarrollados.

Tabla 7
Análisis descriptivo de variables dicotómicas

	Si		No	
	f	%	f	%
Gestión Académica - Presencia de co-asesor	424	5,1%	7900	94,9%
Características Estructurales - Tesis completa en el repositorio	4948	59,4%	3376	40,6%
Enfoque Metodológico - El estudio es de corte transversal	8055	96,8%	269	3,2%
Enfoque Metodológico - Hace uso de muestra	7729	92,9%	595	7,1%
Contenido Temático - Menciona COVID	891	10,7%	7433	89,3%
Contenido Temático - Uso de estadística				
· Descriptiva	8314	99,9%	10	0,1%
· Inferencial	5604	67,3%	2720	32,7%
· Multivariada	1274	15,3%	7050	84,7%
Contenido Temático – Técnica de recopilación de información				
· Documental	7343	88,2%	981	11,8%
· Encuesta	3200	38,4%	5124	61,6%
· Observación	1751	21,0%	6573	79,0%
· Experimento	1630	19,6%	6694	80,4%
· Entrevista	782	9,4%	7542	90,6%
Herramientas Analíticas - Uso de software estadístico				
· SPSS	4398	52,8%	3926	47,2%
· Excel	3261	39,2%	5063	60,8%
· R	1924	23,1%	6400	76,9%
· Stata	249	3,0%	8075	97,0%
· Minitab	207	2,5%	8117	97,5%
· Python	124	1,5%	8200	98,5%
· Matlab	112	1,3%	8212	98,7%
· SAS	106	1,3%	8218	98,7%
Herramientas Analíticas - Uso de técnicas estadísticas				
· Frecuencias/porcentajes	3169	38,1%	5155	61,9%
· Media	1358	16,3%	6966	83,7%
· Prueba T	1145	13,8%	7179	86,2%
· Cronbach	988	11,9%	7336	88,1%
· ANOVA	944	11,3%	7380	88,7%
· Shapiro K-S	856	10,3%	7468	89,7%
· Spearman	834	10,0%	7490	90,0%
· Pearson	704	8,5%	7620	91,5%
· Chi-cuadrado	600	7,2%	7724	92,8%
· Regresión lineal	546	6,6%	7778	93,4%
· Factorial o componentes	395	4,7%	7929	95,3%
· Regresión logística	287	3,4%	8037	96,6%
· Mediana	231	2,8%	8093	97,2%

La Tabla 6 presenta la distribución de variables dicotómicas en las tesis de pregrado de la UNSAAC, evidenciando patrones consistentes en las características metodológicas y estructurales de las investigaciones. En cuanto a la gestión académica, únicamente el 5,1 % de

las tesis cuenta con la presencia de un co-asesor, mientras que el 94,9 % restante prescinde de esta figura, lo que refleja una marcada tendencia hacia la asesoría individual y una limitada colaboración académica en la dirección de los trabajos de investigación. Respecto a la disponibilidad documental, el 59,4 % de las tesis se encuentra registrada en su versión completa en el repositorio institucional, mientras que el 40,6 % restante presenta únicamente el resumen o metadatos, lo que condiciona el acceso integral a los contenidos y limita la replicabilidad de los estudios.

En el ámbito metodológico, predomina ampliamente el diseño transversal (96,8 %) y el uso de muestras (92,9 %) como estrategia de recolección de datos, lo que ratifica la orientación observacional y no experimental identificada previamente. Cabe destacar que solo el 10,7 % de las investigaciones abordan temáticas vinculadas a la pandemia de COVID-19, evidenciando una baja incidencia de estudios contextualizados en coyunturas sanitarias recientes. En lo concerniente al análisis de datos, la estadística descriptiva alcanza una cobertura casi universal (99,9 %), seguida por la estadística inferencial (67,3 %), mientras que la estadística multivariada registra una presencia marginal (15,3 %), lo que limita significativamente la capacidad de los tesisistas para abordar análisis de mayor complejidad y profundidad analítica.

Las fuentes de información más utilizadas son las documentales (88,2 %), seguidas por las encuestas (38,4 %), lo que evidencia una preferencia por métodos de recolección de datos de naturaleza cuantitativa y estructural. En cuanto al software estadístico, SPSS (Statistical Package for the Social Sciences) concentra el mayor porcentaje de uso (52,8 %), seguido por Excel (39,2 %) y R (23,1 %), mientras que herramientas más avanzadas como Python (1,5 %), MATLAB (Matrix Laboratory) (1,3 %) o SAS (Statistical Analysis System) (1,3 %) presentan una incidencia residual. En lo relativo a las técnicas estadísticas, las frecuencias y porcentajes (38,1 %) y la media (16,3 %) constituyen los procedimientos más empleados, en tanto que técnicas de mayor sofisticación como la regresión logística (3,4 %) o el análisis factorial (4,7 %) registran una aplicación limitada. La investigación de pregrado en la UNSAAC se circunscribe predominantemente a análisis descriptivos y al empleo de herramientas estadísticas tradicionales, con una escasa incorporación de metodologías avanzadas que permitan un tratamiento más riguroso de la información.

Tabla 8

Descripción de variables numéricas y conversión a variables ordinales.

	Valores perdidos (N=8324)	Nota	Min	Q1	Me	Q3	Max	Promedio	Desv.	Normalidad K-S (Kolmogorov- Smirnov)
Características Estructurales Numero de palabras en el titulo	0 (0%)	Completo	2	18	21	26	63	21,9	6,6	0,061 sig.=0,000
Características Estructurales Numero de palabras en el resumen	0 (0%)	Completo	3	211	246	270	1508	258,4	99,8	0,214 sig.=0,000
Contenido Temático Numero de palabras clave	0 (0%)	Completo	1	5	6	8	27	6,7	2,4	0,126 sig.=0,000
Gestión Académica Numero de tesis que un asesor llevo desde 2017 a 2025	178 (2,1%)	Tesis donde no se puede identificar asesor	1	11	21	35	161	29,2	30,1	0,183 sig.=0,000
Enfoque Metodológico Tamaño de muestra usada	2162 (26%)	No es identificable la muestra	1	28	75	191	26 355 165	10 114,4	394 195,1	0,490 sig.=0,000
Características Estructurales Peso del PDF en Mb	3376 (40,6%)	Solo de archivos que no son resumen	0,3	2	3,3	5,6	282,7	6,4	14	0,331 sig.=0,000
Características Estructurales Páginas de la tesis	3376 (40,6%)	Solo de archivos que no son resumen	35	108	139	183	1 919	160,7	99,2	0,164 sig.=0,000
Características Estructurales Numero de enlaces	3376 (40,6%)	Solo de archivos que no son resumen	0	0	102	169,3	1 686	115,8	128,1	0,183 sig.=0,000
Características Estructurales Total, de palabras en tesis	3376 (40,6%)	Solo de archivos que no son resumen	0	20 847	27 212	37 454	642 049	34 027,1	29 748,9	0,212 sig.=0,000
Características Estructurales Numero figuras	3376 (40,6%)	Solo de archivos que no son resumen	0	8	18	36	590	29,8	40,6	0,232 sig.=0,000
Características Estructurales Numero tablas	3377 (40,6%)	Solo de archivos que no son resumen	0	12	21	35	714	28	27,9	0,159 sig.=0,000
Calidad y Producción Académica Numero de citas	3561 (42,8%)	Solo de archivos que no son resumen y se puede identificar las citas	5	30	39	58	198	47,7	30,5	0,192 sig.=0,000

La Tabla 7 muestra que las variables numéricas que se estudiaron tienen mucha variedad y no siguen una distribución normal en todos los casos. Esto se vio en la prueba de Kolmogorov-Smirnov, que dio un resultado de $p = 0,000$, en tanto las distribuciones son asimétricas y que hay valores extremos.

Para la estructura de los textos, el número de palabras en el título tiene una mediana de 21 palabras, este resultado menciona que la mitad de los títulos tienen menos de 21 palabras y la otra mitad tiene más de 21 palabras. El rango es muy amplio, desde 2 palabras hasta 63 palabras, en el resumen tiene una mediana de 246 palabras, con un máximo de 1508 palabras, esto muestra que existe bastante variabilidad en la extensión de los resúmenes.

Las palabras clave tienen una mediana de 6, lo que significa que la mayoría de los textos tienen alrededor de 6 palabras clave. Esto es más homogéneo que las otras variables.

Para la producción académica, el número de tesis por asesor tiene una mediana de 21 tesis, no obstante, hay valores extremos de hasta 161 tesis, lo que refleja una carga académica desigual. Hay un 2,1% de datos perdidos, ya que no se identificó al asesor.

El tamaño de muestra es muy variable, con una mediana de 75 y un máximo de 26 355 165, Esto hace que el promedio sea muy alto (10 114,4) y que la desviación estándar sea bastante grande (394 195,1). Esto se debe a la influencia de valores extremos y a que hay un 26% de valores no identificables.

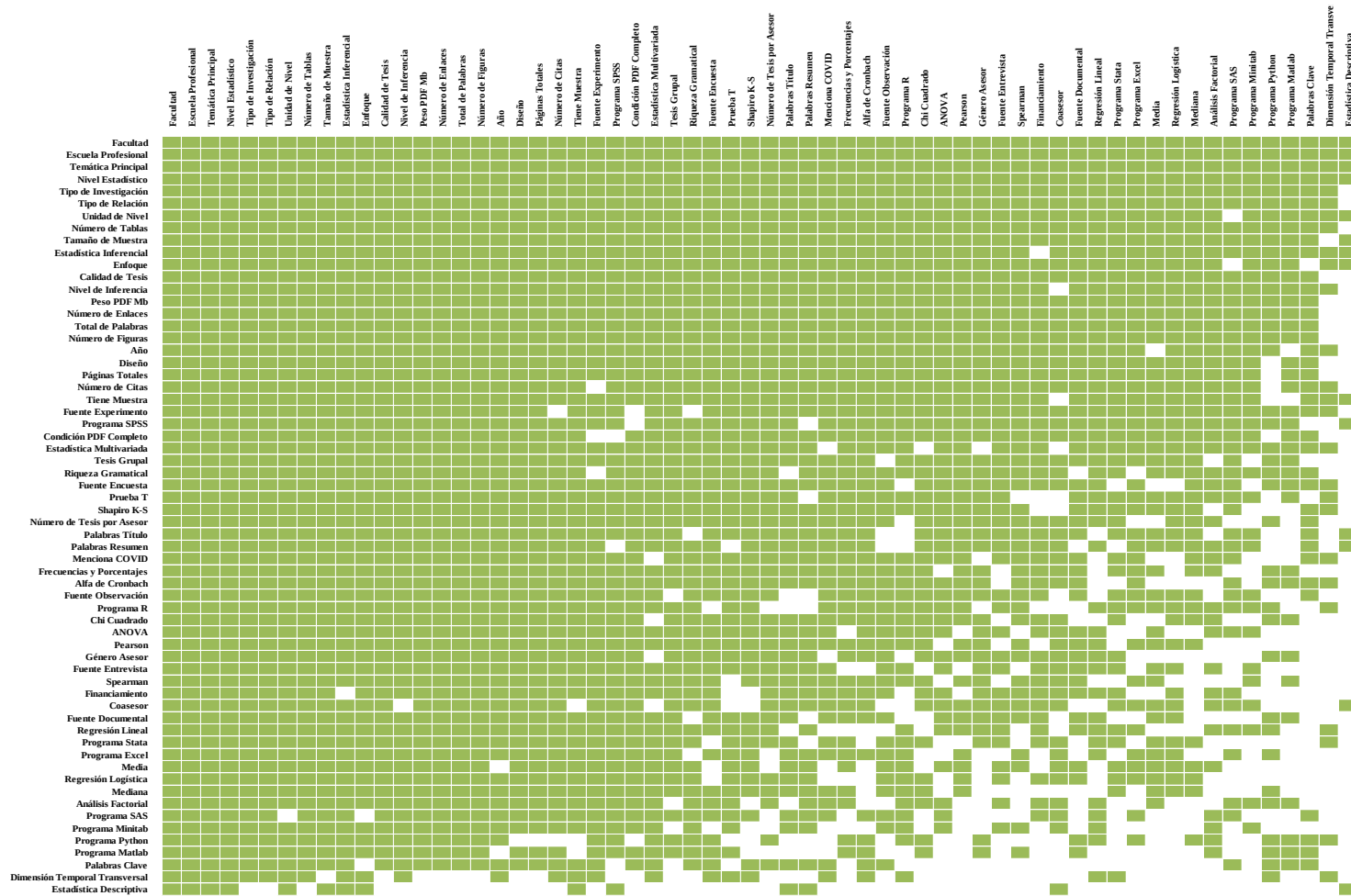
En las variables relacionadas con los documentos completos, hay un 40,6% de valores perdidos, no obstante, se puede ver que el peso del PDF tiene una mediana de 3,3 Mb y que las páginas de la tesis tienen una mediana de 139 páginas, con un máximo de 1919 páginas. Esto muestra que hay mucha variabilidad en la extensión de los documentos. El total de palabras en la tesis tiene una mediana de 27 212 palabras, con valores extremos de más de 600 mil palabras.

El número de enlaces tiene una mediana de 102, con un máximo de 1686. El número de figuras y tablas también es considerable, con medianas de 18 y 21, respectivamente, hay mucha dispersión en estos valores. El número de citas tiene una mediana de 39, con un máximo de 198, pero hay un 42,8% de valores perdidos, estos resultados muestran que la producción académica tiene mucha variabilidad en su estructura y contenido, con valores extremos y una gran proporción de datos faltantes en variables relacionadas con documentos completos. Para análisis más detallado se usará los cuartiles como baremos dado que la técnica de análisis necesitan variables categóricas.

5.2. Análisis bivariado de las variables de estudio

Figura 2

Matriz del valor de significancia (sig.) de la prueba Chi cuadrado a pares de las 62 variables.



El análisis de asociación entre pares de variables tiene como objetivo identificar relaciones significativas dentro del conjunto de variables del estudio. Para ello, se aplicó la prueba de Chi-cuadrado de independencia de manera sistemática a todas las combinaciones posibles. Considerando que se dispone de 62 variables, el número total de comparaciones realizadas asciende a 1 891 pruebas.

En la matriz, cada celda representa el resultado de una de estas pruebas: las celdas de color verde indican asociaciones con un valor de significancia menor a 0,01 ($p < 0,01$), es decir, combinaciones en las que se rechaza la hipótesis nula de independencia y se confirma una relación estadísticamente significativa entre el par de variables; las celdas blancas, en cambio, corresponden a asociaciones no significativas ($p \geq 0,01$), en las que no es posible descartar la independencia entre las variables.

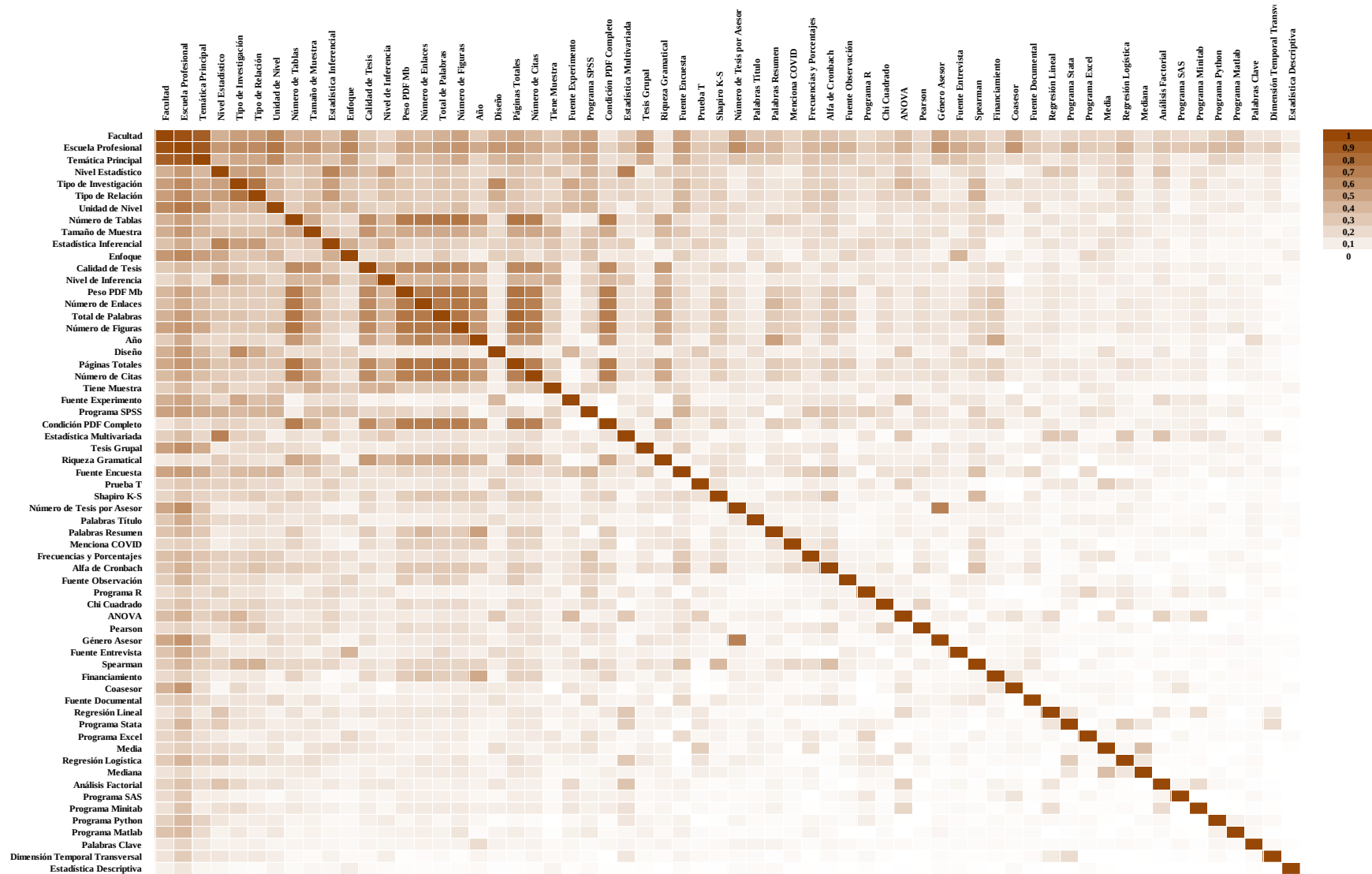
De las 1 891 combinaciones evaluadas, 1 600 relaciones resultaron estadísticamente significativas con un nivel de significancia menor a 0,01, lo que representa el 84,3 % del conjunto analizado. Esta elevada proporción, reflejada visualmente en el predominio de celdas verdes en la matriz, evidencia un alto grado de dependencia entre las variables y confirma la existencia de patrones estructurales en la producción de tesis de pregrado de la UNSAAC.

Desde una perspectiva estructural, las variables más relevantes son aquellas que presentan asociaciones significativas con la mayor parte del sistema, lo que en la matriz se visualiza como filas y columnas casi completamente teñidas de verde. En este núcleo central destacan variables vinculadas a la identidad académica y a la estructura general de la tesis, tales como Facultad, Escuela Profesional, Temática Principal, Nivel Estadístico, Tipo de Investigación y Enfoque metodológico, las cuales desempeñan un papel determinante al condicionar los aspectos metodológicos y formales del trabajo. Variables relacionadas con el tratamiento estadístico, como Estadística Inferencial, Estadística Multivariada, uso del programa SPSS, Chi-cuadrado, ANOVA y Regresión, exhiben altos niveles de conexión, lo que evidencia que la aplicación de herramientas estadísticas se articula con múltiples dimensiones del proceso investigativo. variables de naturaleza cuantitativa como el número de tablas, figuras, palabras, citas, páginas y el peso del archivo PDF presentan un elevado grado de asociación.

Las variables con menor número de asociaciones significativas pueden considerarse periféricas dentro del sistema y se identifican en la matriz como filas y columnas con mayor presencia de celdas blancas. Entre ellas se encuentran aspectos más específicos o de baja frecuencia, como la mención de COVID-19.

Figura 3

Matriz del valor de coeficiente de contingencia a pares de las 62 variables.



La Figura 3 presenta la matriz del coeficiente de contingencia correspondiente a las asociaciones entre las 62 variables categóricas analizadas. A diferencia de la matriz de significancia estadística, este coeficiente permite evaluar la magnitud de las relaciones, es decir, no solo indica si existe dependencia entre dos variables, sino también el grado de asociación conjunta entre sus categorías. La matriz exhibe una estructura heterogénea, con zonas donde se concentran valores elevados, lo que evidencia la presencia de patrones diferenciados y de un núcleo estructural dentro del sistema de variables.

El coeficiente de contingencia se calcula a partir del estadístico Chi-cuadrado. los valores próximos a 0,9 representan asociaciones muy intensas, mientras que los inferiores a 0,2 se consideran débiles o residuales.

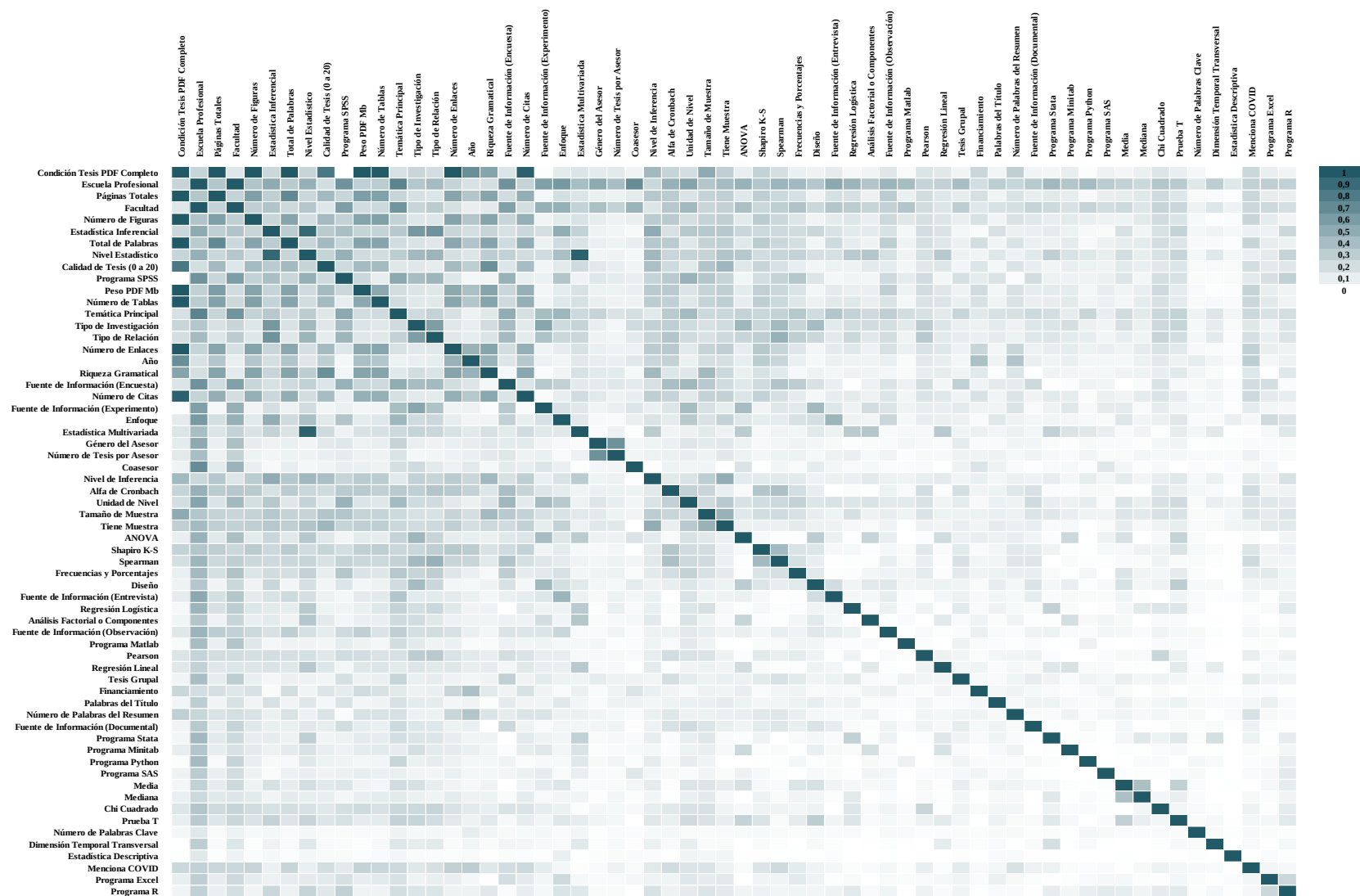
En particular, destacan asociaciones de alta intensidad en variables vinculadas al contexto académico e institucional. La relación entre Facultad y Escuela Profesional alcanza valores próximos a 0,90, lo que refleja una asociación muy fuerte y coherente con la estructura organizacional de la universidad. La asociación entre Facultad y Temática Principal presenta valores en torno a 0,80, lo que indica que las áreas académicas condicionan significativamente la elección de los temas de investigación. Variables como Nivel Estadístico y Tipo de Investigación muestran coeficientes entre 0,60 y 0,70, evidenciando que la complejidad estadística de las tesis guarda una estrecha relación con el diseño metodológico adoptado.

En un segundo nivel, se identifican variables con asociaciones de intensidad moderada, principalmente vinculadas a la estructura y extensión del documento. Variables como número de tablas, número de figuras, total de palabras y número de páginas presentan valores entre 0,30 y 0,50, lo que denota relaciones consistentes, aunque no determinantes. Esto sugiere que, si bien las tesis más extensas tienden a incorporar mayor cantidad de elementos gráficos y contenido textual, dicha relación no es estrictamente proporcional, sino que responde a variaciones propias del enfoque metodológico y del área disciplinar.

Se observa un conjunto de variables con asociaciones de baja intensidad, cuyos valores se sitúan en torno a 0,10 o menores. Entre ellas se encuentran aspectos específicos o poco frecuentes, como la mención de COVID-19, el tipo de financiamiento, la presencia de co-asesor y el uso de programas estadísticos menos comunes (SAS, Minitab o MATLAB). Estas variables mantienen relaciones débiles con el resto del sistema, lo que indica que su presencia responde a situaciones particulares o coyunturales más que a patrones estructurales generalizados. Este hallazgo explica por qué, aun cuando el número de asociaciones estadísticamente significativas es elevado, la mayoría presenta una intensidad baja a moderada.

Figura 4

Matriz del valor de V de cramer a pares de las 62 variables.



La Figura 4 presenta la matriz del coeficiente V de Cramer correspondiente a las relaciones entre las 62 variables categóricas analizadas. A diferencia del coeficiente de contingencia, el V de Cramer ofrece una medida estandarizada de la intensidad de asociación con valores que oscilan entre 0 y 1, lo que permite realizar comparaciones directas entre pares de variables sin verse afectado por las dimensiones de la tabla de contingencia. En este sentido, el V de Cramer resulta metodológicamente más robusto, ya que evita la sobreestimación de la intensidad de las relaciones que puede presentarse con el coeficiente de contingencia.

El patrón visual del mapa de calor revela una estructura heterogénea de asociaciones. Se evidencia que gran parte de las asociaciones se ubican en rangos bajos a moderados (0.10 – 0.30), representados por tonalidades claras en la matriz. Se identifican agrupaciones de celdas con tonalidades más oscuras que evidencian relaciones de alta intensidad, las cuales permiten reconocer un núcleo estructural del sistema, correspondiente principalmente a variables académicas y metodológicas.

Entre las relaciones de mayor magnitud, se identifican las siguientes asociaciones:

Facultad – Escuela Profesional ($V \approx 0.85 - 0.90$)

Facultad – Temática Principal ($V \approx 0.65 - 0.75$)

Escuela Profesional – Temática Principal ($V \approx 0.60 - 0.70$)

Nivel Estadístico – Tipo de Investigación ($V \approx 0.50 - 0.60$)

Tipo de Investigación – Enfoque ($V \approx 0.45 - 0.55$)

Estadística Inferencial – Nivel Estadístico ($V \approx 0.40 - 0.50$)

Estas asociaciones evidencian que la estructura académica (facultad, escuela profesional y temática) constituye el eje principal que organiza las tesis, condicionando tanto el enfoque metodológico como el nivel de complejidad estadística. Se identifica un segundo bloque de relaciones moderadas asociadas a la estructura documental del trabajo, entre las que destacan:

Número de tablas – Número de figuras ($V \approx 0.40 - 0.50$)

Total, de palabras – Páginas totales ($V \approx 0.45 - 0.55$)

Número de citas – Total de palabras ($V \approx 0.30 - 0.40$)

Estas relaciones reflejan que las tesis con mayor extensión tienden a desarrollar un análisis técnico y documental más exhaustivo, incorporando mayor cantidad de elementos gráficos y referencias bibliográficas.

Las relaciones vinculadas al uso de herramientas estadísticas específicas presentan intensidades moderadas a bajas, tales como:

Programa SPSS – Chi-cuadrado / ANOVA ($V \approx 0.30 - 0.40$)

Regresión – Nivel estadístico ($V \approx 0.30 - 0.40$)

Las variables periféricas presentan asociaciones débiles ($V < 0.20$), representadas visualmente por celdas de tonalidad muy clara en la matriz. Entre ellas se encuentran:

Mención de COVID-19

Tipo de financiamiento

Presencia de co-asesor

Programas estadísticos menos frecuentes (SAS, MATLAB, Minitab)

Estas variables no forman parte del núcleo estructural del sistema, sino que responden a condiciones particulares o coyunturales de las investigaciones.

En contraste con el coeficiente de contingencia, donde muchas asociaciones parecían elevadas, el V de Cramer permite evidenciar que solo un grupo reducido de asociaciones presenta realmente una fuerza considerable, mientras que la mayoría corresponde a asociaciones débiles o moderadas.

Las relaciones más fuertes identificadas serán las que ejerzan mayor influencia en el Análisis Factorial de Correspondencias Múltiples, contribuyendo significativamente a la formación de las dimensiones factoriales y a la identificación de tipologías de tesis.

El coeficiente V de Cramer, al ser un estadístico estandarizado que toma valores en el intervalo $[0, 1]$, permite evaluar de forma objetiva la intensidad de asociación entre variables categóricas y facilita la comparación sistemática entre múltiples pares de variables. Su uso, en combinación con el número de relaciones estadísticamente significativas y la cantidad de variables analizadas, permite identificar el punto óptimo de corte para el análisis, es decir, el conjunto de variables que contribuyen de forma más relevante a la estructura del sistema. A diferencia de otras medidas como el coeficiente de contingencia, el V de Cramer permite reconocer con mayor precisión la arquitectura de relaciones significativas, evitando sesgos derivados del tamaño de las tablas de contingencia y brindando una base más sólida para el análisis multivariado posterior. Del tamaño de las tablas y brindando una base más sólida para la explicación multivariada.

Tabla 9*Medida de asociación V de Cramer y frecuencia de pares y variables significativas*

valor V de cramer	número de pares significativos	numero de variables significativas
0,0	1879	62
0,1	969	62
0,2	473	60
0,3	213	56
0,4	116	42
0,5	57	30
0,6	23	25
0,7	13	15
0,8	11	14
0,9	10	13
1,0	0	0
0,57	29	26

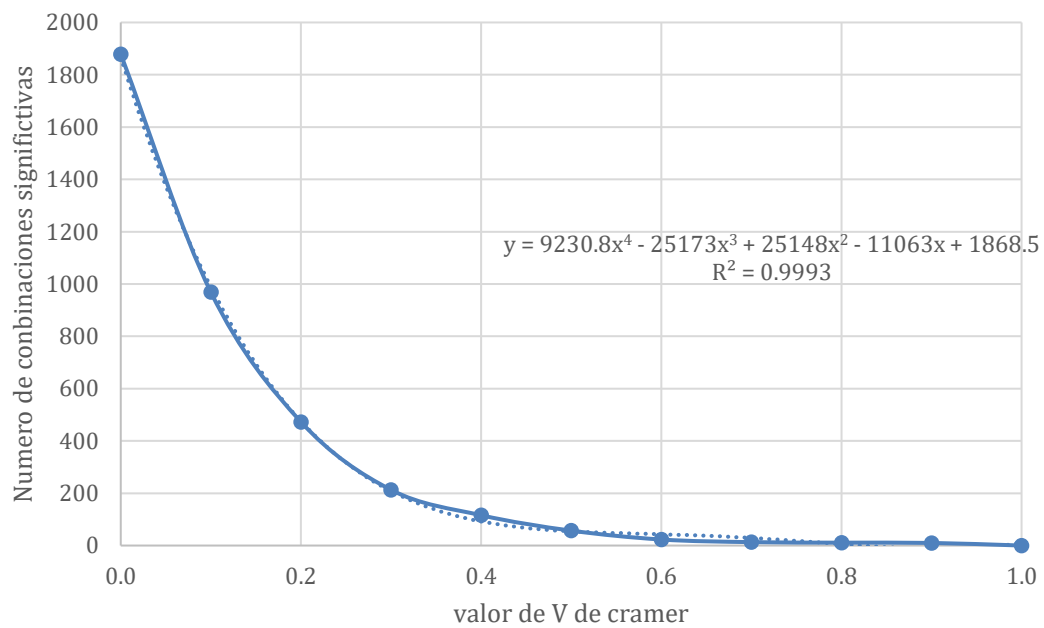
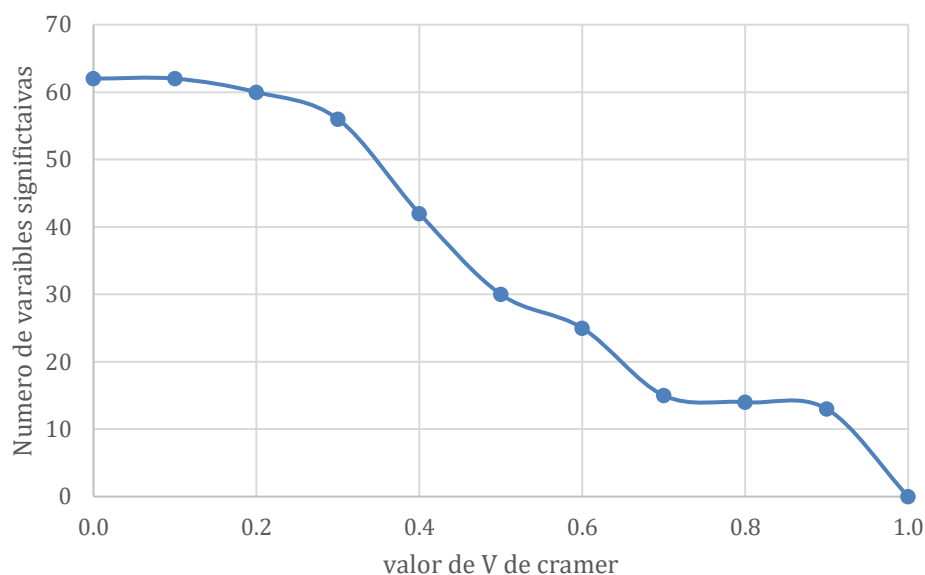
Figura 5*Valor V de cramer y número de combinaciones significativas*

Figura 6
Valor V de cramer y número de variables significativas



La Tabla 8, junto con las Figuras 5 y 6, ilustra la relación entre el valor del coeficiente V de Cramer, el número de pares de variables con asociación significativa y la cantidad de variables implicadas en dichas interacciones. Este análisis permite evaluar cómo evoluciona la estructura del sistema al incrementar el umbral de intensidad de la asociación, transitando desde la consideración de todas las relaciones estadísticamente significativas hasta el aislamiento de aquellas de mayor magnitud.

Se observa que, para umbrales bajos de V de Cramer (0,0 y 0,1), el número de interacciones significativas es elevado (1879 a 969), involucrando prácticamente la totalidad de las variables (62). Esto indica que, aunque existe un alto volumen de relaciones estadísticamente significativas, la mayoría corresponde a asociaciones débiles. Esta situación introduce ruido en el análisis, dificultando la identificación de las estructuras verdaderamente relevantes.

Conforme se incrementa el umbral del V de Cramer, se evidencia una reducción progresiva tanto en el número de pares significativos como en el conjunto de variables implicadas. Para un valor de 0,3 se registran 213 combinaciones significativas que involucran a 56 variables, mientras que para 0,5 estas cifras descienden a 57 combinaciones y 30 variables. Como se aprecia en la Figura 5, esta dinámica sigue una tendencia decreciente no lineal, lo que sugiere que las interacciones de alta intensidad son considerablemente menos frecuentes dentro del sistema.

La Figura 6 muestra que la cantidad de variables significativas también decrece a medida que aumenta el valor de V de Cramer. A partir de umbrales superiores a 0,6, la disminución se vuelve más abrupta, reduciéndose a únicamente 15 variables para $V = 0,7$ y 13 variables para $V = 0,9$. Esto demuestra que solo un subconjunto reducido de variables mantiene relaciones fuertes y consistentes, conformando el núcleo estructural del sistema.

Con el propósito de determinar el punto óptimo de corte, se implementó el criterio de optimización basado en la segunda derivada de la función que relaciona el número de interacciones significativas con el valor de V de Cramer (Figura 5). Se ajustó un modelo polinómico de cuarto grado con un coeficiente de determinación $R^2 = 0.9993$:

$$y = 9230.8x^4 - 25173x^3 + 25148x^2 - 11063x + 1868.5$$

donde (y) representa el número de combinaciones significativas y x el valor del coeficiente V de Cramer. El punto óptimo corresponde al **punto de inflexión** de la curva, el decrecimiento acelerado de combinaciones significativas comienza a estabilizarse en este punto se equilibra la reducción del ruido estadístico con la retención de información estructural relevante.

Para hallarlo, se calcula la primera derivada de la ecuación:

$$y' = 36923.2x^3 - 75519x^2 + 50296x - 11063$$

a partir de esta, la segunda derivada:

$$y'' = 110769.6x^2 - 151038x + 50296$$

El punto de inflexión se obtiene **igualando la segunda derivada a cero**:

$$110769.6x^2 - 151038x + 50296 = 0$$

Aplicando la fórmula general de la ecuación cuadrática:

$$x = \frac{151038 \pm \sqrt{(151038)^2 - 4(110769.6)(50296)}}{2(110769.6)}$$

de donde se obtienen dos raíces:

$$x_1 = \frac{128072.7}{221539.2} \approx 0.578$$

$$x_2 = \frac{174003.3}{221539.2} \approx 0.785$$

La evaluación del signo de y'' confirma que en x_1 la curva transita de convexa ($y'' > 0$, decrecimiento acelerado) a cóncava ($y'' < 0$, estabilización); por ejemplo:

$$y''(0.4) = 7603.9 > 0, \quad y''(0.7) = -1153.5 < 0$$

por lo que x_1 constituye el punto de inflexión buscado. En cambio, $x_2 \approx 0.785$ se localiza en la zona asintótica de la función, donde el sistema ya ha perdido la mayor parte de sus variables (Tabla 8), razón por la cual se descarta.

Como resultado de este análisis, el modelo ubica el punto óptimo en $x_1 \approx 0.578$, coincidente con el umbral de $V = 0.57$ adoptado en el estudio. A partir de este umbral, se identificaron 29 combinaciones significativas que involucran a 26 variables. Se evidencia que la aplicación del coeficiente V de Cramer, complementada con un criterio de optimización basado en la segunda derivada, posibilita la selección objetiva de un subconjunto de variables de alta relevancia. Este procedimiento no solo facilita la interpretación de los resultados, sino que también fortalece la validez de técnicas multivariadas posteriores, como el Análisis Factorial de Correspondencias Múltiples (AFCM), al garantizar el uso de variables con relaciones estructuralmente consistentes.

Las 26 variables seleccionadas, que presentan un nivel de asociación superior a 0,57 según el coeficiente V de Cramer, son consideradas las más relevantes para el análisis multivariado y se detallan en la siguiente matriz (Figura 7)

Figura 7

Matriz del valor de V de cramer a pares de las variables con correlación moderada alta 26 variables

	Documento PDF completo	Escuela profesional	Número total de páginas	Facultad	Número de figuras	Uso de estadística inferencial	Total, de palabras	Nivel estadístico	Calidad de la tesis (0–20)	Uso de SPSS	Peso del archivo PDF (MB)	Número de tablas	Temática principal	Tipo de investigación	Tipo de relación	Número de enlaces	Año de sustentación	Riqueza gramatical	Fuente de información: Número de citas	Fuente de información: Enfoque metodológico	Uso de estadística multivariada	Género del asesor	Número de tesis por asesor	Co-asesor		
Documento PDF completo	0,25	1,00	0,14	1,00	0,24	1,00	0,26	0,82	0,00	1,00	1,00	0,18	0,25	0,20	1,00	0,68	0,55	0,16	0,96	0,02	0,12	0,17	0,12	0,13	0,07	
Escuela profesional	0,25		0,34	1,00	0,36	0,51	0,31	0,48	0,20	0,65	0,29	0,31	0,74	0,34	0,41	0,26	0,18	0,19	0,65	0,27	0,57	0,61	0,40	0,51	0,39	0,69
Número total de páginas	1,00	0,34		0,27	0,58	0,29	0,71	0,24	0,41	0,25	0,59	0,56	0,25	0,18	0,16	0,55	0,35	0,55	0,21	0,49	0,05	0,19	0,17	0,10	0,09	0,12
Facultad	0,14	1,00	0,27		0,29	0,36	0,23	0,34	0,16	0,59	0,21	0,23	0,64	0,25	0,31	0,19	0,10	0,13	0,57	0,20	0,48	0,48	0,27	0,38	0,28	0,45
Número de figuras	1,00	0,36	0,58	0,29		0,25	0,55	0,20	0,41	0,27	0,55	0,57	0,27	0,15	0,16	0,56	0,34	0,55	0,25	0,48	0,07	0,13	0,18	0,11	0,08	0,10
Uso de estadística inferencial	0,24	0,51	0,29	0,36	0,25		0,30	0,91	0,32	0,39	0,26	0,27	0,33	0,62	0,65	0,25	0,20	0,26	0,18	0,24	0,19	0,49	0,28	0,08	0,07	0,04
Total, de palabras	1,00	0,31	0,71	0,23	0,55	0,30		0,25	0,41	0,25	0,56	0,55	0,21	0,17	0,15	0,53	0,35	0,55	0,22	0,49	0,09	0,22	0,17	0,10	0,08	0,11
Nivel estadístico	0,26	0,48	0,24	0,34	0,20	0,91	0,25		0,26	0,48	0,22	0,21	0,30	0,42	0,45	0,20	0,15	0,30	0,28	0,21	0,21	0,38	0,95	0,07	0,08	0,05
Calidad de la tesis (0–20)	0,82	0,20	0,41	0,16	0,41	0,32	0,41	0,26		0,22	0,41	0,41	0,15	0,14	0,12	0,41	0,29	0,67	0,16	0,40	0,06	0,13	0,22	0,13	0,10	0,07
Uso de SPSS	0,00	0,65	0,25	0,59	0,27	0,39	0,25	0,48	0,22		0,24	0,11	0,52	0,40	0,45	0,16	0,05	0,06	0,47	0,10	0,09	0,35	0,08	0,15	0,12	0,10
Peso del archivo PDF (MB)	1,00	0,29	0,59	0,21	0,55	0,26	0,56	0,22	0,41	0,24		0,52	0,21	0,16	0,15	0,52	0,38	0,55	0,25	0,48	0,06	0,12	0,18	0,09	0,09	0,10
Número de tablas	1,00	0,31	0,56	0,23	0,57	0,27	0,55	0,21	0,41	0,11	0,52		0,22	0,14	0,14	0,54	0,34	0,55	0,22	0,48	0,07	0,15	0,19	0,10	0,08	0,08
Temática principal	0,18	0,74	0,25	0,64	0,27	0,33	0,21	0,30	0,15	0,52	0,21	0,22		0,22	0,28	0,18	0,09	0,14	0,50	0,19	0,40	0,44	0,23	0,26	0,17	0,20
Tipo de investigación	0,25	0,34	0,18	0,25	0,15	0,62	0,17	0,42	0,14	0,40	0,16	0,14	0,22		0,59	0,14	0,10	0,20	0,40	0,14	0,54	0,22	0,26	0,10	0,05	0,22
Tipo de relación	0,20	0,41	0,16	0,31	0,16	0,65	0,15	0,45	0,12	0,45	0,15	0,14	0,28	0,59		0,14	0,11	0,17	0,42	0,12	0,37	0,24	0,25	0,08	0,07	0,13
Número de enlaces	1,00	0,26	0,55	0,19	0,56	0,25	0,53	0,20	0,41	0,16	0,52	0,54	0,18	0,14	0,14		0,47	0,55	0,21	0,48	0,08	0,11	0,18	0,10	0,08	0,08
Año de sustentación	0,68	0,18	0,35	0,10	0,34	0,20	0,35	0,15	0,29	0,05	0,38	0,34	0,09	0,10	0,11	0,47		0,46	0,16	0,34	0,08	0,06	0,14	0,11	0,10	0,13
Riqueza gramatical	0,55	0,19	0,55	0,13	0,55	0,26	0,55	0,30	0,67	0,06	0,55	0,55	0,14	0,20	0,17	0,55	0,46		0,05	0,53	0,03	0,07	0,18	0,13	0,13	0,05
Fuente de información: encuesta	0,16	0,65	0,21	0,57	0,25	0,18	0,22	0,28	0,16	0,47	0,25	0,22	0,50	0,40	0,42	0,21	0,16	0,05		0,17	0,34	0,28	0,13	0,13	0,15	0,12
Número de citas	0,96	0,27	0,49	0,20	0,48	0,24	0,49	0,21	0,40	0,10	0,48	0,48	0,19	0,14	0,12	0,48	0,34	0,53	0,17		0,03	0,11	0,21	0,13	0,08	0,11
Fuente de información: experimento	0,02	0,57	0,05	0,48	0,07	0,19	0,09	0,21	0,06	0,09	0,06	0,07	0,40	0,54	0,37	0,08	0,08	0,03	0,34	0,03		0,18	0,14	0,07	0,10	0,16
Enfoque metodológico	0,12	0,61	0,19	0,48	0,13	0,49	0,22	0,38	0,13	0,35	0,12	0,15	0,44	0,22	0,24	0,11	0,06	0,07	0,28	0,11	0,18		0,14	0,04	0,04	0,10
Uso de estadística multivariada	0,17	0,40	0,17	0,27	0,18	0,28	0,17	0,95	0,22	0,08	0,18	0,19	0,23	0,26	0,25	0,18	0,14	0,18	0,13	0,21	0,14	0,14		0,03	0,09	0,01
Género del asesor	0,12	0,51	0,10	0,38	0,11	0,08	0,10	0,07	0,13	0,15	0,09	0,10	0,26	0,10	0,08	0,10	0,11	0,13	0,13	0,13	0,07	0,04	0,03		0,66	0,06
Número de tesis por asesor	0,13	0,39	0,09	0,28	0,08	0,07	0,08	0,08	0,10	0,12	0,09	0,08	0,17	0,05	0,07	0,08	0,10	0,13	0,15	0,08	0,10	0,04	0,09	0,66		0,06
Co-asesor	0,07	0,69	0,12	0,45	0,10	0,04	0,11	0,05	0,07	0,10	0,10	0,08	0,20	0,22	0,13	0,08	0,13	0,05	0,12	0,11	0,16	0,10	0,01	0,06	0,06	

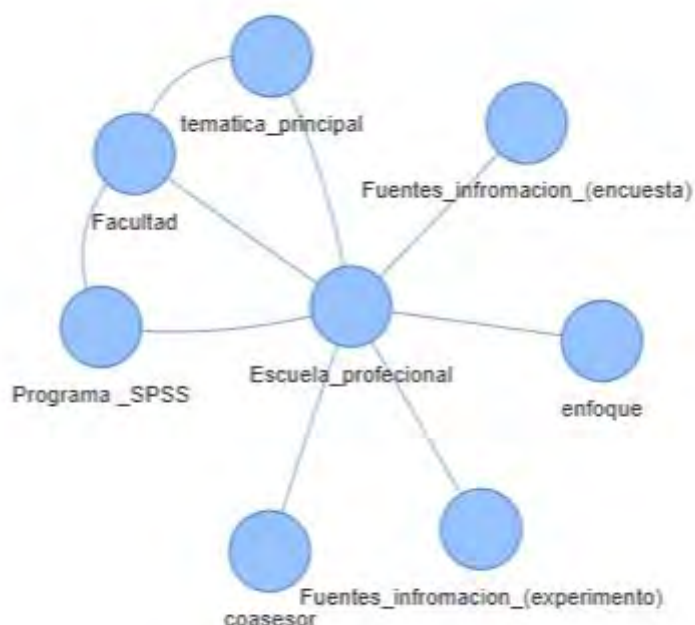
5.3. Análisis multivariado

En esta sección se presenta el análisis multivariado, el cual explora en profundidad la estructura de asociación entre las variables categóricas del estudio. Se empleó el coeficiente V de Cramer como medida de intensidad de la relación, aplicando un umbral de corte de 0,57. Este criterio permitió filtrar y seleccionar únicamente aquellas variables que exhiben una asociación moderada a alta, descartando los vínculos débiles que podrían introducir ruido en el análisis.

Como resultado de este filtrado, se obtuvo la matriz de correlación que se detalla en la Figura 7, la cual muestra los valores del coeficiente V de Cramer a pares para las 26 variables más representativas del sistema. Esta matriz no solo permite visualizar la intensidad de las interacciones más relevantes, sino que también constituye la base estructural para la elaboración de los grafos de red. A través de estas representaciones gráficas, se facilita la identificación de nodos centrales y patrones estructurales dentro del conjunto de datos, lo que posibilita segmentar las variables en bloques cohesivos para su posterior análisis mediante el Análisis Factorial de Correspondencias Múltiples (AFCM)

Figura 8

Grafo 1 de la matriz de correlación



La Figura 8 presenta el grafo de red derivado de la matriz de asociación construida a partir del coeficiente V de Cramer, en el cual se destacan las relaciones más robustas entre

variables categóricas ($V > 0,57$). La variable Escuela profesional emerge como el nodo central y articulador del sistema, conectándose directamente con Facultad, Temática principal, Enfoque, Fuentes de información (encuesta y experimento), Programa SPSS y Co-asesor.

Esta alta centralidad sugiere que la escuela profesional desempeña un rol determinante en la configuración de las características metodológicas y temáticas de las tesis. La conexión estrecha con la facultad y la temática principal revela una lógica estructural que vincula el ámbito académico con las líneas de investigación. El grafo evidencia que las decisiones metodológicas, el uso de herramientas estadísticas y las fuentes de información se encuentran fuertemente condicionadas por la disciplina o unidad académica a la que pertenece el tesista. Figura 8 presenta el grafo derivado de la matriz de asociación construida a partir del coeficiente V de Cramer, en el que se destacan las relaciones más significativas entre variables categóricas con valores superiores a 0,57. Se aprecia que la variable Escuela profesional actúa como nodo central, estableciendo múltiples relaciones con variables como facultad, temática principal, enfoque, fuentes de información (encuesta y experimento), uso de SPSS y co-asesor. Esta centralidad sugiere que la escuela profesional desempeña un papel preponderante en la configuración de las características metodológicas y temáticas de las tesis, la relación con la facultad y la temática principal indica una lógica estructural entre el ámbito académico y las líneas de investigación analizadas, el grafo permite identificar una razón donde las decisiones metodológicas y fuentes de información están fuertemente influenciadas por la unidad académica a la que se relaciona la investigación.

Figura 9

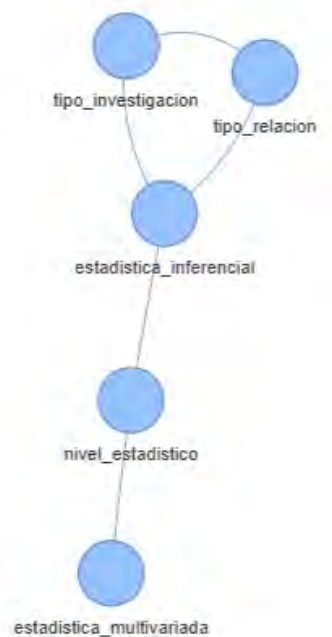
Grafo 2 de la matriz de correlación



La Figura 9 presenta el segundo subgrafo derivado de la matriz de asociación, el cual evidencia un componente aislado compuesto por las variables Género del asesor y Número de tesis por asesor. La arista que conecta ambos nodos supera el umbral de corte establecido ($V > 0,57$), lo que indica una asociación directa y robusta entre la carga de dirección académica y el género del docente.

Figura 10

Grafo 3 de la matriz de correlación

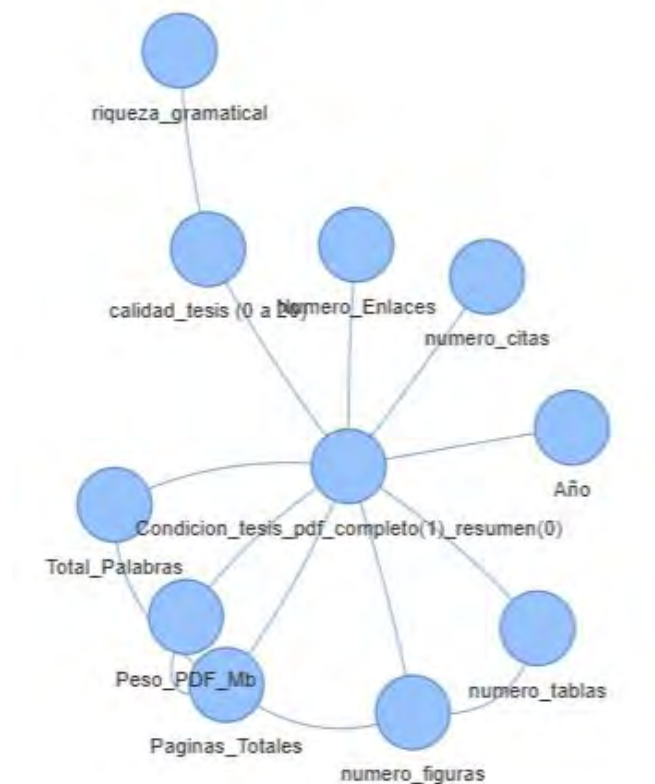


La Figura 10 presenta el tercer subgrafo derivado de la matriz de asociación, el cual evidencia relaciones significativas entre variables metodológicas y estadísticas. En la parte superior, se observa un triángulo de interconexión entre las variables Tipo de investigación, Tipo de relación y Estadística inferencial, lo que denota una alta coherencia entre el diseño metodológico y las técnicas de análisis empleadas.

La estructura se extiende de manera escalonada: la variable Estadística inferencial se conecta con Nivel estadístico, indicando que la complejidad del análisis está intrínsecamente ligada al uso de métodos inferenciales. Nivel estadístico se vincula con Estadística multivariada, sugiriendo una progresión natural hacia una mayor sofisticación en el tratamiento de los datos. El grafo refleja una cadena de dependencias en la que las decisiones metodológicas iniciales condicionan directamente el tipo y el nivel de las herramientas estadísticas utilizadas en las investigaciones.

Figura 11

Grafo 4 de la matriz de correlación



La Figura 11 presenta el cuarto subgrafo derivado de la matriz de asociación, el cual evidencia relaciones significativas entre variables estructurales, cuantitativas y de calidad de las tesis. En esta topología, la variable Condición del documento (PDF completo o solo resumen) emerge como el nodo central y articulador, conectándose directamente con el total de palabras, peso del PDF, número de páginas, figuras, tablas, citas y enlaces. Esta centralidad indica que la integridad y disponibilidad del documento en el repositorio están intrínsecamente ligadas a su nivel de desarrollo y densidad de contenido.

Se observa una arista directa entre la Calidad de la tesis y la Riqueza gramatical, lo que sugiere que la corrección y sofisticación del lenguaje constituyen un predictor relevante de la valoración global del trabajo. La conexión de este núcleo con la variable Año de sustentación refleja posibles cambios temporales en la extensión y estructuración de los documentos a lo largo del tiempo. El grafo demuestra que las características formales, métricas y de calidad no son independientes, sino que conforman un sistema interrelacionado que define los patrones de elaboración de los documentos académicos.

Análisis factorial de correspondencias de las interacciones significativas del estudio

Análisis factorial del grafo 1

Tabla 10

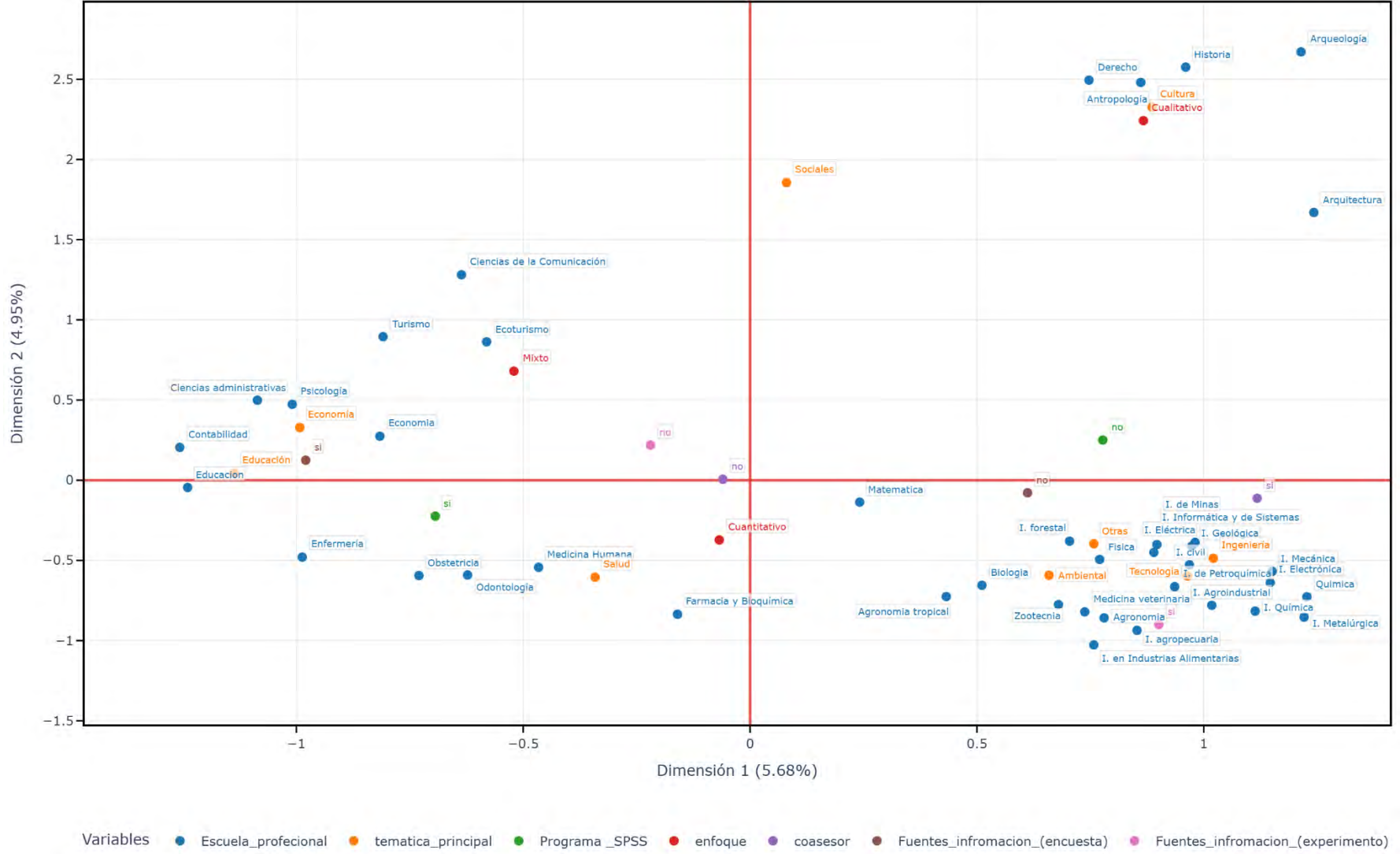
Informe del primer análisis factorial de correspondencia

Elemento	Valor
Dimensión 1	5.68%
Dimensión 2	4.95%
Total, acumulado (Dim1 + Dim2)	10.63%
Interpretación general	Las dos primeras dimensiones resumen de forma limitada la estructura de los datos; conviene revisar más dimensiones.
Categorías clave dimensión 1	Fuentes información_(encuesta) _si, Programa _SPSS__no, Programa _SPSS__si
Categorías clave dimensión 2	enfoque__Cualitativo, tematica_principal__Sociales, tematica_principal__Cultura

Dimensión	% varianza	% acumulado
Dimensión 1	5,68	5,68
Dimensión 2	4,95	10,63
Dimensión 3	3,87	14,50
Dimensión 4	3,59	18,10
Dimensión 5	3,51	21,61

El análisis factorial de correspondencias del Grafo 1 revela que las dos primeras dimensiones explican el 10,63 % de la inercia total (5,68 % y 4,95 %). Si bien este porcentaje acumulado podría parecer bajo en comparación con técnicas aplicadas a variables cuantitativas, en el contexto del Análisis de Correspondencias Múltiples (AFCM) constituye un resultado esperado y metodológicamente válido. Esto se debe a que la inercia total en el AFCM se distribuye de manera más dispersa entre un mayor número de ejes factoriales, dada la naturaleza categórica de los datos y la cantidad de modalidades analizadas. El objetivo de los primeros ejes no es agotar la varianza total, sino capturar las inercias modificadas más relevantes y las asociaciones estructurales de mayor magnitud. La proyección en el plano factorial permite identificar los patrones de asociación más robustos. La dimensión 1 (5,68 %) se consolida como un eje metodológico-instrumental, vinculándose predominantemente con el uso de encuestas como fuente de información y la aplicación del software SPSS. La dimensión 2 (4,95 %) configura un eje temático-epistemológico, al asociarse con el enfoque cualitativo y las temáticas de ciencias sociales y cultura. La contribución progresiva de las dimensiones restantes (que elevan el acumulado a más del 21 % en los primeros cinco ejes) confirma el carácter multidimensional y la complejidad del sistema de variables, validando que la estructura subyacente de las tesis de la UNSAAC.

Figura 12
Mapa factorial del Análisis de Correspondencias Múltiples (AFCM) Grafo 1



La Figura 12 presenta el mapa factorial del Análisis de Correspondencias Múltiples (AFCM) correspondiente al Grafo 1, en el cual se proyectan las asociaciones entre variables categóricas relacionadas con las características académicas, metodológicas y de fuentes de información de las tesis estudiadas.

A lo largo del eje horizontal (Dimensión 1), se observa una clara polarización disciplinar. Hacia el extremo izquierdo se agrupan las escuelas profesionales de ciencias sociales, administrativas y de la salud (como Educación, Contabilidad, Psicología, Enfermería y Medicina Humana), las cuales se asocian predominantemente al uso de encuestas como fuente de información. Hacia el extremo derecho se ubican las carreras de ingeniería y ciencias aplicadas (Ingeniería de Minas, Geológica, Civil, Química, entre otras), vinculadas a enfoques cuantitativos, análisis técnicos y fuentes de información experimentales.

La dimensión 2 (eje vertical) configura un eje temático-epistemológico. En la parte superior del plano se concentran disciplinas humanísticas y sociales (Derecho, Antropología, Historia, Arqueología y Ciencias de la Comunicación), junto con la categoría Cualitativo y la temática Cultura. En la parte inferior se agrupan las disciplinas de carácter experimental, biológico y agropecuario (Agronomía, Zootecnia, Medicina Veterinaria, Biología e Ingenierías), asociadas al enfoque Cuantitativo.

Respecto a las variables metodológicas, el uso del programa SPSS (sí) se distribuye principalmente en los cuadrantes inferiores, siendo más frecuente en áreas que emplean análisis cuantitativos y fuentes experimentales, mientras que su ausencia (no) se asocia con los enfoques cualitativos ubicados en la zona superior. La variable Co-asesor no evidencia una agrupación marcada en los extremos del plano, lo que indica una menor dependencia en la estructuración global de los datos. El mapa factorial revela patrones de asociación coherentes entre la escuela profesional, la temática de investigación, el enfoque metodológico y las fuentes de información, reflejando que las decisiones metodológicas adoptadas en las tesis están fuertemente condicionadas por la disciplina académica de origen.

Análisis factorial del grafo 2

Tabla 11

Informe del segundo grafo análisis factorial de correspondencia

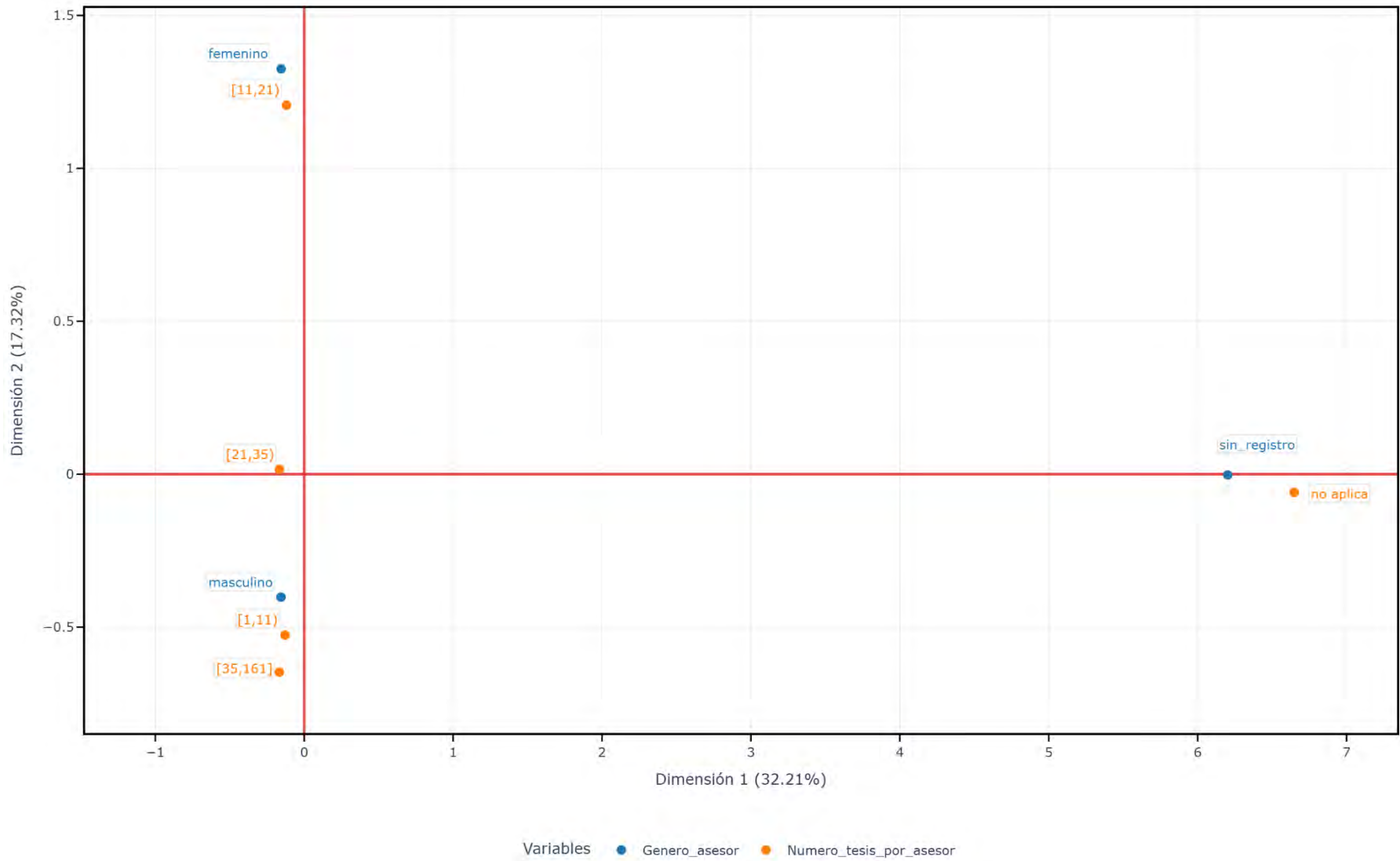
Elemento	Valor
Dimensión 1	32.21%
Dimensión 2	17.32%
Total, acumulado (Dim1 + Dim2)	49.53%
Interpretación general	Las dos primeras dimensiones resumen de manera aceptable la estructura de los datos.
Categorías clave en dimensión 1	Numero_tesis_por_asesor__no aplica, Genero_asesor__sin_registro, Genero_asesor__masculino
Categorías clave en dimensión 2	Genero_asesor__femenino, Numero_tesis_por_asesor__[11,21), Genero_asesor__masculino

Dimensión	% varianza	% acumulado
Dimensión 1	32,21	32,21
Dimensión 2	17,32	49,54
Dimensión 3	16,67	66,21
Dimensión 4	16,67	82,87
Dimensión 5	16,01	98,88

El análisis factorial de correspondencias del segundo grafo revela que las dos primeras dimensiones explican el 49,53 % de la inercia total, lo que permite representar de manera robusta la estructura de los datos en el plano factorial. La dimensión 1 (32,21 %) concentra la mayor proporción de información y se define principalmente por las categorías 'no aplica' (número de tesis por asesor) y 'sin registro' junto con 'masculino' (género del asesor). Esto indica que el primer eje factorial discrimina fundamentalmente en función de la disponibilidad de la información y la presencia predominante de asesores varones. La dimensión 2 (17,32 %) se asocia con el género 'femenino' y con el rango de tesis dirigidas '[11, 21)', configurando un eje que contrasta la participación de las asesoras mujeres con una carga académica intermedia. Cabe destacar que la categoría 'masculino' se proyecta a lo largo de ambas dimensiones, evidenciando su hegemonía y distribución heterogénea en la estructura de asesoría. Las dimensiones restantes contribuyen progresivamente a la explicación del modelo, alcanzando un acumulado del 98,88 % en los primeros cinco ejes, lo que corrobora la alta dimensionalidad del fenómeno. En síntesis, los resultados indican que la estructura de dirección de tesis está condicionada tanto por la carga académica y el género del asesor, como por la calidad y disponibilidad del registro de datos institucionales.

Figura 13

Mapa factorial del Análisis de Correspondencias Múltiples (AFCM) Grafo 2



La Figura 13 presenta el mapa factorial del Análisis de Correspondencias Múltiples (AFCM) correspondiente al Grafo 2, el cual proyecta las asociaciones entre el género del asesor y el número de tesis dirigidas. Las dos primeras dimensiones explican el 49,53 % de la inercia total (Dimensión 1: 32,21 % y dimensión 2: 17,32 %), lo que permite una representación robusta de la estructura de los datos en el plano bidimensional.

En el eje horizontal (Dimensión 1), se observa una polarización marcada por la disponibilidad de la información. Hacia el extremo derecho se ubican las categorías 'sin registro' y 'no aplica', las cuales se asocian fuertemente con la ausencia de datos sobre el asesor y su carga académica. En el lado izquierdo (cerca del origen) se concentran todas las categorías con información válida, evidenciando que este eje actúa principalmente como un discriminante entre la completitud y la falta de los registros institucionales.

La dimensión 2 (eje vertical) revela un patrón de asociación entre el género del asesor y su carga de trabajo. En el cuadrante superior se proyectan la categoría 'femenino' y el rango de tesis '[11, 21]', lo que sugiere un perfil de asesoras con una carga académica intermedia. En el cuadrante inferior, la categoría 'masculino' se asocia con los rangos extremos de carga '[1, 11]' y '[35, 161]', indicando una mayor dispersión y variabilidad en la cantidad de tesis dirigidas por hombres. El rango '[21, 35]' se sitúa cerca del centro del eje, reflejando un comportamiento neutro o transversal a ambos géneros.

El mapa factorial evidencia que la estructura de la dirección de tesis está condicionada en primer lugar por la calidad del registro de datos (Dimensión 1) y, en segundo lugar, por dinámicas de género y carga académica (Dimensión 2), donde las asesoras tienden a concentrarse en rangos medios mientras que los asesores varones presentan una distribución más polarizada.

Análisis factorial del grafo 3

Tabla 12

Informe del tercer grafo mediante análisis factorial de correspondencia

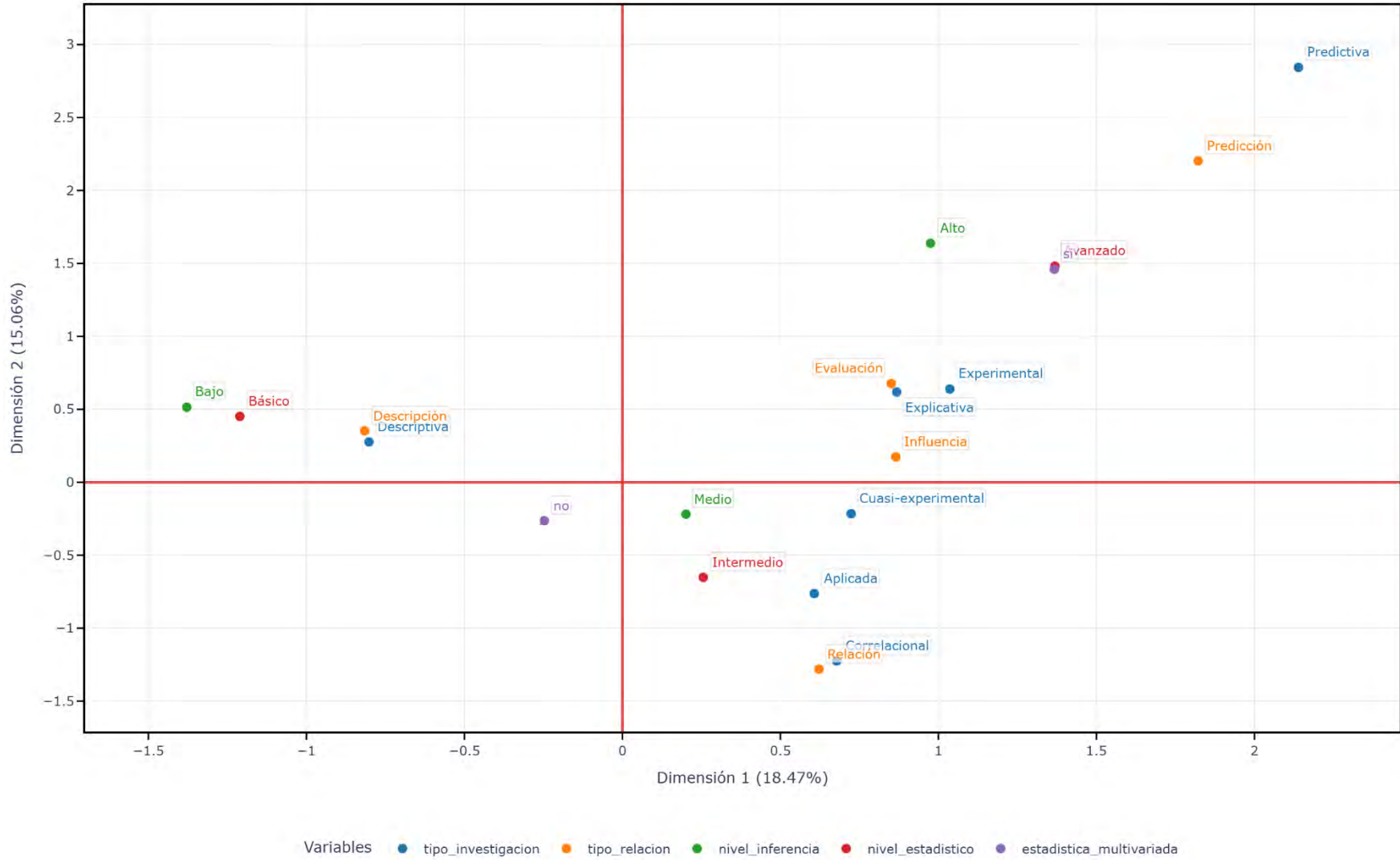
Elemento	Valor
Dimensión 1	18.47%
Dimensión 2	15.06%
Total, acumulado (Dim1 + Dim2)	33.53%
Interpretación general	Las dos primeras dimensiones resumen de forma limitada la estructura de los datos; conviene revisar más dimensiones.
Categorías clave en dimensión 1	nivel_estadistico__Básico, tipo_relacion__Descripción, tipo_investigacion__Descriptiva
Categorías clave en dimensión 2	tipo_relacion__Relación, tipo_investigacion__Correlacional, nivel_estadistico__Avanzado

Dimensión	% varianza	% acumulado
Dimensión 1	18,47	18,47
Dimensión 2	15,06	33,54
Dimensión 3	10,16	43,69
Dimensión 4	9,00	52,69
Dimensión 5	7,35	60,04

El análisis factorial de correspondencias del Grafo 3 revela que las dos primeras dimensiones explican el 33,53 % de la inercia total (Dimensión 1: 18,47 % y dimensión 2: 15,06 %). Si bien este porcentaje acumulado permite visualizar tendencias generales en el plano bidimensional, evidencia que la estructura de los datos es compleja y multidimensional.

La dimensión 1 (18,47 %) se define principalmente por las categorías Nivel estadístico básico, Tipo de relación descriptiva y Tipo de investigación descriptiva, lo que indica que este eje agrupa estudios con menor complejidad analítica, enfocados en la caracterización de fenómenos sin profundizar en las relaciones entre variables. La dimensión 2 (15,06 %) se asocia con las categorías Tipo de relación correlacional, Tipo de investigación correlacional y Nivel estadístico avanzado, configurando un eje que representa investigaciones con mayor rigurosidad metodológica y uso de técnicas estadísticas más sofisticadas. Estas dos dimensiones reflejan una progresión clara desde enfoques simples y descriptivos hacia enfoques correlacionales y avanzados, evidenciando un tránsito en la complejidad del análisis según el tipo de estudio. Las dimensiones restantes contribuyen progresivamente a la explicación de la inercia (alcanzando un acumulado del 60,04 % en los primeros cinco ejes).

Figura 14
Mapa factorial del Análisis de Correspondencias Múltiples (AFCM) Grafo 3



La Figura 14 presenta el mapa factorial del Análisis de Correspondencias Múltiples (AFCM) correspondiente al Grafo 3, el cual examina las asociaciones entre el tipo de investigación, el tipo de relación, el nivel de inferencia, el nivel estadístico y el uso de estadística multivariada

En la dimensión 1 (eje horizontal), se observa una clara polarización metodológica. Hacia el extremo izquierdo se ubican las categorías Descriptiva, Descripción, Básico y Bajo, las cuales configuran un perfil de investigaciones con menor complejidad analítica, enfocadas en la caracterización de fenómenos. Hacia el extremo derecho se proyectan categorías como Predictiva, Predicción, Avanzado y Alto, junto con Experimental y Explicativa, lo que denota un perfil de estudios con mayor rigor metodológico, capacidad de modelamiento y profundidad inferencial. En la zona central se ubican las categorías Correlacional, Relación y Aplicada, representando un nivel intermedio de complejidad.

La dimensión 2 (eje vertical) permite distinguir el grado de sofisticación del análisis. En el cuadrante superior se agrupan las categorías de mayor complejidad (Predictiva, Predicción, Avanzado, Alto y el uso 'sí' de estadística multivariada), reflejando estudios con alta capacidad de modelamiento y evaluación. En el cuadrante inferior se concentran categorías como Correlacional, Relación y Aplicada, que corresponden a investigaciones orientadas al establecimiento de vínculos entre variables. Se evidencia que la categoría 'sí' (estadística multivariada) se asocia estrechamente con los niveles Avanzado y Alto, mientras que la categoría 'no' se proyecta cerca de los niveles Intermedio y Medio.

El mapa factorial revela un gradiente progresivo en la estructura metodológica de las tesis. El tipo de investigación y el tipo de relación se asocian de manera sistemática con el nivel de inferencia y la complejidad estadística, evidenciando un tránsito claro desde enfoques descriptivos y básicos hacia diseños experimentales y predictivos de alta complejidad. Esta configuración refleja la diversidad metodológica de la producción académica analizada y valida la coherencia interna entre las decisiones de diseño y las herramientas analíticas empleadas

Análisis factorial del grafo 4

Tabla 13

Informe del cuarto grafo mediante análisis factorial de correspondencia

Elemento	Valor
Dimensión 1	22.04%
Dimensión 2	7.66%
Total, acumulado (Dim1 + Dim2)	29.70%
Interpretación general	Las dos primeras dimensiones resumen de forma limitada la estructura de los datos; conviene revisar más dimensiones.
Categorías clave en dimensión 1	Numero_Enlaces__no aplica, Peso_PDF_Mb__no aplica, numero_figuras__no aplica
Categorías clave en dimensión 2	Paginas_Totales__[183,1919], Paginas_Totales__[35,108), Total_Palabras__[37453,642049]

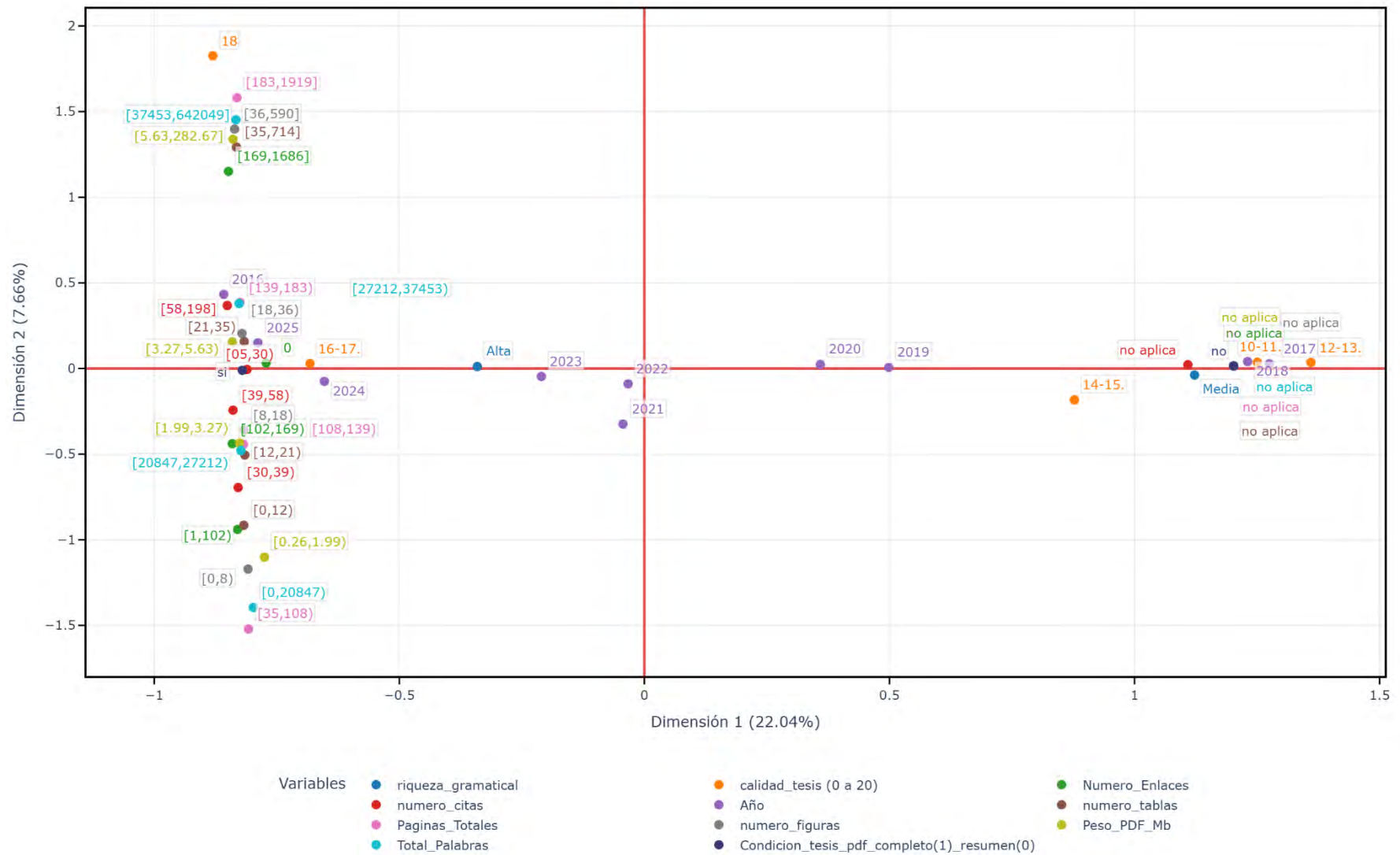
Dimensión	% varianza	% acumulado
Dimensión 1	22,04	22,04
Dimensión 2	7,66	29,71
Dimensión 3	4,98	34,68
Dimensión 4	3,91	38,60
Dimensión 5	3,14	41,73

El análisis factorial de correspondencias del Grafo 4 revela que las dos primeras dimensiones explican el 29,70 % de la inercia total (Dimensión 1: 22,04 % y dimensión 2: 7,66 %). Si bien este porcentaje acumulado permite visualizar tendencias generales en el plano bidimensional, evidencia que la estructura de los datos es compleja y multidimensional.

La dimensión 1 concentra la mayor proporción de inercia y se define principalmente por las categorías 'no aplica' asociadas al número de enlaces, peso del PDF y número de figuras. Esto indica que el primer eje factorial actúa como un discriminante basado en la completitud de los registros, diferenciando de manera marcada las tesis que carecen de información sobre sus características estructurales (generalmente aquellas registradas solo como 'resumen' en el repositorio) de aquellas que presentan el documento completo y detallado.

La dimensión 2 se asocia con los rangos superiores del número total de páginas (por ejemplo, [183, 1919]) y del total de palabras ([37453, 642049]). Este eje configura un gradiente de magnitud o extensión del documento, permitiendo distinguir entre trabajos de mayor volumen y profundidad documental frente a aquellos de menor extensión.

Mapa factorial del Análisis de Correspondencias Múltiples (AFCM) Grafo 4



La Figura 15 presenta el mapa factorial del Análisis de Correspondencias Múltiples (AFCM) correspondiente al Grafo 4, el cual examina las asociaciones entre variables vinculadas a la estructura, extensión y calidad de las tesis, tales como el número de páginas, total de palabras, cantidad de figuras y tablas, número de citas y enlaces, peso del archivo PDF, riqueza gramatical, año de sustentación y condición del documento (PDF completo o solo resumen). Las dos primeras dimensiones explican el 29,70 % de la inercia total (Dimensión 1: 22,04 % y dimensión 2: 7,66 %), lo que indica que, si bien la proyección bidimensional permite visualizar tendencias generales, la estructura de los datos es compleja y requiere la consideración de ejes factoriales adicionales para una interpretación integral.

En la dimensión 1 se observa una clara polarización. Hacia el extremo derecho del plano se proyectan las categorías asociadas a la ausencia de información estructural (modalidades 'no aplica' para número de enlaces, figuras, peso del PDF, etc.), junto con calificaciones de tesis más bajas ('10-11', '12-13', '14-15') y una riqueza gramatical 'Media'. Por el contrario, hacia el extremo izquierdo se ubican las tesis que presentan registros completos de sus características estructurales, asociadas a una calidad '16-17' y '18', y una riqueza gramatical 'Alta'. Esta separación evidencia que la completitud del documento en el repositorio constituye un eje fundamental, diferenciando entre trabajos con menor nivel de desarrollo formal o registro incompleto y aquellos más detallados y de mayor calidad académica.

La dimensión 2 (eje vertical) actúa como un gradiente de magnitud y extensión del documento. En la parte superior se agrupan los rangos superiores de las variables cuantitativas ([183, 1919] páginas, [37453, 642049] palabras y 18 citas), reflejando trabajos de mayor envergadura y densidad documental. En la parte inferior se concentran los rangos inferiores (como [0, 8] figuras o [0, 20847] palabras), correspondientes a documentos más breves. Variables como la calidad de la tesis y la riqueza gramatical se ubican cerca del origen, lo que sugiere que su variación está menos condicionada por la extensión absoluta del documento. El año de sustentación se distribuye a lo largo de este eje, indicando una evolución temporal en las características formales de las tesis.

El mapa factorial evidencia que las características estructurales, la extensión del documento y la calidad académica están fuertemente interrelacionadas, generando perfiles diferenciados de producción científica. La variabilidad moderada explicada por el primer plano factorial subraya la multidimensionalidad del fenómeno, confirmando que la arquitectura de las tesis de pregrado responde a múltiples factores que trascienden una simple representación bidimensional.

DISCUSIÓN

El estudio tuvo como finalidad examinar la estructura de las tesis de pregrado de la Universidad Nacional de San Antonio Abad del Cusco (UNSAAC) entre 2017 y 2025. Los resultados obtenidos, evidenciaron que la mayor parte de los trabajos de investigación se desarrolló bajo un enfoque cuantitativo y además siguieron un alcance descriptivo; asimismo, se observó una escasa colaboración entre los estudiantes en la elaboración de las tesis y una alta dependencia de recursos económicos propios para llevar a cabo sus investigaciones.

Se presentan los resultados alcanzados y se comparan con los resultados de otros estudios similares realizados tanto dentro del país, así como en el extranjero. Los resultados obtenidos evidenciaron que la investigación en la UNSAAC es predominantemente cuantitativa, de nivel descriptivo y correlacional, con un diseño no experimental de tipo transversal (96.8%) y esta concentración en los niveles básicos de investigación guarda similitud con lo reportado por Zúñiga Huamán (2019) que tomó como contexto a la Escuela de Medicina de la Universidad Andina del Cusco, donde el 57.4% de las tesis fueron de tipo descriptivo; asimismo, Enriquez Apaza (2020) pudo hallar que en las tesis de medicina de la UNSAAC el diseño más frecuente empleado es el transversal (56.2%).

Esta tendencia sugiere que el tesista de pregrado opta por diseños que demandan menor tiempo y recursos. Al respecto, Cárdenas Ortega y Gutierrez Blanco (2024) señalan que, en las tesis de comunicación a nivel nacional, el 65.3% se ubica en un nivel descriptivo. la persistencia en estos niveles básicos puede interpretarse, según Mamani (2018), como una priorización por culminar los estudios en corto tiempo por encima de la generación de conocimiento innovador.

En el ámbito estadístico, el presente estudio halló que el 99.9% de las tesis utiliza estadística descriptiva y un 67.3% estadística inferencial, mientras que solo un 15.3% se aventura en la estadística multivariada. Este panorama es globalmente similar al reportado por Debusho et al. (2025), quienes evidenciaron que en una universidad de educación a distancia la estadística descriptiva alcanzó el 80.6% de uso, pero advirtieron que un 41% de los trabajos presentaba errores metodológicos graves. Esto refuerza la necesidad observada en la UNSAAC de fortalecer la enseñanza de métodos avanzados para elevar el rigor científico.

Un hallazgo fundamental de esta investigación es la asociación significativa entre la calidad de la tesis y la riqueza gramatical ($V \approx 0.67$). Esto sugiere que la competencia lingüística es un indicador indirecto del rigor académico. No obstante, esta riqueza contrasta con los hallazgos de Guzmán Mamani y Huamani Muñoz (2024) en la Facultad de Educación

de la UNSAAC, donde se identificó un incumplimiento sistemático de las normas de tildación y ortografía, tanto por parte de los tesisistas como de los revisores.

En cuanto a la gestión académica, nuestro estudio revela que solo un 5.1% de las tesis cuenta con un co-asesor, lo que refleja una estructura de acompañamiento individualista. Cáceres Luna (2021) enfatiza que contar con un docente investigador como asesor influye positivamente en la calidad de los dictámenes y la rigurosidad del trabajo final. Cáceres Luna (2021) y Cárdenas Ortega y Gutierrez Blanco (2024) coinciden en que la calidad de las investigaciones mejora progresivamente cuando los trabajos son elaborados en coautoría (dos o más autores), una tendencia que ha aumentado tras la Ley Universitaria N.º 30220. En nuestra muestra, aunque existe una presencia de tesis grupales, el patrón estructural aún muestra una carga desigual de asesoría, con algunos docentes supervisando hasta 161 tesis.

El hallazgo de que el 90.1% de las tesis en la UNSAAC son financiadas con recursos propios del estudiante es un dato crítico. Esta realidad económica actúa como una barrera que limita el alcance de las investigaciones. Soncco Chullunquia (2017) determinó que, a pesar de que los estudiantes de odontología muestran una actitud positiva y motivación intrínseca (64.4%), la falta de recursos económicos es la principal dificultad personal para concluir la tesis.

Esta precariedad financiera se suma a las barreras institucionales. Cardeña Quispe (2021) reportó que el 85.2% de los internos de odontología de la UNSAAC consideran que el plan de estudios y la labor de los asesores restringen la realización de la tesis. Estos obstáculos alimentan el Síndrome Todo Menos Tesis (TMT), donde factores como la burocracia administrativa y la falta de apoyo docente desmotivan al investigador novel. Los resultados de nuestra investigación corroboran que la extensión del documento (mediana de 139 páginas) y su complejidad métrica están correlacionadas, lo que implica que una tesis de mayor calidad requiere una inversión de tiempo y dinero que la institución no está cubriendo actualmente.

El uso del Análisis Factorial de Correspondencias Múltiples (AFCM) en este estudio permitió identificar que la integridad del documento en el repositorio (PDF completo vs. resumen) organiza el resto de las variables estructurales. Se observó una clara diferenciación entre áreas académicas: las ingenierías se asocian a enfoques cuantitativos y análisis técnicos más complejos, mientras que las ciencias sociales y de la salud recurren mayoritariamente a encuestas.

Este mapeo estructural es comparable al análisis de correspondencia realizado por Arce Rivera (2023), quien identificó tendencias temáticas claras en posgrado, como el predominio del uso de TICs (21%) y enfoques psicológicos (23%). En la UNSAAC, la concentración

temática en torno a Cusco, Educación y Empresa refleja una investigación orientada a lo local, pero que a menudo carece de relevancia científica global. Hidalgo León (2022) advierte que las tesis de pregrado en ingeniería en el sur del Perú no alcanzan los niveles mínimos de relevancia científica y social, debido a la falta de alineación con indicadores bibliométricos internacionales y necesidades sociales verificables.

El estudio revela una escasa producción científica derivada de las tesis. Enriquez Apaza (2020) encontró que solo el 1.6% de las tesis de medicina de la UNSAAC llegaron a publicarse en revistas indexadas. Nuestros resultados estructurales muestran que, aunque el número de tesis en el repositorio ha crecido, la mayoría no reporta versiones publicadas o citas en bases de datos.

Esta falta de visibilidad se vincula con lo planteado por Simahate et al. (2022), quienes sostienen que la investigación debe insertarse en redes de colaboración para superar las desigualdades geográficas en la producción de conocimiento. En la UNSAAC, el patrón de baja colaboración académica (pocos co-asesor y predominancia de autores individuales en ciertas facultades) limita la capacidad de los estudiantes para producir artículos de alto impacto.

Los patrones en las tesis de la UNSAAC indican que los estudiantes cumplen formalmente con los requisitos, pero enfrentan dificultades económicas y limitaciones en sus métodos. La investigación en la universidad se estanca en un nivel básico de análisis cuantitativo y carece de apoyo institucional constante. Para mejorar esta situación, la universidad debe ir más allá de ser solo un lugar donde se archivan documentos. Debe fomentar una cultura de trabajo en equipo en investigación, proporcionar financiamiento y dar importancia a la escritura clara y al análisis estadístico detallado. De esta manera, se podrá pasar de hacer tesis solo por cumplir con un requisito a producir investigaciones que tengan un impacto real y significativo tanto en la región como en el país.

CONCLUSIONES

- Se determinó que las tesis de pregrado de la UNSAAC no constituyen un conjunto de documentos aislados, sino que presentan una estructura subyacente organizada y patrones de asociación sistémicos. La Escuela Profesional actúa como el nodo articulador principal del sistema, condicionando de manera determinante el contenido temático, el enfoque metodológico y las herramientas analíticas. Esto demuestra que la identidad académica y disciplinar dicta la arquitectura completa de la investigación estudiantil, validando el uso del Análisis Factorial de Correspondencias Múltiples (AFCM) para revelar la organización del conocimiento en la universidad.
- Se caracterizó la distribución de la producción académica, evidenciando una concentración temática en Salud (23,1 %), Ingeniería (18,3 %) y Economía (14,6 %). Metodológicamente, existe un predominio abrumador del enfoque cuantitativo (81,1 %) y de diseños descriptivos no experimentales (50,1 %), con picos de producción en 2024 (20,7 %) y 2017 (15,2 %). Prevalece la autoría individual (52,1 %) y la autofinanciación (90,1 %), lo que refleja una cultura de investigación que prioriza la culminación formal del grado mediante diseños de bajo costo y rapidez, en detrimento de la generación de conocimiento colaborativo o de alto impacto.
- Se estableció la existencia de un alto grado de dependencia estadística entre las variables categóricas, donde el 84,3 % de las 1891 combinaciones evaluadas resultaron significativas ($p < 0,01$). Mediante el coeficiente V de Cramer, se identificó un núcleo estructural de 26 variables con asociaciones fuertes ($V > 0,57$), destacando los vínculos entre Facultad-Escuela Profesional ($V \approx 0,85-0,90$) y Facultad-Temática Principal ($V \approx 0,65-0,75$).
- A través de la proyección en el espacio factorial del AFCM, se identificaron cuatro patrones estructurales o tipologías de tesis que explican la arquitectura de la producción científica en la UNSAAC:

Patrón de Determinismo Disciplinar-Metodológico (Grafo 1): Las carreras de ciencias sociales y salud se agrupan en el uso de encuestas y estadística básica (SPSS), mientras que las ingenierías y ciencias aplicadas se polarizan hacia enfoques experimentales, fuentes primarias y modelamiento técnico.

Patrón de Gestión y Asesoría Desconectada (Grafo 2): La dinámica de asesoría (género y carga de tesis) opera como un subgrafo aislado del núcleo metodológico. Está condicionada principalmente por la calidad del registro de datos institucionales y

muestra una distribución polarizada donde las asesoras se concentran en cargas intermedias y los asesores varones en los extremos.

Patrón de Gradiente de Complejidad Metodológica (Grafo 3): Existe una progresión factorial clara desde tesis descriptivas de nivel estadístico básico hacia diseños correlacionales/explicativos con uso de estadística multivariada e inferencial avanzada, evidenciando un tránsito en la madurez analítica.

Patrón de Integridad Estructural y Calidad Documental (Grafo 4): La disponibilidad del PDF completo actúa como el eje que organiza la extensión (páginas, palabras), la densidad de recursos (tablas, figuras, citas) y la calidad formal (riqueza gramatical y calificación)

RECOMENDACIONES

- Se sugiere a las autoridades académicas de la Universidad Nacional de San Antonio Abad del Cusco (UNSAAC) reforzar las políticas institucionales relacionadas con la investigación en pregrado, impulsando una planificación estratégica sustentada en evidencia. Esto permitirá orientar la producción de tesis hacia áreas prioritarias, lograr un mejor equilibrio temático y promover la innovación. Sería conveniente implementar mecanismos de seguimiento continuo del repositorio institucional que faciliten evaluar periódicamente las tendencias de investigación y tomar decisiones con mayor sustento.
- A las escuelas profesionales y comisiones de investigación se les recomienda impulsar una mayor variedad metodológica en las tesis de pregrado, promoviendo el uso de enfoques cualitativos y mixtos, así como la aplicación de técnicas estadísticas más avanzadas. De igual manera, se propone establecer lineamientos claros que orienten a los estudiantes en la elección de temas pertinentes y en el diseño de investigaciones con mayor rigurosidad científica y relevancia social.
- En cuanto a los docentes asesores, se recomienda fortalecer el acompañamiento durante todo el proceso de elaboración de la tesis, no solo en aspectos metodológicos, sino también en la selección de enfoques innovadores y en el uso adecuado de herramientas estadísticas. Sería importante fomentar el trabajo colaborativo entre estudiantes, incentivando la coautoría para enriquecer la calidad del análisis y de los resultados.
- Para los estudiantes de pregrado, se sugiere adoptar una actitud más activa y comprometida en el desarrollo de sus investigaciones, enfocándose en problemáticas actuales y relevantes. También se les recomienda explorar diversas metodologías que fortalezcan la calidad de sus trabajos y aprovechar el repositorio institucional como fuente para identificar vacíos de investigación y generar propuestas innovadoras en sus respectivas áreas.
- Se recomienda a entidades reguladoras como la SUNEDU y otros organismos vinculados a la educación superior promover normas y estándares que impulsen la calidad, la diversidad metodológica y la pertinencia de las tesis universitarias en el país. Sería valioso fomentar la colaboración entre universidades para intercambiar buenas prácticas en investigación y promover el uso de herramientas analíticas avanzadas en el estudio de la producción académica.

REFERENCIAS BIBLIOGRAFÍA

- Arce Rivera, F. R. (2023). Tendencias temáticas de las tesis de maestría de docencia universitaria de una universidad peruana, periodo 2015-2022. *Repositorio Institucional - UCV*. <https://repositorio.ucv.edu.pe/handle/20.500.12692/109189>
- Çakır, T., & Cönk, E. (2025). Bibliometric Analysis of Graduate Theses Containing the Phenomenon of Marketing Communication in the Council of Higher Education (CoHE) National Thesis Center Database. *International Journal of Business and Economic Studies*, 7(1), 19–32. <https://doi.org/10.54821/UIECD.1616516>
- Cardaña Quispe, A. (2021). *Factores que limitan la realización de tesis en internos y egresados de la Escuela Profesional de Odontología de la Universidad Nacional de San Antonio Abad del Cusco en el periodo 2019-2020*. <https://repositorio.unsaac.edu.pe/handle/20.500.12918/5802>
- Cárdenas Ortega, K., & Gutierrez Blanco, N. A. (2024). *Tesis de Comunicación en las universidades públicas licenciadas del Perú (2015-2022). Un estudio sobre las tesis de pregrado* [Universidad Nacional Mayor de San Marcos]. <https://cybertesis.unmsm.edu.pe/backend/api/core/bitstreams/40fcfcb2-3dc4-4fec-af80-30200fecde82/content>
- Catunta Guillen, L. M. (2024). *Aplicaciones de los métodos de análisis de clúster y correspondencia en el estudio de resultados de examen de admisión de la UNSAAC, 2022*. <https://repositorio.unsaac.edu.pe/handle/20.500.12918/8767>
- Chen, X., Chen, J., Wu, D., Xie, Y., & Li, J. (2016). Mapping the Research Trends by Co-word Analysis Based on Keywords from Funded Project. *Procedia Computer Science*, 91, 547–555. <https://doi.org/10.1016/J.PROCS.2016.07.140>
- Debushe, L. K., Mashabela, M. R., Sebatjane, P. N., Sithole, S., Tabo, B., & Rapoo, E. M. (2025). Evaluation of statistical methods applied in theses and dissertations in an Open, Distance and e-Learning University. *PLOS ONE*, 20(3). <https://doi.org/10.1371/JOURNAL.PONE.0319654>
- Enriquez Apaza, W. F. (2020). *Características bibliométricas de las tesis de pregrado de la Escuela Profesional de Medicina Humana, registradas en el Repositorio Institucional Digital de la Universidad Nacional de San Antonio Abad del Cusco, 2013-2020*. <https://repositorio.unsaac.edu.pe/handle/20.500.12918/5414>
- Greenacre, Michael. (2008). *La práctica del análisis de correspondencias*. Fundación BBVA. <https://www.fbbva.es/microsite/multivariate-statistics/practica.html>

- Guzman Mamani, R., & Huamani Muñoz, K. E. (2024). *Tildación en la redacción de las tesis de licenciatura en educación: estudio realizado desde la Universidad Nacional de San Antonio Abad Del Cusco periodo 2019 a 2022*.
<https://repositorio.unsaac.edu.pe/handle/20.500.12918/10260>
- Hernández Sampieri, R., Feránadez Collado, C., & Baptista Lucio, M. D. P. (2014). Metodología de la investigación. In *Metodología de la investigación*. McGraw Hill España.
<https://dialnet.unirioja.es/servlet/libro?codigo=775008&info=resumen&idioma=SPA>
- Hidalgo León, P. V. (2022). *Relevancia social y científica de las tesis de pre grado a través de una propuesta metodológica en las escuelas de ingeniería de sistemas e informática en la región sur de Perú* [Universidad Andina Del Cusco Escuela De Posgrado].
<https://repositorio.uandina.edu.pe/backend/api/core/bitstreams/74065b5b-0316-4ae2-a48e-0555c596e2e1/content>
- Kerlinger, F., & Lee, H. B. (2002). *Investigación del comportamiento* (Cuarta). McGraw-Hill. <https://archive.org/details/investigaciondel0000kerl/page/n9/mode/1up>
- Caceres Luna, G. L. (2021). *Calidad de las tesis de pregrado y rendimiento académico en estudiantes de psicología*. <https://gestion-repo.ucsp.edu.pe/items/5f89c0a3-1697-40afbcb-15b8b478b978/full>
- Pinto Ariza, P. J. (2023). Revisión sistemática de las tesis de Ingeniería en Higiene y Seguridad industrial en una Universidad Peruana, 2000-2022. *Repositorio Institucional - UCV*. <https://repositorio.ucv.edu.pe/handle/20.500.12692/121799>
- Simahate, T., Prawira, D. Y., Nurbaidah, S., Utami, E. F., & Mutia, C. L. (2022). Bibliometric Analysis of Social Sciences. *JUPIIS: JURNAL PENDIDIKAN ILMU-ILMU SOSIAL*, 14(2), 145–157. <https://doi.org/10.24114/JUPIIS.V14I2.39526>
- Soncco Chullunquia, J. L. (2017). *Factores que influyen en la elaboración de la tesis para optar el título profesional de Cirujano Dentista en los estudiantes de internado hospitalario y taller de tesis de la Universidad Alas Peruanas, Lima, 2017*.
https://repositorio.uap.edu.pe/xmlui/bitstream/handle/20.500.12990/1505/Tesis_Internado_Hospitalario.pdf?sequence=1&isAllowed=y
- Sousa, M. J., Suleman, F., Melé, P. M., & Gómez, J. M. (2021). New research and trends in higher education. *Education Sciences*, 11(9).
<https://doi.org/10.3390/EDUCSCI11090456>
- SUNEDU. (2020). *II Informe Bienal sobre la Realidad Universitaria en el Perú - Informes y publicaciones - Superintendencia Nacional de Educación Superior Universitaria -*

- Plataforma del Estado Peruano. <https://www.gob.pe/institucion/sunedu/informes-publicaciones/1093280-ii-informe-bienal-sobre-la-realidad-universitaria-en-el-peru>
- UNSAAC. (2015). *Universidad Nacional de San Antonio Abad del Cusco estatuto asamblea estatutaria ley universitaria N° 30220*.
http://transparencia.unsaac.edu.pe/links/datosgenerales/documentos/EstatutoUniversitario_UNSAAC.pdf
- Viloria, A., López, J. R., Leyva, D. M. G., Vargas-Mercado, C., Hernández-Palma, H., Llinas, N. O., David, M. A., & Rodriguez, J. V. (2019). Data mining techniques and multivariate analysis to discover patterns in university final researches. *Procedia Computer Science*, 155, 581–586. <https://doi.org/10.1016/J.PROCS.2019.08.081>
- Zúñiga Huamán, K. E. (2019). *Calidad de tesis de los egresados de la escuela Profesional de Medicina Humana de la Universidad Andina del Cusco*.
<https://repositorio.uandina.edu.pe/item/0096ef04-a3ec-4649-8e4a-3c7bca792223>
- Zuñiga Rojas, G. (2024). *Modelo de extracción automática de glosario de términos utilizando técnicas de procesamiento de lenguaje natural y clustering*.
<https://repositorio.unsaac.edu.pe/handle/20.500.12918/10014>

ANEXOS

Anexo 1. Ficha de observación de tesis universitarias

I. DATOS GENERALES

Código de tesis: _____

II. IDENTIDAD ACADÉMICA (F1)

Año de publicación: _____

Facultad: _____

☐ Ingeniería ☐ Ciencias de la Salud ☐ Sociales ☐ Educación ☐ Otro: _____

Escuela profesional: _____

III. CONTENIDO TEMÁTICO (F2)

Título de la tesis: _____

Resumen: _____

Palabras clave: _____

Temática principal

☐ Salud ☐ Educación ☐ Ingeniería ☐ Sociales ☐ Ambiental ☐ Otro: _____

¿Menciona COVID-19?

☐ Sí ☐ No

IV. ENFOQUE METODOLÓGICO (F3)

Enfoque de investigación:

☐ Cuantitativo ☐ Cualitativo ☐ Mixto

Tipo de investigación:

☐ Descriptivo ☐ Correlacional ☐ Explicativo ☐ Experimental

Diseño de investigación:

☐ No experimental ☐ Cuasi-experimental ☐ Experimental

Dimensión temporal:

☐ Transversal ☐ Longitudinal

Tipo de relación:

☐ Asociación ☐ Causalidad ☐ Comparación

Nivel estadístico:

☐ Bajo ☐ Medio ☐ Alto

Tipo de estadística utilizada:

☐ Descriptiva ☐ Inferencial ☐ Multivariada

¿Presenta muestra?

☐ Sí ☐ No

Tipo de muestra:

☐ Individuos ☐ Organizaciones ☐ Documentos

Tamaño de muestra: _____

Fuente de información:

☐ Encuesta ☐ Entrevista ☐ Experimento ☐ Documental ☐ Observación

V. HERRAMIENTAS ANALÍTICAS (F4)

Software estadístico utilizado: _____

☐ SPSS ☐ R ☐ Excel ☐ Stata ☐ Minitab ☐ Python ☐ MATLAB ☐ SAS ☐ No menciona

Técnicas estadísticas utilizadas (marcar las que correspondan):

☐ Frecuencias / porcentajes

☐ Media

☐ Mediana

☐ Chi-cuadrado

☐ t de Student

☐ ANOVA

☐ Correlación Pearson

☐ Correlación Spearman

☐ Regresión

☐ Alfa de Cronbach

☐ Shapiro-Wilk / K-S

☐ Análisis factorial /Componentes principales

VI. CARACTERÍSTICAS ESTRUCTURALES (F5)

Condición del documento en repositorio:

☐ Tesis completa ☐ Solo resumen

Número de páginas: _____

Número de palabras del título: _____

Número de palabras del resumen: _____

Número total de palabras del documento: _____

Número de palabras clave: _____

Peso del archivo PDF (MB): _____

Número de tablas: _____

Número de figuras: _____

Número de enlaces: _____

VII. CALIDAD Y PRODUCCIÓN ACADÉMICA (F6)

Número de citas: _____

Calidad de la tesis (0 a 20): _____

☐ Bajo (0–10) ☐ Medio (11–15) ☐ Alto (16–20)

Riqueza gramatical:

☐ Alta ☐ Media ☐ Baja

Nivel de inferencia:

☐ Bajo ☐ Medio ☐ Alto

VIII. GESTIÓN ACADÉMICA (F7)

Género del asesor:

☐ Masculino ☐ Femenino

Número de tesis por asesor: _____

Presencia de co-asesor:

☐ Sí ☐ No

Tipo de tesis según participación:

☐ Individual ☐ Grupal

Tipo de financiamiento:

☐ Institucional ☐ Externo ☐ Propio

Anexo 2. Matriz de consistencia

Problema	Objetivo	Hipótesis	Variables	Metodología
Problema general: ¿Qué asociaciones y patrones estructurales existen entre las características de las tesis de pregrado de la Universidad Nacional de San Antonio Abad del Cusco (UNSAAC), considerando variables como área temática, metodología, año de sustentación, género del tesista, tipo de colaboración y escuela profesional?	Objetivo general: Determinar y analizar las asociaciones y patrones estructurales en las tesis de pregrado de la UNSAAC mediante la prueba Chi-cuadrado y el Análisis Factorial de Correspondencias Múltiples (AFCM).	Hipótesis general: Existe asociación estadísticamente significativa entre las variables categóricas de identidad académica, contenido temático, enfoque metodológico, herramientas analíticas, características estructurales, calidad, producción y gestión académicas que caracterizan las tesis de pregrado de la Universidad Nacional de San Antonio Abad del Cusco (UNSAAC), lo que permite identificar patrones estructurales mediante el Análisis Factorial de Correspondencias Múltiples (AFCM).	Factores y variables: - Identidad académica: Año, facultad, escuela, carrera. - Contenido temático: Título, temática, palabras clave. - Enfoque metodológico: Tipo de metodología, diseño, dimensión temporal, fuentes, programas y técnicas estadísticas. - Composición del trabajo: Tipo (individual/grupal), género, asesoría. - Recursos gráficos y datos: N° tablas, figuras, tamaño de muestra, páginas, financiamiento. - Calidad y originalidad: N° citas, % similitud, riqueza gramatical, errores ortográficos.	Tipo de investigación: Cuantitativa, descriptiva, correlacional, transversal. Diseño muestral: Muestreo probabilístico estratificado proporcional por carrera profesional. Técnicas: Análisis documental. Análisis: Tablas de frecuencias, chi-cuadrado, AFCM.
Problemas específicos: 1. ¿Cómo se distribuyen las tesis de pregrado registradas en el repositorio institucional de la Universidad Nacional de San Antonio Abad del Cusco (UNSAAC), según variables como el área temática, la metodología de investigación, el año de sustentación, el género del tesista, el tipo de colaboración y la escuela profesional? 2. ¿Existe asociación estadísticamente significativa entre las variables categóricas que caracterizan las tesis de pregrado de la Universidad Nacional de San Antonio Abad del Cusco (UNSAAC), evaluadas mediante la prueba de Chi-cuadrado de independencia? 3. ¿Qué dimensiones factoriales subyacentes explican la estructura de asociación entre las variables categóricas presentes en las tesis de pregrado de la Universidad Nacional de San Antonio Abad del Cusco (UNSAAC), a través del Análisis Factorial de Correspondencias Múltiples (AFCM)? 4. ¿Qué patrones estructurales o tipologías de tesis pueden identificarse a partir de la proyección en el espacio factorial obtenido mediante el Análisis Factorial de Correspondencias Múltiples (AFCM) en las tesis de pregrado de la Universidad Nacional de San Antonio Abad del Cusco (UNSAAC)?	Objetivos específicos 1. Describir la distribución de las tesis de pregrado registradas en el repositorio institucional de la Universidad Nacional de San Antonio Abad del Cusco (UNSAAC), considerando variables categóricas como el área temática, la metodología de investigación, el año de sustentación, el género del tesista, el tipo de colaboración y la escuela profesional, con el propósito de caracterizar la producción académica. 2. Evaluar la existencia de asociación estadísticamente significativa entre las variables categóricas que caracterizan las tesis de pregrado de la Universidad Nacional de San Antonio Abad del Cusco (UNSAAC), mediante la aplicación de la prueba de Chi-cuadrado de independencia, a fin de identificar relaciones relevantes entre dichas variables. 3. Determinar las dimensiones factoriales subyacentes que explican la estructura de asociación entre las variables categóricas de las tesis de pregrado de la Universidad Nacional de San Antonio Abad del Cusco (UNSAAC), mediante la aplicación del Análisis Factorial de Correspondencias Múltiples (AFCM), con el fin de reducir la dimensionalidad de los datos y sintetizar la información. 4. Identificar y caracterizar patrones estructurales o tipologías de tesis de pregrado en la Universidad Nacional de San Antonio Abad del Cusco (UNSAAC), a partir de la interpretación del espacio factorial generado por el Análisis Factorial de Correspondencias Múltiples (AFCM), con el propósito de reconocer perfiles diferenciados de producción académica.	Hipótesis específicas: 1. La distribución de las tesis de pregrado de la UNSAAC presenta predominancia en determinadas categorías, tales como enfoque metodológico, tipo de diseño y modalidad de autoría. 2. Existe asociación estadísticamente significativa entre las variables categóricas de identidad académica, contenido temático, enfoque metodológico, herramientas analíticas, características estructurales, calidad, producción y gestión académicas de las tesis de pregrado de la UNSAAC, evaluada mediante la prueba de Chi-cuadrado ($p < 0,01$). 3. Las dimensiones factoriales obtenidas mediante el Análisis Factorial de Correspondencias Múltiples (AFCM) explican una proporción significativa de la variabilidad de las variables categóricas de identidad académica, contenido temático, enfoque metodológico, herramientas analíticas, características estructurales, calidad, producción y gestión académicas de las tesis de pregrado de la UNSAAC, permitiendo interpretar la estructura subyacente de los datos. 4. Se identifican patrones estructurales y tipologías diferenciadas en las tesis de pregrado de la UNSAAC a partir de la interpretación del espacio factorial generado por el Análisis Factorial de Correspondencias Múltiples (AFCM) de las variables categóricas de identidad académica, contenido temático, enfoque metodológico, herramientas analíticas, características estructurales, calidad, producción y gestión académicas.		

Anexo 3. Código

```
# =====
# TESIS - SCRIPT ÚNICO DE RECOLECCIÓN Y ANÁLISIS
# Versión resumida y ordenada para anexos / Google Colab
# =====
#
# Este script realiza, en orden:
# 1) Recolecta metadatos del repositorio UNSAAC mediante OAI-PMH
# 2) Extrae la URL real del PDF desde la página handle
# 3) Descarga masivamente los PDFs
# 4) Convierte PDF a TXT y genera metadatos documentales
# 5) Extrae variables estructuradas con Gemini
# 6) Calcula matrices de Chi-cuadrado y V de Cramer
# 7) Ejecuta Análisis de Correspondencias Múltiples (AFCM)
#
# Está pensado para ejecutarse en Google Colab.
# En anexos de tesis puede presentarse como código principal del estudio.
# =====

# =====
# 0. DEPENDENCIAS
# =====
!pip -q install requests beautifulsoup4 tqdm pandas openpyxl pypdf langdetect scipy prince plotly google-genai
xlswriter

# =====
# 1. IMPORTACIONES
# =====
import os
import re
import csv
import json
import math
import time
import random
from pathlib import Path
from urllib.parse import urljoin
from concurrent.futures import ThreadPoolExecutor, ProcessPoolExecutor, as_completed

import numpy as np
import pandas as pd
import requests
from bs4 import BeautifulSoup
from tqdm import tqdm
from pypdf import PdfReader
from langdetect import detect, LangDetectException
from scipy.stats import chi2_contingency
import prince
import plotly.express as px
from google import genai

# =====
# 2. CONFIGURACIÓN GENERAL
# =====
BASE_DIR = Path("/content/tesis_unsaac")
PDF_DIR = BASE_DIR / "pdfs"
TXT_DIR = BASE_DIR / "txt"
```

```

OUT_DIR = BASE_DIR / "salidas"

BASE_DIR.mkdir(exist_ok=True)
PDF_DIR.mkdir(exist_ok=True)
TXT_DIR.mkdir(exist_ok=True)
OUT_DIR.mkdir(exist_ok=True)

# Repositorio
OAI_BASE_URL = "https://repositorio.unsaac.edu.pe/oai/request"
HANDLE_BASE = "https://repositorio.unsaac.edu.pe"

# Archivos principales
METADATA_XLSX = OUT_DIR / "01_metadatos_repositorio.xlsx"
DOWNLOAD_CSV = OUT_DIR / "02_descargas.csv"
PDF_REPORT_XLSX = OUT_DIR / "03_reporte_pdfs.xlsx"
IA_OUTPUT_XLSX = OUT_DIR / "04_variables_extraidas.xlsx"
CHI_OUTPUT_XLSX = OUT_DIR / "05_matriz_chi_vcramer.xlsx"
MCA_OUTPUT_XLSX = OUT_DIR / "06_resultados_mca.xlsx"
MCA_HTML = OUT_DIR / "06_biplot_mca.html"

# Gemini: guardar en Colab como variable de entorno
# import os
# os.environ["GEMINI_API_KEY"] = "TU_API_KEY"
MODEL_NAME = "gemini-2.5-flash-lite"

# =====
# 3. UTILIDADES GENERALES
# =====
def safe_get(url, params=None, timeout=30, session=None, stream=False):
    try:
        headers = {"User-Agent": "Mozilla/5.0"}
        r = session.get(url, params=params, timeout=timeout, headers=headers, stream=stream) if session else
requests.get(
        url, params=params, timeout=timeout, headers=headers, stream=stream
    )
        r.encoding = r.apparent_encoding or "utf-8"
        return r
    except Exception:
        return None

def clean_text(text):
    if text is None:
        return ""
    text = str(text).replace("\x00", " ")
    text = re.sub(r"\r\n?", "\n", text)
    text = re.sub(r"[\t]+", " ", text)
    text = re.sub(r"\n{3,}", "\n\n", text)
    return text.strip()

def safe_metadata_get(metadata, key):
    return "" if not metadata else str(metadata.get(key, "")) or ""

def unique_keep_order(seq):
    seen = set()
    out = []
    for x in seq:
        x = str(x).strip()
        if x and x not in seen:
            seen.add(x)
            out.append(x)

```

```

return out

# =====
# 4. ETAPA 1: RECOLECCIÓN DE METADATOS DEL REPOSITORIO
# =====
def extract_all(md_soup, field):
    vals = md_soup.find_all(f"dc:{field}")
    return ", ".join(v.text.strip() for v in vals) if vals else ""

def clasificar_subjects(subjects):
    patrones_escuela = [
        "ingeniería", "derecho", "educación", "agronómica", "administración",
        "contabilidad", "enfermería", "odontología", "arquitectura",
        "economía", "medicina", "psicología", "ambiental", "biología",
        "ciencias", "facultad", "escuela", "tecnología"
    ]
    escuelas = []
    keywords = []
    for s in subjects:
        s_clean = s.strip()
        s_lower = s_clean.lower()
        if any(p in s_lower for p in patrones_escuela) or len(s_clean.split()) > 4:
            escuelas.append(s_clean)
        else:
            keywords.append(s_clean)
    return escuelas, keywords

def extract_pdf_from_handle(handle_url, session=None):
    if not handle_url:
        return ""
    if not handle_url.startswith("http"):
        handle_url = urljoin(HANDLE_BASE, handle_url)

    r = safe_get(handle_url, session=session, timeout=20)
    if not r or r.status_code != 200:
        return ""

    soup = BeautifulSoup(r.text, "html.parser")
    for a in soup.find_all("a", href=True):
        href = a["href"]
        if "/bitstream/" in href and ".pdf" in href.lower():
            return urljoin(handle_url, href)
    return ""

def process_record(rec, session=None):
    header = rec.find("header")
    identifier = header.identifier.text.strip() if header and header.identifier else ""
    setSpec = header.setSpec.text.strip() if header and header.setSpec else ""

    md = rec.find("oai_dc:dc") or rec.find("dc")
    if md is None:
        return None

    title = extract_all(md, "title")
    creator = extract_all(md, "creator")
    date = extract_all(md, "date")
    resumen = extract_all(md, "description")
    facultad = extract_all(md, "publisher")
    escuela_guess = extract_all(md, "contributor")
    tipo_doc = extract_all(md, "type") or "No especificado"

```

```

subject_vals = [v.text.strip() for v in md.find_all("dc:subject")] if md.find_all("dc:subject") else []
escuelas, keywords = clasificar_subjects(subject_vals)
escuela_profesional = ", ".join(escuelas) if escuelas else escuela_guess
keywords_str = ", ".join(keywords)

```

```

handle_url = ""
uri = ""
for idtag in md.find_all("dc:identifier"):
    t = idtag.text.strip()
    tl = t.lower()
    if "handle" in tl and not handle_url:
        handle_url = t
    elif tl.endswith(".pdf"):
        continue
    elif t and not uri:
        uri = t

```

```

if not handle_url and "handle" in identifier.lower():
    handle_url = identifier

```

```

pdf_url = extract_pdf_from_handle(handle_url, session=session) if handle_url else ""

```

```

return {
    "identifier": identifier,
    "title": title,
    "creator": creator,
    "date": date,
    "setSpec": setSpec,
    "handle_url": handle_url,
    "uri": uri,
    "tipo_documento": tipo_doc,
    "keywords": keywords_str,
    "pdf_url": pdf_url,
    "resumen": resumen,
    "facultad": facultad,
    "escuela_profesional": escuela_profesional
}

```

```

def harvest_oai_to_excel(output_xlsx):
    session = requests.Session()
    params = {"verb": "ListRecords", "metadataPrefix": "oai_dc"}
    all_rows = []

```

```

while True:
    r = safe_get(OAI_BASE_URL, params=params, session=session)
    if not r or r.status_code != 200:
        print("Error de conexión con OAI. Reintentando...")
        time.sleep(5)
        continue

```

```

soup = BeautifulSoup(r.text, "xml")
records = soup.find_all("record")
if len(records) == 0:
    break

```

```

with ThreadPoolExecutor(max_workers=10) as executor:
    futures = [executor.submit(process_record, rec, session) for rec in records]
    for future in tqdm(as_completed(futures), total=len(futures), desc="Procesando metadatos"):
        row = future.result()

```

```

        if row:
            all_rows.append(row)

    token_tag = soup.find("resumptionToken")
    if token_tag and token_tag.text and token_tag.text.strip():
        params = {"verb": "ListRecords", "resumptionToken": token_tag.text.strip()}
        time.sleep(1)
    else:
        break

df = pd.DataFrame(all_rows)
df.to_excel(output_xlsx, index=False)
print(f"Metadatos guardados en: {output_xlsx}")
return df

# =====
# 5. ETAPA 2: DESCARGA MASIVA DE PDF
# =====
def descargar_pdf(url, save_path):
    try:
        r = requests.get(url, stream=True, timeout=30)
        r.raise_for_status()
        total = int(r.headers.get("content-length", 0))
        descargado = 0

        with open(save_path, "wb") as f, tqdm(
            total=total, unit="B", unit_scale=True, desc=os.path.basename(save_path), ascii=True
        ) as bar:
            for chunk in r.iter_content(chunk_size=1024):
                if chunk:
                    f.write(chunk)
                    descargado += len(chunk)
                    bar.update(len(chunk))

        return True, round(descargado / 1024, 2)
    except Exception as e:
        print(f"Error al descargar {url}: {e}")
        return False, 0

def download_pdfs_from_metadata(metadata_df, output_csv, pdf_dir):
    rows = []
    for i, row in metadata_df.iterrows():
        file_id = str(i + 1).zfill(6)
        url = str(row.get("pdf_url", "")).strip()

        if not url:
            rows.append({"ID": file_id, "url": url, "descargado": "NO", "peso_kb": 0})
            continue

        save_path = pdf_dir / f"{file_id}.pdf"
        ok, size_kb = descargar_pdf(url, save_path)
        rows.append({"ID": file_id, "url": url, "descargado": "SI" if ok else "NO", "peso_kb": size_kb})

    df_desc = pd.DataFrame(rows)
    df_desc.to_csv(output_csv, index=False, encoding="utf-8-sig")
    print(f"Reporte de descargas guardado en: {output_csv}")
    return df_desc

# =====
# 6. ETAPA 3: PDF A TXT + METADATOS DOCUMENTALES

```

```

# =====
TABLA_REGEX_1 = re.compile(r"\btabla\s+\d+", re.IGNORECASE)
TABLA_REGEX_2 = re.compile(r"\bcuadro\s+\d+", re.IGNORECASE)

def process_pdf(pdf_path_str):
    pdf_path = Path(pdf_path_str)
    try:
        reader = PdfReader(str(pdf_path))
        is_encrypted = reader.is_encrypted
        num_pages = len(reader.pages)

        pages_with_images = 0
        total_links = 0
        total_chars = 0
        total_words = 0
        tables_detected = 0
        words_per_page = []
        text_parts = []

        for page in reader.pages:
            text = page.extract_text() or ""
            text_parts.append(text)

            total_chars += len(text)
            total_words += len(text.split())
            words_per_page.append(len(text.split()))

            if TABLA_REGEX_1.search(text) or TABLA_REGEX_2.search(text):
                tables_detected += 1

        try:
            resources = page.get("/Resources")
            if resources and "/XObject" in resources:
                pages_with_images += 1
        except Exception:
            pass

        try:
            annots = page.get("/Annots")
            if annots:
                for annot in annots:
                    try:
                        obj = annot.get_object()
                        if obj.get("/Subtype") == "/Link":
                            total_links += 1
                    except Exception:
                        pass
        except Exception:
            pass

        full_text = "\n".join(text_parts)
        txt_path = TXT_DIR / f"{pdf_path.stem}.txt"
        with open(txt_path, "w", encoding="utf-8") as f:
            f.write(full_text)

        language = "omitido"
        if full_text.strip():
            try:
                language = detect(full_text[:3000])
            except LangDetectException:

```

```

        language = "desconocido"

    metadata = reader.metadata or {}
    file_size_kb = round(pdf_path.stat().st_size / 1024, 2)

    return {
        "Nombre Archivo": pdf_path.name,
        "Paginas Totales": num_pages,
        "Paginas con Imagenes": pages_with_images,
        "Numero de Enlaces": total_links,
        "Protegido": is_encrypted,
        "Idioma Detectado": language,
        "Tamaño (KB)": file_size_kb,
        "Titulo": safe_metadata_get(metadata, "/Title"),
        "Autor": safe_metadata_get(metadata, "/Author"),
        "Asunto": safe_metadata_get(metadata, "/Subject"),
        "Creador": safe_metadata_get(metadata, "/Creator"),
        "Productor": safe_metadata_get(metadata, "/Producer"),
        "Fecha Creacion": safe_metadata_get(metadata, "/CreationDate"),
        "Fecha Modificacion": safe_metadata_get(metadata, "/ModDate"),
        "Total Caracteres": total_chars,
        "Total Palabras": total_words,
        "Conteo Palabras por Pagina": str(words_per_page),
        "Paginas con Posible Tabla": tables_detected
    }
except Exception as e:
    return {"Nombre Archivo": pdf_path.name, "Error": str(e)}

def convert_pdfs_to_txt_and_report(pdf_dir, output_excel):
    pdf_files = [str(p) for p in Path(pdf_dir).glob("*.pdf")]
    data = []

    with ProcessPoolExecutor(max_workers=2) as executor:
        futures = {executor.submit(process_pdf, pdf): pdf for pdf in pdf_files}
        for future in as_completed(futures):
            result = future.result()
            data.append(result)

    df = pd.DataFrame(data)
    df.to_excel(output_excel, index=False)
    print(f"Reporte PDF/TXT guardado en: {output_excel}")
    return df

# =====
# 7. ETAPA 4: EXTRACCIÓN ESTRUCTURADA CON GEMINI
# =====
SCHEMA_MIN = {
    "type": "object",
    "properties": {
        "archivo": {"type": "string"},
        "F1_url_tesis": {"type": "string"},
        "F1_anio_sustentacion": {"type": "string"},
        "F1_facultad": {"type": "string"},
        "F1_escuela_profesional": {"type": "string"},
        "F2_titulo": {"type": "string"},
        "F2_tematica_principal": {"type": "string"},
        "F2_palabras_clave": {"type": "array", "items": {"type": "string"}},
        "F3_enfoque": {"type": "string"},
        "F3_tipo_investigacion": {"type": "string"},
        "F3_diseno": {"type": "string"},
    }
}

```

```

    "F3_dimension_temporal": {"type": "string"},
    "F3_tecnicas_estadisticas": {"type": "array", "items": {"type": "string"}},
    "F3_software": {"type": "array", "items": {"type": "string"}},
    "F4_numero_autores": {"type": "integer"},
    "F4_asesor": {"type": "string"},
    "F5_numero_tablas": {"type": "integer"},
    "F5_numero_figuras": {"type": "integer"},
    "F5_numero_paginas": {"type": "integer"},
    "F5_tamano_muestra": {"type": "integer"},
    "F6_numero_citas": {"type": "integer"},
    "F6_similitud": {"type": "number"},
    "F6_calidad_tesis": {"type": "string"},
    "F6_riqueza_gramatical": {"type": "string"},
    "F6_errores_ortograficos": {"type": "integer"}
},
"required": [
    "archivo", "F1_url_tesis", "F1_anio_sustentacion", "F1_facultad", "F1_escuela_profesional",
    "F2_titulo", "F2_tematica_principal", "F2_palabras_clave",
    "F3_enfoque", "F3_tipo_investigacion", "F3_diseño", "F3_dimension_temporal",
    "F3_tecnicas_estadisticas", "F3_software",
    "F4_numero_autores", "F4_asesor",
    "F5_numero_tablas", "F5_numero_figuras", "F5_numero_paginas", "F5_tamano_muestra",
    "F6_numero_citas", "F6_similitud", "F6_calidad_tesis", "F6_riqueza_gramatical",
    "F6_errores_ortograficos"
]
}

```

SYSTEM_PROMPT = """

Actúa como experto en metodología de investigación y minería de texto académico.

Debes extraer variables de una tesis universitaria y devolver exclusivamente JSON válido.

Usa el título, resumen y texto como evidencia.

Mantén coherencia entre enfoque, diseño, muestra, técnicas estadísticas y software.

"""

```

def get_gemini_client():
    api_key = os.getenv("GEMINI_API_KEY", "").strip()
    if not api_key:
        raise ValueError("No se encontró GEMINI_API_KEY en variables de entorno.")
    return genai.Client(api_key=api_key)

def detect_title(text, default_name):
    lines = [x.strip() for x in text.splitlines() if x.strip()]
    for line in lines[:20]:
        if len(line.split()) >= 5:
            return line[:300]
    return default_name

def detect_abstract(text):
    m = re.search(r"(resumen|abstract)s*[:\n](.{0,4000})", text, re.IGNORECASE | re.DOTALL)
    if m:
        return clean_text(m.group(2)[:2000])
    return ""

def gemini_json_call(prompt, schema):
    client = get_gemini_client()
    response = client.models.generate_content(
        model=MODEL_NAME,
        contents=prompt,
        config={
            "system_instruction": SYSTEM_PROMPT,

```

```

        "temperature": 0,
        "response_mime_type": "application/json",
        "response_json_schema": schema
    }
)
return json.loads(response.text)

def process_single_txt(path):
    raw_text = Path(path).read_text(encoding="utf-8", errors="ignore")
    text = clean_text(raw_text)
    title = detect_title(text, Path(path).stem)
    abstract = detect_abstract(text)

    prompt = f"""
ARCHIVO: {Path(path).name}

TÍTULO DETECTADO:
{title}

RESUMEN DETECTADO:
{abstract}

TEXTO DE LA TESIS:
{text[:120000]}

Completa las variables solicitadas y devuelve solo JSON.
"""

    data = gemini_json_call(prompt, SCHEMA_MIN)
    data["archivo"] = Path(path).name

    # limpieza ligera
    for key in ["F2_palabras_clave", "F3_tecnicas_estadisticas", "F3_software"]:
        data[key] = unique_keep_order(data.get(key, []))

    return data

def run_gemini_on_txt_folder(txt_dir, output_xlsx):
    txt_files = list(Path(txt_dir).glob("*.txt"))
    rows = []

    for path in tqdm(txt_files, desc="Extrayendo variables con Gemini"):
        try:
            rows.append(process_single_txt(path))
        except Exception as e:
            rows.append({"archivo": path.name, "error": str(e)})

    df = pd.DataFrame(rows)
    df.to_excel(output_xlsx, index=False)
    print(f"Variables extraídas guardadas en: {output_xlsx}")
    return df

# =====
# 8. ETAPA 5: CHI-CUADRADO + V DE CRAMER
# =====
def cramers_v(chi2, n, r, c):
    k = min(r - 1, c - 1)
    if n <= 0 or k <= 0:
        return np.nan
    return math.sqrt(chi2 / (n * k))

```

```

def safe_crosstab(s1, s2):
    temp = pd.DataFrame({"x": s1, "y": s2}).dropna()
    if temp.empty:
        return pd.DataFrame()
    return pd.crosstab(temp["x"].astype(str), temp["y"].astype(str))

def build_chi_vcramer_matrices(df, output_xlsx):
    columnas = list(df.columns)

    mat_p = pd.DataFrame(np.nan, index=columnas, columns=columnas)
    mat_v = pd.DataFrame(np.nan, index=columnas, columns=columnas)
    mat_min = pd.DataFrame(np.nan, index=columnas, columns=columnas)
    mat_n = pd.DataFrame(np.nan, index=columnas, columns=columnas)

    for i, c1 in enumerate(columnas):
        for j in range(i, len(columnas)):
            c2 = columnas[j]

            if i == j:
                mat_n.loc[c1, c2] = df[[c1]].dropna().shape[0]
                continue

            tabla = safe_crosstab(df[c1], df[c2])

            if tabla.shape[0] < 2 or tabla.shape[1] < 2:
                continue

            obs = tabla.to_numpy()
            chi2, p, _, _ = chi2_contingency(obs)
            n = int(obs.sum())

            mat_p.loc[c1, c2] = mat_p.loc[c2, c1] = p
            mat_v.loc[c1, c2] = mat_v.loc[c2, c1] = crammers_v(chi2, n, *obs.shape)
            mat_min.loc[c1, c2] = mat_min.loc[c2, c1] = float(obs.min())
            mat_n.loc[c1, c2] = mat_n.loc[c2, c1] = n

    with pd.ExcelWriter(output_xlsx, engine="openpyxl") as writer:
        mat_p.to_excel(writer, sheet_name="Sig_p_valor")
        mat_v.to_excel(writer, sheet_name="V_Cramer")
        mat_min.to_excel(writer, sheet_name="Min_conteo_tabla")
        mat_n.to_excel(writer, sheet_name="N_validos")

    print(f"Matrices Chi/V de Cramer guardadas en: {output_xlsx}")
    return mat_p, mat_v

# =====
# 9. ETAPA 6: AFCM
# =====
def preparar_datos_mca(df):
    df = df.copy()
    df.columns = [clean_text(c) for c in df.columns]
    for col in df.columns:
        df[col] = (
            df[col]
            .astype(str)
            .str.strip()
            .replace({"": np.nan, "nan": np.nan, "None": np.nan})
            .fillna("Missing")
            .astype("category")
        )

```

```

    return df

def separar_variable_categoria(valor):
    valor = str(valor)
    if "_" in valor:
        partes = valor.split("_", 1)
        return partes[0], partes[1]
    if "=" in valor:
        partes = valor.split("=", 1)
        return partes[0], partes[1]
    return "Variable", valor

def run_mca(df_input, output_xlsx, output_html):
    df_cat = preparar_datos_mca(df_input)

    mca = prince.MCA(
        n_components=2,
        n_iter=10,
        copy=True,
        check_input=True,
        engine="sklearn",
        random_state=42
    )
    mca = mca.fit(df_cat)

    eigen = pd.DataFrame({
        "Dimensión": [f"Dimensión {i+1}" for i in range(len(mca.eigenvalues_))],
        "Autovalor": mca.eigenvalues_,
        "% varianza": np.array(mca.percentage_of_variance_),
        "% acumulado": np.cumsum(np.array(mca.percentage_of_variance_))
    })

    col_coords = mca.column_coordinates(df_cat).copy()
    col_coords.columns = [f"Dimensión {i+1}" for i in range(col_coords.shape[1])]

    coords_excel = col_coords.iloc[:, :2].copy().reset_index()
    coords_excel.columns = ["Categoría_completa", "Dimensión 1", "Dimensión 2"]

    variables = []
    categorias = []
    for valor in coords_excel["Categoría_completa"]:
        var, cat = separar_variable_categoria(valor)
        variables.append(var)
        categorias.append(cat)

    coords_excel["Variable"] = variables
    coords_excel["Categoría"] = categorias
    coords_excel = coords_excel[["Variable", "Categoría", "Dimensión 1", "Dimensión 2"]]

    with pd.ExcelWriter(output_xlsx, engine="xlsxwriter") as writer:
        eigen.to_excel(writer, sheet_name="Inercia", index=False)
        coords_excel.to_excel(writer, sheet_name="Coordenadas", index=False)

    fig = px.scatter(
        coords_excel,
        x="Dimensión 1",
        y="Dimensión 2",
        text="Categoría",
        color="Variable",
        title="Biplot de categorías - AFCM"
    )

```

```

)
fig.update_traces(textposition="top center")
fig.write_html(output_html)

print(f"Resultados MCA guardados en: {output_xlsx}")
print(f"Gráfico HTML guardado en: {output_html}")

# =====
# 10. PROCESO PRINCIPAL
# =====
def main():
    print("=" * 70)
    print("INICIO DEL PROCESO COMPLETO DE TESIS")
    print("=" * 70)

    # 1) Metadatos del repositorio
    df_meta = harvest_oai_to_excel(METADATA_XLSX)

    # 2) Descarga masiva
    df_desc = download_pdfs_from_metadata(df_meta, DOWNLOAD_CSV, PDF_DIR)

    # 3) PDF -> TXT + metadatos documentales
    df_pdf = convert_pdfs_to_txt_and_report(PDF_DIR, PDF_REPORT_XLSX)

    # 4) Extracción con Gemini
    df_ia = run_gemini_on_txt_folder(TXT_DIR, IA_OUTPUT_XLSX)

    # 5) Preparar matriz categórica para Chi y MCA
    cols_modelo = [
        c for c in df_ia.columns
        if c.startswith("F1_") or c.startswith("F2_") or c.startswith("F3_")
        or c.startswith("F4_") or c.startswith("F5_") or c.startswith("F6_")
    ]
    df_modelo = df_ia[cols_modelo].copy()

    # Convertir listas a texto
    for col in df_modelo.columns:
        df_modelo[col] = df_modelo[col].apply(lambda x: ", ".join(x) if isinstance(x, list) else x)

    # 6) Chi-cuadrado y V de Cramer
    build_chi_vcramer_matrices(df_modelo, CHI_OUTPUT_XLSX)

    # 7) AFCM
    run_mca(df_modelo, MCA_OUTPUT_XLSX, MCA_HTML)

    print("=" * 70)
    print("PROCESO FINALIZADO")
    print("=" * 70)

if __name__ == "__main__":
    main()

```

Anexo 3. Imágenes

